

HAND BOOKS

IN ECONOMICS

Gilles Duranton
J. Vernon Henderson
William Strange

Regional and Urban Economics

VOLUME 5A

NORTH-HOLLAND

Handbook of
**REGIONAL AND
URBAN ECONOMICS**

This page intentionally left blank

Handbook of **REGIONAL AND URBAN ECONOMICS**

Volume 5A

Edited by

GILLES DURANTON

Wharton School, University of Pennsylvania,
Philadelphia, PA, USA, and CEPR

J. VERNON HENDERSON

Department of Geography, London School of Economics,
London, UK

WILLIAM C. STRANGE

Rotman School of Management, University of Toronto,
Toronto, ON, Canada



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
North-Holland is an imprint of Elsevier



North-Holland is an imprint of Elsevier

Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Copyright © 2015 Elsevier B.V. All rights reserved.

Chapter 15, How Mortgage Finance Affects the Urban Landscape, Copyright © 2015 Elsevier B.V. and FRBNY. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting Obtaining permission to use Elsevier material.

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-59517-1 (Vol. 5A)

ISBN: 978-0-444-59531-7 (Vol. 5B)

For information on all North-Holland publications
visit our website at <http://store.elsevier.com/>



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Typeset by SPi Global, India

Printed and bound in the UK

Publisher: Nikki Levy

Acquisition Editor: J. Scott Bentley

Editorial Project Manager: Joslyn Chaiprasert-Paguio

Production Project Manager: Nicky Carter

Designer: Alan Studholme

INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

Kenneth J. Arrow and Michael D. Intriligator

This page intentionally left blank

CONTENTS

<i>Foreword</i>	xv
<i>Contributors</i>	xvii

Volume 5A

Section I. Empirical Methods	1
1. Causal Inference in Urban and Regional Economics	3
Nathaniel Baum-Snow, Fernando Ferreira	
1.1. Introduction	4
1.2. A Framework for Empirical Investigation	6
1.3. Spatial Aggregation	20
1.4. Selection on Observables	23
1.5. IV Estimators	43
1.6. Regression Discontinuity	53
1.7. Conclusion	62
References	63
2. Structural Estimation in Urban Economics	69
Thomas J. Holmes, Holger Sieg	
2.1. An Introduction to Structural Estimation	70
2.2. Revealed Preference Models of Residential Choice	74
2.3. Fiscal Competition and Public Good Provision	79
2.4. The Allocation of Economic Activity Across Space	96
2.5. Conclusions	110
Acknowledgments	111
References	111
3. Spatial Methods	115
Steve Gibbons, Henry G. Overman, Eleonora Patacchini	
3.1. Introduction	116
3.2. Nonrandomness in Spatial Data	120
3.3. Spatial Models	124
3.4. Identification	136
3.5. Treatment Effects When Individual Outcomes Are (Spatially) Dependent	152
3.6. Conclusions	157

Appendix A: Biases with Omitted Spatial Variables	158
Appendix B: Hypothetical RCT Experiments for Identifying Parameters in the Presence of Interactions Within Spatial Clusters	161
References	164
Section II. Agglomeration and Urban Spatial Structure	169
4. Agglomeration Theory with Heterogeneous Agents	171
Kristian Behrens, Frédéric Robert-Nicoud	
4.1. Introduction	172
4.2. Four Causes and Two Moments: A Glimpse at the Data	175
4.3. Agglomeration	187
4.4. Sorting and Selection	211
4.5. Inequality	234
4.6. Conclusions	239
Acknowledgments	240
References	241
5. The Empirics of Agglomeration Economies	247
Pierre-Philippe Combes, Laurent Gobillon	
5.1. Introduction	248
5.2. Mechanisms and Corresponding Specifications	252
5.3. Local Determinants of Agglomeration Effects	270
5.4. Estimation Strategy	282
5.5. Magnitudes for the Effects of Local Determinants of Productivity	298
5.6. Effects of Agglomeration Economies on Outcomes Other Than Productivity	314
5.7. Identification of Agglomeration Mechanisms	328
5.8. Conclusion	338
Acknowledgments	340
References	341
6. Agglomeration and Innovation	349
Gerald Carlino, William R. Kerr	
6.1. Introduction	350
6.2. What is Innovation?	352
6.3. Patterns of Agglomeration and Innovation	358
6.4. Formal Theories Linking Agglomeration and Innovation	366
6.5. Additional Issues on Innovation and Agglomeration	390
6.6. Conclusions	396
Acknowledgments	397
References	398

7. Cities and the Environment	405
Matthew E. Kahn, Randall Walsh	
7.1. Introduction	406
7.2. Incorporating Local and Global Environmental Externalities into Locational Equilibrium Models	409
7.3. Global Externalities Exacerbated by the Intrametro Area Locational Choice of Households and Firms	423
7.4. Environmental Amenities in a System of Cities	427
7.5. The Urban Building Stock's Energy Consumption	445
7.6. Conclusion	457
Acknowledgment	458
References	458
 8. Urban Land Use	 467
Gilles Duranton, Diego Puga	
8.1. Introduction	468
8.2. Modeling Urban Land Use: The Monocentric Model	472
8.3. Extending the Monocentric Model	483
8.4. Agglomeration and Commercial Land Use: Modeling Polycentric Cities	503
8.5. Land Use Regulation	515
8.6. Empirical Price and Development Gradients	522
8.7. Patterns of Residential Sorting Within Cities	530
8.8. Patterns of Residential Land Development	537
8.9. Employment Decentralization and Patterns of Business Location Changes Within Cities	544
8.10. Conclusion	551
Acknowledgments	553
References	553
 9. Neighborhood and Network Effects	 561
Giorgio Topa, Yves Zenou	
9.1. Introduction	562
9.2. Neighborhood Effects	566
9.3. Network Effects	578
9.4. Neighborhood and Network Effects	599
9.5. Concluding Remarks	615
Acknowledgments	617
References	617

10. Immigration and the Economy of Cities and Regions	625
Ethan Lewis, Giovanni Peri	
10.1. Introduction	626
10.2. Immigrants' Distribution and Native Exposure	632
10.3. Theoretical Framework: The Skill Cells Approach at the National and Local Level	637
10.4. Empirical Approaches to Identify Causal Effects on Local Economies	657
10.5. Estimates of Native Responses and Effects on Outcomes	661
10.6. Recent Evolutions: Employer–Employee Panel Data and Historical Data	675
10.7. Conclusions	680
References	681
<i>Index</i>	687
 Volume 5B	
Section III. Housing and Real Estate	699
11. Housing Bubbles	701
Edward L. Glaeser, Charles G. Nathanson	
11.1. Introduction	702
11.2. The Linear Asset Pricing Model and the Idiosyncrasies of Housing	705
11.3. Empirical Regularities of Housing Dynamics	715
11.4. Rationalizing the Seemingly Irrational: Search, Heterogeneity and Agency Problems in Credit Markets	722
11.5. A Menagerie of Modest Madness: Bounded Rationality and Housing Markets	732
11.6. Public Policy and Bubbles	743
11.7. Conclusion	747
Acknowledgment	748
References	748
12. Housing, Finance, and the Macroeconomy	753
Morris A. Davis, Stijn Van Nieuwerburgh	
12.1. Introduction	754
12.2. Stylized Facts	756
12.3. Housing and the Business Cycle	767
12.4. Housing over the Life Cycle and in the Portfolio	773
12.5. Housing and Asset Pricing	787
12.6. The Housing Boom and Bust and the Great Recession	792
12.7. Housing Policy	800
12.8. Conclusion	804
Acknowledgments	805
References	805

13. The Microstructure of Housing Markets: Search, Bargaining, and Brokerage	813
Lu Han, William C. Strange	
13.1. Introduction	815
13.2. One-Sided Search	819
13.3. Random Matching	825
13.4. Pre-search, Focused Search, and Segmented Search	835
13.5. Directed Search for Housing	839
13.6. Auctions	845
13.7. Real Estate Brokers: Fundamentals	850
13.8. Competition in the Residential Real Estate Brokerage Industry	855
13.9. Incentive Issues in Real Estate Brokerage	865
13.10. Conclusions	878
Acknowledgments	879
References	879
14. US Housing Policy	887
Edgar O. Olsen, Jeffrey E. Zabel	
14.1. Introduction	888
14.2. Methods and Data	890
14.3. US Low-Income Rental Housing Policy	892
14.4. US Homeownership Policy	938
14.5. Conclusion	977
References	978
15. How Mortgage Finance Affects the Urban Landscape	987
Sewin Chan, Andrew Haughwout, Joseph Tracy	
15.1. Mortgage Finance in the United States	989
15.2. How Mortgage Finance Affects the Market for Owner-Occupied Housing	997
15.3. The Distribution of Mortgage Credit	1005
15.4. Negative Equity	1022
15.5. Foreclosures	1034
15.6. Conclusion	1039
Acknowledgments	1040
References	1040
16. Change and Persistence in the Economic Status of Neighborhoods and Cities	1047
Stuart S. Rosenthal, Stephen L. Ross	
16.1. Introduction	1048
16.2. Neighborhood Economic Status	1054

16.3. City Dynamics	1088
16.4. Conclusions and Future Research	1106
Appendix Supplemental Figures	1108
Acknowledgments	1114
References	1114

Section IV. Applied Urban Economics 1121

17. Taxes in Cities 1123

Marius Brühlhart, Sam Bucovetsky, Kurt Schmidheiny

17.1. Introduction	1124
17.2. Institutional Background	1126
17.3. Tax Setting Across Asymmetric Jurisdictions	1145
17.4. Taxation and Urban Population Sorting	1161
17.5. Taxation and Agglomeration Economies	1171
17.6. Concluding Remarks	1178
Appendix	1179
Acknowledgments	1191
References	1191

18. Place-Based Policies 1197

David Neumark, Helen Simpson

18.1. Introduction	1198
18.2. Theoretical Basis for Place-Based Policies	1206
18.3. Evidence on Theoretical Motivations and Behavioral Hypotheses Underlying Place-Based Policies	1215
18.4. Identifying the Effects of Place-Based Policies	1221
18.5. Evidence on Impacts of Policy Interventions	1230
18.6. Unanswered Questions and Research Challenges	1279
Acknowledgments	1282
References	1282

19. Regulation and Housing Supply 1289

Joseph Gyourko, Raven Molloy

19.1. Introduction	1290
19.2. Data: Old and New	1294
19.3. Determinants of Regulation	1304
19.4. Effects of Regulation	1316
19.5. Welfare Implications of Regulation	1327
19.6. Conclusion	1330
Acknowledgments	1333
References	1333

20. Transportation Costs and the Spatial Organization of Economic Activity	1339
Stephen J. Redding, Matthew A. Turner	
20.1. Introduction	1340
20.2. Stylized Facts About Transportation	1343
20.3. Theoretical Framework	1355
20.4. Reduced-Form Econometric Framework	1366
20.5. Reduced-Form Empirical Results	1372
20.6. Discussion	1383
20.7. Conclusion	1393
Acknowledgments	1394
References	1394
 21. Cities in Developing Countries: Fueled by Rural–Urban Migration, Lacking in Tenure Security, and Short of Affordable Housing	 1399
Jan K. Brueckner, Somik V. Lall	
21.1. Introduction	1400
21.2. The Empirical Aspects of Rural–Urban Migration	1402
21.3. Models of Migration and City Sizes in Developing Countries	1409
21.4. Tenure Insecurity: A Hallmark of Housing Markets in Developing Countries	1422
21.5. Provision of Affordable Housing in Developing Countries	1439
21.6. Conclusion	1448
Appendix	1450
Acknowledgments	1451
References	1451
 22. The Geography of Development Within Countries	 1457
Klaus Desmet, J. Vernon Henderson	
22.1. Introduction	1458
22.2. Development and the Aggregate Spatial Distribution	1459
22.3. Development, Space, and Industries	1475
22.4. The Urban Sector	1482
22.5. Concluding Remarks	1512
References	1513
 23. Urban Crime	 1519
Brendan O’Flaherty, Rajiv Sethi	
23.1. Introduction	1521
23.2. Criminogenic Characteristics	1522
23.3. Incentives and Deterrence	1536
23.4. Interactions	1552

23.5. Incarceration	1567
23.6. Big Swings in Crime	1583
23.7. Where are Crimes Committed?	1604
23.8. Conclusions	1612
Acknowledgments	1613
References	1613

<i>Index</i>	<i>1623</i>

FOREWORD

The fields of *Regional and Urban Economics* have evolved remarkably since 2004 when the last volume of the Handbook series (Volume 4) was published.

The emphasis of Volume 4 was very much on agglomeration at various spatial scales (neighborhood, urban, and regional). Much of the content was theoretical, with a large proportion of theoretical chapters and a clear separation between theory and empirics. Volume 4 also arrived as Krugman's New Economic Geography had reached its peak. This emphasis on agglomeration meant that many traditional urban issues were not covered. As such, policy discussions were limited to agglomeration issues, such as regional inequalities and the effect of market integration (following worries associated with "globalization" and deeper economic integration within Europe and North America). The decade since Volume 4 has seen continued progress on agglomeration and related areas, but it has also seen a significant broadening in both the areas of study and the methods of inquiry.

This volume is in part a return to more traditional urban topics that were covered in Volumes 1–3 of the Handbook series. One example of this is housing, a research topic which has seen major advances in the last 10 years. A major housing crisis in the United States and much of the developed world is certainly part of the explanation for revival of research on housing. In addition, there are also important ongoing debates about urban sprawl and its effects and how land use regulations are shaping cities in the United States and elsewhere. Technology and sometimes legislation are also changing the way we buy and sell houses. This raises some interesting questions about the microstructure of the housing market. Thus, Volume 5 of the *Handbook of Regional and Urban Economics* has a significant emphasis on housing and property markets.

Housing is not the only new focus for urban research. There is also renewed interest in the effects of transportation on cities, neighborhood and city dynamics, urban amenities, urban environmental issues, urban crime, urban costs, land use, migration, and a range of other topics. These issues are considered in both developed and developing world settings. Volume 5 reflects this intellectual broadening as well.

Another important shift in urban and regional economics is in methods. For the first time in the *Handbook of Regional and Urban Economics* series, explicit chapters on methodology are included. The greater availability of data and the gradual adaption of "modern" methodologies have profoundly changed the nature of empirical work. These approaches (structural and quasi-experimental) are becoming more widely adopted. The chapters in this volume acknowledge this, but they also point out that a lot urban and regional research remains in need of a methodological upgrade. In addition, the chapters point to a range of unique methodological challenges arising from the spatial data that is

used in urban and regional research. The direct application of methodologies borrowed from labor economics or industrial organization is, thus, often not enough. Fortunately, both the chapters focusing primarily on methods and those that consider individual topics offer numerous suggestions of how to move forward. In most instances, this involves forging closer links between theory and empirical research.

All of these issues have significant implications for public policy. Volume 5 includes chapters focusing on policy topics that have had little coverage in previous volumes, such as mortgages, place-based policies, and urban crime. The volume also includes chapters on more traditional issues such as tax competition, neighborhood effects, and housing policy. These traditional issues are still extremely important but are now explored using more credible empirical approaches. And although these chapters are particularly oriented toward policy, the applied nature of Urban and Regional Economics means that most chapters are policy relevant at least to some degree.

Ultimately, we see the chapters included in the volume as making a strong case for research that appropriately combines theory and empirics, that embraces the many elements of urban economies, and that is policy relevant. Of course, as the volume has come together, it has become apparent that there are gaps in the volume just as there are gaps in the fields of regional and urban economics. For instance, too much of the empirical evidence on urban issues comes from American cities. While the volume does contain two chapters focused on issues in developing countries, more work on urban phenomena in developing countries is needed. As another example, while there is a chapter on transportation focused on evaluation of major inter regional transport networks, there is no coverage of traditional and evolving topics such as modal choice, peak pricing, the use of incurred transport costs to value urban amenities, and the like. We hope that these and other gaps will motivate young (and less young) researchers to expand our knowledge.

We are grateful to many people and organizations for helping to make this project happen. The contribution of the authors is obvious. These contributions were sharpened by the participants at conferences sponsored by the Wharton Real Estate Department and the Centre for Real Estate at the Rotman School of Management at the University of Toronto. Several papers were also presented at the Urban Economics Association sessions at the North American Regional Science Council meetings and at the National Meetings of the American Real Estate and Urban Economics Association. We are grateful to the people and organizations who have made these interactions possible. We also are grateful to various people at Elsevier for their helpfulness and professionalism, especially Joslyn Chaiprasert-Paguio and Scott Bentley. Finally, we are all grateful to all those who are close to us for their patience and support.

Gilles Duranton
Vernon Henderson
William Strange
November 4, 2014

CONTRIBUTORS

Nathaniel Baum-Snow

Department of Economics, Brown University, Providence, RI, USA

Kristian Behrens

Department of Economics; CIRPÉE, Université du Québec à Montréal, Montréal, QC, Canada;
National Research University, Higher School of Economics, Moscow, Russia, and CEPR,
London, UK

Marius Brühlhart

University of Lausanne, Lausanne, Switzerland, and Centre for Economic Policy Research
(CEPR), London, UK

Jan K. Brueckner

Department of Economics, University of California, Irvine, CA, USA

Sam Bucovetsky

York University, Toronto, ON, Canada

Gerald Carlino

Federal Reserve Bank of Philadelphia, Philadelphia, PA, USA

Sewin Chan

Robert F. Wagner School of Public Service, New York University, NY, USA

Pierre-Philippe Combes

Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Marseille;
Economics Department, Sciences Po, Paris, France, and Centre for Economic Policy
Research (CEPR), London, UK

Morris A. Davis

Department of Finance and Economics, Rutgers Business School, Rutgers University, Newark,
NJ, USA

Klaus Desmet

Department of Economics, Southern Methodist University, Dallas, TX, USA

Gilles Duranton

Wharton School, University of Pennsylvania, Philadelphia, PA, USA, and CEPR, London, UK

Fernando Ferreira

The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Steve Gibbons

London School of Economics, London, UK

Edward L. Glaeser

Harvard University and NBER, Cambridge, MA, USA

Laurent Gobillon

Centre for Economic Policy Research (CEPR), London, UK; Institut National d'Etudes Démographiques; Paris School of Economics, Paris, France, and The Institute for the Study of Labor (IZA), Bonn, Germany

Joseph Gyourko

The Wharton School, University of Pennsylvania, Philadelphia, PA, and NBER, Cambridge, MA, USA

Lu Han

Rotman School of Management, University of Toronto, Toronto, ON, Canada

Andrew Haughwout

Federal Reserve Bank of New York, NY, USA

J. Vernon Henderson

Department of Geography, London School of Economics, London, UK

Thomas J. Holmes

University of Minnesota and Federal Reserve Bank of Minneapolis, Minneapolis, MN, USA

Matthew E. Kahn

Department of Economics, UCLA and NBER and IZA, Los Angeles, CA, USA

William R. Kerr

Harvard University, Bank of Finland, and NBER, Boston, MA, USA

Somik V. Lall

Urban Development and Resilience Unit, Sustainable Development Network, World Bank, USA

Ethan Lewis

Dartmouth College, Hanover, NH, and NBER, Cambridge, MA, USA

Raven Molloy

Board of Governors, Federal Reserve System, Washington, DC, USA

Charles G. Nathanson

Northwestern University, Evanston, IL, USA

David Neumark

UCI, NBER, and IZA, Irvine, CA, USA

Brendan O'Flaherty

Department of Economics, Columbia University, NY, USA

Edgar O. Olsen

Department of Economics, University of Virginia, Charlottesville, VA, USA

Henry G. Overman

London School of Economics, London, UK

Eleonora Patacchini

Cornell University, Ithaca, NY, USA

Giovanni Peri

University of California-Davis, CA, and NBER, Cambridge, MA, USA

Diego Puga

CEPR, London, UK, and Centro de Estudios Monetarios y Financieros (CEMFI), Madrid, Spain

Stephen J. Redding

Economics Department and WWS, Princeton University Fisher Hall, Princeton, NJ, USA

Frédéric Robert-Nicoud

CEPR; SERC, The London School of Economics and Political Science, London, UK, and Geneva School of Economics and Management, Université de Genève, Genève, Switzerland

Stuart S. Rosenthal

Maxwell Advisory Board Professor of Economics, Department of Economics, Syracuse University, Syracuse, NY, USA

Stephen L. Ross

Department of Economics, University of Connecticut, Storrs, CT, USA

Kurt Schmidheiny

Centre for Economic Policy Research (CEPR), London, UK; University of Basel, Basel, Switzerland, and CESifo, Munich, Germany

Rajiv Sethi

Department of Economics, Barnard College, Columbia University, NY, USA, and Santa Fe Institute, Santa Fe, NM, USA

Holger Sieg

University of Pennsylvania, Philadelphia, PA, USA

Helen Simpson

University of Bristol, CMPO, OUCBT and CEPR, Bristol, UK

William C. Strange

Rotman School of Management, University of Toronto, Toronto, ON, Canada

Giorgio Topa

Federal Reserve Bank of New York and IZA, NY, USA

Joseph Tracy

Federal Reserve Bank of New York, NY, USA

Matthew A. Turner

Economics Department, Brown University, Providence, RI, USA

Stijn Van Nieuwerburgh

Department of Finance, Stern School of Business, New York University, NY, USA

Randall Walsh

Department of Economics, University of Pittsburgh and NBER, Pittsburgh, PA, USA

Jeffrey E. Zabel

Department of Economics, Tufts University, Medford, MA, USA

Yves Zenou

Stockholm University, IFN, and CEPR, Stockholm, Sweden

This page intentionally left blank

SECTION I

Empirical Methods

This page intentionally left blank

CHAPTER 1

Causal Inference in Urban and Regional Economics

Nathaniel Baum-Snow^{*}, Fernando Ferreira[†]

^{*}Department of Economics, Brown University, Providence, RI, USA

[†]The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Contents

1.1. Introduction	4
1.2. A Framework for Empirical Investigation	6
1.2.1 A binary treatment environment	9
1.2.2 A taxonomy of treatment effects	11
1.2.3 Continuous treatments	15
1.2.4 Randomization	15
1.3. Spatial Aggregation	20
1.4. Selection on Observables	23
1.4.1 Fixed effects methods	24
1.4.2 Difference in differences methods	30
1.4.3 Matching methods	37
1.5. IV Estimators	43
1.5.1 Foundations	45
1.5.2 Examples of IV in urban economics	47
1.6. Regression Discontinuity	53
1.6.1 Basic framework and interpretation	54
1.6.2 Implementation	56
1.6.3 Examples of RD in urban economics	59
1.7. Conclusion	62
References	63

Abstract

Recovery of causal relationships in data is an essential part of scholarly inquiry in the social sciences. This chapter discusses strategies that have been successfully used in urban and regional economics for recovering such causal relationships. Essential to any successful empirical inquiry is careful consideration of the sources of variation in the data that identify parameters of interest. Interpretation of such parameters should take into account the potential for their heterogeneity as a function of both observables and unobservables.

Keywords

Causal inference, Urban economics, Regional economics, Research design, Empirical methods, Treatment effects

JEL Classification Code

R1

1.1. INTRODUCTION

The field of urban and regional economics has become much more empirically oriented over recent decades. In 1990, 49% of publications in the *Journal of Urban Economics* were empirical, growing to 71% in 2010. Moreover, the set of empirical strategies that are most commonly employed has changed. While most empirical papers in 1990 only used cross-sectional regressions, articles in 2010 were more likely to use instrumental variables (IV), panel data, and nonlinear models. Furthermore, special attention is now paid to the employment of research designs that can plausibly handle standard omitted variable bias problems. While only a handful of papers attempted to deal with these problems in 1990, more than half of the empirical publications in 2010 used at least one research design that is more sophisticated than simple ordinary least squares (OLS), such as difference in differences (DD), matching, and IV, to recover causal parameters. However, the credibility of estimates generated with these more sophisticated techniques still varies. While, in general, the credibility of empirical work in urban economics has improved markedly since 1990, many studies continue to mechanically apply empirical techniques while omitting important discussions of the sources of identifying variation in the data and of which treatment effects, if any, are being recovered. [Table 1.1](#) details the percentages of publications in the *Journal of Urban Economics* that were empirical and the distribution of empirical methods used for the years 1980, 1990, 2000, and 2010.

This chapter discusses the ways that researchers have successfully implemented empirical strategies that deliver the most credible treatment effect estimates from data sets that describe urban and regional phenomena. Our treatment emphasizes the importance of randomization, which has been more broadly recognized in other fields, most notably development economics. Randomized trials are an important tool to recover treatment effects, especially those of interest for policy evaluation ([Duflo et al., 2008](#)). However, it is typically more challenging and expensive to implement field

Table 1.1 Prevalence of empirical methods in the *Journal of Urban Economics*, 1980–2010
As percentages of empirical papers

Year	Empirical	OLS	IV	Logit/ probit	Panel data	Difference in differences	Randomization	Matching
1980	57%	87%	10%	3%	0%	0%	0%	0%
1990	49%	79%	17%	13%	4%	0%	0%	0%
2000	62%	64%	32%	36%	14%	4%	0%	0%
2010	71%	77%	46%	26%	62%	8%	3%	5%

Notes: Authors calculations from all published articles in the *Journal of Urban Economics* in the indicated years.

experiments in settings of interest to urban and regional economists, as it is in other fields such as labor economics. General equilibrium effects, which contaminate control groups with influences of treatment, are more likely to arise in urban settings. Moreover, the nature of such general equilibrium effects is more likely to be the object of inquiry by urban and regional researchers. Labor economists have typically adopted higher standards for evaluating the credibility of estimated causal effects in research that uses nonexperimental data. Here we explore identification strategies that have been successfully used to recover credible estimates of treatment effects, typically in the absence of experimental variation. These include DD, various fixed effects methods, propensity score matching, IV, and regression discontinuity (RD) identification strategies. We also discuss treatment effect heterogeneity and how differences in results across identification strategies may simply reflect different causal relationships in the data. We emphasize that especially without experimental variation (and even often with experimental variation), no one identification strategy is ever perfect. Moreover, when considering causal effects of treatments, it is useful to think in the context of a world in which a distribution of treatment effects exists. Selection into treatment (on both observable and unobservable characteristics) and treatment effect heterogeneity makes empirical work complicated.

One recurring theme of this chapter is the following principle, which applies to all empirical strategies: it is crucial to consider the sources of variation in the treatment variables that are used to recover parameters of interest. Distinguishing this “identifying variation” allows the researcher to consider two central questions. First, could there be unobserved variables that both influence the outcome and are correlated with this identifying variation in the treatment variable? If such omitted variables exist, coefficients on the treatments are estimated as biased and inconsistent. We typically label such situations as those with an “endogeneity problem.” Second, how representative of the population is the subset of the data for which such identifying variation exists? If clean identification exists only in a small unrepresentative subset of the population, coefficients on treatment variables apply only narrowly and are unlikely to generalize to other populations.

Throughout the chapter, we discuss the key properties of various identification strategies mostly assuming a simple linear data-generating process which allows for heterogeneous treatment effects. Each section cites articles from the literature for readers interested in the details of more complex applications. This structure allows us to easily explain the relationships between different empirical strategies while leaving space to cover applications in urban and regional economics. In each section, we illustrate best practices when implementing the research design by discussing several recent examples from the literature.

Given the importance of the use of economic models to aid in the specification of empirical models and interpret treatment effect estimates, we view the material on structural empirical modeling in [Chapter 2](#) as complementary to the material discussed

in this chapter. [Chapter 2](#) also considers the recovery of causal relationships in urban and regional data, but through making use of model formulations that are more involved than those considered in this chapter. The advantage of the structural approach is that it allows for the recovery of parameters that could never be identified with observational or experimental data alone. Estimates of a model's "deep" parameters facilitate evaluation of more sophisticated counterfactual simulations of potential policy changes than is possible with the less specific treatment effect parameters considered in this chapter. However, structural models are by their very natures full of assumptions that are most often stronger than the assumptions needed to make use of randomization to recover treatment effects. Additionally, because models can always be misspecified, such theory-derived treatment effects may be less credible than those whose data-based identification we discuss here. When possible, we present a unified treatment of causal relationships that can be interpreted in the context of an economic model or as stand-alone parameters.

While the field of urban economics has made considerable progress recently in improving its empirical methods, we hope that this chapter promotes further advances in the credibility of our empirical results by encouraging researchers to more carefully consider which particular treatment effects are being identified and estimated. In defense of our field, it is fortunately no longer acceptable to report regression results without any justification for the econometric identification strategy employed. Nonetheless, we hope we can go beyond this admittedly low bar. This includes dissuading ourselves from simply trying several instruments and hoping for the best without careful thought about the conditions under which each instrument tried is valid or the different causal effects (or combinations thereof) that each instrument may be capturing.

This chapter proceeds as follows. [Section 1.2](#) develops an empirical framework as a basis for discussion, defines various treatment effects, and considers the importance of randomization. [Section 1.3](#) briefly considers some of the consequences of using spatially aggregated data. [Section 1.4](#) considers methods for recovering causal effects from purely observational data. [Section 1.5](#) considers various ways of handling nonrandom sorting on unobservables leading up to a discussion of IV estimators. [Section 1.6](#) describes the use of various types of RD designs. Finally, [Section 1.7](#) concludes the chapter.

1.2. A FRAMEWORK FOR EMPIRICAL INVESTIGATION

In this section, we lay out an empirical framework that we use throughout this chapter as a basis for discussion and development. Our specification of the nature of the data-generating process facilitates consideration of the fundamental problem of causal inference. In particular, we emphasize the importance of determining the sources of variation in treatment variables that identify causal relationships of interest. Making use of explicit or pseudo random sources of variation in treatment variables is essential to credible

identification of causal relationships in any data set. We also consider the implications of the potential existence of heterogeneous causal effects of treatment variables on outcomes of interest.

In general, we are interested in causal relationships between a vector of “treatment” variables T and an outcome y . A flexible data-generating process for the outcome y can be represented by the following linear equation which holds for each observation i :

$$y_i = T_i\beta_i + X_i\delta_i + U_i + e_i. \quad (1.1)$$

For now, we think of observations as individuals, households, or firms rather than geographic regions. There is a vector of “control” variables X , which are observed. The vector U incorporates all unobserved components that also influence the outcome of interest. One can think of U as $W\rho$, where W is a vector of unobserved variables, and ρ is a set of coefficients that are never identified under any circumstances. We collapse $W\rho$ into U for ease of exposition. Given the existence of U , any remaining stochasticity e in the outcome y can be thought of as classical (uncorrelated) measurement error or, equivalently for statistical purposes, as fundamental stochasticity which may come from an underlying economic model and is uncorrelated with T , X , and U . We are also not interested in recovery of the coefficients δ_i on X_i , but it is useful for expositional purposes to define these coefficients separately from the coefficients of interest β_i .

Note that we express the relationships between predictors and the outcome of interest in a very general way by allowing coefficients to be indexed by i . In order to make progress on recovering the parameters of interest β_i for each individual, some further assumptions will be required. The linearity of (1.1) may incorporate nonlinear relationships by including polynomials of treatment variables and treatment–control interactions in T and polynomials of control variables in X .

It is often useful to think of (1.1) as being the “structural” equation describing the outcome of interest y , generated from an economic model of individual or firm behavior. For some outcomes such as firms’ output or value added, this structural equation may result from a mechanical model such as a production function. More often for urban and regional questions, (1.1) can be thought of as an equilibrium condition in a theoretical model of human or firm behavior. In either type of model, we typically treat T , X , and U as “exogenous.” This means that these variables are determined outside the model and do not influence each other through the model.

While the linearity in (1.1) may come from additive separability in the equilibrium condition, typically after a log transformation, we can more generally justify linearity in the empirical representation of a static model’s equilibrium condition through implicit differentiation with respect to time. That is, if some model of individual behavior generates the equilibrium condition $y = f(T, X, U, e)$, differentiation yields an equation resembling (1.1) as an approximation, with partial derivatives of f represented by coefficients and each variable measured in first differences. That is,

$$\begin{aligned}\Delta\gamma_i \approx & \Delta T_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial T} + \Delta X_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial X} \\ & + \Delta U_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial U} + \Delta e_i \frac{\partial f(T_i, X_i, U_i, e_i)}{\partial e},\end{aligned}$$

in which Δ indicates differences over time. Note that this expression can be equivalently stated in semilog or elasticity form depending on the context. If the treatment status for every agent is the same in the base period and $\tilde{X}_i$ includes 1, ΔX_i , X_i in the base period, and various interactions, this expression thus reduces to

$$\Delta\gamma_i = \Delta T_i B(X_i, U_i) + \tilde{X}_i D(U_i) + \varepsilon_i. \quad (1.2)$$

(1.2) closely resembles (1.1), with appropriate reinterpretation of γ , T , and X , and can in principle form the basis for estimation.¹ Note that the error term ε incorporates both changes in unobservables U and changes in residual stochasticity e . Because it includes changes in unobservables, ε is likely to be correlated with ΔT . Moreover, we see that ε is likely to exhibit heteroskedasticity. As we explore further in [Section 1.4](#), this “first difference” formulation has the advantage of differencing out any elements of U that are fixed over time, but has the potential disadvantage of increasing the variance of the error term.

There are a few important practical general implications of the exercise of deriving (1.2). First, first-differencing data is valuable as it allows the researcher to linearize non-linear relationships, at least for small changes in γ , T , and X . Second, it is really useful to have information from an initial period when the treatment variable is the same for all agents. Third, all but the simplest models deliver coefficients that are heterogeneous as functions of both observables and unobservables. If the model being estimated is sure to be the true data-generating process (which it never actually is), then coefficients in the linear (1.2) may allow for recovery of estimates of some or all of the model’s parameters. Even if individual model parameters cannot be identified, $B(x, u)$ represents the causal effect of T on γ for an agent with characteristics (x, u) . Regardless of the true underlying data-generating process, this is an object which is often of inherent interest to researchers. Finally, the exact specification of the control set $\tilde{X}$ depends crucially on the underlying economic model; thus, this object can very easily be misspecified. For this reason, there are distinct advantages to using estimators that permit elements of $\tilde{X}$ to be dropped.

Our discussion of the recovery of treatment effects in this chapter primarily examines “total effects” of treatments on outcomes, or full derivatives $\frac{dy}{dT}$. Of course, the decomposition of these total effects into direct and indirect effects, in which causal links from the

¹ In some contexts, it may be appropriate to differentiate over space rather than time. We leave a more complete discussion of this issue to the [Chapter 3](#) on spatial methods by Gibbons et al. and our discussion of the RD research design in [Section 1.6](#).

treatment to the outcome operate both independently and through the treatment's influence on other predictor variables, is also interesting (Pearl, 2009). The distinction between total effects versus direct and indirect effects is a statistical restatement that the generic economic model with the equilibrium condition $y = f(T, X, U, e)$ used as a starting point above includes only exogenous variables on the right-hand side. Decomposition into direct and indirect effects of treatment is often recovered in economics applications by using some model structure, since indirect effects by definition operate through some endogenous channel. In Sections 1.4 and 1.5, we return to discussions of direct and indirect effects in the contexts of considerations of properties of particular estimators.

1.2.1 A binary treatment environment

Though urban and regional applications often involve more complicated environments, we begin by considering the case in which the treatment is binary. Analysis of this simple case is a straightforward point of departure as it is well understood in the statistics literature going back to the classic treatment of Rubin (1974), and discussed extensively in Holland (1986), and in the economics literature going back to Roy (1951). Because the recovery of causal relationships in environments with binary treatment environments is also discussed at length by DiNardo and Lee (2011), we leave the development of many details to them. Indeed, much of our mission in this chapter is to extend their discussion of various empirical identification strategies to environments in which the treatment is continuous and the data are spatially indexed. The simplicity of the binary treatment environment is important, however, as properties of the various estimators we discuss in this chapter are well known for the binary treatment case.

On the basis of the setup in (1.1), a binary treatment variable yields the following equation for each treatment level, in which treated observations receive $T = 1$ and untreated (control) observations receive $T = 0$:

$$\begin{aligned} y_i^0 &= X_i\delta_i + U_i + e_i, \\ y_i^1 &= \beta_i + X_i\delta_i + U_i + e_i. \end{aligned}$$

These two equations describe the potential outcome for each agent i if that agent were not treated and if that agent were treated, respectively. The resulting causal effect of treatment for agent i is thus β_i . When all agents in the population are considered, the result is two separate distributions of outcomes y , one for each treatment status. In evaluating the effects of the treatment, we typically aim to characterize differences between elements of these two distributions.

It should be immediately evident from this example with binary treatments that it is impossible to recover each particular β_i without further assumptions on the data-generating process, even with ideal data. This is the fundamental problem of causal inference: no agent can simultaneously be in both the treated group and the untreated group.

That is, there is no counterfactual available for individual members of any population or sample, since each agent is either treated or not treated. In the language of [Holland \(1986\)](#), there is not “unit homogeneity” if each observation has its own treatment effect. Even if we had panel data such that we could observe individuals before and after treatment, the contextual environment of “before treatment” versus “after treatment” is collinear with the treatment itself. That is, the context can be thought of as an element of X (or U if not accounted for). Each individual and time period combination would have its own observation index, and therefore its own treatment effect.²

To make progress on recovering information about causal effects of treatment, we need to limit ourselves to considering how to identify elements of the distribution of treatment effects over the population. This recognition brings up the fundamental issue that we address in this chapter: how to identify groups of agents that are similar on both observables and unobservables but who have received different levels of treatment. If the treatment effect is different for each agent, then the agents are so fundamentally different by definition that recovering any information about the distribution of β_s is a hopeless endeavor. To make progress on identification of treatment effects, we must put restrictions on the coefficients in the above equations such that they are not unique across individuals, but instead may be unique only across individuals with different observables and unobservables. One general formulation for doing so is the following:

$$\begin{aligned} y_i^0 &= X_i D(U_i) + U_i + e_i, \\ y_i^1 &= B(X_i, U_i) + X_i D(U_i) + U_i + e_i. \end{aligned}$$

Because the distribution of treatment effects captured in the $B(\cdot)$ function depends on the characteristics of the treated agent only and not on the identity of each agent itself, we can imagine finding another agent with the same observable and unobservable characteristics with whom the treated agent can be compared. In practice, since we do not by definition know the unobservable characteristics of any agent, we do not have any way to recover the “marginal” treatment effect (MTE) for any particular unobserved type U without the imposition of an economic model, as in [Heckman and Vytlačil \(2005\)](#). Instead, depending on how the treatment is assigned, we are potentially able to recover various model-agnostic statistics about the distribution of $B(X, U)$ over the population. Note that we restrict the coefficients on observables X to be functions only of U . To account for potential nonlinear impacts of X (that interact with U), one can define X to include polynomial terms and interactions.

² In a few cases, researchers have assumed that unobservables do not differ over time and have attempted to estimate individual treatment effects by causing individual fixed effects to interact with a treatment variable. The work of [De La Roca and Puga \(2014\)](#) is an example in the context of estimating causal effects of city sizes in labor market histories on individuals’ wage profiles. [Section 1.3](#) discusses in detail the assumptions needed for fixed effects identification strategies like this to deliver credible estimates of causal effects.

1.2.2 A taxonomy of treatment effects

Before returning to an empirical model with continuous treatments, it is useful to consider the various treatment effects that may be of interest in the context of the binary treatment environment. These treatment effect definitions generalize with minor modifications to the continuous treatment case, as explained below. In the following sections, we carefully consider which treatment effects can be identified with each of the estimators that we consider.

One way of conceptualizing the binary treatment environment is as the existence of two counterfactual distributions in the population y^0 and y^1 which differ only because of treatment status. The restrictions on the empirical model formulated above force the difference between these two distributions for agents of a given type (x, u) to be $B(x, u)$.

The most closely related causal effect is the MTE. As in [Heckman and Vytlačil \(2005\)](#), we define $\text{MTE}(x, u)$ as the causal effect of treating an individual with characteristics $X = x$ and $U = u$:

$$\text{MTE}(x, u) \equiv E[y^1 - y^0 | X = x, U = u] = B(x, u).$$

While the MTE is a useful construct, it is only possible to recover any particular MTE within the context of a specified economic model. This is because the MTE is indexed by unobservable U , which is an object that the researcher can never know directly, but can only assign to individuals through the structure of a model. [Heckman and Vytlačil \(2005\)](#) consider a simple generalized Roy-type sorting model ([Roy, 1951](#)) on the basis of which they identify the full distribution of MTEs. All other treatment effects can be viewed as weighted averages of various combinations of MTEs.

Unconditional quantile treatment effects (QTEs) of [Abadie et al. \(2002\)](#) provide information about the distribution of treatment effects, as indexed by the realization of outcome variables. The QTE for quantile τ is the difference in the τ th quantile of the y^1 and y^0 distributions, which in this case is the τ th quantile of the distribution $f(B(X, U))$. QTEs are informative about whether the treatment differentially influences different parts of the distribution of the outcome of interest. [Athey and Imbens \(2006\)](#) show how to estimate the full counterfactual distributions y^1 and y^0 without any functional form assumptions assuming treatment randomization, thereby allowing for calculation of all QTEs. The difficulty with QTEs is that their recovery typically requires randomization to apply very broadly to the distribution of potential outcomes, which rarely occurs. QTEs do not provide information about the unobserved characteristics of agents to whom they apply, though one can similarly define QTEs over the conditional distributions of unobservables only $f_x(B(x, U))$ given $X = x$.

Perhaps the commonest treatment effect of interest is the average treatment effect (ATE). The ATE describes the mean treatment effect averaged over all members of the population with a particular set of observed characteristics x and is represented as follows:

$$\text{ATE}(x) \equiv E(y^1 - y^0 | X = x) = \int B(x, U) dF(U | X = x).$$

Often, rather than being interested in the ATE for a particular subpopulation, researchers may be interested in the ATE for the full population:

$$\text{ATE} \equiv E(y^1 - y^0) = \int B(X, U) dF(X, U).$$

As with QTEs, it is important to recognize that the ATE is not easily recovered in most empirical contexts without strong model assumptions. The reason is that in the absence of widespread randomization, there are some groups which either always receive the treatment or never receive the treatment. Since calculation of the ATE requires knowing the MTE for the full joint distribution of (X, U) , the portions of the support of $f(X, U)$ which are in only the treated state or the untreated state must have their MTE distributions inferred by model assumption. Depending on the approach, the model used to recover these MTE distributions may be statistical or economic.

The local average treatment effect (LATE), first defined by [Imbens and Angrist \(1994\)](#) and also discussed by [Bjorklund and Moffitt \(1987\)](#), is the average effect of treating the subset of the joint distribution of X and U that has been induced into (or out of) treatment through explicit or pseudo randomization. Suppose that an “instrument” Z allows the researcher to manipulate the probability that agents end up in the treatment group or the control group. Imagine manipulating Z from values z to z' , where $\Pr(D = 1 | Z = z) > \Pr(D = 1 | Z = z')$ for all combinations of X and U .³ The resulting LATE is defined as

$$\text{LATE}(z, z') \equiv \frac{E[y | Z = z] - E[y | Z = z']}{\Pr(D = 1 | Z = z) - \Pr(D = 1 | Z = z')}. \quad (1.3)$$

That is, the LATE captures the change among those newly treated in the mean of y for a change in the fraction treated. This definition can be interpreted as a simple weighted average of all MTEs:

$$\text{LATE}(z, z') = \frac{\int B(X, U) [\Pr(D = 1 | Z = z, X, U) - \Pr(D = 1 | Z = z', X, U)] dF(X, U)}{\Pr(D = 1 | Z = z) - \Pr(D = 1 | Z = z')}$$

Here we see that the weights depend on the relative probability of being induced into the treatment group rather than the control group by the change in the instrument Z . In principle, this manipulation of the instrument could cause some increase in the

³ It is also possible to define the LATE for cases in which the variation in Z induces movement into treatment for some types and out of treatment for other types. However, to the extent that such bidirectional flows are unobserved, the resulting object is very difficult to interpret as it conflates positive treatment effects for some agents with negative treatment effects for others.

probability of treatment for all observed and unobserved types. Heckman and Vytlačil (2005) consider LATE's interpretation in the context of a structural model in which each value of U explicitly determines the choice into or out of treatment. That is, the range of U for which there is identification is the range over which the manipulation of the instrument Z induces membership in the treated group that would not otherwise have occurred.

Unlike the MTE, QTE, and ATE, the LATE is defined on the basis of the empirical context because the empirical context determines (z, z') . The LATE is an important concept because it is often the only treatment effect that can be identified when there exists randomization over only some subset of the support of the joint distribution of X and U .⁴

The intention to treat (ITT) is the average effect of offering the treatment. This is a policy-relevant treatment effect for many program evaluations since many of those offered the opportunity to participate in government programs do not accept it. Suppose that agents in the group offered treatment have $Z = 1$ and those in the group not offered treatment (the "control" group) have $Z = 0$. Those who would accept the offer of treatment if available have $D = 1$ and others have $D = 0$. We assume that those in the control group cannot under any circumstances procure the treatment. That is, if $Z = 0$, D necessarily equals 0. However, those in the treatment group may refuse treatment, such that $Z = 1$ and $D = 0$ for some agents. Given this environment and assuming that membership in the group offered treatment is randomized, we have

$$\begin{aligned} \text{ITT} &\equiv E(y|Z=1) - E(y|Z=0) \\ &= E(y^1|Z=1, D=1) \Pr(D=1|Z=1) - E(y^0|Z=0, D=1) \Pr(D=1|Z=0) \\ &= E(y^1 - y^0|D=1) \Pr(D=1) \\ &= \int B(X, U) \Pr(D=1|X, U) dF(X, U). \end{aligned}$$

This simple expression for ITT assumes that because of treatment randomization, $E(y^0|Z=1, D=0) = E(y^0|Z=0, D=0)$. Like other treatment effects considered above, ITT can be conditioned on X .

The treatment on the treated (TT) is the average effect of the treatment for those who would choose to accept an offer for treatment. This can be expressed as

$$\begin{aligned} \text{TT} &\equiv E(y^1 - y^0|D=1) \\ &= \frac{\int B(X, U) \Pr(D=1|X, U) dF(X, U)}{\int \Pr(D=1|X, U) dF(X, U)}. \end{aligned}$$

Notice that TT is typically greater in magnitude than ITT, because it is defined only for those with $D = 1$. In the above expression TT is written as the MTE weighted by the probability of treatment for each combination of X and U , with high values of U

⁴ LATE can also be conditioned on values of X provided that there is some variation in Z for $X = x$.

presumably being more likely to select agents into treatment, normalized by the mass of the portion of the distribution $f(X, U)$ that selects agents into treatment. The closely related treatment on the untreated is the average effect of the treatment for those who choose not to accept the treatment offer. Notice that if every agent were to accept the offer of treatment, $ITT = TT = ATE$.

To be more concrete about the differences between these various treatment effects, we compare them in the context of the Moving to Opportunity (MTO) experiment, which randomized Section 8 housing vouchers to two treatment groups of public housing residents in five cities in the mid 1990s. Data on a control group that was not offered vouchers were also collected. Households in the “Section 8” treatment group received only a housing voucher, which subsidized rent in any apartment whose landlord would accept the voucher. The “experimental” treatment group was additionally provided with counseling and was required to move to a neighborhood with a poverty rate below 10% for at least 1 year. Baseline information about households in the treatment and control groups was collected prior to randomization and in various posttreatment periods. Let us consider labor market earnings as an example outcome for the Section 8 treatment group.

Each household in the population of public housing residents has some particular observed and unobserved characteristics (x, u) . $MTE(x, u)$ is the causal effect on earnings of moving a household with characteristics (x, u) out of public housing into a Section 8 apartment of its choice. Because the MTE is conceptualized such that a different value of U is assigned to each household with a different treatment effect, there is only one possible MTE per (x, u) combination. The QTE for quantile τ is the comparison of earnings quantile τ in the treatment group relative to the control group in an environment in which all treated households comply with the treatment. $ATE(x)$ is the average difference in earnings for the treatment group versus the control group for those households with characteristics x assuming all treated households comply. ITT is the average difference in earnings between treatment and control groups, whether or not those in the treatment group accepted the voucher. TT is the average difference in earnings between those in the treatment group that accepted the offer of the voucher and those in the control group who would have accepted the voucher if it had been offered. In the binary treatment context, LATE is identical to TT , since the housing voucher offer manipulates the probability of leaving public housing for a Section 8 subsidized apartment. As we discuss further in [Section 1.5](#), LATE terminology is most commonly invoked when IV estimation is used to recover causal links from a continuous treatment to an outcome. For example, since the offer of the housing voucher caused treated households to move to lower-poverty neighborhoods at a higher rate than control households, one can conceptualize the LATE of neighborhood poverty on household earnings. This LATE applies only to the types of households induced by the treatment to move to lower-poverty neighborhoods.

1.2.3 Continuous treatments

With continuous treatments, instead of imagining two counterfactual states for each agent in the population, y_i^0 and y_i^1 , we imagine a continuum of counterfactual states, which we denote y_i^T . To be consistent with the literature and allow parameters of the data-generating process to be tractably estimated using standard techniques, we restrict our attention to the following linear model which puts only a few additional restrictions on (1.1):

$$y_i = T_i B(X_i, U_i) + X_i D(U_i) + U_i + e_i. \quad (1.4)$$

While it is commonly implemented as a linear equation, there is no need to interpret (1.4) as strictly linear since T could be formulated as a vector of treatments which are a polynomial in one continuous treatment variable, just as X can incorporate higher-order terms. Note that we typically do not consider the possibility that $B(X_i, U_i)$ and $D(U_i)$ can be functions of the treatments themselves.

Each of the treatment effects discussed above applies to the continuous case as well with only slight modification (Heckman et al., 2006). In general, treatment effects for a continuous treatment must also be indexed by the specific values of the treatment variables to which they refer. For example, the prior subsection defines the ATE for moving from treatment value 0 to treatment value 1, which could be written as $ATE_{0,1}(x)$. Because of the linearity assumption in (1.4), (or that $B(\cdot)$ is not itself a function of T), any treatment effects in the continuous case are identical regardless of which unit iteration of the treatment variable is considered. That is, $ATE_{0,1}(x) = ATE_{q,q+1}(x)$ for all q . Therefore, each of the treatment effects defined above maintains its definition in the continuous case with minimal adjustment for any arbitrary unit iteration in T , understanding, of course, that this comes by assumption and may not hold beyond the support of T .

It is important to emphasize that while we sometimes consider the case $B(X_i, U_i) = \beta$, most empirical research should recognize the possibility that there exists some “essential” heterogeneity across agents in the causal effects of treatment variables of interest. If that is the case, the assumption of a homogeneous treatment effect can lead to invalid interpretations of estimation results. In the course of this chapter, we lay out which elements of the distribution of β can be recovered with various estimators commonly applicable to recovering causal relationships of interest to urban and regional economists.

1.2.4 Randomization

One difficulty that comes out of this section’s motivation for using an economic model of behavior as a starting point for empirical investigation is that as researchers we can never be sure what the “correct” empirical specification is for an estimating equation because we never know the true data-generating process for y . Even if we did know what variables belong in X and W , it is often the case that different particular economic models

have the same such exogenous variables as inputs into the data-generating process. Structural parameters are informative only in the context of the structural model within which they are defined. Therefore, rather than concerning ourselves with recovering structural parameters, we often find it fruitful to concentrate empirical work on recovery of particular treatment effects, which then may also have interpretations in the context of specific structural models. The main challenge in doing so is that there are almost always unobservables that influence y yet may be correlated with the treatment variables of interest. This is the classic econometric identification problem.

One path toward a solution to this identification problem is to recognize that if there is randomization in treatment variables T , it is unnecessary to observe X and U to recover some information about $B(X, U)$. The role of randomization is that it assigns different values of T to agents with the same X and U . That is, it creates comparable treated and untreated populations. Of course, the reason that we need randomization to achieve this, rather than simply some assignment rule based on observables, is that U is unobserved. By its very nature, pure randomization of T over the population balances the joint distribution of X and U for all treatment levels.

With pure randomization of T over the population and a data-generating process described by (1.4), it is straightforward to see that the OLS estimate of β in a simple regression of y on T yields the ATE. In particular,

$$p\lim(\hat{\beta}_{OLS}) = E[B(X, U)] = ATE,$$

which is simply a difference in means between treatment and control groups. Intuitively, this result comes about because randomization ensures that the full distribution of individuals in the population receives each level of treatment. One may wish to control for X in this regression in order to reduce the variance of the error term, and as a result, the standard error of $\hat{\beta}_{OLS}$. By extension, it is also straightforward to estimate a series of specific ATEs $ATE(x)$ by regressing y on T interacting with dummy variables capturing various portions of the support of X . For example, if a researcher is interested in knowing the ATE among those with observable attributes in sets A and B , which partition the full support of X , the researcher could estimate the following regression equation by OLS:

$$y = T1(X \in A)\beta_A + T1(X \in B)\beta_B + X\delta + \varepsilon.$$

In this equation, $1(\cdot)$ is the indicator function. The result is that $p\lim(\hat{\beta}_{AOLS}) = E[B(X, U)|X \in A]$. That is, $\hat{\beta}_A$ as estimated by OLS captures the ATE for the portion of the X distribution in set A . It is important to recognize here that the distributions of unobservables in sets A and B may be quite different. There is no way to know whether the reason that OLS estimates of β_A and β_B may be different is because set A contains individuals with a distribution of observables (on which they have been partitioned) or unobservables correlated with these observables different from those in set B . One can extend this procedure to estimate a broader set of ATEs.

Recovery of treatment effects with simple OLS regression typically requires explicit treatment randomization. However, implementation of randomized controlled trials (RCTs) can be quite challenging and expensive. [Duflo et al. \(2008\)](#) provide a practical guide and toolkit for researchers wishing to introduce randomization as a part of the research design in a field experiment.⁵ A general issue with all experiments is that it is rarely possible or practical to randomize a treatment over the full population. Small sample sizes often make inference about treatment effects which apply to subpopulations difficult. For this reason, estimation of treatment effect heterogeneity is often limited to simple interactions of T and X in a regression model.⁶

Individual participation in randomized trials is rarely mandatory. This means that those participating in an experiment may differ on unobservables from other populations of interest. Randomization of treatment thus often occurs over only a subset of the population of interest. For example, in the MTO experiment, housing vouchers were offered only to those who had the motivation and initiative to show up to an initial meeting at which the program was described. While it is possible to see whether these MTO subjects differ on some observables from remaining public housing residents, they may differ more markedly on unobserved attributes that also influence well-being measures and labor market outcomes of interest. That is, because the sample over which the treatment is randomized is almost always self-selected on some unobservables, any results necessarily only apply to this self-selected group. As a result, there is likely to be some portion of the support of the distribution of U for which treatment effects cannot be recovered without extrapolation. Equally important is that it is common for many agents offered treatment not to accept it. That is, even though the treatment and control groups have the same distribution of unobservables, those who do and those who do not actually get treated do not. In these contexts, it is typically infeasible to recover the full distribution of treatment effects, and researchers focus on estimating ITT and TT.

[Ludwig et al. \(2013\)](#) summarize estimated treatment effects of MTO using data from 10–15 years after program implementation. They find that the program had no detectable effect on economic outcomes, youth schooling, or physical health. However, they do find some positive effects on mental health and measures of subjective well-being. This evidence follows up the study of [Kling et al. \(2007\)](#), which reports positive effects of MTO on behavioral outcomes for girls but negative effects for boys 5–8 years after implementation. [Galiani et al. \(2012\)](#) leverage the MTO randomization to estimate a structural model of neighborhood choice. They use their estimates to recover counterfactual

⁵ Most RCTs conducted by American researchers can be found at the AEA RCT Registry website. Even though this is a voluntary registry, the AEA encourages the registration of all new RCTs.

⁶ When researchers are interested in recovering treatment effects for certain subpopulations, these groups are typically oversampled relative to their share of the full population. When using data for these groups to recover other treatment effects or parameters, one should apply sampling weights to ensure that these oversampled groups do not contribute disproportionately to the estimates.

distributions of poverty rates in neighborhoods chosen by voucher recipients given alternative voucher assignment policies that were never actually implemented. They find that take-up of the voucher offer is severely reduced by restricting destination neighborhoods to the point of being counterproductive if such restrictions limit destination choice too much. This is a good example of a study that uses clean identification to recover parameters of a structural model, and ultimately a broader set of treatment effects than could be recovered using atheoretical methods alone.

There are many potential concerns about extrapolating the causal impacts of the MTO experiment from program effects to neighborhood effects. Indeed, the neighborhood improvements caused by housing voucher randomization are conflated with the disruption of moving, changes in neighborhood quality may not have been sufficiently large to generate statistically measurable effects, voucher recipients select particular destination neighborhoods of their choice, and MTO results may not generalize to other settings. Moreover, the MTO experiment reveals little about the effects of moving the approximately 50% of households who chose not to leave public housing despite receiving the offer of a housing voucher. Despite those caveats, the MTO experiment has produced among the most convincing estimates of the impacts of changes in neighborhood quality on individual outcomes. In particular, these results have weakened the “spatial mismatch hypothesis” view that low neighborhood quality and poor job access promote high rates of unemployment in poor neighborhoods (Kain, 1992).

Explicit treatment randomization has also generated data that are informative about the internal and external effects of improved housing conditions. Galiani et al. (2013) examine effects of the randomized provision of prefabricated homes for slum dwellers in El Salvador, Mexico, and Uruguay. They find that beneficiaries exhibited no improvement in labor market outcomes but improved general well-being and housing conditions relative to a control group. Freedman (2014) finds that tax credits for home improvements that were allocated to applicants by lottery in St Louis, Missouri slightly increase the value of neighboring homes.

As with treatment effect estimation in most settings, one important general consideration about using data with treatments allocated by lottery is the potential existence of general equilibrium effects. Interpretation of average differences in outcomes between treatment and control groups as treatment effects requires that the stable unit treatment value assumption (SUTVA) (Cox, 1958) of no direct or indirect influence of the treatment of one observation on outcomes of control observations must hold. For example, if in the MTO environment some control group households were to hear about neighborhood relocation options from experimental group households and act on this information, the SUTVA would be violated. To avoid this problem, many RCTs in development economics randomize treatment at the village level rather than the household level. However, since many questions of interest to urban and regional economists are fundamentally about the operation of cities rather than villages, this strategy may be of limited use in our field.

Nonetheless, RCTs for answering urban and regional questions will likely become commoner as evaluating the impacts of urban policy interventions becomes more important in developing countries, where urbanization is rapidly occurring.

One additional setting in which explicit randomization has been used to learn about causal effects is in analysis of peer effects. Without randomization, it is very difficult to get around the problem that people very likely sort into peer groups, including classes in school and friendship networks, on correlated unobservables. [Sacerdote \(2001\)](#) uses the random assignment of freshman roommates at Dartmouth College to recover estimates of peer effects in college performance. [Bayer et al. \(2009\)](#) use the random allocation of juvenile prisoners to cells to recover information about peer effects in recidivism. However, using data collected about experimentally manipulated peer groups among freshmen at the Air Force Academy, [Carrell et al. \(2013\)](#) find negative peer effects on the lowest-ability group members, perhaps partly because of endogenous subgroup formation which separated them from their highest-ability peers. The randomization of students into classrooms in the first year of the Project Star program in Tennessee has also been used to recover estimates of peer effects; see [Graham \(2008\)](#), for example.

Much of the remainder of this chapter considers strategies for recovering treatment effects for settings in which explicit treatment randomization is not available. [Section 1.4](#) essentially considers various strategies for indirectly controlling for unobservables U . [Section 1.5](#) considers strategies for identifying and effectively making use of pseudorandom variation in treatments. [Section 1.6](#) considers how best to make use of discontinuities in treatment intensity. As a general principle, we reiterate that whatever the empirical strategy used, it is critical for the researcher to understand the source of variation that is providing identification for parameters of interest. Thinking through such identification arguments often reveals the existence of potential endogeneity problems in which the treatment variable may be correlated with elements in W and/or the extent to which the treatment effects being estimated apply only to certain narrow subpopulations.

While perhaps not ideal, there are many contexts in which neither randomization nor credible strategies for controlling for unobservables are available to recover treatment effects of interest. The main alternative viable strategy is to explicitly model the heterogeneity and sorting equilibrium and recover treatment effects through model simulation. Holmes and Sieg discuss such structural options at length in [Chapter 2](#). It should be emphasized that making use of model structure requires much stronger assumptions than are needed for a randomized treatment to yield credible treatment effects. Moreover, because no model completely describes the data-generating process, the credibility of model-derived results still requires careful consideration of the sources of variation in the data that are identifying estimates, and whether these sources of variation are random (unlikely), or at least plausibly uncorrelated with mechanisms that could be important but are not explicitly modeled.

1.3. SPATIAL AGGREGATION

Before delving into the specifics of various identification strategies and econometric estimators, we briefly explore the implications of having a data structure that is spatially aggregated above the individual, household, or firm level. Such a data structure may be imposed by a data provider, be chosen by the researcher because the treatment is administered to regions rather than individual agents, or be chosen by the researcher in order to strengthen the empirical strategy. When imposed by the researcher, spatial aggregation of data is often carried out to alleviate concerns about SUTVA violations, in which spillovers occur between spatially proximate geographic units with different levels of treatment. Researchers often aggregate data to the local labor market or metropolitan area level in order to avoid this potential problem.

Suppose that the treatment and outcomes are observed at some level of spatial aggregation such as census tracts or zip codes, indexed by j . In the case of a binary treatment that is applied to the same fraction of the measure of each (x, u) in each location, a strong assumption, the equation of the data-generating process becomes

$$\tilde{y}_j = S_j \tilde{B}(X_j, U_j) + \frac{1}{N_j} \sum_{i(j)} X_i D(U_i) + \tilde{U}_j + \tilde{e}_j.$$

In this equation, tildes ($\sim$) indicate sample means over all observations in j . N_j is the total number of observations in j , S_j is the fraction of observations in region j that were treated, and $\tilde{B}(X_j, U_j) = \int B(X, U) dF_j(X, U)$, where $F_j(X, U)$ is the joint cumulative distribution function of X and U in unit j . Notice that because of the heterogeneous coefficients $D(U_i)$, $\frac{1}{N_j} \sum_{i(j)} X_i D(U_i)$ cannot in general be simplified into some simple function of

means $\tilde{X}_j$. Therefore, controlling for mean values of each element of X does not appropriately control for observables about individual agents unless $D(U_i) = \delta$. Instead, the full distribution of X within each j shows up in the aggregate equation. Therefore, in this sort of aggregation environment it makes sense to control not just for the mean but also for the full distribution of each observable characteristic if possible. Therefore, if regional means of X are all that is observed about control variables, we can think of other elements of the within- j distributions of X as being part of $\tilde{U}_j$.⁷

In the case of a more general continuous set of treatments and heterogeneous treatment effects, aggregation gives rise to the nonseparable treatment terms $\frac{1}{N_j} \sum_{i(j)} T_i B(X_i, U_i)$ replacing $S_j \tilde{B}(X_j, U_j)$ above. Estimation of statistics about $B(X, U)$ is

⁷ If the goal is to recover the treatment effect averaged across individuals (rather than regions j), one should weight any estimation by N_j . Doing so allows the more populous regions to influence the estimates more than the regions that have few agents. If, however, the goal is to recover the treatment effect averaged across regions, one should not weight such an estimation.

thus quite difficult without further assumptions about the underlying data-generating process. One common simplifying assumption is that of perfect sorting across regions. This assumption can be justified to an approximation as the equilibrium in a [Tiebout \(1956\)](#) sorting model like that specified by [Epple and Platt \(1998\)](#). With this structural assumption, which applies more accurately to finer levels of spatial aggregation, we have a resulting data-generating process given by

$$\tilde{y}_j = T_j B(X_j, U_j) + X_j D(U_j) + U_j + \tilde{u}_j.$$

Because of homogeneity within each region j in X and U , we need only index these elements by j to represent any and all quantiles of their distributions in j . Without this sort of homogeneity assumption, it becomes clear that while perhaps some progress can be made with spatially aggregate data in recovering information about $B(X, U)$, making use of micro data or the structure of a sorting model would be preferable for recovering treatment effects, even in a context with explicit treatment randomization.

Rather than having an underlying data-generating process described by (1.4), in some contexts we determine the treatment itself at the local area level. For example, the federal Empowerment Zone (EZ) program treated certain census tracts with various forms of government subsidies, and the Clean Air Act treated certain counties with pollution reductions. Often with these sorts of policies, we are interested in the effects on local residents or firms. At the local area (e.g., census tract) level, the data-generating process is thus

$$\tilde{y}_j = T_j \tilde{B}(X_j, U_j) + \frac{1}{N_j} \sum_{i(j)} X_i D(U_i) + \tilde{U}_j + \tilde{u}_j. \quad (1.5)$$

As above, in this equation, $\tilde{B}(X_j, U_j)$ denotes the average effect of the treatment in each region j given the distribution of X and U in unit j . In this case we do not need assumptions about homogeneity of populations in local areas or homogeneity of treatment effects to make some progress in recovering information about $B(X_j, U_j)$. In particular, given global randomization in T_j and no changes in location that is related to receiving the treatment, an OLS regression of mean outcomes on the treatment dummy weighted by the population of each region j yields a coefficient on the treatment with a probability limit of the ATE, by the law of iterated expectations.

One key assumption here is that the composition of the population of each region j does not respond to the treatment. This assumption is a strong one. If the treatment changes the amenity value of certain locations, we may expect certain types of people to move out of untreated locations into treated locations, thereby changing the joint distribution of the population in each location $f_j(X, U)$ and breaking the orthogonality between T and $\tilde{U}$ needed to identify $E[\tilde{B}(X_j, U_j)]$, even with initial treatment randomization across space. While one can look in the data for such resorting on observables X , including such intermediate outcomes as controls may bias treatment effect estimates since such intermediate outcomes are now endogenous. [Cellini et al. \(2010\)](#) provide

an alternative strategy to deal with such situations in the context of a dynamic model. Once again, making use of an economic model of behavior that takes sorting into account would aid econometric identification.

The final aggregation structure that we consider here is one in which each metropolitan area or other large spatial aggregation is an observation, potentially at different points in time. The sorts of questions that lend themselves to be answered with such highly aggregated data are those for which the full data-generating process must be described at the local labor market level and subsumes a set of complicated micro level interactions. One can conceptualize this by aggregating (1.4) to the local labor market level while recognizing that (1.4) incorporates the simultaneous existence of heterogeneous treatment effects, heterogeneous treatments across agents within each local labor market, and spatial lags. For example, measuring the size of agglomeration within local labor markets (Glaeser et al., 1992; Henderson et al., 1995) and measuring the effects of highways on urban decentralization (Baum-Snow, 2007) or urban growth (Duranton and Turner, 2012) lend themselves to be considered using aggregate data structures. Sorting difficulties or other general equilibrium effects that would make econometric identification difficult when examining micro data are aggregated away in these examples. For these types of applications, we typically think of the treatment as occurring at the metropolitan area level because even those metropolitan area subregions that were not explicitly treated are indirectly influenced by the treatment through general equilibrium effects. For this sort of empirical strategy to be successful, it is essential that the data be at a sufficient level of spatial aggregation that there are minimal links across observations. If the data are not sufficiently aggregated, the endogeneity problem caused by spillovers across spatial units of observation may be very difficult to handle.

The following equation captures the data-generating process for some local labor market aggregate statistic y such as population or GDP:

$$y_k = T_k B(X_k, U_k) + X_k D(U_k) + U_k + u_k. \quad (1.6)$$

In this equation, k indexes local labor markets or other highly aggregated spatial units such as states, which are spatial aggregates of j . Depending on the context, the coefficients may be heterogeneous as a function of the distribution of household or firm characteristics in k or other summary attributes of k , either of which we denote as the couple (X_k, U_k) . If the treatment effect of interest concerns effects on individuals, this equation is analogous to (1.5), and one thus may need to consider any potential resorting of the population across k in response to the treatment. If instead the goal is to recover treatment effects on metropolitan area aggregate measures, this equation is perfectly analogous to (1.4), and exhibits all of the same challenges with respect to econometric identification and the interpretation of estimates, though the mechanisms may be subtle owing to sorting. One difference from more micro analyses which in practice is often important is that typically the number of observations is quite small. For example, historical data on

metropolitan areas in the United States sometimes include information for only 100 regions nationwide. With such a limited number of observations, statistical power becomes weak very quickly if treatment variables are defined too nonparametrically. Therefore, little statistical power may be available to recover a lot of information about the $B(\cdot)$ function in (1.6).

One word of general caution about estimation of empirical models with spatially indexed data is that standard errors are likely to be understated without implementation of an appropriate correction. This is because common elements of unobservables U in nearby observations manifest themselves as correlated errors. Spatially and/or temporally correlated unobservables W (or, equivalently, unexplained components of y) is why such spatially correlated errors ensue. Bertrand et al. (2004) discuss block bootstrap (Efron and Tibishirani, 1994) and clustering (Moulton 1990, 1986) methods to account for these problems in environments in which there is a fixed number of observations per cluster and the number of clusters increases toward infinity. Cameron et al. (2008) compare various procedures for calculating standard errors with a small number of clusters using Monte Carlo simulation. Their results indicate that the “clustered wild bootstrap- t ” procedure generates the most accurate test statistics when clusters are independent and the number of clusters is small. Bester et al. (2011) discuss estimation of heteroskedasticity-autocorrelation consistent standard errors and generalized cluster procedures for conducting inference with spatially correlated errors when clusters are not independent and the number of clusters is fixed but the number of observations within each cluster goes to infinity.

Now that we have specified the possibilities for the types of data-generating processes that show up most often in urban and regional empirical applications, we consider various empirical strategies for recovering treatment effects.

1.4. SELECTION ON OBSERVABLES

While having a source of explicit or pseudo randomization is typically the preferred way to recover the most credible causal relationships in data, there are many important questions that do not lend themselves easily to this sort of empirical strategy. As such, in this section we consider options for recovering causal parameters of interest in the absence of such randomization. It should be clear that estimating (1.4) by simple OLS recovers only the ATE, $E[B(X, U)]$, in the unlikely event that T is uncorrelated with U , or that T is fully randomized. This section thus explores alternatives to simple OLS that do not involve explicit or implicit randomization, and therefore may not account for the influence of unobserved variables in the economic relationship of interest. These other methods are fixed effects, DD, and matching estimators. We emphasize that these methods can sometimes most successfully be used in tandem with each other and/or with other empirical strategies discussed elsewhere in this chapter. Key decisions in implementing nonexperimental estimators in

many contexts are the choices of treatment and particularly control groups. The primary goal in choosing a control group is to choose a set of observations for which the distribution of unobservables is likely to be similar to that in the treatment group. Below we present some formal options for doing this by examining the distribution of observables, though it is standard to assign all untreated observations to the control group in a robustness check while explicitly accounting for differences in observables. For example, the final subsection discusses estimators that reweight observations in the control group to match its distribution of observables with that in the treatment group.

We emphasize that it is almost as much an art as a science to determine the most convincing identification strategy. This determination depends crucially on the setting and the structure of the available data. For example, if the available data include an individual level panel, fixed effects methods are feasible. If the data are structured as two repeated cross sections, DD may be most feasible. Even within the identification strategies that we explore, the details of implementation require many decisions. As such, we hope this section provides a general guide to the available options, along with their advantages and pitfalls and examples of their use in published research, rather than specific recipes for carrying out empirical work.

1.4.1 Fixed effects methods

Fixed effects and panel methods can be used when there are multiple observations per agent or spatial unit. Inclusion of fixed effects in a regression is intended to remove all unobserved characteristics that are fixed over time, or across space if multiple agents are observed in the same spatial unit, from the error term. This means that any components of unobservables that are fixed over time are controlled for through inclusion of fixed effects. DD, whose discussion we delay to the following subsection, is a particular identification strategy which typically incorporates fixed effects. We consider the use of panel data on individuals or firms, homes, and spatial units at various levels of aggregation, respectively.

A generic fixed effects regression specification, for individuals i at times t , is as follows:

$$y_{it} = T_{it}\beta + X_{it}\delta + \alpha_i + \varepsilon_{it}. \quad (1.7)$$

In the absence of the fixed effects α_i , β is identified by comparing outcomes at different levels of T both between and within agents i . Inclusion of fixed effects is equivalent to differencing y , T , and X relative to sample means within each i . Therefore, β in a fixed effects regression such as (1.7) is identified by comparing changes in y for different changes in T (or first derivatives) within agents. Variation in T between agents is not used to recover information about β . With more than two time periods, one can also estimate (1.7) on first-differenced data, which identifies β by comparing DD (or second derivatives) within agents.

Because β in the above regression is identified from variation in T over time within agents, those agents with more variation in T influence the estimate of β more.

Therefore, if treatment effects are heterogeneous at β_i across agents, $\hat{\beta}_{FE}$ does not capture the ATE, but rather captures some combination of individual treatment effects weighted by each individual's contribution to econometric identification. Indeed, [Gibbons et al. \(2013\)](#) derive that the fixed effects estimator for β is

$$\hat{\beta}_{FE} = \sum_{i=1}^I \left(\frac{N_i}{N} \hat{\beta}_i \frac{\widehat{\text{Var}}(\tilde{T}_i)}{\widehat{\text{Var}}(\tilde{T})} \right).$$

In this expression, $\tilde{T}$ is the residual after projecting T onto other covariates, including fixed effects. $\text{Var}(\tilde{T}_i)$ is the variance of this object within i , while $\text{Var}(\tilde{T})$ is its variance overall in the data. Commensurate with the intuition given above, this coefficient is a particularly weighted combination of individual treatment effects. If such treatment effect heterogeneity is important, one can instead estimate individual treatment effects β_i in the following interacted regression equation, in which $\tilde{\alpha}_i$ are fixed effects that are distinct from α_i in (1.7):

$$y_{it} = T_{it}\beta_i + X_{it}\delta + \tilde{\alpha}_i + \varepsilon_{it}.$$

Then, these individual treatment effects can be averaged at will. For example, [Wooldridge \(2005\)](#) suggests the “sample-weighted” treatment effect, which is identical to the ATE if each agent is sampled the same number of times, as $\sum_{i=1}^I \left(\frac{N_i}{N} \hat{\beta}_i \right)$. Unfortunately, in many applications there is no variation in T across time for some agents, making it impossible to identify their individual treatment effects, nor the sample-weighted treatment effect nor the ATE.

In the urban economics literature, regression models with individual fixed effects have been extensively employed to try to understand the effects of city size or density on wages, and by extension productivity, through agglomeration economies. [Glaeser and Maré \(2001\)](#), [Combes et al. \(2008\)](#), [Baum-Snow and Pavan \(2012\)](#), and [De La Roca and Puga \(2014\)](#) among others estimate Mincerian regressions of log wages on experience, some measure of city size, and individual fixed effects that resemble the following formulation:

$$\ln w_{it} = \beta[\text{citysize}]_{it} + X_{it}\delta + \alpha_i + \varepsilon_{it}. \quad (1.8)$$

Identification of the city size coefficient β comes from individuals' moves across cities of different sizes. Note that citysize can be specified as a vector of treatment dummy variables or as a more continuous measure of city size or density. In the context of the data-generating process (1.4), the role of the individual fixed effects α_i is to control for the time-invariant component of U_i . As a consequence, one interpretation of α_i is as indicators of time-invariant ability or skill. These studies consistently find strong relationships between wages and city size even after controlling for individual fixed effects, though inclusion of individual fixed effects typically reduces the coefficient on city size or density

by about one-third to one-half. The *prima facie* implication of this result is that while there is a causal effect of city size or density on wages, there is also important positive sorting of high fixed effect (unobserved ability) individuals into larger cities that must be accounted for in any evaluation of agglomeration economies through wages.

The greatest threat to identification in such studies is that some unobservable that may predict wages and labor market attachment is correlated with decisions to move across cities of different sizes. Individuals with positive unobserved personal productivity shocks may be more likely to move to larger cities. Potential omitted variables could be marital status, home foreclosure, winning the lottery, moving to care for a sick relative, losing one's job, or moving to start a better job. These unobserved variables are time-varying components of U_i , though one could argue that variation in job offer or separation rates across cities should be counted as part of the variation in city productivity.⁸ If such endogeneity of the move decision is important, making use of only the within-individual variation in city size may actually introduce more bias to the estimate of β than not including fixed effects and making use of comparisons between individuals. Fixed effects models make no use whatsoever of any potential information in the "control" group of individuals who never moved but who may have unobservables similar to those of individuals who are located in cities of different sizes.⁹

Heterogeneous treatment effects are also of first-order importance for consideration for two reasons. First, those who move more frequently are weighted more heavily in the calculation of the city size effect β . If more able people with higher wage growth potential move more often, they receive higher weight in the estimation of β . If this is the case, their types U are oversampled from the MTE distribution $B(X, U)$, and β may thus highly overstate the ATE. Moreover, the fact that moves are more prevalent soon after labor force entry means that the fixed effect estimator recovers the causal effect of city size primarily for those early in their working lives and not for the average stage in one's career. In the language of [Section 1.2](#), we can think of labor market experience as an element of X and the MTE $B(X, U)$ as being larger at certain values of X than at others. Therefore, even without an omitted variables problem, the fixed effects estimator in this case recovers a particular LATE which may overstate the ATE because of both oversampling of high-ability individuals and moves early in the life cycle. Failure to incorporate this treatment effect heterogeneity into the empirical specification can bias the fixed effects estimates, in which case

⁸ While differences across cities of different sizes in the arrival rate of job offers and separations are typically considered one mechanism for agglomeration economies, this data-generating process is inherently dynamic with the job match as an important state variable. Therefore, in the context of an estimation equation such as (1.8) which could never capture such a data-generating process, it is more straightforward to treat search and matching as showing up in U_i rather than as part of the coefficient on citysize. [Baum-Snow and Pavan \(2012\)](#) consider how to recover estimates of the importance of search and matching in agglomeration economies using a dynamic structural model.

⁹ Observations about individuals that remain in the same location during the sample period do help increase the precision of the estimates of δ .

they would not be good measures of individual ability. These observations are made by [De La Roca and Puga \(2014\)](#) using Spanish data and [Baum-Snow and Pavan \(2012\)](#) using US data in their assessments of the effects of city size on wages.

Absent some source of randomization in treatment, the literature has heretofore been only partially successful at handling the potential endogeneity of moves without the use of a structural model, as in [Baum-Snow and Pavan \(2012\)](#). [De La Roca and Puga \(2014\)](#) have made some progress in recovering information about heterogeneity in treatment effects and in the amount of selective migration by allowing β and δ to differ by individual fixed effects α_i . They estimate their empirical model iteratively by first capturing fixed effects and then interactions until a stable set of fixed effects is estimated. They find that returns to experience are larger for higher-ability individuals in larger cities, but wage level differences do not depend much on ability. By examining the distributions of fixed effects in different locations, [Combes et al. \(2012\)](#) argue that selective migration is not a big enough phenomenon in French data to drive a large wedge between the true ATE and OLS estimates of city size coefficients, a conclusion that [Baum-Snow and Pavan \(2012\)](#) and [De La Roca and Puga \(2014\)](#) share.

Another context in which fixed effects methods are standard is in hedonic models. With use of data on home prices from transactions and home characteristics, fixed effects remove time-invariant unobserved home characteristics that contribute to home value. Repeat sales hedonic models (which originally excluded observable home characteristics) are the basis of housing price indices going back to [Bailey et al. \(1963\)](#), including the S&P Case–Schiller index ([Case and Shiller 1987, 1989](#)). Repeat sales indices are constructed using a regression model such as the following, typically with some adjustment for potential heteroskedasticity in the errors:

$$\ln p_{ijt} = \beta_{jt} + X_{ijt}\delta + \alpha_i + \varepsilon_{ijt}.$$

In this equation, $\ln p_{ijt}$ is the log transaction price of home i in market j at time t . The fixed effects α_i account for unobserved fixed home characteristics, β_{jt} captures the home price index for market j at time t , and X_{ijt} includes time-variant home characteristics. [Rosenthal \(2014\)](#) uses a similar specification with homeowner's log income on the left-hand side to account for fixed unobserved home characteristics in his investigation of filtering.

This repeat sales specification also forms the basis for several studies which evaluate the willingness to pay for various local public goods and services, including various aspects of actual and perceived school quality. For example, [Figlio and Lucas \(2004\)](#) examine how housing prices and mobility changed when new school report cards in Florida provided the public with condensed information about local public school quality. To achieve this, they partition $\beta_{jt} = \mu_{jt} + T_{jt}\beta + X_{jt}^s\gamma$. In this expression, T_{jt} is a vector of dummy variables for the locally zoned elementary school's state-assigned grades in attendance zone j and X_{jt}^s is a vector of school characteristics that go into construction of the grade. The estimated treatment effect β reflects a causal effect of school grades on local housing values.

Econometric identification comes from the assertion that reported grades were a surprise and involve considerable random noise, and therefore are unlikely to be correlated with neighborhood unobservables. Moreover, all time-varying observable attributes about local schools are controlled for in X^s and there is no possible correlation between better school grades and time-invariant influences on home prices because of controls for home fixed effects α_i . The interpretation of the β vector is thus the average effects of changing neighborhood school grades on local home prices. It is important to recognize that the hedonic valuation of an A grade is likely identified mostly from variation in homes in quite wealthy neighborhoods with a strong taste for school quality, because these are the locations in which schools had variation in the A grade dummy, whereas the hedonic valuation of an F grade is identified primarily from poor neighborhoods. Therefore, these are local treatment effects which apply only for the subset of the full distribution of homes that experienced variation in relevant grades.

Beyond the local nature of such β estimates, clear interpretation of hedonic regression results requires careful consideration of the data-generating process for home prices. Hedonic models starting with that of Rosen (1974) indicate that shifts in the quality of one attribute of a product may induce a shift in the composition of buyers of that product. In addition, the elasticity of housing supply determines the extent to which such quality increases may be reflected in prices versus quantities. In this context, an increase in perceived local school quality and the resulting outward shift in local housing demand may be driven by wealthier residents looking to move into the neighborhood. These wealthier residents may seek higher quantities of housing services, and the demand shift may spur developers to increase the housing stock. Therefore, even if a regression such as that specified above is well identified and β is a causal effect of school grades on home prices, it is not straightforward to interpret it as the marginal willingness to pay by any particular potential buyer for this increase in local public goods. Indeed, Figlio and Lucas (2004) demonstrate that A grades induced sorting of higher-achieving students into the schools' attendance zones—students whose parents are likely willing to pay more for school quality than the families they replaced. Greenstone and Gallagher (2008) consider how to recover estimates of welfare consequences of toxic waste cleanups using home price data aggregated to the census tract level. In general, because neighborhoods with different attributes have different household compositions, β in the standard hedonic equation above recovers only the marginal willingness to pay under the strong assumption that all households have homogeneous preferences over neighborhood attributes.¹⁰

¹⁰ Recovery of heterogeneity in marginal willingness to pay for neighborhood attributes typically requires additional economic modeling. The article by Bayer et al. (2007), which we discuss in Section 1.6, shows how to recover the distribution of willingness to pay for school quality and sociodemographic characteristics of neighborhoods using a structural model married with an RD identification strategy to control for unobserved neighborhood characteristics. Kuminoff et al. (2013) present a review of the many structural models of supply and demand equilibrium in housing markets that can be used to recover willingness to pay for public goods.

Another setting in which fixed effects have been effectively used is to control for unobserved neighborhood characteristics in cross-sectional or repeated cross-sectional data with geographic identifiers. A typical specification is as follows, in which j indexes local units such as census tracts or block groups:

$$y_{ijt} = b_{jt} + T_{ijt}\beta + X_{ijt}\delta + \varepsilon_{ijt}.$$

[Campbell et al. \(2011\)](#) use this sort of specification to examine the effects of forced sales, through foreclosure or resident death, for example, on home prices. In their context, the treatment is a dummy that equals 1 if a home transaction was a forced sale or 0 otherwise. Census tract-period fixed effects b_{jt} control for the possibility that homes may be more likely to be force sold in lower socioeconomic status neighborhoods. [Autor et al. \(2014\)](#) use a similar specification to measure the effects of rent decontrol in Cambridge, Massachusetts, on housing values and [Ellen et al. \(2013\)](#) do so for examining the effects of foreclosures on crime. [Bayer et al. \(2008\)](#) use census block group fixed effects to control for sorting and unobserved job options in their evaluation of job referral networks in which each observation is set up as a worker pair. Their basic identifying assumption is that those looking for a home can at best find one in a particular block group rather than a particular block, yet they find that living on the same block is strongly related to working on the same block conditional on individual and block fixed effects.

One somewhat arbitrary feature of the standard use of spatial unit fixed effects is the assignment of each observation to only one particular spatial region fixed effect, even though observations typically differ in their centrality in such regions. That is, those observations on the edge of a census tract or block group may receive some spillover from neighboring tracts' unobserved characteristics and not all locations within spatial unit j are likely to have exactly the same set of unobservables. To the extent that the treatment differs as a function of location (e.g., because of spillovers from nearby regions) in a way that is correlated with subregion level unobservables, estimates of β would be biased and inconsistent. One way of accounting for microgeographic fixed effects that alleviates this problem is by using a spatial moving average specification. We replace b_{jt} in the above regression equation with

$$b_{ijt} = \sum_k \left[W[\text{dist}(i, k)] \tilde{b}_{kt} \right].$$

Assuming knowledge of the exact location of each i and indexing spatial units by k , one can take a weighted average of nearby fixed effects. In this expression, $W(\cdot)$ is a weighting function that equals 1 when the distance between observations is 0 and declines with distance or adjacency. This weighting function could have one estimated parameter ρ and could take a standard form with exponential or linear decay, as in $W(d) = e^{-\rho d}$ or $W(d) = \max[1 - \frac{d}{\rho}, 0]$. Estimation of the fixed effects and $\tilde{b}_{kt}$ and decay parameter ρ could be implemented by nonlinear least squares or the generalized method of moments (GMM). One could also generalize this specification to incorporate a separate

individual fixed effect for smaller spatial aggregations. This is a particular case of the spatial moving average model which is discussed at greater length in [Chapter 3](#) by Gibbons et al. and in which the endogenous portion of the error term is controlled for.

We delay our discussion of fixed effects estimators applied to data aggregated to the local labor market level to the following subsection.

1.4.2 Difference in differences methods

The DD identification strategy is a particularly common application of fixed effects. To be viable, it typically requires a data structure in which “treatment” and “control” groups are observed in multiple treatment environments, at least one of which is the same for the two groups. Typically, one difference is over time such that in initial periods the treatment has not yet been implemented, though in some studies treatment and control groups are instead compared in different locations or contexts other than time periods. Differencing over time (or across contexts), often implemented by including group or subgroup fixed effects, purges from the error term any time-invariant unobservables U that may differ between treatment and control groups. Differencing across groups, typically implemented by including time fixed effects, purges from the error term time-varying elements of unobservables U that are the same in the treatment and control groups. The primary identification assumption in DD estimators is that there are no time-varying differences in unobservables that are correlated with the treatment. The DD strategy can be generalized to the case in which the treatment is given to different observations at different points in time and/or to incorporate additional “differences.”

Implementation of the DD identification strategy is straightforward. With data in levels, one can think of the coefficient of interest as that on the interaction between the treatment group and a posttreatment dummy. One can equivalently calculate a simple DD in mean outcomes for the treatment group versus the control group in the posttreatment period versus the pretreatment period. The following regression equation, which can be estimated by OLS, incorporates the standard DD specification for panel data, in which β is the coefficient of interest. It includes period fixed effects ρ_t , individual fixed effects κ_i (which can be constrained to be the same within entire treatment and control groups, or subsets thereof), and the treatment variable of interest T_{it} , which is only nonzero for the posttreatment period:

$$y_{it} = \rho_t + \kappa_i + T_{it}\beta + X_{it}\delta + \varepsilon_{it}. \quad (1.9)$$

One may also wish to control for X . However, if unobservables are differenced out by the DD estimator, observable controls X should be differenced out as well. Therefore, in most cases controlling for X will not matter for estimating β since X is orthogonal to T conditional on the fixed effects. Below we consider the consequences of controlling for X in cases in which X is correlated with T . At least one period of data in both the

pretreatment environment and the posttreatment environment is required in order to recover a DD estimate. To ease exposition, we denote period 0 as the pretreatment period and period 1 as the posttreatment period.

Depending on the context, the DD estimator may consistently recover different treatment effects or no treatment effect at all. In the context of the data-generating process described by (1.5), consistent estimation of any treatment effect requires that any shocks to U are not correlated with the treatment. Put another way, any differences in the composition of the treatment and control groups in period 1 versus period 0 must be random. In mathematical terms, the key identification assumption is

$$(E[U|T_1 = 1, t = 1] - E[U|T_1 = 1, t = 0]) - (E[U|T_1 = 0, t = 1] - E[U|T_1 = 0, t = 0]) = 0. \quad (1.10)$$

This assumption is valid as long as there are no time-varying unobservables that differ across treatment and control groups and predict the outcome of interest. Differencing between treatment and control groups over time (or, equivalently, including group fixed effects κ_i) purges all fixed differences between the treatment and control groups, even if the distribution of unobservables is different in these two groups. Differencing across groups at each point in time (or, equivalently, including time fixed effects ρ_t) controls for differences in the pretreatment and posttreatment environments. The comparison between these two differences thus recovers a treatment effect averaged over the distribution of observables and unobservables in the treatment group provided that any differences in unobservables between the treatment and control groups are not time varying.

It is straightforward to derive that $\hat{\beta}_{OLS}$ only consistently estimates $ATE = E[B(X, U)]$ if all of those in the treatment group receive a full treatment, none in the control group do, and the treatment is fully randomized, meaning that the treatment and control groups have the same joint distribution of observables and unobservables. However, because the DD estimator is typically applied in settings in which some selection into treatment can occur, it is unlikely that an ATE is recovered. This selection into treatment can be conceptualized as existing for spatial units or for individuals within spatial units. Because spatial units cannot select out of treatment, a well-identified DD estimator recovers the TT for data-generating processes such as (1.6), in which the object of interest is at the level of spatial units rather than individual agents. If we think of the treatment as being applied to spatial units but individual agents to be the objects of interest as in (1.5), we can also think of the DD identification strategy as delivering TT for spatial units. However, if those for whom $T_{it} = 1$ can refuse treatment (as is typical) and the set of agents offered treatment is representative of the overall population, the DD estimator at best recovers ITT as defined at the individual agent level. If the researcher has information about the probability that agents who received the offer of treatment accept it, this ITT estimate can be rescaled to produce an agent-level estimate of TT.

It is common to use the DD identification strategy to analyze situations in which a treatment is applied to specific regions and outcomes of interest are at the individual level. Though the researcher may care about such individual-level outcomes, outcomes may only be reported at spatially aggregated levels such as census tracts or counties, as in (1.5). In this context, the treatment group is in practice identified as treated locations, in which individuals are presumably more likely to be treated. An important threat to identification in such a setting in which aggregate data are used is the potential resorting of individuals (on unobservables) between the treatment and control groups. If the treatment is valuable to some people in untreated areas, they may migrate to treated areas, thereby displacing some that do not benefit as much from the treatment. Such sorting on unobservables that is correlated with (and happens because of) the treatment would violate a version of the identification condition (1.10) with aggregate data (which looks exactly the same because of the law of iterated expectations), thereby invalidating the DD identification strategy.

One indicator pointing to a high likelihood of differing distributions in unobservables in the treatment and control groups existing before treatment versus after treatment is differing pretreatment trends in outcomes for the two groups. For example, if the control group experienced a positive shock in period 0 and is reverting toward its long-run mean between periods 0 and 1, that would cause the DD estimator to overstate the true effect of the treatment. Similarly, if the treatment group received a negative shock prior to treatment, this would similarly make it look like the treatment had a causal effect when all that is different is simply mean reversion. Indeed, in some settings agents are selected for treatment because of low values of observables, some of which may be transitory. This threat to identification is colloquially known as the “Ashenfelter dip” (Ashenfelter, 1978).

As empirical researchers, we often have access to a data set with some observables that are available to be included as controls. It is not clear that these variables should always be used. Indeed, one should think of most elements of X as analogous to the W variables that make up U , except that they are observed. Including these elements of X should thus not influence the estimate of β in (1.9) if the DD strategy is sound, though they may reduce its estimated standard error. However, in some settings there may be elements of X that describe attributes of agents on which they sort in response to the treatment. This phenomenon may arise, for example, in cases in which the treatment and control groups are defined as geographic units rather than individuals. If such sorting across treatment/control groups is fully predicted by attributes, then controlling for X is appropriate as it rebalances the treatment and control groups in both periods. That is, the two identification requirements on conditional expectations of U listed above may be true conditional on X even if not unconditionally. However, if inclusion of X in (1.9) influences the estimate of β for this reason, and sorting on observables exists, it is likely that sorting on unobservables also exists, thereby invalidating the identification assumptions listed above. Therefore, comparison of estimates of β including and

excluding controls for X is some indication as to whether sorting on unobservables may be biasing the coefficient of interest.

In some settings, it may be the case that some elements of X respond directly to the treatment. For example, it may be that incomes increased in areas that received federal EZ funding at the same time that income influences the outcome of interest y such as the home ownership rate. In this example, controlling for income changes the estimate of β because absent controls for income and assuming $E(T\epsilon) = 0$, β measures a full derivative, whereas controlling for income, β captures a partial derivative. However, controlling for an endogenous variable such as income runs the risk of violating the basic identification condition $E[X\epsilon] = 0$, thereby rendering $\hat{\beta}_{OLS}$ inconsistent. This violation would occur if, in this example, income were a function of T and some unobservable in ϵ , thereby making T correlated with ϵ as well. Therefore, a less fraught approach for recovering the partial effect of T on y holding income constant is to directly estimate the treatment's effect on income (by making it an outcome), and then separating out that effect directly to recover the residual effect of the treatment on the real outcome of interest y , which does require knowledge of $\frac{\partial y}{\partial X}$ from elsewhere. Note that a standard robustness check in DD estimators is to control for pretreatment X variables interacting with time. These are exogenous to the treatment because the treatment is 0 in all pretreatment observations.

Ham et al. (2011) use several flavors of the DD estimator to evaluate various impacts of several local economic development programs, including the federal EZ program. This program's first round started in 1994 and provided tax credits to businesses to hire local residents, reduced borrowing costs for community development, and committed billions of dollars in community development block grants to these communities. EZ status was awarded to a group of poor census tracts in each of 11 cities selected for the program. Ham et al. (2011) use census tract data to evaluate the effects of EZ status on poverty, labor earnings, and employment, and argue that EZs improved all of these outcomes. Their initial analysis uses data from the 1990 and 2000 censuses, with nearby tracts acting as a control group for EZ tracts. One may be concerned that tracts with negative economic shocks prior to 1990 were selected to be EZ tracts because of this, violating the assumption of common pretreatment trends. To handle this, the authors introduce a third difference—between 1980 and 1990—making this a differences in differences in differences (DDD) estimator. In practice, one can implement a DDD estimator by carrying out the DD estimator exactly as laid out above on first-differenced data for each of two time spans. The advantage of the DDD estimator in this context is that any common linear trends in unobservables in treatment and control groups are differenced out, eliminating any potential bias because of an “Ashenfelter dip.” However, any higher-order (e.g., quadratic) trends are not accounted for, nor is the possibility that the treatment status itself changed tract compositions. That is, if the treated tracts and control tracts have a different composition of residents and firms in 1990 and 2000 that is partly unobserved, part of any estimate recovered may reflect this composition shift.

The evaluation of the EZ program by [Busso et al. \(2013\)](#) also employs DD and DDD strategies but instead uses census tracts in areas that were barely rejected for inclusion in EZs in other cities as the control group. As with the [Ham et al. \(2011\)](#) study, the disadvantage of using this control group is that these locations were likely rejected for inclusion in the first round of the EZ program because they were slightly less distressed than those that ended up being included. The advantage of the [Busso et al. \(2013\)](#) approach is that they use an estimator that reweights the control group on observables to be more comparable than the equal weighting given by standard OLS. This study is further discussed in the following subsection, along with the use by [Kline and Moretti \(2014\)](#) of the same estimator in tandem with a DD identification strategy to evaluate the effects of the Tennessee Valley Authority on long-run outcomes.

[Greenstone et al. \(2010\)](#) use a DD estimator to recover the effects of large new industrial plants on incumbent plants' total factor productivity. Their treatment group is the set of counties which received new industrial plants and their control group is the set of counties that were barely rejected for the siting of an industrial plant. The idea is that counties chosen for these new plants should be similar on unobservables to those barely rejected, and indeed the paper shows evidence that the treatment and control groups of counties have similar pretreatment observable characteristics and pretreatment trends. Incumbent plant outcomes in treatment and control counties are compared before and after the arrival of new industrial plants, as are differential posttreatment trends in these outcomes. Their results indicate that these large new industrial plants had significant spillovers of about 5% on average to incumbent plant total factor productivity, with larger effects in closely related industries. This is direct evidence of positive agglomeration spillovers.

[Figure 1.1](#), taken from [Greenstone et al. \(2010\)](#), is an instructive illustration of how the DD strategy can be implemented. The top panel shows the average total factor productivity in incumbent manufacturing plants in treatment and control counties each year from 7 years before to 5 years after the arrival of the new large industrial plant in each treatment county, normalized to zero in the year prior to entry. This plot shows that pretreatment trends were very similar for treatment and control groups, with these trends diverging starting at period 0. The bottom panel shows the differences between treatment and control groups in each period, and a marked shift up in these differences after period 0. The simplest DD estimator, which could be estimated with a specification such as (1.9), is indicated in the lower panel as the gap in average gaps between treatment and control groups after treatment relative to before treatment. The authors extend the simplest DD specification (1.9) to recover information about dynamic responses to the treatment. [Greenstone and Gallagher \(2008\)](#) use a similar strategy to argue that cleaning up hazardous waste sites had negligible effects on housing prices, housing quantities, population, and population composition in nearby census tracts. These can be thought of as special cases of the RD estimator discussed in detail in [Section 1.6](#).

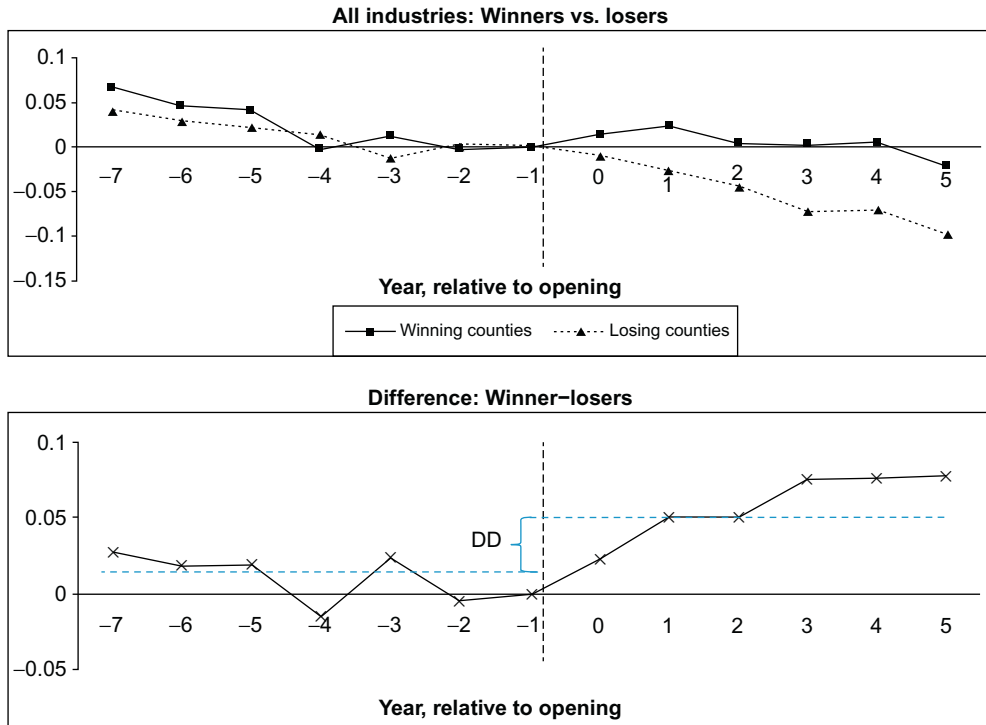


Figure 1.1 TFP of incumbent firms in “Winning” and “Losing” Counties from [Greenstone et al. \(2010\)](#).

A nonexhaustive list of other prominent empirical studies in urban and regional economics which make use of DD or DDD identification strategies follows. [Field \(2007\)](#) examines the labor supply effects of land titling in Peru by comparing squatters to those with land title in areas with recent title provision. [Costa and Kahn \(2000\)](#) examine the extent to which large cities better foster “power couple” location or formation by examining differences between large and small cities and various demographic groups who have more versus fewer constraints on forming a dual-worker couple over time. [Linden and Rockoff \(2008\)](#) show that home values declined nearer to the homes of sex offenders moving into neighborhoods relative to those further away. In a similar vein, [Schwartz et al. \(2006\)](#) demonstrate that new subsidized housing developments in New York City increased values of nearby homes more than those further away. These two spatial DD studies employ more flexible specifications than in (1.9) because they allow for full spatial variation in responses to treatment to be captured in the regression specification.

The DD identification strategy has also been applied in settings with data-generating processes that operate at the metropolitan area or county levels. For example, [Redding and Sturm \(2008\)](#) show that after the division of German, population growth rates in

areas near the West German border were less rapid, whereas after reunification they were more rapid than elsewhere in the country. This study uses differences over time and between border and nonborder regions. [Baum-Snow and Lutz \(2011\)](#) evaluate the effects of school desegregation on residential location patterns by race. Identification in this study comes from comparing metropolitan areas that had recently been treated with those that had been not by treated by 1970 or 1980. The years 1960 and 1990 bookend their study, in which all metropolitan areas in the sample were untreated and treated, respectively. This is implemented as regressions of the form of (1.9) in which i indexes metropolitan areas and T_{it} is a binary for whether the central school district in the metropolitan area is under court-ordered desegregation at time t . Because of variation in the timing of treatment, the compositions of the treatment and control groups depend on the year. Identification in this case depends on there not being unobservables that are correlated with the timing of treatment. Because all metropolitan areas go from being untreated to treated during the sample period exactly once, the resulting treatment effect estimates apply broadly within the sample used and can be interpreted as ATEs for the set of metropolitan areas considered.

[Abadie et al. \(2014\)](#) describe how to implement a method of “synthetic controls” as a way to construct the control group in DD-type estimation environments. This method is often applied when the treatment group is very small or consists of just one unit but there are many candidate control units. Instead of cherry-picking a few particular units for the control group that may or may not represent good counterfactuals for treated units, the authors show how to use a weighted combination of all available control observations, with weights set to represent how close they are to treated observations. The resulting

treatment effect estimate is $\hat{\beta} = Y_{1t} - \sum_{j=2}^{J=1} w_j^* Y_{jt}$, where Y_{1t} is the outcome at time t for

the treated unit (or an average among treated units), Y_{jt} are the outcomes for the control units, and w_j^* is a set of weights. These weights are chosen in a way that minimizes some distance criteria between predetermined characteristics of the treated units and the predetermined characteristics of the control units. For example, [Abadie and Gardeazabal \(2003\)](#) and [Abadie et al. \(2010\)](#) choose the vector W^* as the value of W that minimizes

$$\sum_{m=1}^k v_m (X_{1m} - X_{0m} W)^2.$$

In this expression, X_{1m} denotes the average value characteristic m for treated observations, while X_{0m} is the vector of the same characteristic for control observations, all calculated prior to treatment. Further, v_m is a measure of the importance of characteristic m , which can be chosen to be proportional to the predictive power of characteristic m for the outcome. The problem with the synthetic controls approach is that the choice of predetermined characteristics and distance criteria can be ad hoc, and one may end up giving too

much weight to control units that are not appropriate counterfactuals owing to differential pretrends or other unobserved components. But the interesting characteristic of this approach is that it allows for simple construction of generalized control groups. In the following subsection, we analyze matching methods that more directly use this idea.

1.4.3 Matching methods

The DD and fixed effects identification strategies discussed thus far are only credible if the treatment group is observed prior to treatment and there are no time-varying unobservables correlated with the treatment. However, there are many settings in which either a pretreatment period is not observed or time-varying unobservables that are different in the treatment and control groups and may influence outcomes are likely to exist. One potential solution to such problems is to use an estimator that makes use of information about observables to try to infer information about unobservables. We focus on cases in which the treatment is binary.

As a starting point, consider trying to recover information about the causal effect of treatment in the constant coefficient version of the data-generating process in (1.1) using cross-sectional data. That is, suppose the true data-generating process is as follows:

$$y_i = T_i\beta + X_i\delta + W_i\rho + u_i.$$

Note that because this is a constant coefficient model by assumption and if W and T are uncorrelated, the OLS estimate of β is the ATE. Trying to estimate this equation by OLS leads to biased estimates of β if some unobservables W are correlated with the treatment. One common heuristic method for addressing such potential bias is to estimate this equation by varying the set variables in the control set X . The idea is that as variables are moved from unobservables W to observables X , any reductions in estimates of β indicate omitted variables bias is influencing these estimates. If β is stable with inclusion of additional controls, there is more confidence that omitted variables bias is not a problem. Crucial for this method to be informative is for the R^2 of the model to increase as variables are moved from W to X . If R^2 does not increase, these are irrelevant variables with true coefficients of 0. As crucial is that the set of controls in X is in some sense representative of the full set of possible control variables $[XW]$. At the end of this subsection, we consider how examples in the literature have attempted to correct the bias using a proportional selection bias assumption, formalizing this intuition.

Standard practice for attempting to estimate causal effects in the absence of implicit randomization is to employ a propensity score matching estimator. The idea of such estimators, originally proposed by [Rosenbaum and Rubin \(1983\)](#), is to compare outcomes of individuals with the same propensity to be treated, some of whom receive treatment and others of whom do not. The underlying “propensity score” $P(X)$ is the probability of being treated, and depends on observables only. This score can be estimated by a probit or logit with a flexible specification.

The main difficulty with matching estimators is that they assume that selection into or out of treatment is fully predicted either by observables or by unobservables that do not predict the outcome of interest. If unobservables influence both outcomes and whether agents receive treatment, treated and untreated observations are not comparable for any given propensity score, and matching estimators are not informative about any treatment effect. If unobservables influence outcomes but not the probability of treatment, matching estimators are still informative about treatment effects. This intuition is the same intuition about potential threats to identification in OLS regression, so it is not surprising that OLS is a particular form of a propensity score matching estimator. [Heckman and Navarro-Lozano \(2004\)](#) demonstrate that matching estimators can be quite sensitive to the conditioning sets used and argue that control function methods in which choices are more explicitly modeled are more robust. We briefly consider such methods at the beginning of the following section.

Formally, the following conditions must hold in order for a propensity score estimator to produce consistent treatment effect estimates ([Wooldridge, 2002](#)):

$$E(y^0|X, T) = E(y^0|X), E(y^1|X, T) = E(y^1|X). \quad (1.11)$$

These conditions say that those receiving the treatment have the same mean outcomes whether they are treated or not as those who do not receive the treatment. That is, actually receiving treatment cannot predict outcomes in either the treated or untreated counterfactual states of the world. These assumptions are sometimes called “selection on observables” because they allow selection into treatment to be fully predicted by X , but not by U . This assumption implies $TT(x) = ATE(x)$, but not necessarily that $TT = ATE$.

Provided that the data set being used is rich with observables, there is information in the propensity score coupled with treatment status about whether unobservables correlated with the treatment may be an important source of bias. If there is very little overlap in the range of the propensity score in which both treated and untreated observations exist, this indicates that since treatment and control groups differ on observables, they may be more likely to differ on unobservables as well. Consequently, the range of the propensity score for which there is overlap is the region of the data for which the propensity score matching estimator is providing more convincing identification. As a result, it is often informative to graph the density of treated and untreated observations against the propensity score, plus the implied treatment effect at each level of the propensity score, to get a sense of the treatment effect over the range of the propensity score for which unobservables are less likely to be driving selection into treatment. To calculate such a treatment effect, one can nonparametrically estimate the conditional expectations $E(y|P(X), T = 1)$ and $E(y|P(X), T = 0)$ and then take the difference for every value of $P(X)$. This uses the argument that unobservables act in some sense like observables.

[Figure 1.2](#) provides two schematic diagrams which match these suggested graphs. Panel (a) shows the density of treatment and control group observations as a function

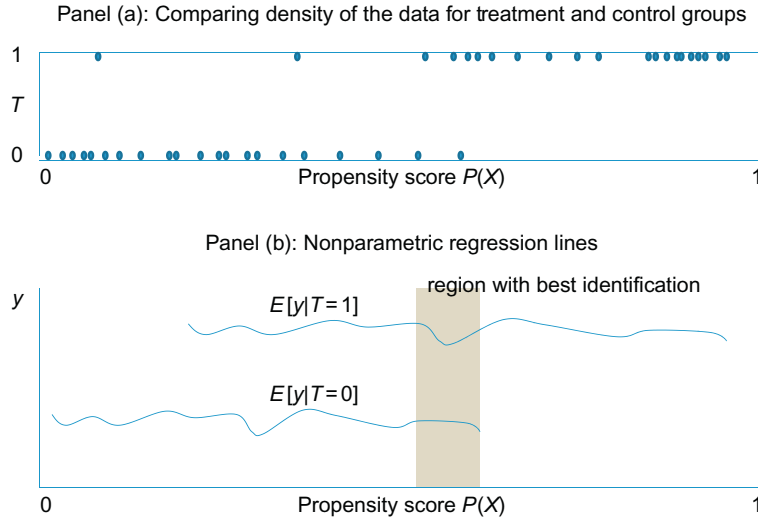


Figure 1.2 Schematic diagrams for matching estimators.

of the propensity score. In this example, there is very little overlap between the treatment and control groups. Indeed, just a few observations from both groups have similar propensity scores. Panel (b) presents nonparametric plots of some fictional outcome against the propensity score for treatment and control groups. Standard error bands are not included to make the figure less busy. However, it should be clear that standard error bands must be tighter at values of $P(X)$ near which there are more data. That is, even though it may be possible to calculate a nonparametric regression line for the treatment group at low values of the propensity score, it will be very imprecisely estimated because of the thin data in this region. The main message from Fig. 1.2 is that there are very few comparable observations across treatment and control groups at most propensity scores. Comparability between these two groups typically exists at propensity scores near 0.5, but may not exist for other regions. As a result, it may make sense to limit considerations of treatment effects to treated observations with control observations that have comparable propensity scores.¹¹

As discussed by Dehejia and Wahba (2002), identifying “matched” observations in propensity score neighborhoods of treated observations is a fruitful way of identifying a reasonable control group if not many observations have been treated relative to the number of candidate controls. They suggest choosing a propensity score window and only making use of control observations within this window of each treated observation.

¹¹ While we would have liked to use an example from the urban economics literature to depict graphs such as those in Fig. 1.2, this depiction has hardly ever been used in our field.

Given that the resulting control group observations are sufficiently close on observables to the treated observations, one can calculate TT as follows:

$$\widehat{TT} = \frac{1}{N_{T=1}} \sum_{T_i=1} (y_i - \frac{1}{J_i} \sum_{j(i)} y_j).$$

In this expression, $N_{T=1}$ is the total number of treated observations and J_i is the number of “matched” control observations for treated observation i . Those control observations matched to i are indexed by $j(i)$. Treated observations without a match are discarded, with appropriate reinterpretation of TT to apply only to the remaining treated observations.

Standard implementation of the propensity score estimator, which strictly assumes the conditions in (1.11), uses all available data. Given first-step estimation of the propensity score $P(X)$, the following equation can be estimated in a second step by OLS regression:

$$y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 P(X_i) + \alpha_3 T_i(P(X_i) - E[P(X)]) + \varepsilon_i.$$

In this regression, α_1 is the ATE provided that $E[y^1|P(X)]$ and $E[y^0|P(X)]$ are both linear in $P(X)$. A related but more nonparametric procedure that allows for direct recovery of $ATE(x)$ and $TT(x)$ is to estimate a regression such as the following:

$$y_i = b_0 + b_1 T_i + X_i B_2 + T_i(X_i - \bar{X}) B_3 + u_i.$$

Here, $ATE(x) = TT(x) = b_1 + (x - \bar{x}) B_3$ and $ATE = b_1$. If there is no treatment effect heterogeneity and $ATE(x) = ATE$, then this equation reduces to a standard linear regression of y on T and X . Calculation of the propensity score using a linear probability model and no treatment effect heterogeneity reduces the first equation to standard OLS as well. Therefore, we can interpret the OLS as a propensity score matching estimator that incorporates no treatment effect heterogeneity.

Some prominent recent applications of matching estimators have adopted a variant due to Kline (2011) which can be implemented in two steps. First, estimate regressions of the form

$$y_i = c_0 + c_1 T_i + (1 - T_i) X_i C_2 + e_i.$$

Here, X is accounted for in the control group only and not in the treatment group. The purpose is to determine Oaxaca–Blinder-type weights C_2 which serve as inputs into the following treatment effect calculation:

$$\widehat{TT} = \widehat{c}_1 - \frac{1}{N_{T=1}} \sum_{i=1}^N T_i X_i \widehat{C}_2.$$

This procedure compares the average outcome in treated observations with the average outcome in observations with the same distribution of X but that did not receive the treatment. Information from untreated observations in the first step is used to determine

the counterfactual mean for the treated set of observations absent treatment. [Kline \(2011\)](#) shows that this is equivalent to a propensity score reweighting estimator.

The best use of matching and propensity score methods is when there is a good reason to believe that conditional on X , treatment and control groups are similar on unobservables. In recent successful applications, this often involves marrying a matching estimator with a DD-type estimator, which is intended to make the treatment and control groups similar on unobservables. In addition, some observations in the untreated group are typically omitted from the control group in order to make the treatment and control groups as comparable as possible. Such use of propensity score matching estimators is a slightly more sophisticated version of the DD estimator, as they reweight control group observations to look like those in the treatment group on observables.

[Busso et al. \(2013\)](#) use the Oaxaca–Blinder estimator to compare outcomes in census tracts in federal EZs with those in areas that were rejected for inclusion in the program. They find that EZ tracts experienced 12–21% increases in total employment and 8–13% increases in weekly wages, but little change in rents or the composition of the population, though housing values and the percentage of residents with a college degree do increase. They carry out a placebo exercise that compares tracts that are similar on pretreatment observables but not assigned to EZs in EZ counties with the same control group and find no significant effects. [Kline and Moretti \(2014\)](#) use the same estimator in their evaluation of the Tennessee Valley Authority program, for which they trim counties adjacent to the Tennessee Valley Authority region and potential remaining control counties with propensity scores in the lowest 25% and from the control group. Their estimates indicate long-run significant positive effects on manufacturing employment, incomes, and land values and negative effects on agricultural employment.

[Gobillon et al. \(2012\)](#) employ a standard propensity score reweighting estimator to evaluate the effects of the French enterprise zone program, which provides wage subsidies for firms to hire local workers. They find that the program had a small significant effect on the rate at which unemployed workers find a job. [McMillen and McDonald \(2002\)](#) use such an estimator to examine how the type of zoning in Chicago influenced land values immediately after zoning was introduced in 1923. Using the propensity score to match prezoning characteristics between plots zoned for residential versus commercial use, they find that residential plots experienced greater price appreciation. As with the other studies discussed above, the propensity score estimator may be more defensible for this study since the treatment was presumably assigned on the basis of observables and so there is less opportunity for plots of land to sort in or out of treatment on the basis of unobservable characteristics. When individuals are analyzed such sorting concerns are more serious.

In addition to recovering treatment effects in cases of selection on observables, propensity scores can be useful to identify a control group of matched observations for cases in which a specific set of observations has been treated and a very large set of potential

control group observations must be pared down to include just close matches. [Alesina et al. \(2004\)](#) employ such an approach for evaluating the effects of racial heterogeneity on the number of jurisdictions. They identify “treatment” counties as those in northern states which experienced at least a 2 percentage point increase in the black population share during 1910–1920 (during World War I) or 1940–1950 (during World War II). Their challenge is to identify “control” counties that look as similar as possible on observables, and therefore (hopefully) unobservables. To achieve this goal, they first estimate a propensity score for all counties in affected states through a probit regression of treatment status on state fixed effects and various baseline county demographic characteristics and polynomials thereof. As in [Dehejia and Wahba \(2002\)](#), they identify propensity score windows around treated counties in which no significant difference in any observable exists. Then, these treatment and control groups were analyzed both descriptively and in a regression context. The results indicate that greater increases in racial heterogeneity were strong predictors of smaller declines in the number of school districts in the county.

Rather than using propensity score matches to identify a control group that look similar on observables to the treatment group, another strategy that also works with continuous treatments is to think of X as a representative set of potential control variables. [Altonji et al. \(2005\)](#) use this idea to evaluate the magnitude of omitted variables bias in the context of evaluating the causal effects of Catholic schools on high school graduation rates, college attendance, and test scores. Their basic assumption is that including an additional randomly chosen unobservable variable would have the same effect in reducing selection bias as including an additional available observable in X in an OLS regression. [Oster \(2013\)](#) reformulates this assumption as the following proportional selection relationship:

$$\nu \frac{\text{Cov}(T, X\delta)}{\text{Var}(X\delta)} = \frac{\text{Cov}(T, W\rho)}{\text{Var}(W\rho)}.$$

That is, the correlation between observables and the treatment is proportional to the correlation between the unobservables and the treatment.

To implement the resulting estimator, consider the following two regression equations, which can be estimated by OLS, yielding β' and β'' in addition to R^2 of R' and R'' , respectively:

$$\begin{aligned} \gamma &= \alpha' + T\beta' + \varepsilon', \\ \gamma &= \alpha'' + T\beta'' + X\delta'' + \varepsilon''. \end{aligned}$$

Having estimated these regressions and capturing their coefficients and R^2 , the only remaining required objects are the constant of proportionality ν and the maximum R^2 that would be recovered by estimating the full model, $R_{\max}$. These can be used in the following relationship, which incorporates the bias adjustment to the OLS regression from the full model:

$$\beta \xrightarrow{p} \beta'' - \nu \frac{(\beta' - \beta'')(R_{\max} - R'')}{(R'' - R')}.$$

Of course, the main difficulty is that ν and $R_{\max}$ are unknown. But one can get an idea of how large the bias could be by determining what ν would need to be for $\beta = 0$ given $R_{\max} = 1$. Standard errors need to be bootstrapped to conduct inference on the resulting bias-corrected coefficient.

The key obstacle to the use of matching, DD, and fixed effects estimators is the lack of any source of randomization. In some sense, all of these estimators end up in an environment in which we must assume that T is allocated in a way that is as good as random conditional on the other observed elements of the estimation equation. The following section's exploration of IV estimators instead focuses on environments in which there is some randomization in T , which is usually implicit.

1.5. IV ESTIMATORS

IV estimators are used to recover consistently estimated coefficients on treatment variables of interest when treatments are endogenous. One way of conceptualizing such an endogeneity problem is that a treatment variable is generated by a second linear equation which includes some unobservables that are correlated with unobservables which appear in the main estimation equation of interest. This makes the treatment T be correlated with the U part of the error term in the primary estimation equation, rendering the OLS estimate of the coefficient on the treatment biased and inconsistent. In the language of structural systems, there needs to be an “exclusion restriction” in which at least one observed variable must be excluded from one equation in order to identify coefficients of both equations without making ad hoc distributional assumptions. In the language of single-equation linear regression, there needs to be an “instrument” which isolates variation in T that is not correlated with any part of the error term in the main estimating equation. We sometimes label such variation “pseudorandom” because the role of the instrument is essentially to pick out random variation in T .

Consideration of how to estimate the classic [Roy \(1951\)](#) model by [Gronau \(1974\)](#) and [Heckman \(1979\)](#) is informative about the more structural background of the IV estimator. In this model, there is a binary treatment T into which individuals may self-select because it is presumably valuable for them. This self-selection generates a correlation between T and the error term in a linear regression of some outcome of interest on T and control variables X because of sorting on unobservables into the treatment. In particular, the underlying data-generating process is assumed to be

$$y^0 = X\delta_0 + U_0; \quad y^1 = X\delta_1 + U_1.$$

[Heckman \(1979\)](#) shows that if U_0 and U_1 are jointly normal, one can identify δ_1 and evidence of selection into treatment. The key insight is that the choice of whether to accept

treatment can be recovered explicitly using the fact that only those for whom $\gamma^1 > \gamma^0$ select into treatment. Operationally, one way of estimating the model is by estimating the model as a “Heckman two-step.” First, predict the probability of treatment as a function of X using a probit regression. Second, estimate the equation

$$\gamma^1 = X\delta_1 + \rho\sigma_u\lambda(X\gamma) + \varepsilon.$$

In this equation, $\lambda(\cdot)$ is the inverse Mills ratio constructed from the first step, which controls for selection into treatment. Because γ^0 was never observed in the original application, the standard treatment does not have a second step equation for γ^0 , though one could be constructed using analogous logic. The sign and magnitude of estimated ρ indicate the nature of selection into treatment on unobservables. One important insight of this work is thus that one can treat nonrandom selection into treatment as an omitted variables problem. The difficulty is that if the errors are not truly jointly normal, the model is misspecified and coefficients in the second step equation are inconsistently estimated unless an exclusion restriction is imposed.

Altonji et al. (2005) also consider a two-equation structural system in their exploration of evaluating the effects of attending Catholic schools on college attendance. They consider a bivariate probit model in which a set of demographic characteristics predict both Catholic school attendance and college attendance, such that Catholic school attendance is an explicitly endogenous treatment variable. They demonstrate how the estimate of the coefficient on T (Catholic school attendance) depends crucially on the magnitude of the correlation between the errors in the two equations. Higher correlations between the error terms mean that there are more similar unobservables driving both Catholic school attendance and success in school. As a consequence, the causal effect on Catholic school attendance declines because this variable simply reflects more positively selected students as the error correlation increases.¹² In the context of a data-generating process such as (1.4), one way to make progress in breaking a potential correlation between T and U , which renders OLS or probit estimates inconsistent, is to find variables that predict T but are not correlated with U . These are instruments, or exclusion restrictions.

In summary, the IV estimator is used to break a potential correlation between T and U . This correlation could exist because individuals with high values of U are sorting into the treatment at higher rates than others, as in the classic two-equation structural selection model in which T is “endogenous” because it is generated by a second equation. Or this correlation could exist because, regardless of where T comes from, there are variables correlated with T for which the researcher cannot control that end up in U as a result.

¹² Neal (1997) considers a similar bivariate probit setup to address the same questions except that he excludes religious affiliation and local Catholic population density from the graduation equation. These exclusion restrictions allow for recovery of estimates of the covariance of the errors between the two equations and the coefficient on Catholic schooling in the estimation equation of primary interest.

This is an omitted variables problem. These two ways of thinking about why $E(TU) \neq 0$ have distinct intellectual histories but many of the same implications.

1.5.1 Foundations

To be mathematically precise, we can think of IV estimators as those that recover β in the following system of equations:

$$y_i = T_i\beta + X_i\delta + \varepsilon_i, \quad (1.12)$$

$$T_i = Z_i^1\zeta_1 + X_i\zeta_2 + \omega_i. \quad (1.13)$$

In the second equation, Z^1 is the set of excluded instruments, of which there must be at least one per treatment variable for this econometric model to be identified. These additional Z^1 variables are “excluded” from the first equation. In the first equation, recall that $\varepsilon_i = U_i + e_i$ from (1.4). Denote the set of exogenous variables as $Z = [Z^1 X]$. IV estimators recover consistent estimates of β if $E(Z\varepsilon) = 0$ and the coefficients on the excluded instruments ζ_1 in (1.13) are sufficiently different from 0 in a statistical sense. We sometimes use the “reduced form” of this two-equation system, which is as follows:

$$y_i = Z_i^1\phi_1 + X_i\phi_2 + \psi_i.$$

If there is just one excluded instrument per endogenous variable, one simple way to estimate β is through indirect least squares (ILS): $\hat{\beta}_{\text{ILS}} = \frac{\hat{\phi}_{\text{1OLS}}}{\hat{\zeta}_{\text{1OLS}}}$. This is an intuitive object which shows how the first-stage coefficient rescales the reduced form effect of the instrument on the outcome.

Another simple intuitive way to estimate β is by substituting (1.13) into (1.4) and then explicitly including a proxy for ω_i in the estimation of the resulting (1.14):

$$y_i = T_i\beta + X_i\delta + \hat{\omega}_i\zeta + e_i. \quad (1.14)$$

This proxy acts as a “control function” for unobservables correlated with T_i . In the linear case above, β can be properly estimated by using $\hat{\omega}_i$ consistently recovered as residuals from OLS estimation of the first-stage (1.13). This method is closely related to the two-stage least squares (2SLS) estimator in which $\hat{T}_i$ is predicted from the first stage and inserted in place of T_i in (1.12), which can then be estimated by OLS to recover $\hat{\beta}_{\text{2SLS}}$.¹³ However, as discussed in Imbens and Wooldridge (2007), the control function approach sometimes provides additional flexibility when dealing with nonlinear models. Moreover, the coefficient ζ has a useful economic interpretation. ω_i is positive for those observations which were treated more than expected as predicted by Z^1 and X . One could thus interpret those agents as having higher than predicted returns from receiving treatment. Therefore, the sign of ζ indicates whether the type of agent who had a higher

¹³ For 2SLS estimation, it is important that the standard errors use estimates of ε_i calculated using the actual rather than the predicted T_i .

return from the treatment had better or worse outcomes y than the types of agents who had lower treatment returns. That is, ζ tells us about the nature of selection into treatment, much like the coefficient on the inverse Mills ratio does in Heckman (1979), as is fleshed out further in the development by Heckman and Honoré (1990) of the empirical content of Roy's model (Roy, 1951).

In addition to ILS, 2SLS, and control function methods, GMM, which makes use of the moment condition $E[Z^1 \varepsilon] = 0$, and limited information maximum likelihood are options for estimating β in the two-equation econometric model specified in (1.12) and (1.13). All of the various estimators of β in (1.12) suffer from weak small sample properties, though limited information maximum likelihood has been found to be most robust in small samples. All of these estimators are identical if the model is just identified, meaning that there is the same number of excluded variables as there are endogenous variables. Recent work has found that 2SLS can be more robust in some instances with many instruments if they predict not only T but also an element of X (Kolesar et al., 2013).

Most important for successful implementation of IV is the choice of good excluded instruments. One fruitful way of conceptualizing an instrument is as a source of random variation in T conditional on X . That is, a good instrument generates variation in T conditional on X that is not correlated with any unobservables in U . However, each element of X must also be exogenous. Therefore, the best instruments are those that generate truly random variation in T and therefore require no conditioning on X in the first equation. With such ideal instruments, which typically are only found with explicit randomization, the prudent researcher can avoid having to control for any elements of X and facing the associated danger of introducing a potential endogeneity problems. We discuss using IV estimators as a means to make use of explicit randomization in the context of RD in the following section.

The more typical situation is that a researcher is concerned about the endogeneity of some treatment T and there is no explicit randomization available. The following is one strategy for selecting good candidate instruments: Consider all of the possible sources of variation in T . From this list, select the ones that are least likely to be correlated with variables that directly predict y or are correlated only with observables that predict y that are very likely exogenous. Coming up with this list typically requires both creativity and a detailed investigation into the process by which the treatment was assigned. There is no direct test for instrument exogeneity, only a set of exogeneity arguments that are unique to each setting, though there are various standard auxiliary tests, some of which are suggested below in the context of examples from the literature. The next step is to estimate the first stage, (1.13), and to evaluate whether the instruments are sufficiently strong predictors of T . If they are not, the researcher has to keep looking. If multiple strong instruments are identified, special care is needed, as is also discussed below.

If the partial F statistic from the test of whether coefficients on excluded instruments are each significantly different from 0 is above about 9, then the instruments are strong enough predictors of T such that the estimated standard errors on β can be used.¹⁴ Otherwise, standard errors on β must be adjusted upward to account for a “weak instrument” problem. [Stock and Yogo \(2005\)](#) provide standard critical values for F tests for evaluating instrument strength. When implementing the primary specification of an IV estimator, one should control only for those predictors of γ that may be correlated with the instrument so as to avoid controlling for endogenous variables.

While the exposition thus far assumes a common coefficient β , in general we expect there to be heterogeneous coefficients on T of $B(X, U)$. Crucial to understanding IV estimates is to recognize that IV recovers a LATE, which is the average effect of the treatment for the subpopulation whose behavior was influenced by the excluded instrument, conditional on X ([Imbens and Angrist, 1994](#)). It typically requires further investigation to gather information about the particular LATE that is recovered from any given instrument. Continuous instruments and treatments in particular usually require some detective work to determine for whom the treatment effect being estimated by IV applies. With multiple instruments, it becomes even more complicated. Indeed, [Heckman et al. \(2006\)](#) lament that with many instruments it is often virtually impossible to determine which combination of MTEs is being estimated by IV.

Because of the fact that IV recovers a LATE, and that in typical urban economics applications it is difficult enough to find one valid instrument let alone many, it is prudent to stick to using only one excluded instrument at a time in most settings, with additional candidate instruments possibly used for robustness. The only reason to use multiple instruments at once is if one instrument by itself is too weak. Though it is possible to test for stability in β when switching between different instruments as a test of instrument validity, this process crucially assumes that the data are generated by a process with a constant coefficient. If instead there are heterogeneous coefficients, it may well be the case that multiple instruments generate different legitimate treatment effect estimates, all of which are different LATEs.

1.5.2 Examples of IV in urban economics

In the urban and regional economics literature, the IV empirical strategy has been most commonly used when the unit of observation is aggregated to the local labor market level. That is, the data-generating processes that have best lent themselves to IV estimation are either fully conceptualized at the aggregate level, as in (1.6), or are agent based but involve a treatment that operates at some aggregate geographic level, as in (1.5). Here we review examples of how IV has been used to successfully isolate exogenous components of local labor demand and labor supply shocks, construction of infrastructure, the

¹⁴ This is equivalent to evaluating if the t statistic is above 3 if there is just one excluded instrument.

implementation of local economic development policies, and the prevalence of various drivers of local agglomeration spillovers.

The classic use of IV in economics is to isolate exogenous supply or demand shifters in some particular market. Since supply and demand functions are fundamentally theoretical constructs, use of IV to isolate demand or supply shocks is probably most effective when an economic model is incorporated into the analysis in some way in order to organize thoughts about the most important forces buttressing equilibrium prices and quantities. Given the centrality of the demand–supply paradigm in economics, use of IV to isolate exogenous variation in demand and supply has a strong tradition. For example, [Angrist et al. \(2000\)](#) use weather variables as a source of exogenous variation in supply shifts to recover demand system parameters using the well-known Fulton Street Fish Market data ([Graddy, 1995](#)).

Following in this tradition, one of the commonest uses of IV estimation in the urban and regional economics literature is to isolate sources of exogenous variation in local labor demand. The commonest instruments for doing so are attributed to [Bartik \(1991\)](#) and [Blanchard and Katz \(1992\)](#). The idea is to isolate shifts in local labor demand that come only from national shocks in each sector of the economy, thereby purging potentially endogenous local demand shocks driving variation in employment or wages. While this type of instrument has been used to help recover parameters of local labor supply functions, it has more often been used to isolate exogenous variation in metropolitan area wages or employment levels.

There are two ways that “Bartik” instruments are most commonly constructed. A quantity version of the instrument is constructed by fixing each local labor market’s industry composition of employment at some base year and calculating the employment growth that would have occurred in each market had the industry composition not changed but employment in each industry had grown at the national rate for that industry. The price version of the instrument instead calculates the wage growth that would have occurred in each market had wages in each industry grown at the national rate for that industry, again holding the employment composition in each local labor market fixed to a base year. In order to allay potential concerns of a mechanical relationship between base year industry composition and unobservables driving an outcome of interest, researchers typically take industry composition from a year that predates measurements of any other variables used for estimation.¹⁵

A host of papers make use of such instruments for identification. [Notowidigdo \(2013\)](#) uses exogenous variation from Bartik instruments to demonstrate that positive local labor

¹⁵ To allay the potential concern that any particular local labor market influences national employment or wage growth, many studies exclude the own local labor market or state in the calculation of national growth rates by sector. This means that this growth component of the instrument is slightly different for each observation.

demand shocks increase the population more than negative demand shocks reduce it, and that this asymmetry is more pronounced for less skilled workers. However, he finds that housing prices, wages, and rents do not exhibit the same asymmetric responses. Through the structure of a [Roback \(1982\)](#) style spatial equilibrium model, these results are interpreted as indicating low mobility costs for everyone and a concave local housing supply function. Leveraging the same exogenous variation in local labor demand for identification, GMM estimates of the full model reveal that less skilled workers are more highly compensated through various transfers for negative local labor demand shocks than highly skilled workers, which accounts for the different mobility rates of these two groups. In a precursor to [Notowidigdo \(2013\)](#), [Bound and Holzer \(2000\)](#) examine the general equilibrium population responses by skill to exogenous local labor demand shocks. Through GMM estimation of a spatial equilibrium model, [Diamond \(2013\)](#) uses the identifying variation available from Bartik instruments to recover how local labor demand shocks lead to knock-on shifts in local skill composition and skill-specific amenities. [Boustan et al. \(2013\)](#) use Bartik instruments to help demonstrate that jurisdictions with greater increases in income inequality collected more local government revenues and had higher expenditures. [Luttmer \(2005\)](#) uses Bartik instruments in a reduced form specification to control for changes in average area incomes in showing that people whose incomes fall behind those of their neighbors are less happy, even if everyone's incomes are increasing. [Gould et al. \(2002\)](#) use Bartik shocks as an instrument for income in examining the causal effects of income on local crime rates.

In an important study, [Saiz \(2010\)](#) uses Bartik instruments to isolate exogenous local housing demand shocks interacted with a measure of land unavailable for development and an index of housing market regulation to recover an estimate of the housing supply elasticity for each US metropolitan area. He estimates inverse housing supply regression equations of the form

$$\Delta \ln P_k = \alpha_0 + \alpha_1 \Delta \ln Q_k + \alpha_2 \text{unavailable\_land}_k \Delta \ln Q_k + \alpha_3 \text{WRI}_k \Delta \ln Q_k + u_k,$$

in which k indexes metropolitan area, P denotes housing price, Q denotes housing quantity, and WRI is an index of local housing market regulation. Differences are taken for the 1970–2000 period. Bartik quantity instruments provide exogenous variation in all terms which include $\Delta \ln Q_j$.¹⁶ Housing supply elasticity estimates from this study have been widely used. In the work of [Beaudry et al. \(2014\)](#), such estimates interact with Bartik instruments to form a series of instruments in the estimation of a spatial equilibrium model which incorporates unemployment and wage bargaining frictions. The works

¹⁶ [Saiz \(2010\)](#) also makes use of hours of January sun and immigration inflows as additional sources of exogenous variation in $\Delta \ln Q_k$ and the prevalence of evangelical Christians as a source of exogenous variation in WRI_k .

of [Mian and Sufi \(2009\)](#) and [Chaney et al. \(2012\)](#) are two prominent examples from the finance literature that use these [Saiz \(2010\)](#) housing elasticity measures.

The main source of identifying variation in Bartik instruments comes from differing base year industry compositions across local labor markets. Therefore, validity of these instruments relies on the assertion that neither industry composition nor unobserved variables correlated with it directly predict the outcome of interest conditional on controls. As with any IV, the credibility of this identification assumption depends on the context in which the IV is being applied. Generically, one may be concerned that base year industrial composition may be correlated with fundamentals related to trends in labor supply. For example, it may be the case that manufacturing-intensive cities have declined not only because the demand for skill has declined more in these locations, but also because they have deteriorated more in relative amenity values with the increasing blight and decay generated by obsolete manufacturing facilities. That is, negative labor supply shifts may be correlated with negative labor demand shifts. Indeed, when Bartik instruments are implemented using one-digit industry classifications, as is often done, the initial manufacturing share tends to drive a lot of the variation in the instrument. In these cases, one can conceptualize this IV as generating a comparison between manufacturing-heavy and nonmanufacturing-heavy local labor markets. Finally, depending on how it is implemented, the Bartik instrument may isolate variation in different components of labor demand depending on the skill composition of the workforce in the industry mix in the base year. For example, two local labor markets may be predicted to have similar employment growth because of the prevalence of retail and wholesale trade in one of them and the prevalence of business services in the other. In fact, the latter likely would have experienced a much greater outward shift in labor demand if measured in efficiency units terms, which may be the more appropriate quantity measure depending on the application.

Another common use of IV is to isolate exogenous variation in local labor supply. Following [Card \(2001\)](#), one common strategy for doing so is to make use of immigration shocks. As is discussed in more detail in [Chapter 10](#) by Lewis and Peri, this variation has been used extensively in the immigration literature as an instrument for the flow of immigrants to domestic local labor markets. This instrument is typically constructed by multiplying the fraction of immigrants to the United States from various regions of origin worldwide that reside in each metropolitan area in a base year with the total flow of immigrants into the United States from each region over some subsequent time period, and then summing over all regions of origin.¹⁷ As in [Lewis \(2011\)](#), an analogous exercise can be carried out by observed skill to generate variation across local labor markets in the relative supply of skill, though this exercise has a stronger first stage for less skilled groups.

¹⁷ As with Bartik instruments, some studies leave out the own local labor market or state when calculating national immigrant flows from each world region of origin.

Boustan (2010) uses a similar historical pathways instrument for the size of the African American population in northern metropolitan areas after World War II.

IV has also been widely used to isolate exogenous variation in infrastructure treatments. The commonest types of instruments used for transportation infrastructure variables are historical plans and networks. For example, Baum-Snow (2007) estimates the impacts of the construction of radial limited access highways serving central cities in US metropolitan areas on population decentralization. He finds that each radial highway emanating from a central city decentralized about 9% of the central city's population to the suburbs. He uses the highways laid out in a 1947 federal plan for a national highway system as a source of exogenous variation. The validity of this empirical strategy rests on the fact that the 1947 highway plan delineated routes that were chosen because they would facilitate military transportation and intercity trade. Local travel demand was not considered in making this highway plan. The 90% federal funding commitment for highway construction ensured that virtually all planned highways were built, with considerable additions to the interstate system to serve local travel demand. The primary analysis in Baum-Snow (2007) involves estimating 1950–1990 differenced regressions of the central city population on radial highways, controlling for metropolitan area population, in order to subsume the full time period during which the interstate system was constructed. Central to successful identification is to control for variables that may be correlated with planned highways and drive decentralization. Controls for central city size, 1950 metropolitan area population, and industrial structure in various specifications serve this purpose, though only the central city size control matters. Baum-Snow (2007) also reports estimates from a DD-type specification using data from decades between 1950 and 1990 and including metropolitan area and year fixed effects. For this empirical strategy, 1990 radial highways interacted with the fraction of federally funded mileage completed by the year of the observation enters as the highways instrument. Michaels (2008) uses a similar 1944 plan as an instrument for highways serving rural counties in his investigation of how better market integration changed the demand for skill. Though they turn out to be insufficiently strong, he also tries using the existence of nearby cities on the north–south or east–west axes relative to each county in question as instruments, since the interstate system is oriented in this way.

Duranton and Turner (2011, 2012) and Duranton et al. (2014) also use the 1947 plan as an instrument for highways, but supplement it with 1898 railroads and an index of continental exploration routes during the 1528–1850 period. These papers evaluate the effects of highways on the amount of intracity travel, urban growth, and the composition of interregional trade, respectively. Baum-Snow et al. (2014) similarly use aspects of historical urban road and railroad networks as an instrument for their modern counterparts in their investigation of changes in urban form in post-1990 Chinese cities. The idea of using historical infrastructure as instruments is that though such infrastructure is obsolete today, its rights of way are likely to be preserved, allowing for lower cost

modern construction. [Dinkelman \(2011\)](#) uses land gradient as an instrument for the prevalence of rural electrification in South Africa. She finds that much like new highways, electrification led to employment growth. As discussed further in [Chapter 20](#) by Redding and Turner in this handbook, how to distinguish between the effects of infrastructure on growth versus redistribution is still very much an open question. Whatever their interpretation, however, well identified IV regressions can recover some causal effects of infrastructure.

[Hoxby \(2000\)](#) is one of the earlier users of IV estimation in the local public finance literature. This paper attempts to recover the effects of public school competition, as measured by the number of public school districts in metropolitan areas, on student test scores. To account for the potential endogeneity of the number of school districts, Hoxby uses the prevalence of rivers and streams in the metropolitan area as an instrument. The idea is that metropolitan areas with more rivers and streams had more school districts because historically it was difficult for students to cross rivers to get to school, but these natural features do not directly influence levels or accumulation of human capital today. Potentially crucial for identification, of course, is to control for factors that might be correlated with rivers and streams but predict test scores. For example, metropolitan areas with more rivers and streams may be more likely to be located in more productive parts of the country such as the Northeast and Midwest, so controlling for parents' education and outcomes may be important.¹⁸ More recently, [Serrato et al. \(2014\)](#) have used city population revisions because of decennial censuses to isolate exogenous variation in federal transfers to recover that the local income multiplier is 1.57 per federal dollar and the fiscal cost per additional job is \$30,000 per year.

One additional common type of instrument uses variation in political power and incentives. For example, [Levitt \(1997\)](#) uses mayoral election cycles as an instrument for the number of police deployed in cities in a given month in his investigation of the effects of police on crime. The idea is that mayors up for reelection expand the police force during this time in an attempt to reduce crime. Consistent with the intuition of ILS, this study essentially compares crime rates during election cycles with those at other times, scaling by the difference in the numbers of police in these two environments. Of course, isolating a causal effect of police requires controlling for other policy changes implemented during election cycles.¹⁹ [Hanson \(2009\)](#) and [Hanson and Rohlin \(2011\)](#) use congressional representation on the Ways and Means Committee as an instrument for selection of proposed EZs for federal funding.

We hope that this incomplete survey of the use of IV in the urban and regional literature has shown that credible implementation of IV is far from a mechanical process. As with any empirical strategy, the successful use of IV requires careful thought about the

¹⁸ [Rothstein \(2007\)](#) provides additional analysis of the question using additional data.

¹⁹ See [McCrary \(2002\)](#) for a reanalysis of the same data set.

identifying variation at play. A convincing logical argument must be made for exogeneity of each instrument conditional on exogenous control variables, or equivalently that remaining variation in the instrument is uncorrelated with unobservables that drive the outcome of interest. In addition, ideally some idea should be given of which LATEs IV estimates using each instrument return.

One can use the mechanics of the IV estimator to recover TT in environments in which the treatment is explicitly randomized, as in the MTO studies discussed in [Section 1.2.4](#). [Katz et al. \(2001\)](#) walk through this process in detail. In the MTO context, assign $Z = 1$ to households in the Section 8 treatment group and $Z = 0$ to households in the control group. $D = 1$ if a household moves out of public housing with a Section 8 voucher and $D = 0$ if the household does not. One can think of Z as being a valid instrument for D . Households receiving a voucher choose whether or not to use it, making D endogenous. Recall from [Section 1.2.2](#) the definition of LATE, which in this binary treatment context becomes
$$\text{LATE} \equiv \frac{E[y|Z=1] - E[y|Z=0]}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}.$$
 The numerator is the coefficient on Z in a “reduced form” regression of y on Z . The denominator is the coefficient on Z in a “first-stage” regression of D on Z . That is, we see in this simple context how LATE is a restatement of the ILS IV estimator. Additionally, recall from [Section 1.2.2](#) the definition
$$\text{TT} \equiv E(y^1 - y^0 | D = 1) = \frac{E[y|Z=1] - E[y|Z=0]}{\Pr(D=1|Z=1)}.$$
 Therefore, $\text{TT} = \text{LATE}$ if $\Pr(D=1|Z=0) = 0$, or no members of the control group use a Section 8 voucher to move out of public housing.

It is also typical to use the IV estimator to implement the RD empirical strategy. The following section details how this is done.

1.6. REGRESSION DISCONTINUITY

Use of the RD research design in economics has dramatically increased in the past decade, as attested in recent reviews by [Lee and Lemieux \(2010\)](#) and [Imbens and Lemieux \(2008\)](#). Our interpretation of RD estimates has also changed in this period. Initially thought of as another method to deal with selection on observables, RD was subsequently motivated as a type of local IV, and then finally defined as a creative way of implementing random assignment in a nonexperimental setting. In this section, we discuss the different interpretations of the RD framework, the relevant details on how to implement the approach, and some of its notable uses in urban and regional economics. Even though RD designs have been quite rare in urban economics papers until recently,²⁰ the approach shows much promise for future research, and we expect its use in urban economics to grow over time in the same way experienced by other applied economics fields. This section can be thought of as a first gateway to the approach; more detailed discussions are presented in [Lee and Lemieux \(2010\)](#) and [Imbens and Lemieux \(2008\)](#).

²⁰ For example, zero papers used the RD design as recently as 2010 in the *Journal of Urban Economics*.

1.6.1 Basic framework and interpretation

There are two main prerequisites for RD to apply as a potential empirical strategy. First, the researcher needs to know the selection into treatment rule, and there should be a discontinuity in how the treatment is assigned. For example, US cities often promote referenda that ask local citizens if they would approve raising extra funds through bond issuances that will be used to invest in local infrastructure. The selection rule in this case is based on the vote share needed to approve the bond issue, let us say two-thirds of the local vote. The discontinuity in treatment is obvious: cities whose referenda got less than two-thirds of the votes will not raise the funds, while cities whose referenda achieved the two-thirds mark will be able to issue the bonds and subsequently invest the proceeds in local infrastructure. The second prerequisite is that agents are not able to sort across the selection threshold. Such “selection” would by definition invalidate the ability to compare similar individuals in the control and treatment groups on either side of the threshold. In the referenda example, this no endogenous sorting condition means that cities are not able to manipulate the referendum in order to influence their ability to get one additional vote to reach the two-thirds threshold. At the end of the section we will discuss how researchers can potentially deal with violations of this condition, such as in boundary-type applications in which sorting is expected to happen over time.

If both conditions above are met, the RD estimate will provide a comparison of individuals in treatment and control groups that were “matched” on a single index—that is, the selection rule. This single index is usually referred to as the running variable or the assignment variable.

To formalize those concepts, define y_i as the outcome of interest and T_i as the relevant binary treatment status, and assume $\beta_i = \beta$ and X_i is a vector of covariates:

$$y_i = \alpha + T_i\beta + X_i\delta + U_i + e_i, \quad (1.15)$$

where $T_i = 1(Z_i \geq z_0)$. Z_i is the single index for selection into treatment, and z_0 is the discontinuity threshold. Individuals with $Z_i \geq z_0$ are assigned to the treatment group, while the remaining individuals are assigned to the control group. Such a setup is usually referred to as the “sharp” RD design because there is no ambiguity about treatment status given the known and deterministic selection rule. In this setting, the ATE of T_i on y_i around the threshold is

$$\begin{aligned} E[y_i|Z_i = z_0 + \Delta] - E[y_i|Z_i = z_0 - \Delta] &= \beta + \{E[X_i\delta|Z_i = z_0 + \Delta] - E[X_i\delta|Z_i = z_0 - \Delta]\} \\ &\quad + \{E[U_i + e_i|Z_i = z_0 + \Delta] - E[U_i + e_i|Z_i = z_0 - \Delta]\}. \end{aligned}$$

Note that this ATE applies only to the agents with characteristics of those near the threshold. Two key assumptions allow for the identification of ATE. First, continuity of the joint distribution of X_i and Z_i . This assumption makes the term $\{E[X_i\delta|Z_i = z_0 + \Delta] - E[X_i\delta|Z_i = z_0 - \Delta]\}$ in the equation above negligible, and guarantees that both the control group and the treatment group will have similar observed characteristics

around the discontinuity threshold. This assumption is easily tested in the data, and it is one of the reasons for interpreting RD as a selection on observables type of framework. The second assumption is that the joint distribution of the unobserved component $(U_i + e_i)$ and Z_i is continuous, which makes the term $\{E[U_i + e_i|Z_i = z_0 + \Delta] - E[U_i + e_i|Z_i = z_0 - \Delta]\}$ also negligible. This assumption can never be tested. This type of sharp RD is analogous to random assignment in the sense that, around the threshold, the assignment of individuals to control and treatment groups is exogenous given the two assumptions above.

In some circumstances, however, the selection rule may not be deterministic. For example, even when local citizens approve a bond issue, overall market conditions may prevent the municipality from raising the funds. Or US cities in which a bond referendum failed today may try to pass other bond measures in the near future. Those events may turn the selection rule into a probabilistic equation, leading to the so-called fuzzy RD design. Formally, the treatment status T_i can be rewritten as

$$T_i = \theta_0 + \theta_1 G_i + u_i,$$

where $G_i = 1(Z_i \geq z_0)$, and u_i corresponds to the other unobserved components that determine treatment status. Plugging in the new equations for T_i and G_i in the outcome equation generates

$$y_i = \alpha + \beta\theta_0 + G_i\beta\theta_1 + u_i\beta + X_i\delta + U_i + e_i,$$

and the new treatment effect around the threshold becomes

$$\begin{aligned} E[y_i|Z_i = z_0 + \Delta] - E[y_i|Z_i = z_0 - \Delta] &= \beta\theta_1 + \beta\{E[u_i|Z_i = z_0 + \Delta] - E[u_i|Z_i = z_0 - \Delta]\} \\ &+ \{E[X_i\delta|Z_i = z_0 + \Delta] - E[X_i\delta|Z_i = z_0 - \Delta]\} + \{E[U_i + e_i|Z_i = z_0 + \Delta] \\ &- E[U_i + e_i|Z_i = z_0 - \Delta]\}. \end{aligned}$$

In order to estimate the parameter β we first need to back out the parameter θ_1 , which establishes the relationship between G_i and T_i ,

$$E[T_i|Z_i = z_0 + \Delta] - E[T_i|Z_i = z_0 - \Delta] = \theta_1 + \{E[u_i|Z_i = z_0 + \Delta] - E[u_i|Z_i = z_0 - \Delta]\},$$

and a LATE can be recovered using the ratio of the reduced form impact of the single index Z_i on outcome y_i , and of the first stage described above:

$$\beta = \frac{E[y_i|Z_i = z_0 + \Delta] - E[y_i|Z_i = z_0 - \Delta]}{E[T_i|Z_i = z_0 + \Delta] - E[T_i|Z_i = z_0 - \Delta]}. \quad (1.16)$$

This expression closely resembles the definition of LATE in (1.3). The reason the fuzzy RD design can be thought of as delivering a LATE is that the treatment effect is recovered only for some agents. If the set of agents induced into treatment by having an assignment variable value that is beyond the critical threshold is random, then this coincides with the same ATE estimated in the sharp RD environment. However, if the fuzzy RD occurs

because a group of agents do not comply with the “treatment” of being beyond the threshold, presumably because they differ from compliers on some observables or unobservables, then the fuzzy RD design allows the researcher to recover only a LATE, which can also be thought of as a particular version of treatment on the treated (TT).

The validity of the fuzzy RD design relies on the following assumptions: (1) there is random assignment of control and treatment groups around the threshold; (2) there is a strong first stage, allowing the estimation of θ_1 ; (3) there is an exclusion restriction, so that the term $\{E[u_i|Z_i = z_0 + \Delta] - E[u_i|Z_i = z_0 - \Delta]\}$ also becomes negligible.²¹ This setup is very similar to the IV approach covered in the previous section, and the fuzzy RD is sometimes interpreted as a local IV.

As emphasized in DiNardo and Lee (2011), the simplistic IV interpretation misses the most important characteristic of the RD design: the random assignment of treatment and control groups. Even though the fuzzy design resembles the mechanics of an IV approach, the key characteristic of the design is the ability of mimicking random assignment in a nonexperimental setting. In fact, the fuzzy RD design could be more properly designated as a locally randomized IV.

An important issue in RD designs is external validity, as one potential interpretation of the approach is that “it only estimates treatment effects for those individuals close to the threshold.” DiNardo and Lee (2011) clarify the interpretation of those estimates by using the idea that individuals do not get to choose where they locate with respect to the RD threshold. If that is the case, RD estimates can be viewed as a weighted average effect, where the weights are proportional to the ex ante likelihood that the value of the individual’s assignment variable would lie in a neighborhood of the threshold.

Independent of using a sharp or fuzzy design, the RD approach provides a method of approximating the empirical estimation to a randomization setting. As discussed in earlier sections, randomization is the Holy Grail of empirical work, and any method that allows nonexperimental approaches to replicate the characteristics of a experimental design is bound to be welcomed by researchers.

1.6.2 Implementation

The popularity of the RD approach is explained not only by its relationship with randomized experiments, but also because of the transparency of the framework. RD estimation can be transparently shown in a graphical format. The standard RD figure plots conditional or unconditional means of the treatment and/or outcome of interest by bins of the assignment variable. Following the bond issue example, Cellini et al. (2010) show average expenditures and average capital outlays per pupil by the vote share in a bond referendum (see Fig. 1.3). This simple figure first shows that a treatment

²¹ This approach also relies on a monotonicity assumption, similar to the one used to cleanly interpret LATE in an IV setting. It means that as one moves across the assignment variable threshold, the probability of treatment for every combination of observables X and unobservables U increases.

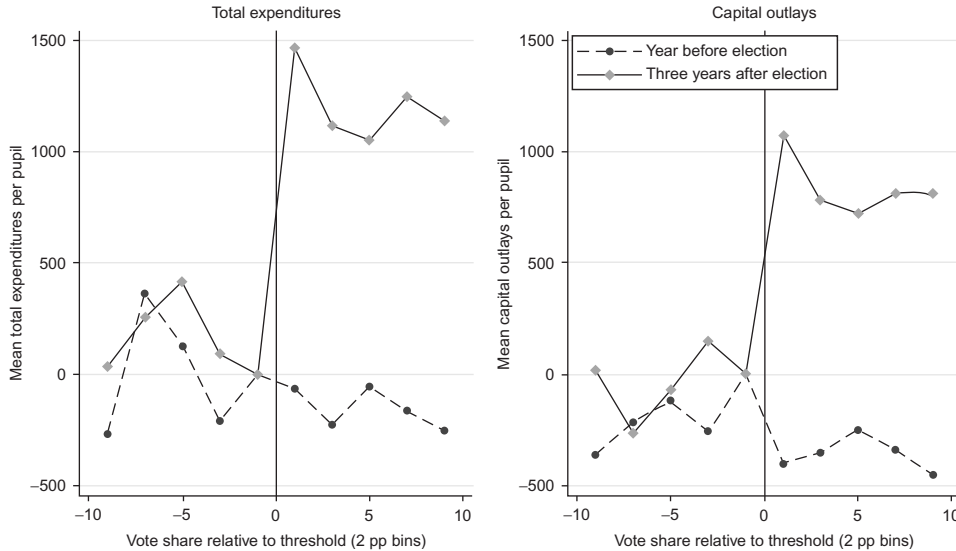


Figure 1.3 Total spending and capital outlays per Pupil, by vote share, 1 year before and 3 years after Election (Cellini et al., 2010). Graph shows average total expenditures (left panel) and capital outlays (right panel) per pupil, by the vote share in the focal bond election. Focal elections are grouped into bins 2 percentage points wide: measures that passed by between 0.001% and 2% are assigned to the 1 bin; those that failed by similar margins are assigned to the -1 bin. Averages are conditional on year fixed effects and the -1 bin is normalized to zero.

exists: total expenditures and capital outlays increased for school districts that had vote shares above the threshold, and only in the 3 years after the bond measure was approved. It also tests the sharpness of the research design: school districts whose referenda had vote shares below the threshold had similar expenditures and capital outlays in the year before and in the 3 years after the referendum. The combination of these results for treatment and control groups is a clear discontinuity of a given magnitude around the threshold.

A similar graphical approach should be used to test the validity of the research design. All relevant covariates should be displayed in unconditional plots by bins of the assignment variable, and the statistical test of a discontinuity for each covariate should be presented. This is the main test of the assumption that control and treatment groups have balanced characteristics around the discontinuity threshold. An additional test of sorting around the discontinuity can be performed by plotting the total number of observations in each bin against the running variable. That will test whether there is a disproportional number of individuals on each side of the threshold, which could potentially indicate the ability of individuals to manipulate their treatment status and therefore invalidate the research design—see McCrary (2008). In practice though, such sorting would usually show up as differences in other covariates as well. Finally, other common robustness tests, including testing for a discontinuity in predetermined covariates (in the case of a

treatment that has a time component), testing if the outcome variable presents a discontinuity at a fake discontinuity threshold, meaning that a discontinuity only happens at the true threshold, and testing whether other unrelated outcomes, have a similarly discontinuous relationship with the running variable, which would indicate that the treatment may not be the only mechanism impacting outcomes.

Many RD applications also plot parametric or nonparametric estimates of the ATE along the unconditional means of the assignment variable. When a parametric estimate is used, the graphical analysis can also help with the choice of the functional form for the RD single index. As mentioned earlier, the assignment variable Z_i can be interpreted as a single index of the sources of observed bias in the relationship between outcome and treatment status. If the single index is smooth at the RD threshold z_0 , that indicates that any discontinuity in y_i would be due to T_i . In the easiest case, there is no correlation between the outcome y_i conditional on treatment status and the running variable Z_i , and a simple regression such as $y_i = \alpha_0 + T_i\beta + \epsilon_i$ would generate proper estimates of the ATE. A commoner situation is where y_i is also some function of Z_i , with similar slopes on either side of the threshold. A more general empirical model that allows for different functions of Z_i above and below z_0 which is commonly used to implement sharp RD estimation is

$$y_i = \alpha_0 + T_i\alpha_1 + f_1(z_0 - Z_i)1(Z_i < z_0) + f_2(Z_i - z_0)1(Z_i \geq z_0) + X_i\delta + \epsilon_i, \quad (1.17)$$

where $T_i = 1(Z_i \geq z_0)$ in the sharp RD case. Many researchers implement $f_1(\cdot)$ and $f_2(\cdot)$ as cubic or quadratic polynomials with estimated coefficients, imposing the constraints that $f_1(0) = f_2(0) = 0$ by excluding intercept terms from the polynomials. The inclusion of α_0 in (1.17) allows the level of y_0 at $Z = z_0 - \Delta$ to be nonzero. This equation can be estimated by OLS. The underlying idea, again, is to compare treatment and control units near the threshold z_0 . The role of the $f_1(\cdot)$ and $f_2(\cdot)$ control functions in (1.17) is to control for (continuous) trends in observables and unobservables moving away from the assignment variable threshold. Though not necessary if the RD empirical strategy is sound, it is common to additionally control for observables X in order to reduce the variance of the error term and more precisely estimate α_1 . As with our discussion of including observables in the DD estimators, it is important not to include any observables that may respond to the treatment, meaning they are endogenous. Moreover, it is common not to utilize data beyond a certain distance from the threshold z_0 for estimation because such observations do not contribute to identification yet they can influence parametric estimates of the control functions.

The empirical model in (1.17) can also be used as a basis for estimating a LATE in environments that lend themselves to using a fuzzy RD research design. Here, however, the researcher must also consider the following auxiliary treatment equation:

$$T_i = \gamma_0 + D_i\rho + g_1(z_0 - Z_i)1(Z_i < z_0) + g_2(Z_i - z_0)1(Z_i \geq z_0) + X_i\nu + u_i,$$

where $D_i = 1(Z_i \geq z_0)$, and T_i in (1.17) is simply a treatment indicator. As this is now a simultaneous equations model, the fuzzy RD LATE can thus be estimated using any IV estimator. Commensurate with (1.16), the ILS estimate of the fuzzy RD LATE is $\frac{\alpha}{\hat{\rho}}$.

Nonparametric estimation can also be used to recover the ATE at the discontinuity threshold—see [Hahn et al. \(2001\)](#). The randomization nature of the RD design implies that most estimation methods should lead to similar conclusions. If ATE estimates from different methods diverge, that is usually a symptom of a more fundamental problem, such as a small number of observations near z_0 . In fact, the main practical limitation of nonparametric methods is that they require a large number of observations near the threshold, especially since nonparametric estimators are quite sensitive to bandwidth choice at boundaries.

To this point, we have assumed that we know the critical value z_0 of the assignment variable at which there is a discontinuous change in treatment probability. In some contexts, that critical value is unknown. It is possible to estimate the “structural break” z_0 jointly with the treatment effect at z_0 . This can be done by estimating (1.17) by OLS for every candidate z_0 , and then choosing the $\hat{z}_0$ that maximizes R^2 . The work of [Card David and Rothstein \(2008\)](#) is one notable example in the urban economics literature that carries out this procedure. This paper recovers estimates of the critical fraction of the population that is black in neighborhoods at which they “tip,” meaning they lose a large number of white residents. Jointly estimated with these tipping points are the magnitudes of this tipping.

1.6.3 Examples of RD in urban economics

There are various examples of RD applications in urban economics. [Ferreira and Gyourko \(2009\)](#) study the impacts of local politics on fiscal outcomes of US cities. [Chay and Greenstone \(2005\)](#) recover hedonic estimates of willingness to pay for air quality improvements in US counties. [Baum-Snow and Marion \(2009\)](#) estimate the impacts of low income housing subsidies on surrounding neighborhoods. [Ferreira \(2010\)](#) studies the impact of property taxes on residential mobility, and [Pence \(2006\)](#) studies the impact of mortgage credit laws on loan size. In this subsection we first discuss the bond referenda example that was mentioned above in detail. We then discuss the use of the “boundary discontinuity” research design, which is a particular application of RD that comes with its own challenges.

[Cellini et al. \(2010\)](#) investigate the importance of capital spending in education. There are two central barriers to identification in this setting. First, resources may be endogenous to local outcomes. Spending is usually correlated with the socioeconomic status of students. Second, even causal estimates of the impact of school investments may not be able to capture all measured benefits to students, such as nonacademic benefits. To deal with this second issue, they look at housing markets. Given standard theory ([Oates, 1969](#)), if home buyers value a local project more than they value the taxes they

pay to finance it, spending increases should lead to higher housing prices—also implying that the initial tax rate was inefficiently low.

In order to isolate exogenous variation in school investments, they create control and treatment groups based on school districts in California that had very close bond referenda. The logic is that a district where the proposal for a bond passes by one vote is likely to be similar to one where the proposal fails by the same margin. They test and confirm this assumption using three methods: they show that control and treatment groups have balanced covariates around the margin of victory threshold, they show that the prebond outcomes and trends of those outcomes are also balanced, and they show that the distribution of bond measures by vote share is not discontinuous around the threshold.

They also test whether the design is sharp or fuzzy by looking at the future behavior of districts after a bond referendum. Districts in which a bond referendum failed were more likely to pass and approve another bond measure within the next 5 years. The authors deal with the dynamic nature of bond referenda by developing two estimators of ITT and TT. The estimates indicate that passage of a bond measure causes house prices to rise by about 6%, with this effect appearing gradually over 2–3 years following the referendum, and the effect persists for about a decade. Finally, the authors convert their preferred TT estimates of the impact of bond passage on investments and prices into the willingness to pay for marginal home buyers. They find a marginal willingness to pay of \$1.50 or more for each \$1 of capital spending. Even though several papers in the public choice literature emphasize the potential for “Leviathan” governments, those estimates suggest the opposite for this California case.

We now consider the boundary discontinuity research design. Many researchers have used geographic boundaries to construct more comparable treatment and control groups that are likely to mitigate omitted variable biases. [Holmes \(1998\)](#), for example, aspires to disentangle the effects of state policies from other state-specific characteristics. As discussed in [Section 1.4.2](#), a DD approach is often less than ideal when applied to large geographic areas such as states. Holmes’s strategy is to zoom in on state borders at which one state has right-to-work laws and the other state does not. Geography, climate, fertility of soil, access to raw materials, and access to rivers, ports, etc., may be the same for cities on either side of the border. Such a design thus mitigates potential biases arising from differences in omitted factors. Looking across these borders, [Holmes \(1998\)](#) finds that manufacturing activity is much higher on the “probusiness” sides of the borders.

But borders are usually not randomly assigned. They may follow certain geographic features, such as rivers, or they may be the result of a political process, such as when states choose boundaries for congressional districts. The lack of randomization implies that there might be more than one factor that is not similar across geographic areas separated by boundaries. For example, some boundaries may be used to separate multiple jurisdictions, such as cities, school districts, counties, states, and perhaps countries. Even if

borders were randomly assigned, there is ample opportunity for sorting of agents or policies across borders on unobservable characteristics.

These issues can be illustrated in the example of valuation of school quality. [Black \(1999\)](#) compares house prices on either side of school attendance boundaries in order to estimate valuation of school quality on the high-quality side versus the low-quality side. Attendance zones rather than school district boundaries are used because no other local service provision is different on either side of these boundaries. School district boundaries would have two problems: they may also be city or county boundaries, and different districts may have very different systems of school financing. School attendance zones, on the other hand, have similar financing systems, and are unlikely to be used to separate other types of jurisdictions. Black also shows that the distance to the boundary matters. Only small distances, within 0.2 miles, are likely to guarantee similarity in local features.

However, even those precise local attendance zones may not deal with the issue of endogenous sorting of families. Given a discontinuity in local school quality at the boundary, one might expect that residential sorting would lead to discontinuities in the characteristics of the households living on opposite sides of the same boundary—even when the housing stock was initially identical on both sides. [Bayer et al. \(2007\)](#) empirically report those discontinuities for the case of the San Francisco Bay Area. High income, high education level, and white households are more likely to be concentrated on the high school quality side of the attendance zone boundaries. Those differences are noticeable even within very small distances to the boundary. Given these sorting patterns, it becomes important to control for neighborhood demographic characteristics when estimating the value of school quality, since the house price differences may reflect the discontinuities in school quality and also the discontinuities in sociodemographics. As in [Black \(1999\)](#), [Bayer et al. \(2007\)](#) find that including boundary fixed effects in standard hedonic regressions reduces the estimated valuation of school quality. But they also find that such valuation is reduced even further, by approximately 50%, when precise sociodemographic characteristics are added.

Additional caveats are that even the best data sets will not have all of the sociodemographic characteristics that may influence house prices. Also, most data sets have limited information about detailed characteristics of houses, such as type of floor and views. Biases may arise if such unobserved housing features or unobserved demographic characteristics differ across boundaries used for identification. These problems could be mitigated in settings where boundaries were recently randomly assigned, and therefore families or firms still did not have enough time to re-sort.

In another use of the boundary discontinuity empirical setup, [Turner et al. \(2014\)](#) examine land prices across municipal borders to decompose the welfare consequences of land use regulation into own lot, external, and supply components. The idea is that as long as land use regulation is enforced evenly over space up to municipal borders, one can recover the direct costs of regulation by comparing across borders. Indirect

(spillover) costs of regulation can be found with a spatial differencing type estimator within jurisdictions adjacent to those with regulatory changes. Supply effects of regulation are reflected in differences across municipal borders in the share of land that is developed. Results indicate strong negative effects of land use regulations on the value of land and welfare that operate through all three channels.

Recent developments in labor economics and public finance have also uncovered many discontinuities in slopes, using the so-called regression kink (RK) design ([Card David and Weber, 2012](#)). These kinks are a common feature of many policy rules, such as the formulas that establish the value of unemployment insurance benefits as a function of previous earnings. Card et al. explain that the basic intuition of the RK design is similar to that of the RD design and is based on a comparison of the relationship between the outcome variable (e.g., duration of unemployment) and the treatment variable (e.g., unemployment benefit levels) at the point of the policy kink. However, in contrast to an RD design, which compares the levels of the outcome and treatment variables, the estimated causal effect in an RK design is given by the ratio of the changes in the slope of the outcome and treatment variables at the kink point. As with RD, one threat to identification is sorting at the kink. This type of sorting often results in visible bunching in the distribution of the running variable at the kink point and invalidates the assumptions underlying the RK design. However, though such bunching may invalidate RD and RK designs, many researchers in public economics—such as [Saez \(2010\)](#) and [Chetty et al. \(2011\)](#)—have been able to leverage this type of bunching to recover estimates of the behavioral responses to various public policies such as income taxes. The idea in such “bunching designs” is to compare the actual bunching observed in the data with the predictions from a behavioral model that does not have the policy kink. Assuming everything else is constant, any differences between the amount of bunching observed in the data and the amount that would be implied by the model in the absence of the policy kink can be attributed directly to the policy variation around the kink. Recent applications of this approach to housing markets include [Best and Kleven \(2014\)](#), [Kopczuk and Munroe \(2014\)](#), and [De Fusco and Pacioret \(2014\)](#).

Finally, in some situations one may observe both an RD and an RK at the same threshold—see [Turner \(2012\)](#). New developments in these areas of research may arise in the coming years, as researchers strive to understand the underlying sources of variation in the data that allow for identification of treatment effects that are difficult to credibly estimate with nonexperimental data.

1.7. CONCLUSION

This chapter has laid out some best practices for recovering causal empirical relationships in urban and regional economics contexts. We hope that we have successfully conveyed the idea that carrying out quality empirical work requires creativity and careful thought.

Beyond basic decisions about the general empirical strategy to be used are always many smaller decisions that are inherently particular to the question at hand and available data. In general, however, two central considerations should permeate all empirical work that aspires to recover causal relationships in data. The first is to consider the sources of variation in treatment variables that identify these relationships of interest. The second is to recognize which treatment effect, if any, is being estimated.

We see a bright future for empirical research in urban and regional economics. The wide integration of tractable economic theory and empirical inquiry among those working on urban and regional questions in economics positions our field well to make convincing progress on important questions. The wide range of detailed spatially indexed data available to us provides many opportunities for the beginnings of serious investigations of new topics. Indeed, while recovery of treatment effects is important, a descriptive understanding of important patterns in the data is perhaps more important for new questions. Particularly in our field, which is finding itself overwhelmed with newly available data, the first step should always be to get a handle on the facts. Doing so often leads to ideas about convincing identification strategies that can be used to recover causal relationships of interest.

REFERENCES

- Abadie, A., Angrist, J., Imbens, G., 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70, 91–117.
- Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: estimating the effect of california's tobacco control program. *J. Am. Stat. Assoc.* 105, 493–505.
- Abadie, A., Diamond, A., Hainmueller, J., 2014. Comparative politics and the synthetic control method. *Am. J. Polit. Sci.* (Online, forthcoming).
- Abadie, A., Gardeazabal, J., 2003. The economic costs of conflict: a case study of the basque country. *Am. Econ. Rev.* 93, 113–132.
- Alesina, A., Baqir, R., Hoxby, C., 2004. Political jurisdictions in heterogeneous communities. *J. Polit. Econ.* 112, 348–396.
- Altonji, J., Elder, T., Taber, C., 2005. Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. *J. Polit. Econ.* 113, 151–184.
- Angrist, J., Graddy, K., Imbens, G., 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Rev. Econ. Stud.* 67, 499–527.
- Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. *Rev. Econ. Stat.* 60, 47–57.
- Athey, S., Imbens, G., 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74, 431–497.
- Autor, D., Palmer, C., Pathak, P., 2014. Housing market spillovers: evidence from the end of rent control in Cambridge Massachusetts. *J. Polit. Econ.* 122, 661–717.
- Bailey, M., Muth, R., Nourse, H., 1963. A regression method for real estate price index construction. *J. Am. Stat. Assoc.* 58, 933–942.
- Bartik, T., 1991. Who Benefits from State and Local Economic Development Policies? Upjohn Institute, Kalamzoo, MI.
- Baum-Snow, N., 2007. Did highways cause suburbanization? *Q. J. Econ.* 122, 775–805.
- Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M., Zhang, Q., 2014. Roads, Railroads and Decentralization of Chinese Cities (manuscript).

- Baum-Snow, N., Lutz, B., 2011. School desegregation, school choice and changes in residential location patterns by race. *Am. Econ. Rev.* 101, 3019–3046.
- Baum-Snow, N., Marion, J., 2009. The effects of low income housing tax credit developments on neighborhoods. *J. Publ. Econ.* 93, 654–666.
- Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage gap. *Rev. Econ. Stud.* 79, 88–127.
- Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. *J. Polit. Econ.* 115, 588–638.
- Bayer, P., Hjalmarsson, R., Pozen, D., 2009. Building criminal capital behind bars: peer effects in juvenile corrections. *Q. J. Econ.* 124, 105–147.
- Bayer, P., Ross, S., Topa, G., 2008. Place of work and place of residence: informal hiring networks and labor market outcomes. *J. Polit. Econ.* 116, 1150–1196.
- Beaudry, P., Green, D., Sand, B., 2014. Spatial equilibrium with unemployment and wage bargaining: theory and estimation. *J. Urban Econ.* 79, 2–19.
- Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? *Q. J. Econ.* 119, 249–275.
- Best, M.C., Kleven, H.J., 2014. Housing Market Responses to Transaction Taxes: Evidence from Notches and Stimulus in the UK. Mimeo.
- Bester, A., Conley, T., Hansen, C., 2011. Inference with dependent data using cluster covariance estimators. *J. Econometr.* 165, 137–151.
- Bjorklund, A., Moffitt, R., 1987. The estimation of wage gains and welfare gains in self-selection models. *Rev. Econ. Stat.* 69, 42–49.
- Black, S., 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 114, 577–599.
- Blanchard, O.J., Katz, L.F., 1992. Regional evolutions. *Brook. Pap. Econ. Act.* 1, 1–69.
- Bound, J., Holzer, H.J., 2000. Demand shifts, population adjustments and labor market outcomes during the 1980's. *J. Labor Econ.* 18, 20–54.
- Boustan, L., Ferreira, F., Winkler, H., Zolt, E.M., 2013. The effect of income inequality on taxation and public expenditures: evidence from U.S. municipalities and school districts, 1970–2000. *Rev. Econ. Stat.* 95, 1291–1302.
- Boustan, L.P., 2010. Was postwar suburbanization “white flight”? Evidence from the black migration. *Q. J. Econ.* 125, 417–443.
- Busso, M., Gregory, J., Kline, P., 2013. Assessing the incidence and efficiency of a prominent place based policy. *Am. Econ. Rev.* 103, 897–947.
- Cameron, A.C., Gelbach, J.B., Miller, D.L., 2008. Bootstrap-based improvements for inference with clustered errors. *Rev. Econ. Stat.* 90, 414–427.
- Campbell, J., Giglio, S., Pathak, P., 2011. Forced sales and house prices. *Am. Econ. Rev.* 101, 2108–2131.
- Card, D., 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *J. Labor Econ.* 19, 22–64.
- Card David, A.M., Rothstein, J., 2008. Tipping and the dynamics of segregation. *Q. J. Econ.* 123, 177–218.
- Card David, David Lee, Z.P., Weber, A., 2012. Nonlinear policy rules and the identification and estimation of causal effects in a generalized regression kink design, NBER Working paper No. 18564.
- Carrell, S., Sacerdote, B., West, J., 2013. From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica* 81, 855–882.
- Case, K., Shiller, R., 1987. Prices of Single Family Homes Since 1970: New Indexes for Four Cities. *New England Economic Review*, Boston, MA September/October.
- Case, K., Shiller, R., 1989. The efficiency of the market for single-family homes. *Am. Econ. Rev.* 79, 125–137.
- Cellini, S., Ferreira, F., Rothstein, J., 2010. The value of school facility investments: evidence from a dynamic regression discontinuity design. *Q. J. Econ.* 125, 215–261.
- Chaney, T., Sraer, D., Thesmar, D., 2012. The collateral channel: how real estate shocks affect corporate investment. *Am. Econ. Rev.* 102, 2381–2409.
- Chay, K., Greenstone, M., 2005. Does air quality matter? Evidence from the housing market. *J. Polit. Econ.* 113, 376–424.

- Chetty, R., Friedman, J.N., Hilger, N., Saez, E., Schanzenbach, D., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. *Q. J. Econ.* 126, 1593–1660.
- Combes, P.P., Duranton, G., Gobillon, L., 2008. Spatial wage disparities: sorting matters! *J. Urban Econ.* 63, 723–742.
- Combes, P.P., Duranton, G., Gobillon, L., Roux, S., 2012. Sorting and local wage and skill distributions in France. *Reg. Sci. Urban Econ.* 42, 913–930.
- Costa, D., Kahn, M., 2000. Power couples: changes in the locational choice of the college educated, 1940–1990. *Q. J. Econ.* 115, 1287–1315.
- Cox, D.R., 1958. Some problems connected with statistical inference. *Ann. Math. Stat.* 29, 357–372.
- De La Roca, J., Puga, D., 2014. Learning by Working in Big Cities (manuscript).
- Dehejia, R., Wahba, S., 2002. Propensity score-matching methods for nonexperimental causal studies. *Rev. Econ. Stat.* 84, 151–161.
- Diamond, R., 2013. The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000 (manuscript).
- DiNardo, J., Lee, D., 2011. Program evaluation and research designs. In: Orley, A., David, C. (Eds.), *Handbook of Labor Economics. Part A, Vol 4*. Elsevier, Amsterdam, pp. 463–536.
- Dinkelman, T., 2011. The effects of rural electrification on employment: new evidence from South Africa. *Am. Econ. Rev.* 101, 3078–3108.
- Duflo, E., Glennerster, R., Kremer, M., 2008. Using randomization in development economics research: A toolkit. In: Srinivasan, T.N., Behrman, J. (Eds.), *Handbook of Development Economics. Volume 4*. Elsevier, Amsterdam, pp. 3895–3962.
- Duranton, G., Morrow, P., Turner, M.A., 2014. Roads and trade: evidence from the U.S. *Rev. Econ. Stud.* 81, 681–724.
- Duranton, G., Turner, M., 2011. The fundamental law of road congestion: evidence from the US. *Am. Econ. Rev.* 101, 2616–2652.
- Duranton, G., Turner, M., 2012. Urban growth and transportation. *Rev. Econ. Stud.* 79, 1407–1440.
- Efron, B., Tibshirani, R., 1994. *An Introduction to the Bootstrap*. Monograph in Applied Statistics and Probability, No 57, Chapman & Hall, New York, NY.
- Ellen, I., Laco, J., Sharygin, C., 2013. Do foreclosures cause crime? *J. Urban Econ.* 74, 59–70.
- Epple, D., Platt, G., 1998. Equilibrium and local redistribution in an urban economy when households differ in both preferences and incomes. *J. Urban Econ.* 43, 23–51.
- Ferreira, F., 2010. You can take it with you: proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities. *J. Publ. Econ.* 94, 661–673.
- Ferreira, F., Gyourko, J., 2009. Do political parties matter? Evidence from U.S. cities. *Q. J. Econ.* 124, 399–422.
- Field, E., 2007. Entitled to work: urban property rights and labor supply in Peru. *Q. J. Econ.* 122, 1561–1602.
- Figlio, D., Lucas, M., 2004. What's in a grade? School report cards and the housing market. *Am. Econ. Rev.* 94, 591–605.
- Freedman, M., 2014. Tax Incentives and Housing Investment in Low Income Neighborhoods (manuscript).
- Fusco, De, Anthony, A., Paciorek, A., 2014. The interest rate elasticity of mortgage demand: evidence from bunching at the conforming loan limit. *Fin. Econ. Disc. Ser.* 2014–11.
- Galiani, S., Gertler, P., Cooper, R., Martinez, S., Ross, A., Undurraga, R., 2013. Shelter from the Storm: Upgrading Housing Infrastructure in Latin American Slums. NBER Working paper 19322.
- Galiani, S., Murphy, A., Pantano, J., 2012. Estimating Neighborhood Choice Models: Lessons from a Housing Assistance Experiment (manuscript).
- Gibbons, C., Serrato, J.C.S., Urbancic, M., 2013. Broken or Fixed Effects? Working paper.
- Glaeser, E., Hedi Kallal, J.S., Shleifer, A., 1992. Growth in cities. *J. Polit. Econ.* 100, 1126–1152.
- Glaeser, E., Maré, D., 2001. Cities and skills. *J. Labor Econ.* 19, 316–342.
- Gobillon, L., Magnac, T., Selod, H., 2012. Do unemployed workers benefit from enterprise zones? The French experience. *J. Publ. Econ.* 96, 881–892.
- Gould, E., Weinberg, B., Mustard, D., 2002. Crime rates and local labor market opportunities in the United States: 1979–1997. *Rev. Econ. Stat.* 84, 45–61.

- Graddy, K., 1995. Testing for imperfect competition at the fulton fish market. *Rand J. Econ.* 26, 75–92.
- Graham, B., 2008. Identifying social interactions through conditional variance restrictions. *Econometrica* 76, 643–660.
- Greenstone, M., Gallagher, J., 2008. Does hazardous waste matter? Evidence from the housing market and the superfund program. *Q. J. Econ.* 123, 951–1003.
- Greenstone, M., Hornbeck, R., Moretti, E., 2010. Identifying agglomeration spillovers: evidence from winners and losers of large plant openings. *J. Polit. Econ.* 118, 536–598.
- Gronau, R., 1974. Wage comparisons. a selectivity bias. *J. Polit. Econ.* 82, 1119–1143.
- Hahn, J., Todd, P., van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69, 201–209.
- Ham, J., Swenson, C., Imbroglu, A., Song, H., 2011. Government programs can improve local labor markets: evidence from state enterprise zones, federal empowerment zones and federal enterprise community. *J. Publ. Econ.* 95, 779–797.
- Hanson, A., 2009. Local employment, poverty, and property value effects of geographically-targeted tax incentives: an instrumental variables approach. *Reg. Sci. Urban Econ.* 39, 721–731.
- Hanson, A., Rohlin, S., 2011. The effect of location based tax incentives on establishment location and employment across industry sectors. *Publ. Financ. Rev.* 39, 195–225.
- Heckman, J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–162.
- Heckman, J., Honoré, B., 1990. The empirical content of the roy model. *Econometrica* 58, 1121–1149.
- Heckman, J., Navarro-Lozano, S., 2004. Using matching, instrumental variables, and control functions to estimate economic choice models. *Rev. Econ. Stat.* 86, 30–57.
- Heckman, J., Urzua, S., Vytlacil, E., 2006. Understanding instrumental variables in models with essential heterogeneity. *Rev. Econ. Stat.* 88, 389–432.
- Heckman, J., Vytlacil, E., 2005. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73, 669–738.
- Henderson, V., Kuncoro, A., Turner, M., 1995. Industrial development in cities. *J. Polit. Econ.* 103, 1067–1090.
- Holland, P., 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81, 945–960.
- Holmes, T., 1998. The effects of state policies on the location of industry: evidence from state borders. *J. Polit. Econ.* 106, 667–705.
- Hoxby, C., 2000. Does competition among public schools benefit students and taxpayers? *Am. Econ. Rev.* 90, 1209–1238.
- Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 467–475.
- Imbens, G., Lemieux, T., 2008. Regression discontinuity designs: a guide to practice. *J. Econometr.* 142, 615–635.
- Imbens, G., Wooldridge, J., 2007. Control function and related methods. In: *What's New In Econometrics?* NBER Lecture Note 6.
- Kain, J.F., 1992. The spatial mismatch hypothesis: three decades later. *Hous. Pol. Debate* 3, 371–462.
- Katz, L.F., Kling, J.R., Liebman, J.B., 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. *Q. J. Econ.* 116, 607–654.
- Kline, P., 2011. Oaxaca-blinder as a reweighting estimator. *Am. Econ. Rev.* 101, 532–537.
- Kline, P., Moretti, E., 2014. Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee valley authority. *Q. J. Econ.* 129, 275–331.
- Kling, J., Liebman, J., Katz, L., 2007. Experimental analysis of neighborhood effects. *Econometrica* 75, 83–119.
- Kolesar, M., Chetty, R., Friedman, J., E.G., 2013. Identification and Inference with Many Invalid Instruments (manuscript).
- Kopczuk, W., Munroe, D.J., 2014. Mansion tax: the effect of transfer taxes on the residential real estate market. *Am. Econ. J. Econ. Pol.* (forthcoming).
- Kuminoff, N.V., Smith, V.K., Timmins, C., 2013. The new economics of equilibrium sorting and policy evaluation using housing markets. *J. Econ. Liter.* 51, 1007–1062.
- Lee, D., Lemieux, T., 2010. Regression discontinuity designs in economics. *J. Econ. Liter.* 48, 281–355.

- Levitt, S., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *Am. Econ. Rev.* 87, 270–290.
- Lewis, E., 2011. Immigration, skill mix, and capital skill complementarity. *Q. J. Econ.* 126, 1029–1069.
- Linden, L., Rockoff, J., 2008. Estimates of the impact of crime risk on property values from Megan's laws. *Am. Econ. Rev.* 98, 1103–1127.
- Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. *Am. Econ. Rev.* 103, 226–231.
- Luttmer, E., 2005. Neighbors as negatives: relative earnings and well-being. *Q. J. Econ.* 130, 963–1002.
- McCrary, J., 2002. Using electoral cycles in police hiring to estimate the effect of police on crime: comment. *Am. Econ. Rev.* 92, 1236–1243.
- McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: a density test. *J. Econometr.* 142, 698–714.
- McMillen, D., McDonald, J., 2002. Land values in a newly zoned city. *Rev. Econ. Stat.* 84, 62–72.
- Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the U.S. mortgage default crisis. *Q. J. Econ.* 124, 1449–1496.
- Michaels, G., 2008. The effect of trade on the demand for skill—evidence from the interstate highway system. *Rev. Econ. Stat.* 90, 683–701.
- Moulton, B., 1986. Random group effects and the precision of regression estimates. *J. Econometr.* 32, 385–397.
- Moulton, B., 1990. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Rev. Econ. Stat.* 72, 334–338.
- Neal, D., 1997. The effects of catholic secondary schooling on educational achievement. *J. Labor Econ.* 15, 98–123.
- Notowidigdo, 2013. The Incidence of Local Labor Demand Shocks (manuscript).
- Oates, W.E., 1969. The effects of property taxes and local public spending on property values: an empirical study of tax capitalization and the tiebout hypothesis. *J. Polit. Econ.* 77, 957–971.
- Oster, E., 2013. Unobservable Selection and Coefficient Stability: Theory and Validation. Working paper.
- Pearl, J., 2009. Causal inference in statistics: an overview. *Stat. Surv.* 3, 96–146.
- Pence, K.M., 2006. Foreclosing on opportunity: state laws and mortgage credit. *Rev. Econ. Stat.* 88, 177–182.
- Redding, S., Sturm, D., 2008. The costs of remoteness: evidence from German division and reunification. *Am. Econ. Rev.* 98, 1766–1797.
- Roback, J., 1982. Wages, rents and the quality of life. *J. Polit. Econ.* 90, 1257–1278.
- Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *J. Polit. Econ.* 82, 34–55.
- Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55.
- Rosenthal, S., 2014. Are private markets and filtering a viable source of low-income housing? Estimates from a “repeat income” model. *Am. Econ. Rev.* 104, 687–706.
- Rothstein, J., 2007. Does competition among public schools benefit students and taxpayers? A comment on Hoxby (2000). *Am. Econ. Rev.* 97, 2026–2037.
- Roy, A.D., 1951. Some thoughts on the distribution of earnings. *Oxf. Econ. Pap. New Ser.* 3, 135–146.
- Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701.
- Sacerdote, B., 2001. Peer effects with random assignment: results for Dartmouth roommates. *Q. J. Econ.* 116, 681–704.
- Saez, E., 2010. Do taxpayers bunch at kink points? *Am. Econ. J. Econ. Pol.* 2, 180–212.
- Saiz, A., 2010. The geographic determinants of housing supply. *Q. J. Econ.* 125, 1253–1296.
- Schwartz, A.E., Ellen, I.G., Voicu, I., Schill, M., 2006. The external effects of place-based subsidized housing. *Reg. Sci. Urban Econ.* 36, 679–707.
- Serrato, S., Carlos, J., Wingender, P., 2014. Estimating Local Fiscal Multipliers (manuscript).

- Stock, J., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Stock, J., Andrews, D. (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas J. Rothenberg*. Cambridge University Press, Cambridge, pp. 109–120.
- Tiebout, C., 1956. A pure theory of local expenditures. *J. Polit. Econ.* 64, 416–424.
- Turner, M.A., Haughwout, A., van der Klaauw, W., 2014. Land use regulation and welfare. *Econometrica* 82, 1341–1403.
- Turner, N., 2012. Who benefits from student aid? The economic incidence of tax based federal student aid. *Econ. Educ. Rev.* 31, 463–481.
- Wooldridge, J., 2002. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.
- Wooldridge, J., 2005. Inverse probability weighted M-estimators for sample selection, attrition, and stratification. *Port. Econ. J.* 1, 117–139.

CHAPTER 2

Structural Estimation in Urban Economics

Thomas J. Holmes*, Holger Sieg†

*University of Minnesota and Federal Reserve Bank of Minneapolis, Minneapolis, MN, USA

†University of Pennsylvania, Philadelphia, PA, USA

Contents

2.1. An Introduction to Structural Estimation	70
2.1.1 Model selection and development	70
2.1.2 Identification and estimation	71
2.1.3 Policy analysis	73
2.1.4 Applications	74
2.2. Revealed Preference Models of Residential Choice	74
2.3. Fiscal Competition and Public Good Provision	79
2.3.1 Theory	80
2.3.1.1 <i>Preferences and heterogeneity</i>	80
2.3.1.2 <i>Household sorting</i>	81
2.3.1.3 <i>Community size, housing markets, and budgets</i>	82
2.3.1.4 <i>Equilibrium</i>	84
2.3.1.5 <i>Properties of equilibrium</i>	86
2.3.1.6 <i>Computation of equilibrium</i>	86
2.3.1.7 <i>Extensions</i>	86
2.3.2 Identification and estimation	88
2.3.2.1 <i>The information set of the econometrician</i>	88
2.3.2.2 <i>Predictions of the model</i>	88
2.3.2.3 <i>Household sorting by income</i>	89
2.3.2.4 <i>Public good provision</i>	91
2.3.2.5 <i>Voting</i>	92
2.3.2.6 <i>Identifying and estimating housing supply functions</i>	92
2.3.3 Policy analysis	93
2.3.3.1 <i>Evaluating regulatory programs: the Clean Air Act</i>	93
2.3.3.2 <i>Decentralization versus centralization</i>	95
2.4. The Allocation of Economic Activity Across Space	96
2.4.1 Specialization of regions	96
2.4.1.1 <i>Model development</i>	97
2.4.1.2 <i>Estimation and identification</i>	99
2.4.2 Internal structure of cities	100
2.4.2.1 <i>Model development</i>	100
2.4.2.2 <i>Estimation and identification</i>	101

2.4.3 Policy analysis	103
2.4.4 Relation to entry models in the industrial organization literature	106
2.5. Conclusions	110
Acknowledgments	111
References	111

Abstract

Structural estimation is a methodological approach in empirical economics explicitly based on economic theory, in which economic modeling, estimation, and empirical analysis are required to be internally consistent. This chapter illustrates the structural approach with three applications in urban economics: (1) discrete location choice, (2) fiscal competition and local public good provision, and (3) regional specialization. For each application, we first discuss broad methodological principles of model selection and development. Next we treat issues of identification and estimation. The final step of each discussion is how estimated structural models can be used for policy analysis.

Keywords

Structural estimation, Fiscal competition, Public good provision, Regional specialization

JEL Classification Codes

R10, R23, R51

2.1. AN INTRODUCTION TO STRUCTURAL ESTIMATION

Structural estimation is a methodological approach in empirical economics explicitly based on economic theory. A requirement of structural estimation is that economic modeling, estimation, and empirical analysis be internally consistent. Structural estimation can also be defined as theory-based estimation: the objective of the exercise is to estimate an explicitly specified economic model that is broadly consistent with observed data. Structural estimation, therefore, differs from other estimation approaches that are either based on purely statistical models or based only implicitly on economic theory.¹ A structural estimation exercise typically consists of the following three steps: (1) model selection and development, (2) identification and estimation, and (3) policy analysis. We discuss each step in detail and then provide some applications to illustrate the key methodological issues that are encountered in the analysis.

¹ For example, the most prominent approach in program evaluation is based on work by [Neyman \(1923\)](#) and [Fisher \(1935\)](#), who suggested evaluating the impact of a program by using potential outcomes that reflect differences in treatment status. The objective of the exercise, then, is typically to estimate average treatment effects. This is a purely statistical model, which is sufficiently flexible such that it has broad applications in many sciences.

2.1.1 Model selection and development

The first step in a structural estimation exercise is the development or selection of an economic model. These models can be simple static decision models under perfect information or complicated nonstationary dynamic equilibrium models with asymmetric information.

It is important to recognize that a model that is suitable for structural estimation needs to satisfy requirements that are not necessarily the same requirements that a theorist would typically find desirable. Most theorists will be satisfied if an economic model captures the key ideas that need to be formalized. In structural estimation, we search for models that help us understand the real world and are consistent with observed outcomes. As a consequence, we need models that are not rigid, but are sufficiently flexible to fit the observed data. Flexibility is not necessarily a desirable property for a theorist, especially if the objective is to analytically characterize the properties of a model.

Theorists are typically reluctant to work with parameterized versions of their model, since they aim for generality. An existence proof is, for example, considered to be of limited usefulness by most theorists if it crucially depends on functional form assumptions. Flexible economic models often have the property that equilibria can only be computed numerically—that is, there are no analytical solutions. Numerical computations of equilibria require a fully parameterized and numerically specified model. The parametric approach is, therefore, natural to structural modeling in microeconomics as well as to much of modern quantitative macroeconomics. Key questions, then, are how to determine the parameter values and whether the model is broadly consistent with observed outcomes. Structural estimation provides the most compelling approach to determine plausible parameter values for a large class of models and to evaluate the fit of the model.

2.1.2 Identification and estimation

Structural estimation also requires that we incorporate a proper error structure into the economic model. Since theory and estimation must be internally consistent, the model under consideration needs to generate a well-specified statistical model.² Any economic model is, by definition, an abstraction of the real world. As a consequence, it cannot be an exact representation of the “true” data-generating process. This criticism is not specific to structural estimation, since it also applies to any purely statistical modeling and estimation approach. We are interested in finding economic models that, in the best-case scenario, cannot be rejected by the data using conventional statistical hypothesis or specification tests. Of course, models that are rejected by the data can also be very helpful and improve our knowledge. These models can provide us with guidance on how to improve our modeling approach, generating a better understanding of the research questions that we investigate.

² Notice that this is another requirement that is irrelevant from a theorist’s perspective.

A standard approach for estimating structural models requires the researcher to compute the optimal decision rules or the equilibrium of a model to evaluate the relevant objective function of an extremum estimator. It is a full-solution approach, since the entire model is completely specified on the computer. In many applications, it is not possible to use canned statistical routines to do this. Rather, the standard approach involves programming an economic model, though various procedures and routines can be pulled off the shelf to use in solving the model.³ The step of obtaining a solution of an economic model for a given set of parameters is called the “inner loop” and often involves a fixed point calculation (i.e., taking as given a vector of endogenous variables, agents in the model make choices that result in the same vector of endogenous variables, satisfying the equilibrium conditions). There is also an “outer loop” step in which the parameter vector is varied and a maximization problem is solved to obtain the parameter vector that best fits the data according to a given criterion. The outer/inner loop approach is often called a “nested fixed point” algorithm.

Whenever we use nested fixed point algorithms, the existence and uniqueness of equilibrium are potentially important aspects of the analysis. Uniqueness of equilibrium is not a general property of most economic models, especially those that are sufficiently flexible to be suitable for structural estimation. Moreover, proving uniqueness of equilibrium can be rather challenging.⁴ Nonuniqueness of equilibrium can cause a number of well-known problems during estimation and counterfactual comparative static analysis. Sometimes we may want to condition on certain observed features of the equilibrium and only impose a subset of the equilibrium conditions. By conditioning on observed outcomes, we often circumvent a potential multiplicity of equilibria problems.

Another potential drawback of the full-solution estimation approach is that it is computationally intensive. We are likely to hit the feasibility constraints quickly because of the well-known curses of dimensionality that are encountered, for example, in dynamic programming.⁵

It is, therefore, often desirable to derive estimation approaches that do not rely on full-solution approaches. Often we can identify and estimate the parameters of a model using necessary conditions of equilibrium, which can take the form of first-order conditions, inequality constraints, or boundary indifference conditions. We call these “partial solution” approaches.⁶ These approaches are often more elegant than brute force

³ A useful reference for algorithms to solve economic models is Judd (1998). Another standard reference for numerical recipes in C programming is Press et al. (1988).

⁴ For example, the only general uniqueness proofs that we have for the Arrow–Debreu model rely on high-level assumptions about the properties of the excess demand function.

⁵ See Rust (1994) for a discussion of computational complexity within the context of dynamic discrete choice models.

⁶ Some of the most compelling early applications of partial solution methods in structural estimation are those of Heckman and MaCurdy (1980) and Hansen and Singleton (1982). See Holmes (2011) for a recent example of an application of an inequality constraint approach used to estimate economies of density.

approaches, but they are more difficult to derive, since they typically exploit specific idiosyncratic features of the model. Finding these approaches requires a fair bit of creativity.

A parametric approach is not necessary for identification or estimation. It can be useful to ask the question whether our model can be identified under weak functional form assumptions. Those approaches, then, typically lead us to consider nonparametric or semiparametric approaches for identification or estimation. Notice that identification and estimation largely depend on the available data—that is, the information set of the econometrician. Thus, identification and estimation are closely linked to the data collection decisions made by the researchers.

Once we have derived and implemented an estimation procedure, we need to determine whether our model fits the data. Goodness of fit can be evaluated on the basis of moments used in estimation or moments that are not used in estimation. We would also like to validate our model—that is, we would like to use some formal testing procedures to determine whether our model is consistent with the data and not seriously misspecified. A number of approaches have been proposed in the literature. First, we can use specification tests that are typically based on overidentifying conditions. Second, we can evaluate our model on the basis of out-of-sample predictions. The key idea is to determine whether our model can predict the observed outcomes in a holdout sample. Finally, we sometimes have access to experimental data that may allow us to identify certain treatment or causal effects. We can then study whether our theoretical model generates treatment effects that are of similar magnitude.⁷

2.1.3 Policy analysis

The third and final step of a structural estimation exercise consists of policy analysis. Here, the objective is to answer the policy questions that motivated the empirical analysis. We can conduct retrospective or prospective policy analysis.

Retrospective analysis evaluates an intervention that happened in the past and is observed in the sample period. One key objective is to estimate treatment effects that are associated with the observed policy intervention. Not surprisingly, structural approaches compete with nonstructural approaches. As pointed out by [Lucas \(1976\)](#), there are some compelling reasons for evaluating a policy change within an internally consistent framework. The structural approach is particularly helpful if we are interested in nonmarginal or general equilibrium effects of policies.

Prospective analysis focuses on new policies that have not been enacted. Again, evaluating the likely impact of alternative policies within a well-defined and internally consistent theoretical framework has some obvious advantages. Given that large-scale

⁷ Different strategies for model validation are discussed in detail in [Keane and Wolpin \(1997\)](#) and [Todd and Wolpin \(2006\)](#).

experimental evaluations of alternative policies are typically expensive or not feasible in urban economics, the structural approach is the most compelling one in which to conduct prospective policy analysis.

2.1.4 Applications

Having provided an overview of the structural approach, we now turn to the issue of applying these methods in urban and regional economics. We focus on three examples that we use to illustrate broad methodological principles. Given our focus on methodology, we acknowledge that we are not able to provide a comprehensive review of various articles in the field that take a structural estimation approach.⁸ Our first application is location choice. This is a classic issue, one that was addressed in early applications of McFadden's Nobel Prize-winning work on discrete choice (McFadden, 1978). As noted earlier, structural estimation projects typically require researchers to write original code. The literature on discrete choice is well developed, practitioner's guides are published, and reliable computer code is available on the Web.

Our second application considers the literature on fiscal competition and local public good provision. One of the key functions of cities and municipalities is to provide important public goods and services such as primary and secondary education, protection from crime, and infrastructure. Households are mobile and make locational decisions based, at least in part, on differences in public goods, services, and local amenities. This analysis combines the demand side of household location choice with the supply side of what governments offer. Since the focus is on positive analysis, political economy models are used to model the behavior of local governments. In this literature, one generally does not find much in the way of canned software, but we provide an overview of the basic steps for working in this area.

The third application considers recent articles related to the allocation of economic activity across space, including the Ahlfeldt et al. (2014) analysis of the internal structure of the city of Berlin and the Holmes and Stevens (2014) analysis of specialization by industry of regions in the United States. We use the discussion to highlight (1) the development of the models, (2) identification and the basic procedure for estimation, and (3) how the models can be used for policy analysis.

2.2. REVEALED PREFERENCE MODELS OF RESIDENTIAL CHOICE

A natural starting point for a discussion of structural estimation in urban and regional economics is the pioneering work by Daniel McFadden on estimation of discrete choice

⁸ For example, we do not discuss a number of articles that are squarely in the structural tradition, such as those of Holmes (2005), Gould (2007), Baum-Snow and Pavan (2012), Kennan and Walker (2011), or Combes et al. (2012).

models. One of the main applications that motivated the development of these methods was residential or locational choice. In this section, we briefly review the now classic results from McFadden and discuss why urban economists are still struggling with some of the same problems that McFadden studied in the early 1970s.

The decision-theoretical framework that underlies modern discrete choice models is fairly straightforward. We consider a household i that needs to choose among different neighborhoods that are indexed by j . Within each neighborhood there are a finite number of different housing types indexed by k . A basic random utility model assumes that the indirect utility of household i for community j and house k is given by

$$u_{ijk} = x_j' \beta + z_k' \gamma + \alpha(\gamma_i - p_{jk}) + \epsilon_{ijk}, \quad (2.1)$$

where x_j is a vector of observed characteristics of community j , z_k is a vector of observed housing characteristics, γ_i is household income, and p_{jk} is the price of housing type k in community j . Each household chooses the neighborhood-housing pair that maximizes utility. One key implication of the behavioral model is that households make deterministic choices—that is, for each household there exists a unique neighborhood-house combination that maximizes utility.

McFadden (1974) showed how to generate a well-defined econometric model that is internally consistent with the economy theory described above. Two assumptions are particularly noteworthy. First, we need to assume that there is a difference in information sets between households and econometricians. Although households observe all key variables, including the error terms (ϵ_{ijk}), econometricians observe only x_j , z_k , γ_i , and p_{jk} , and a set of indicators, denoted by d_{ijk} , where $d_{ijk} = 1$ if household i chooses neighborhood j and house type k and $d_{ijk} = 0$ otherwise. Integrating out the unobserved error terms then gives rise to well-behaved conditional choice probabilities that provide the key ingredient for a maximum likelihood estimator of the parameters of the model.

Second, if the error terms are independent and identically distributed across i , j , and k and follow a type I extreme value distribution, we obtain the well-known conditional logit choice probabilities:

$$\Pr\{d_{ijk} = 1 | x, z, p, \gamma_i\} = \frac{\exp\{x_j' \beta + z_k' \gamma + \alpha(\gamma_i - p_{jk})\}}{\sum_{n=1}^J \sum_{m=1}^K \exp\{x_n' \beta + z_m' \gamma + \alpha(\gamma_i - p_{nm})\}}. \quad (2.2)$$

A key advantage of the simple logit model is that conditional choice probabilities have a closed-form solution. The only problem encountered in estimation is that the likelihood function is nonlinear in its parameters. The estimates must be computed numerically. All standard software packages will allow researchers to do that. Standard errors can be computed using the standard formula for maximum likelihood estimators.

One unattractive property of the logit model is the independence of irrelevant alternatives property. It basically says that the ratio of conditional choice probabilities of two products depends only on the relative utility of those two products. Another (related)

unattractive property of the simple logit model is that it generates fairly implausible substitution patterns for the aggregate demand. Own and cross-price elasticities are primarily functions of a single parameter (α) and are largely driven by the market shares and not by the proximity of two products in the characteristic space.

One way to solve this problem is to relax the assumption that idiosyncratic tastes are independent across locations and houses. [McFadden \(1978\)](#) suggested modeling the distribution of the error terms as a generalized extreme value distribution, which then gives rise to the nested logit model. In our application, we may want to assume that idiosyncratic shocks of houses within a given neighborhood are correlated owing to some unobserved joint neighborhood characteristics. A main advantage of the nested logit model is that conditional choice probabilities still have closed-form solutions, and estimation can proceed within a standard parametric maximum likelihood framework. Again, most major software packages will have a routine for nested logit models. Hence, few technical problems are involved in implementing this estimator and computing standard errors. The main drawback of the nested logit is that the researcher has to choose the nesting structure before estimation. As a consequence, we need to have strong beliefs about which pairs of neighborhood-house choices are most likely to be close substitutes. We, therefore, need to have detailed knowledge of the neighborhood structure within the city that we study in a given application.

An alternative approach, one that avoids the need to impose a substitution structure prior to estimation and can still generate realistic substitution patterns, is based on random coefficients.⁹ Assume now that the utility function is given by

$$ijk = x'_j \beta_i + z'_k \gamma_i + \alpha_i (\gamma_i - p_{jk}) + \epsilon_{ijk}, \quad (2.3)$$

where γ_i , β_i , and α_i are random coefficients. A popular approach is based on the assumption that these random coefficients are normally distributed. It is fairly straightforward to show that substitutability in the random coefficient logit model is driven by observed housing and neighborhood characteristics. Households that share similar values of random coefficients will substitute between neighborhood-housing pairs that have similar observed characteristics.

A key drawback of the random coefficient model is that the conditional choice probabilities no longer have closed-form solutions and must be computed numerically. This process can be particularly difficult if there are many observed characteristics, and hence high-dimensional integrals need to be evaluated. These challenges partially led to the development of simulation-based estimators (see [Newey and McFadden, 1994](#) for some basic results on consistency and asymptotic normality of simulated maximum likelihood estimators). As discussed, for example, in [Judd \(1998\)](#), a variety of numerical algorithms have been developed that allow researchers to solve these integration

⁹ For a detailed discussion, see, for example, [Train \(2003\)](#).

problems. A notable application of these methods is that of [Hastings et al. \(2006\)](#), who study sorting of households among schools within the Mecklenburg Charlotte school district. They evaluate the impact of open enrollment policies under a particular parent choice mechanism.¹⁰

Demand estimation has also focused on the role of unobserved product characteristics ([Berry, 1994](#)). In the context of our application, unobserved characteristics may arise at the neighborhood level or the housing level. Consider the case of an unobserved neighborhood characteristic. The econometrician probably does not know which neighborhoods are popular. More substantially, our measures of neighborhood or housing quality (or both) may be rather poor or incomplete. Let ξ_j denote an unobserved characteristic that captures aspects of neighborhood quality that are not well measured by the researcher. Utility can now be represented by the following equation:

$$u_{ijk} = x'_j \beta_i + z'_k \gamma_i + \alpha_i (y_i - p_{jk}) + \xi_j + \epsilon_{ijk}. \quad (2.4)$$

This locational choice model is then almost identical in mathematical structure to the demand model estimated in [Berry et al. \(1995\)](#). The key insight of that article is that the unobserved product characteristics can be recovered by matching the observed market shares of each product. The remaining parameters of the model can be estimated by using a generalized method of moments estimator that uses instrumental variables to deal with the correlation between housing prices and unobserved neighborhood characteristics. Notice that the Berry–Levinsohn–Pakes estimator is a nested fixed point estimator. The inner loop inverts the market share equations to compute the unobserved product characteristics. The outer loop evaluates the relevant moment conditions and searches over the parameter space.

Estimating this class of models initially required some serious investment in programming, since standard software packages did not contain modules for this class of models. Now, however, both a useful practitioner's guide ([Nevo, 2000](#)) and a variety of programs are available and openly shared. This change illustrates an important aspect of structural estimation. Although structural estimation may require some serious initial methodological innovations, subsequent users of these techniques often find it much easier to modify and implement these techniques.¹¹ Notable articles that introduced this empirical approach to urban economics are those of [Bayer \(2001\)](#), [Bayer et al. \(2004\)](#), and [Bayer et al. \(2007\)](#), who estimate models of household sorting in the Bay Area.

¹⁰ Bayesian estimators can also be particularly well suited for estimating discrete choice models with random coefficients. [Bajari and Kahn \(2005\)](#) adopt these methods to study racial sorting and peer effects within a similar framework.

¹¹ Computation of standard errors is also nontrivial, as discussed in [Berry et al. \(2004\)](#). Most applied researchers prefer to bootstrap standard errors in these models.

Extending these models to deal with the endogenous neighborhood characteristics or peer effects is not trivial. For example, part of the attractiveness of a neighborhood may be driven by the characteristics of neighbors. Households may value living, for example, in neighborhoods with a large fraction of higher-income households because of the positive externalities that these families may provide. Three additional challenges arise in these models. First, peer effects need to be consistent with the conditional choice probabilities and the implied equilibrium sorting. Second, endogenous peer effects may give rise to multiplicity of equilibria, which creates additional problems in computation and estimation. Finally, the standard Berry–Levinsohn–Pakes instrumentation strategy, which uses exogenous characteristics of similar house–neighborhood pairs, is not necessarily feasible anymore, since we are dealing with endogenous neighborhood characteristics that are likely to be correlated with the unobserved characteristics.¹² Finding compelling instruments can be rather challenging. Some promising examples are given by [Ferreira \(2009\)](#), who exploits the impact of property tax limitations (Proposition 13) in California on household sorting. [Galliani et al. \(2012\)](#) exploit random assignment to vouchers to construct instruments in their study of the effectiveness of the Moving to Opportunity housing assistance experiment.

Researchers have also started to incorporate dynamic aspects into the model specification. Locational choices and housing investments are inherently dynamic decisions that affect multiple time periods. As a consequence, adopting a dynamic framework involves some inherent gains. In principle, we can follow [Rust \(1987\)](#), but adopting a dynamic version of the logit model within the context of locational choice is rather challenging. Consider the recent article by [Murphy \(2013\)](#), who estimates a dynamic discrete choice model of land conversion using data from the Bay Area. One key problem is measuring prices for land (and housing). In a dynamic model, households must also forecast the evolution of future land and housing prices to determine whether developing a piece of land is optimal. That creates two additional problems. First, we need to characterize price expectations based on simple time series models. Second, we need one pricing equation for each location (assuming land or housing (or both) within a neighborhood is homogeneous), which potentially blows up the dimensionality of state space associated with the dynamic programming problem.¹³ Some user guides are available for estimating dynamic discrete choice models, most notably the chapter by [Rust \(1994\)](#). Estimation and inference is fairly straightforward as long as one stays within the parametric maximum likelihood framework.

¹² [Bayer and Timmins \(2005\)](#) and [Bayer et al. \(2007\)](#) provide a detailed discussion of these issues in the context of the random utility model above. See also the survey articles on peer effects and sorting in this handbook. [Epple et al. \(2014\)](#) estimate a game of managing school district capacity, in which school quality is largely defined by peer effects.

¹³ Other promising examples of dynamic empirical approaches are those of [Bishop \(2011\)](#), who adopts a Hotz–Miller conditional choice probabilities estimator, and [Bayer et al. \(2012\)](#). [Yoon \(2012\)](#) studies locational sorting in regional labor markets, adopting a dynamic nonstationary model.

Thanks to the requirement to disclose estimation codes by a variety of journals, some software programs are also available that can be used to understand the basic structure of the estimation algorithms. However, each estimation exercise requires some coding.

Finally, researchers have worked on estimating discrete choice models when there is rationing in housing markets. [Geyer and Sieg \(2013\)](#) develop and estimate a discrete choice model that captures excess demand in the market for public housing. The key issue is that simple discrete choice models give rise to biased estimators if households are subject to rationing and, thus, do not have full access to all elements in the choice set. The idea of that article is to use a fully specified equilibrium model of supply and demand to capture the rationing mechanism and characterize the endogenous (potentially latent) choice set of households. Again, we have to use a nested fixed point algorithm to estimate these types of models. The key finding of this chapter is that accounting for rationing implies much higher welfare benefits associated with public housing communities than simple discrete choice estimators that ignore rationing.

2.3. FISCAL COMPETITION AND PUBLIC GOOD PROVISION

We next turn to the literature on fiscal competition and local public good provision. As noted above, one key function of cities and municipalities is to provide important public goods and services. Households are mobile and make locational decisions based on differences in public goods, services, and local amenities. The models developed in the literature combine the demand side of household location choice, which are similar to the ones studied in the previous section, with political economy models that are used to model the behavior of local governments.

We start [Section 2.3.1](#) by outlining a generic model of fiscal competition that provides the basic framework for much of the empirical work in the literature. We develop the key parts of the model and define equilibrium. We also discuss existence and uniqueness of equilibrium and discuss key properties of these models. We finish by discussing how to numerically compute equilibria for more complicated specifications of the model, and we discuss useful extensions.

In [Section 2.3.2](#), we turn to an empirical issue. We start by broadly characterizing the key predictions of this class of models and then develop a multistep approach that can be used to identify and estimate the parameters of the model. We finish this section by discussing alternative estimators that rely less on functional form assumptions.

In [Section 2.3.3](#), we turn to policy analysis. We consider two examples. The first example considers the problem of estimating the willingness to pay for improving air quality in Los Angeles. We discuss how to construct partial and general equilibrium measures that are consistent with the basic model developed above. Our second application considers the potential benefits of decentralization and compares decentralized with centralized outcomes within a general equilibrium model.

2.3.1 Theory

The starting point of any structural estimation exercise is a theoretical model that allows us to address key research questions. In this application, we consider fiscal competition and public good provision within a system of local jurisdictions.¹⁴ This literature blends the literature on demand for public goods and residential choice with the literature on political economy models of local governments that characterize the supply of public goods and services.

2.3.1.1 Preferences and heterogeneity

We consider an urban or metropolitan area that consists of J communities, each of which has fixed boundaries. Each community has a local housing market, provides a (congestible) public good g , and charges property taxes, t . There is a continuum of households that differ by income, γ . Households also differ by tastes for public goods, denoted by α . Note that unobserved heterogeneity in preferences is a key ingredient in any empirical model that must be consistent with observed household choices, since households that have the same observed characteristics typically do not make the same decisions.

Households behave as price takers and have preferences defined over a local public good, housing services, h , and a composite private good, b . Households maximize utility with respect to their budget constraint:

$$\begin{aligned} \max_{(h,b)} \quad & U(\alpha, g, h, b) \\ \text{s.t.} \quad & (1+t)p^h h = \gamma - b, \end{aligned} \quad (2.5)$$

which yields housing demand functions $h(p, \gamma; \alpha, g)$. The corresponding indirect utility function is given by

$$V(\alpha, g, p, \gamma) = U(\alpha, g, h(p, \gamma, \alpha), \gamma - ph(p, \gamma, \alpha, g)), \quad (2.6)$$

where $p = (1+t)p^h$. Consider the slope of an indirect indifference curve in the (g, p) -plane:

$$M(\alpha, g, p, \gamma) = - \frac{\partial V(\alpha, g, p, \gamma) / \partial g}{\partial V(\alpha, g, p, \gamma) / \partial p}. \quad (2.7)$$

If $M(\cdot)$ is monotonic in γ for given α , then indifference curves in the (g, p) -plane satisfy the single-crossing property. Likewise, monotonicity of $M(\cdot)$ in α provides a single crossing for given γ . As we will see below, the single-crossing properties are key to characterizing both the sorting and the voting behavior of households. One challenge encountered in structural

¹⁴ Our theoretical model builds on previous work by Ellickson (1973), Westhoff (1977), Epple et al. (1984), Goodspeed (1989), Epple and Romer (1991), Nechyba (1997), Fernandez and Rogerson (1996), Benabou (1996a,b), Durlauf (1996), Fernandez and Rogerson (1998), Epple and Platt (1998), Glomm and Lagunoff (1999), Henderson and Thisse (2001), Benabou (2002), Rothstein (2006), and Ortalo-Magne and Rady (2006).

estimation is to find a flexible parameterization of the model that is not overly restrictive.¹⁵ A promising parameterization of the indirect utility function is given below:

$$V(g, p, \gamma, \alpha) = \left\{ \alpha g^\rho + \left(e^{\frac{\gamma^{1-\nu}-1}{1-\nu}} e^{-\frac{Bp^{\eta+1}-1}{1+\eta}} \right)^\rho \right\}^{1/\rho}, \quad (2.8)$$

where α is the relative weight that a household assigns to the public goods. Roy's identity implies that the housing demand function is given by

$$h = B p^\eta \gamma^\nu. \quad (2.9)$$

Note that η is the price elasticity of housing and ν is the income elasticity. This demand function is a useful characterization of the demand, since it does not impose unitary income or price elasticities.¹⁶ Note that this utility function satisfies the single-crossing property if $\rho < 0$.

2.3.1.2 Household sorting

One objective of the model is to explain household sorting among the set of communities. There are no mobility costs, and hence households choose j to maximize

$$\max_j V(\alpha, g_j, p_j, \gamma). \quad (2.10)$$

Define the set C_j to be the set of households living in community j :

$$C_j = \{(\alpha, \gamma) \mid V(\alpha, g_j, p_j, \gamma) \geq \max_{i \neq j} V(\alpha, g_i, p_i, \gamma)\}. \quad (2.11)$$

Figure 2.1 illustrates the resulting sorting in the (p, g) -space. It considers the case of three communities denoted by $j-1$, j , and $j+1$. It plots the indifference curve of a household that is indifferent between $j-1$ and j , denoted by $\gamma_{j-1}(\alpha)$. Similarly, it plots the indifference curve of a household that is indifferent between j and $j+1$, denoted by $\gamma_j(\alpha)$. Note that for a given level of α , the household that is indifferent between j and $j+1$ must have higher income than the household that is indifferent between $j-1$ and j , and as a consequence, we have $\gamma_j(\alpha) > \gamma_{j-1}(\alpha)$. Single crossing then implies that the household with higher income levels must have steeper indifference curves than the household with lower income levels. Finally, Figure 2.1 also plots the indifference curve of a household with income given by $\gamma_j(\alpha) > \gamma > \gamma_{j-1}(\alpha)$. This household will strictly prefer to live in community j .

¹⁵ We will discuss nonparametric or semiparametric identification below.

¹⁶ To avoid stochastic singularities, we can easily extend the framework and assume that the housing demand or expenditures are subject to an idiosyncratic error that is revealed to households after they have chosen the neighborhood. This error term thus enters the housing demand, but does not affect the neighborhood choice. Alternatively, we can assume in estimation that observed housing demand is subject to measurement error. We follow that approach in our application.

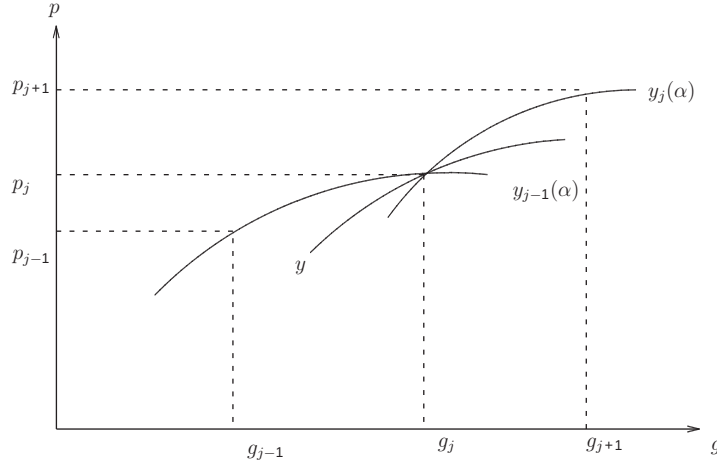


Figure 2.1 Sorting in the (p, g) -space.

Alternatively, we can characterize household sorting by deriving the boundary indifference loci $\alpha_j(\gamma)$, which are defined as

$$V(\alpha_j(\gamma), g_j, p_j, \gamma) = V(\alpha_j(\gamma), g_{j+1}, p_{j+1}, \gamma), \quad (2.12)$$

and are hence the inverse of $y_j(\alpha)$. Given our parameterization, these boundary indifference conditions can be written as

$$\ln \alpha - \rho \left(\frac{\gamma^{1-\nu} - 1}{1-\nu} \right) = \ln \left(\frac{Q_{j+1} - Q_j}{g_j^\rho - g_{j+1}^\rho} \right) \equiv K_j, \quad (2.13)$$

where

$$Q_j = e^{-\frac{\rho}{1+\eta}} (B p_j^{\eta+1} - 1). \quad (2.14)$$

Figure 2.2 illustrates the resulting sorting of households across communities in equilibrium in the $(\ln \gamma, \ln \alpha)$ -space. The loci passing through the K -intercepts characterize the boundary indifference conditions. The loci passing through the L -intercepts characterize the set of decisive voters within each community (as explained in detail below).

2.3.1.3 Community size, housing markets, and budgets

A measure of the size (or market share) of community j is given by

$$n_j = P(C_j) = \int_{C_j} f(\alpha, \gamma) d\gamma d\alpha. \quad (2.15)$$

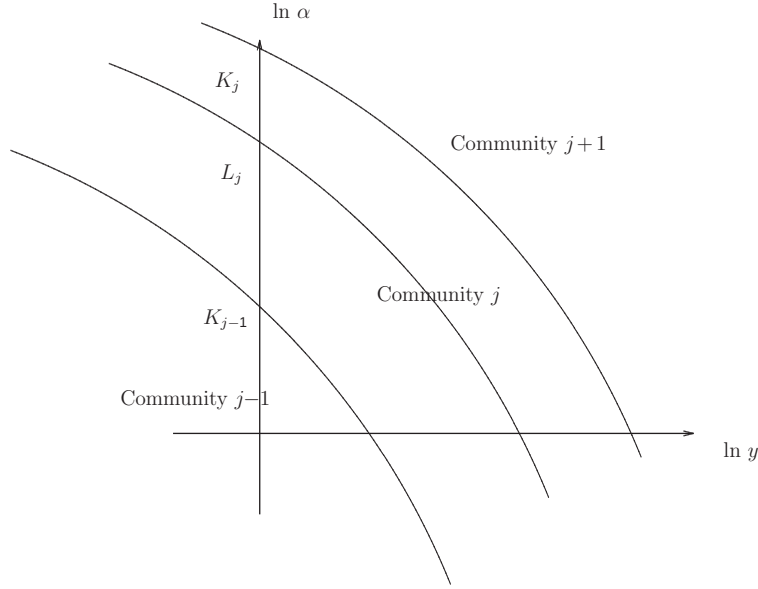


Figure 2.2 The distribution of households across and within communities.

Aggregate housing demand is defined as

$$H_j^d = \int_{C_j} h(p_j, \alpha, \gamma) f(\alpha, \gamma) d\gamma d\alpha. \quad (2.16)$$

Housing is owned by absentee landlords, and the aggregate housing supply in community j depends on the net-of-tax price of housing p_j^h and a measure of the land area of community j denoted by l_j . Hence, we have that

$$H_j^s = H(l_j, p_j^h). \quad (2.17)$$

A commonly used housing supply function is given by $H_j^s = l_j [p_j^h]^\tau$. Note that τ is the price elasticity and l_j is a measure of the availability of land. Housing markets need to clear in equilibrium for each community.

The budget of community j must be balanced. This implies that

$$t_j p_j^h \int_{C_j} h(p_j, \alpha, \gamma) f(\alpha, \gamma) d\gamma d\alpha / P(C_j) = c(g_j), \quad (2.18)$$

where $c(g)$ is the cost per household of providing g .¹⁷

Next we endogenize the provision of local public goods, assuming that residents vote on fiscal and tax policies in each community. [Fernandez and Rogerson \(1996\)](#) suggest the following timing assumptions:

1. Households choose a community of residence having perfect foresight of equilibrium prices, taxes, and spending in all communities.

¹⁷ A linear cost function is commonly used in quantitative work—that is, $c(g) = c_0 + c_1 g$.

2. The housing markets clear in all communities.
3. Households vote on feasible tax rates and levels of public goods in each community. Hence, the composition of each community, the net-of-tax price of housing, and the aggregate housing consumption are determined prior to voting. Voters treat the population boundaries of each community and the housing market outcomes as fixed when voting. This timing assumption then implies that the set of feasible policies at the voting stage is given by the following equation:

$$p_j(g) = p_j^h + \frac{c(g_j)}{H_j/P(C_j)}. \quad (2.19)$$

This set is also sometimes called the government-services possibility frontier (GPF) in the literature.

Consider a point (g^*, p^*) on the GPF. We say that (g^*, p^*) is a majority rule equilibrium if there is no other point on the GPF $(\hat{g}, \hat{p})$ that would beat (g^*, p^*) in a pairwise vote.¹⁸

A voter's preferred level of g is then obtained by maximizing the indirect utility function $V(\alpha, g_j, p_j, \gamma)$ subject to the feasibility constraint derived above. Single crossing implies that for any level of income γ , the single-crossing properties imply that households with higher (lower) values of α will have higher (lower) demands for local public goods. As a consequence, there exists a function $\tilde{\alpha}_j(\gamma)$ that characterizes the set of pivotal voters. This function is implicitly defined by the following condition:

$$\int_0^\infty \int_{\alpha_{j-1}(\gamma)}^{\tilde{\alpha}_j(\gamma)} f(\alpha, \gamma) d\alpha d\gamma = \frac{1}{2} P(C_j). \quad (2.20)$$

Given our parameterization, the locus of decisive voters is given by

$$\ln \alpha - \rho \left(\frac{\gamma^{1-\nu} - 1}{1-\nu} \right) = L_j = \ln \left(\frac{B e^{-\rho} \frac{B p_j^{\eta+1} - 1}{1+\eta} p_j^\eta p_j'(g)}{g_j^{\rho-1}} \right). \quad (2.21)$$

See Figure 2.2 for an illustration of this locus.

2.3.1.4 Equilibrium

Definition 2.1

An *intercommunity equilibrium* consists of a set of communities, $\{1, \dots, J\}$; a continuum of households, C ; a distribution, P , of household characteristics α and γ ; and a partition of C across communities $\{C_1, \dots, C_J\}$, such that every community has a positive population—that is, $0 < n_j^* < 1$; a vector of prices and taxes, $(p_1^*, t_1^*, \dots, p_J^*, t_J^*)$; an

¹⁸ Note that in this model, sincere voting is a dominant strategy.

allocation of public good expenditures, $(g_1^*, \dots, g_f^*)$; and an allocation, (h^*, b^*) , for every household (α, γ) , such that the following hold:

1. Every household, (α, γ) , living in community j maximizes its utility subject to the budget constraint¹⁹

$$\begin{aligned} (h^*, b^*) &= \arg \max_{(h, b)} U(\alpha, g_j^*, h, b) \\ \text{s.t. } p_j^* h &= \gamma - b. \end{aligned}$$

2. Each household lives in one community and no household wants to move to a different community—that is, for a household living in community j , the following holds:

$$V(\alpha, g_j^*, p_j^*, \gamma) \geq \max_{i \neq j} V(\alpha, g_i^*, p_i^*, \gamma). \quad (2.22)$$

3. The housing market clears in every community:

$$\int_{C_j} h^*(p_j^*, \gamma, \alpha) f(\alpha, \gamma) d\gamma d\alpha = H_j^s \left(\frac{p_j^*}{1 + t_j^*} \right). \quad (2.23)$$

4. The population of each community, j , is given by

$$n_j^* = P(C_j^*) = \int_{C_j} f(\alpha, \gamma) d\gamma d\alpha. \quad (2.24)$$

5. The budget of every community is balanced:

$$\frac{t_j^*}{1 + t_j^*} p_j^* \int_{C_j} h^*(p_j^*, \gamma, \alpha) f(\alpha, \gamma) d\gamma d\alpha / n_j = c(g_j^*). \quad (2.25)$$

6. There is a voting equilibrium in each community: Over all levels of (g_j, t_j) that are perceived to be feasible allocations by the voters in community j , at least half of the voters prefer (g_j^*, t_j^*) over any other feasible (g_j, t_j) .

Existence of equilibrium can be shown under a number of regularity conditions discussed in [Epple et al. \(1984, 1993\)](#). In general, there are no uniqueness proofs, and there is some scope for nonuniqueness in these types of models. Multiple equilibria can arise, since it is possible that different endogenous levels of public good provision are consistent with optimal household decisions and market clearing conditions. As a consequence, these equilibria will have different endogenous housing prices and sorting patterns across communities. However, [Calabrese et al. \(2006\)](#) prove that there can only be one equilibrium that is consistent with a given distribution of community sizes and community ranking; that is, different equilibria will result in different size distributions and (p, g) orderings.

¹⁹ Strictly speaking, all statements only have to hold for almost every household; deviations of behavior of sets of households with measure zero are possible.

2.3.1.5 Properties of equilibrium

Given that we have defined an equilibrium for our model, it is desirable to characterize the properties of equilibria. From the perspective of structural estimation, these properties are interesting, since they provide (a) some predictions that can potentially be tested and (b) necessary conditions that can be exploited to form orthogonality conditions for an estimator.²⁰

Epplé and Platt (1998) show that for an allocation to be a locational equilibrium, there must be an ordering of community pairs, $\{(g_1, p_1), \dots, (g_J, p_J)\}$, such that we have the following:

1. Boundary indifference. The set of border individuals are indifferent between the two communities: $I_j = \{(\alpha, \gamma) \mid V(\alpha, g_j, p_j, \gamma) = V(\alpha, g_{j+1}, p_{j+1}, \gamma)\}$.
2. Stratification. Let $\gamma_j(\alpha)$ be the implicit function defined by the equation above. Then, for each α , the residents of community j consist of those with income, γ , given by $\gamma_{j-1}(\alpha) < \gamma < \gamma_j(\alpha)$.
3. Increasing bundles. Consider two communities i and j such that $p_i > p_j$. Then, $g_i > g_j$ if and only if $\gamma_i(\alpha) > \gamma_j(\alpha)$.
4. Majority voting equilibrium exists for each community and is unique.
5. The equilibrium is the preferred choice of households (γ, α) on the downward-sloping locus $\tilde{\gamma}_j(\alpha)$ satisfying $\int_{\alpha}^{\tilde{\alpha}} \int_{\gamma}^{\tilde{\gamma}_j(\alpha)} f_j(\gamma, \alpha) d\gamma d\alpha = 0.5P(C_j)$.
6. Households living in community j with (γ, α) to the northeast (southwest) of the $\tilde{\gamma}_j(\alpha)$ locus in the (α, γ) -plane prefer a tax that is higher (lower) than the equilibrium.

We will show below how to exploit these properties to estimate the parameters of the model.

2.3.1.6 Computation of equilibrium

Since equilibria can only be computed numerically, we need an algorithm to do so. Note that an equilibrium is characterized by a vector $(t_j, p_j, g_j)_{j=1}^J$. To compute an equilibrium, we need to solve a system of $J \times 3$ nonlinear equations: budget constraints, housing market equilibria, and voting conditions. We also need to check second order conditions once we have found a solution to the system of equations.

Computing equilibria is essential to conducting counterfactual policy analysis, especially if we have strong reasons to believe that policy changes can have substantial general equilibrium effects. It is also important if we want to use a nested fixed point approach to estimation. We will discuss these issues in the next sections in detail.

2.3.1.7 Extensions

Peer effects and private schools

Calabrese et al. (2006) develop an extended model with peer effects. The quality of local public good provision, denoted by q , depends on expenditures per household, g , and a measure of peer quality, denoted by $\bar{y}$:

²⁰ We will show in Section 2.3.2 how to use spatial indifference loci and voting loci to construct an estimator for key parameters of the model.

$$q_j = g_j \left(\frac{\bar{\gamma}_j}{\bar{\gamma}} \right)^\phi, \quad (2.26)$$

where peer quality can be measured by the mean income in a community,

$$\bar{\gamma}_j = \int_{C_j} \gamma f(\alpha, \gamma) d\gamma d\alpha / n_j. \quad (2.27)$$

Ferreyra (2007) also introduced peer effects as well as private school competition within a model with a fixed housing stock to study the effectiveness of different school voucher programs.

Amenities and heterogeneity

One key drawback of the model above is that it assumes that households only sort on the basis of local public good provisions. It is possible to account for exogenous variation in amenities without having to change the structure of the model, as discussed in Epple et al. (2010a). Allowing for more than one endogenous public good is difficult, however, because it is hard to establish the existence of voting equilibrium when voting over multi-dimensional policies. As a consequence, the empirical literature in fiscal competition has primarily considered the model discussed above.

Dynamics

Benabou (1996b), Benabou (2002), and Fernandez and Rogerson (1998) reinterpret the model above using an overlapping generations approach to study fiscal competition. In their models, young individuals do not make any decisions. Hence, individuals make decisions only at one point in time. Epple et al. (2012) then extend the approach and develop an overlapping generations model in which individuals make decisions at different points during the life cycle. This model captures the differences in preferred policies over the life cycle and can be used to study the intergenerational conflict over the provision of public education. This conflict arises because the incentives of older households without children to support the provision of high-quality educational services in a community are weaker than the incentives of younger households with school-age children.

Epple et al. show that the observed inequality in educational policies across communities not only is the outcome of stratification by income, but also is determined by the stratification by age and a political process that is dominated by older voters in many urban communities with low-quality educational services. The mobility of older households creates a positive fiscal externality, since it creates a larger tax base per student. This positive tax externality can dominate the negative effects that arise because older households tend to vote for lower educational expenditures. As a consequence, sorting by age can reduce the inequality in educational outcomes that is driven by income sorting.²¹

²¹ Only a few studies have analyzed voting in a dynamic model. Coate (2011) models forward-looking behavior in local elections that determine zoning policies. He is able to use a more general approach to voting by adopting an otherwise simpler structure in which there is limited housing choice and heterogeneity and housing prices are determined by construction costs.

2.3.2 Identification and estimation

The second step involved in structural estimation is to devise an estimation strategy for the parameters of the model. At this stage, a helpful approach is to check whether the model that we have written down is broadly consistent with the key stylized facts that we are trying to explain. In the context of this application, we know that community boundaries rarely change (Epplé and Romer, 1989). As a consequence, we do not have to deal with the entry or exit of communities. We also know that there is a large amount of variation in housing prices, mean income, expenditures, and property taxes among communities within most US metropolitan areas. Our model seems to be well suited for dealing with those sources of heterogeneity. At the household level, we observe a significant amount of income and housing expenditure heterogeneity both within and across communities. Again, our model is broadly consistent with these stylized facts.

2.3.2.1 The information set of the econometrician

Before we develop an estimation strategy, an essential step is to characterize the information set of the econometrician. Note that this characterization largely depends on the available data sources. If we restrict our attention to publicly available aggregate data, then we can summarize the information set of the econometrician for this application as follows. For all communities in a single metropolitan area, we observe tax rates and expenditures; the marginal distribution of income and community sizes; and a vector of locational amenities, denoted by x . Housing prices are strictly speaking not observed, but can be estimated as discussed in Sieg et al. (2002). Alternatively, they need to be treated as latent.²²

2.3.2.2 Predictions of the model

Next, it is useful to summarize the key predictions of the model:

1. The model predicts that households will sort by income among the set of communities.
2. The model predicts that household sorting is driven by differences in observed tax and expenditure policies, which are, at least, partially capitalized in housing prices.
3. The model predicts that observed tax and expenditure policies must be consistent with the preferences of the decisive voter in each community.

We need to develop a strategy to test the predictions of the model in an internally consistent way.

²² Microdata that contain locational identifiers at the local level are available only through census data centers.

2.3.2.3 Household sorting by income

More formally, the model predicts the *distribution of households by income* among the set of communities. Intuitively speaking, we can test this prediction of the model by matching the predicted marginal distribution of income in each community, $f_j(y)$, to the distribution reported in the US census.

To formalize these ideas, recall that the size of community j is given by

$$P(C_j) = \int_{-\infty}^{\infty} \int_{K_{j-1} + \rho \frac{y^{1-\nu} - 1}{1-\nu}}^{K_j + \rho \frac{y^{1-\nu} - 1}{1-\nu}} f(\ln \alpha, \ln y) d \ln \alpha d \ln y. \quad (2.28)$$

One key insight that facilitates estimation is that we can (recursively) express the community-specific intercepts, $(K_0, \dots, K_J)$, as functions of the community sizes, $(P(C_1), \dots, P(C_J))$, and the parameters of the model:

$$\begin{aligned} K_0 &= -\infty, \\ K_j &= K_j(K_{j-1}, P(C_j) \mid \rho, \mu_\gamma, \sigma_\gamma, \mu_\alpha, \sigma_\alpha, \lambda, \nu), \quad j = 1, \dots, J-1, \\ K_J &= \infty. \end{aligned} \quad (2.29)$$

The intuition for this result is simple.²³ By definition, $K_0 = -\infty$, which establishes the lower boundary for community 1. As we increase the value of K_1 , we push the boundary locus that characterizes the indifference between communities 1 and 2 to the northwest in Figure 2.2. We keep increasing the value of K_1 until the predicted size of the population of community 1 corresponds to the observed population size. This step of the algorithm then determines K_1 . To determine K_2 , we push the boundary locus that characterizes the indifference between communities 2 and 3 to the northwest by increasing the value of K_2 . We continue in this way until all values of K_j have been determined.²⁴ Finally, note that one could also start with the richest community and work down.

Let q be any given number in the interval $(0, 1)$, and let $\zeta_j(q)$ denote the q th quantile of the income distribution—that is, $\zeta_j(q)$ is defined by $F_j[\zeta_j(q)] = q$. We observe the empirical income distribution for each community. An estimator of $\zeta_j(q)$ is given by

$$\zeta_j^N(q) = F_{j,N}^{-1}(q), \quad (2.30)$$

where $F_{j,N}^{-1}(\cdot)$ is the inverse of the empirical distribution function.

The q th quantile of community j 's income distribution predicted by the model is defined by the following equation:

²³ For a formal proof, see [Epple and Sieg \(1999\)](#).

²⁴ Note that this algorithm is similar to the share inversion algorithm proposed in [Berry \(1994\)](#) for random utility models.

$$\int_{-\infty}^{\ln(\zeta_j(q))} \int_{K_{j-1} + \rho \frac{y^{1-\nu}-1}{1-\nu}}^{K_j + \rho \frac{y^{1-\nu}-1}{1-\nu}} f(\ln \alpha, \ln y) d \ln \alpha d \ln y = q P(C_j). \quad (2.31)$$

Given the parameterization of the model, the income distributions of the J communities are completely specified by the parameters of the distribution function, $(\mu_\gamma, \mu_\alpha, \lambda, \sigma_\gamma, \sigma_\alpha)$, the slope coefficient, ρ , the curvature parameter, ν , and the community-specific intercepts, $(K_0, \dots, K_J)$.

[Epple and Sieg \(1999\)](#) use estimates of the 25% quantile, the median, and the 75% quantiles. For notational simplicity, we combine the $3 \times J$ restrictions into one vector:

$$e_N(\theta_1) = \begin{Bmatrix} \ln(\zeta_1(0.25, \theta_1)) - \ln(\zeta_1^N(0.25)) \\ \ln(\zeta_1(0.50, \theta_1)) - \ln(\zeta_1^N(0.50)) \\ \ln(\zeta_1(0.75, \theta_1)) - \ln(\zeta_1^N(0.75)) \\ \dots \\ \ln(\zeta_J(0.25, \theta_1)) - \ln(\zeta_J^N(0.25)) \\ \ln(\zeta_J(0.50, \theta_1)) - \ln(\zeta_J^N(0.50)) \\ \ln(\zeta_J(0.75, \theta_1)) - \ln(\zeta_J^N(0.75)) \end{Bmatrix}, \quad (2.32)$$

where θ_1 is the vector of parameters identified at this stage. [Epple and Sieg \(1999\)](#) show that we can identify and estimate only the following parameters at this stage: $\mu_{\ln y}$, $\sigma_{\ln y}$, λ , $\rho/\sigma_{\ln \alpha}$, and ν .

If the model is correctly specified, the difference between the observed and the predicted quantiles will vanish as the number of households in the sample goes to infinity. The estimation is simplified, since the quantiles of the income distribution of community j depend on (p_j, g_j) only through K_j , which can be computed recursively using the observed community sizes. We can, therefore, estimate a subset of the underlying structural parameters of the model using the following minimum distance estimator:

$$\begin{aligned} \theta_1^N &= \arg \min_{\theta_1 \in \Theta_1} \{e_N(\theta_1)' A_N e_N(\theta_1)\} \\ \text{s.t. } K_j &= K_j(K_{j-1}, P(C_j) \mid \theta_1), \quad j = 1, \dots, J-1, \end{aligned}$$

where θ_1 is the unknown parameter vector, and A_N is the weighting matrix. This is a standard nonlinear parametric estimator. Standard errors can be computed using the standard formula described in [Newey and McFadden \(1994\)](#). Note that we need the number of households and not necessarily the number of communities to go to infinity in order to compute asymptotic standard errors.

[Epple and Sieg \(1999\)](#) find that the estimates have plausible values and high precision. The overall fit of the income quantiles is quite remarkable, especially given the fact that the model relies on only a small number of parameters. The model specification is rejected using conventional levels of significance. Rejection occurs largely because we cannot match the lower quantiles for the poor communities very well.

Epplé et al. (2010c) show that it is possible to nonparametrically identify and estimate the joint distribution of income and tastes for public goods.²⁵ More important, the analysis in Epplé et al. (2010c) shows that the rejection of the model reported in Epplé and Sieg (1999) is primarily driven by the parametric log-normality assumptions. If one relaxes this assumption while maintaining all other parametric assumptions made above, one cannot reject the model above solely on the basis of data that characterize community sizes and local income distributions. By construction of the semiparametric estimator developed in Epplé et al. (2010c), we obtain a perfect fit of the observed income distribution for each community. We, therefore, conclude that the type of model considered above is fully consistent with the observed income distributions at the community level.

2.3.2.4 Public good provision

The first stage of the estimation yields a set of community-specific intercepts, K_j . Given these intercepts, the levels of public good provision that are consistent with observed sorting by income are given by the following recursive representation:

$$g_j = \left\{ g_1^\rho - \sum_{i=2}^j (Q_i - Q_{i-1}) \exp(-K_i) \right\}^{1/\rho}. \quad (2.33)$$

To obtain a well-defined econometric model, we need to differentiate between observed and unobserved public good provision. A natural starting point would be to assume that observed public good provision, measured by expenditures per capita, is a noisy measure of the true public good provision.

A slightly more general model specification assumes that the level of public good provision can be expressed as an index that consists of observed characteristics of community j denoted x_j and an unobserved characteristic denoted ϵ_j :

$$g_j = x_j' \gamma + \epsilon_j, \quad (2.34)$$

where γ is a parameter vector to be estimated. The first component of the index $x_j' \gamma$ is local government expenditures with a coefficient normalized to be equal to 1. The characteristic ϵ_j is observed by the households, but is unobserved by the econometrician. We assume that $E(\epsilon_j | z_j) = 0$, where z_j is a vector of instruments. Define

$$m_j(\theta) = g_j - x_j' \gamma. \quad (2.35)$$

²⁵ Technically speaking, the marginal distribution of income is identified. In addition, one can identify only a finite number of points on the distribution of tastes conditional on income. These points correspond to the points on the boundary between adjacent neighborhoods. For points that are not on the boundary loci, we can provide only lower and upper bounds for the distribution. These bounds become tighter as the number of differentiated neighborhoods in the application increases.

We can estimate the parameters of the model using a generalized method of moments estimator, which is defined as follows:

$$\hat{\theta} = \arg \min_{\theta \in \Theta} \left\{ \frac{1}{J} \sum_{j=1}^J z_j m_j(\theta) \right\}' V^{-1} \left\{ \frac{1}{J} \sum_{j=1}^J z_j m_j(\theta) \right\}, \quad (2.36)$$

where z_j is a set of instruments. [Epple and Sieg \(1999\)](#) suggest using the functions of the rank of the community as instruments. Hence, we can identify and estimate the following additional parameters: γ , $\mu_{\ln \alpha}$, $\sigma_{\ln \alpha}$, ρ , and η . [Epple and Sieg \(1999\)](#) find that the estimates are reasonable and that the fit of the model is good. Standard errors can be approximated using the standard formula described in [Newey and McFadden \(1994\)](#). Note that we need the number of communities to go to infinity to compute asymptotic standard errors.

2.3.2.5 Voting

The model determines tax rates, expenditures on education, and mean housing expenditures for each community in the metropolitan area. We need to determine whether these levels are consistent with optimal household sorting and voting in equilibrium. Again, we can take a partial-solution approach and use necessary conditions that voting imposes on observed tax and expenditure policies. This approach was taken in [Epple et al. \(2001\)](#). They find that the simple voting model discussed above does not fit the data. More sophisticated voting models perform better.

Alternatively, we can take a full-solution approach and estimate the remaining parameters of the model using a nested fixed point algorithm. The latter approach is taken in [Calabrese et al. \(2006\)](#). They modify the equilibrium algorithm discussed in [Section 2.3.1.7](#) and compute equilibrium allocations that satisfy (a) optimal household sorting, (b) budget balance, and (c) majority rule equilibrium, and that are consistent with the observed community sizes. These allocations are an equilibrium in the sense that a housing supply function exists for each community that generates a housing market equilibrium. We can then match the equilibrium values for expenditures, tax rates, and average housing consumption to the observed ones using a simulated maximum likelihood estimator. That article confirms the results in [Epple et al. \(2001\)](#) that the simple model does not fit the data. However, an extended model, in which the quality of public goods depends not only on expenditures, but also on local peer effects, significantly improves the fit of the model.

2.3.2.6 Identifying and estimating housing supply functions

Finally, we briefly discuss how to estimate the housing supply function. If one treats the prices of land and structures as known, few methodological problems arise. However, the key problem encountered in estimating the supply function of housing is that the quantity of housing services per dwelling and the price per unit of housing services are not

observed by the econometrician. Instead, we observe the value (or rental expenditures) of a housing unit, which is the product of the price per unit of housing services and the quantity of housing services per dwelling.²⁶

Epple et al. (2010b) provide a new flexible approach for estimating the housing production function that treats housing quantities and prices as latent variables. Their approach to identification and estimation is based on duality theory. Assuming that the housing production function satisfies constant returns to scale, one can normalize output in terms of land use. Although we do not observe the price or quantity of housing, we often observe the value of housing per unit of land. The key insight of that article is that the price of housing is a monotonically increasing function of the value of housing per unit of land. Since the price of housing is unobserved, the attention thus focuses on the value of housing per unit of land instead. Constant returns to scale and free entry also imply that profits of land developers must be zero in equilibrium. One can exploit the zero profit condition and derive an alternative representation of the indirect profit function as a function of the price of land and value of housing per unit of land. Differentiating the alternative representation of the indirect profit function with respect to the (unobserved) price of housing gives rise to a differential equation that implicitly characterizes the supply function per unit of land. Most important, this differential equation depends only on functions that can be consistently estimated by the econometrician. Using a comprehensive database of recently built properties in Allegheny County, Pennsylvania, they found that this new method provides reasonable estimates for the underlying production function of housing and the implied housing supply function.

2.3.3 Policy analysis

Once we have found a model that fits the data well and passes the standard specification tests, we can use the model to perform counterfactual policy analysis. Here, we consider two applications. The first one estimates welfare measures for air quality improvements. The second application focuses on the benefits of decentralization.

2.3.3.1 Evaluating regulatory programs: the Clean Air Act

An important need is to evaluate the efficiency of public regulatory programs such as the Clean Air Act. Most methods commonly used in cost–benefit analyses are designed to consider relatively small projects that can be evaluated within a partial equilibrium framework. Sieg et al. (2004) show how to use the methods discussed above to develop an approach for evaluating the impact of large changes in spatially delineated public goods

²⁶ This problem is similar to the omitted price problem that is encountered in the estimation of production functions. That problem arises because researchers typically observe only revenues and not prices and quantities. If there is a large local or regional variation in product prices, revenues are not a good proxy for quantity.

or amenities on economic outcomes. They study Los Angeles, which has been the city in the United States with the worst air quality. As a consequence, we have access to high-quality data because southern California has a good system of air quality monitors. Between 1990 and 1995, southern California experienced significant air quality improvements. Ozone concentrations were reduced by 18.9% for the study area as a whole. Ozone changes across communities ranged from a 2.7% increase to a 33% decline. In Los Angeles County, the number of days that exceeded the federal 1 h ozone standard dropped by 27% from 120 to 88 days. We want to estimate welfare measures for these improvements in air quality.

One important distinction is to differentiate between partial and general equilibrium welfare measures. As pointed out by [Scotchmer \(1986, pp. 61–62\)](#), “an improvement to amenities will induce both a change in property values and a change in the population of the improved area. Short-run benefits of an improvement are those which accrue before the housing stock, or distribution of population, adjusts. Long-run benefits include the benefits which accrue when the housing stock and distribution of population change. The literature has not dwelled on the distinction between benefits in the short run and long run, probably because the value of *marginal* improvements is the same in both cases.” Consider the case in which we exogenously change the level of public good provision in each community from g_j to $\bar{g}_j$. In our application, the change in public good provision arises from improvements in air quality that are due to federal and state air pollution policies. The conventional partial equilibrium Hicksian willingness to pay, WTP_{PE} , for a change in public goods is defined as follows:

$$V(\alpha, \gamma - \text{WTP}_{\text{PE}}, \bar{g}_j, p_j) = V(\alpha, \gamma, g_j, p_j). \quad (2.37)$$

Households will adjust their community locations in response to these changes. Such an analysis implies that housing prices can change as well. An evaluation of the policy change should reflect the price adjustments stemming from any changes in community-specific public goods. We can define the general equilibrium willingness to pay as follows:

$$V(\alpha, \gamma - \text{WTP}_{\text{GE}}, \bar{g}_k, \bar{p}_k) = V(\alpha, \gamma, g_j, p_j), \quad (2.38)$$

where $k(j)$ indexes the community chosen in the new (old) equilibrium. Since households may adjust their location, the subscripts for $(\bar{g}_k, \bar{p}_k)$ need not match (g_j, p_j) .

Using data from Los Angeles in 1990, [Sieg et al. \(2004\)](#) estimate the parameters of a sorting model that is similar to the one discussed in the previous sections. They find that willingness to pay ranges from 1% to 3% of income. The model predicts significant price increases in communities with large improvements in air quality and price decreases in communities with small air quality improvements. Partial equilibrium gains are thus often offset by price increases. At the school district level, the ratio of general to partial equilibrium measures ranges from 0.28 to 8.81, with an average discrepancy of nearly 50%. Moreover, there are large differences between the distributions of gains in partial versus general equilibrium.

Sieg et al. (2004) use the projected changes in ozone concentrations for 2000 and 2010, together with the estimates for household preferences for housing, education, and air quality, to conduct a prospective analysis of policy changes proposed by the Environmental Protection Agency. They measure general equilibrium willingness to pay for the policy scenarios developed for the prospective study as they relate to households in the Los Angeles area. Estimated general equilibrium gains from the policy range from \$33 to \$2400 annually at the household level (in 1990 dollars).²⁷

2.3.3.2 Decentralization versus centralization

One of the key questions raised in the seminal article of Tiebout (1956) is whether decentralized provision of local public goods, together with sorting of households among jurisdictions, can result in an efficient allocation of resources. It is not difficult to construct some simple examples in which allocations are not efficient in Tiebout models (Bewley, 1981). However, this question is more difficult to answer once we consider more realistic models. Moreover, we would like to have some idea about the quantitative magnitude of potential inefficiencies.

Calabrese et al. (2012) attempt to answer both sets of questions. First, they derive the optimality conditions for a model that is similar to the one developed in Section 2.3.1. They show that an efficient differentiated allocation must satisfy a number of fairly intuitive conditions. First, the social planner relies on lump-sum taxes and sets property taxes equal to zero. The planner does not rely on distortionary taxes. Second, the level of public good provision in each community satisfies the Samuelson condition. Finally, each household is assigned to a community that maximizes the utility of the household. The last condition is not obvious because of the fiscal externalities that households provide.

The second step of the analysis, then, is to try to quantify the potential efficiency losses that arise in equilibria. They calibrated the model and compared welfare in property tax equilibria, both decentralized and centralized, with the efficient allocation. Inefficiencies with decentralization and property taxation are large, dissipating most if not all of the potential welfare gains that efficient decentralization could achieve. In property tax equilibria, centralization is frequently more efficient! An externality in community choice underlies the failure to achieve efficiency with decentralization and property taxes: poorer households crowd richer communities and free ride by consuming relatively little housing, thereby avoiding taxes. They find that the household average compensating variation for adopting the multijurisdictional equilibrium is \$478. The per household

²⁷ Tra (2010) estimates a random utility model using a similar data set for Los Angeles. His findings are comparable to the ones reported in Sieg et al. (2004). Wu and Cho (2003) also study the role of environmental amenities in household sorting. Walsh (2007) estimates a model that differentiates between publicly and privately provided open space to study policies aimed at preventing urban sprawl in North Carolina.

compensating variation for land owners is $-\$162$. Hence, the decentralized Tiebout equilibrium implies a welfare loss equal to $\$316$ per household. This equals 1.3% of 1980 per household income.

2.4. THE ALLOCATION OF ECONOMIC ACTIVITY ACROSS SPACE

Understanding how economic activity is allocated across space is a core subject in urban and regional economics. This section considers two applications related to the topic: the regional specialization of industry and the internal structure of cities. We begin by developing models used in the two applications and discuss identification and estimation. Finally, we address various issues that need to be confronted when using the estimated models to evaluate the effects of counterfactual policies.

Although the focus is on methodology, we want to emphasize the interesting questions that can be addressed with structural models along the lines that we discuss. The first application is a model in which locations specialize in industries. With a successful quantitative model, we can evaluate questions such as how investments in transportation infrastructure affect the pattern of regional specialization. The second application is a model of where people live and work in a city, and it takes into account economies of density from concentrating workers and residents in particular locations. If we succeed in developing a computer-generated quantitative model of the city, we can evaluate how regulations, subsidies, or investments in infrastructure affect where people live and work, and how these policies affect levels of productivity and welfare. Note that, befitting its importance for the field, other chapters in this handbook delve into various aspects of the allocation of economic activity across space. In particular, [Chapter 5](#), by Combes and Gobillon, reviews empirical findings in the literature on agglomeration, including results from structural approaches.²⁸ And [Chapter 8](#), by Duranton and Puga, reviews the theoretical and empirical literature on urban land use. Although the other chapters focus primarily on results, again, the focus here is on methodology.

2.4.1 Specialization of regions

The first application is based on articles that apply the [Eaton and Kortum \(2002\)](#) model of trade to a regional context, with regions the analog of countries. Note that in our second application on the internal structure of cities that follows, we will assume that workers are mobile across different locations in a city. In contrast, here in our first application, there is no factor mobility across locations; only goods flow. [Donaldson \(forthcoming\)](#) applies the framework to evaluate the regional impact of investments in transportation infrastructure. [Holmes and Stevens \(2014\)](#) apply the framework to evaluate the effects of increased imports from China on the regional distribution of manufacturing within the United States. In the exposition, we focus on the [Holmes and Stevens \(2014\)](#) version.

²⁸ See also [Combes et al. \(2011\)](#) and [Rosenthal and Strange \(2004\)](#).

2.4.1.1 Model development

Suppose there is a continuum of different goods in an industry, with each good indexed by $\omega \in [0, 1]$. There are J different locations indexed by j . For expositional simplicity, assume for now there is a single firm at location j that is capable of producing good ω . Let $z_{\omega,j}$ be the firm's productivity, defined as output per unit input, and let w_j be the cost of one input unit at location j . Let $\mathbf{z}_\omega \equiv (z_{\omega,1}, z_{\omega,2}, \dots, z_{\omega,J})$ denote the vector of productivity draws across all firms, and let $F(\mathbf{z}_\omega)$ be the joint distribution. There is a transportation cost to ship goods from one location to another. As is common in the literature, we assume iceberg transportation costs. Specifically, to deliver one unit from j to k , $d_j^k \geq 1$ units must be delivered. Assume $d_j^j = 1$ and $d_j^k > 1$, $k \neq j$ —that is, there is no transportation cost for same-location shipments, but there are strictly positive costs for shipments across locations. The cost for firm j to deliver one unit to k is then

$$c_{\omega,j}^k = \frac{w_j d_j^k}{z_{\omega,j}}. \quad (2.39)$$

The minimum cost of serving k over all J source locations is

$$\underline{c}_\omega^k = \min_j c_{\omega,j}^k, \quad (2.40)$$

and let j^k be the firm solving (2.40), the firm with the lowest cost to sell to k . If the joint distribution $F(\mathbf{z}_\omega)$ is continuous, the lowest-cost firm j^k is unique except for a set of measure zero. If firms compete on prices in a Bertrand fashion in each market k , the most efficient firm for k , firm j^k , gets the sale. For a given product ω , the likelihood the firm at j is the most efficient for k depends on the joint distribution of productivity draws, transportation costs d_j^k , and input costs $(w_1, w_2, \dots, w_J)$.

Eaton and Kortum (2002) make a particular assumption on the joint distribution $F(\mathbf{z}_\omega)$ that yields an extremely tractable framework. Specifically, productivity draws of individual firms are assumed to come from the Fréchet distribution. The draws across firms are independent, and the cumulative distribution function (c.d.f.) for a firm at location j is given by

$$F_j(z) = e^{-T_j z^{-\theta}}. \quad (2.41)$$

The *shape* parameter θ governs the curvature of the distribution and is constant across locations; the lower θ , the greater the variation in productivity draws across firms. The *scale* parameter T_j allows locations to differ in mean productivity; the higher T_j , the higher the average productivity drawn by a firm at location j . Let $G_j^k(c)$ be the c.d.f. of the cost c_j^k of firm j to ship goods to k . This can be derived by plugging (2.39) into (2.41). It is convenient to write the equation in terms of the complement of the c.d.f. (the probability of drawing above c_j^k):

$$1 - G_j^k(c_j^k) = e^{-T_j (w_j d_j^k)^{-\theta} (c_j^k)^\theta}. \quad (2.42)$$

This equation has the same functional form as (2.41), only now the scale parameter takes wages and transportation costs into account. Consider the c.d.f. $\underline{G}^k(\underline{c}^k)$ of $(\underline{c}^k)$, the lowest cost across all sources. Writing the equation in terms of its complement, we calculate the probability that the cost is higher than $\underline{c}^k$ at all locations—that is,

$$\begin{aligned} 1 - \underline{G}^k(\underline{c}^k) &= \prod_{j=1}^J \left[1 - G_j^k(\underline{c}^k) \right] \\ &= e^{-\sum_{j=1}^J T_j \left(w_j d_j^k \right)^{-\theta} (\underline{c}^k)^\theta}. \end{aligned} \quad (2.43)$$

Note that the shape of the functional form of (2.43) is the same as (2.42), only now the scale factor is the sum of the scale factors of the cost distributions across the different locations. This is a convenient property of the Fréchet. Moreover, straightforward calculations yield the following expression for the probability that the firm at j is the lowest-cost source for serving location k :

$$\pi_j^k = \frac{T_j \left(w_j d_j^k \right)^{-\theta}}{\sum_{s=1}^J T_s \left(w_s d_s^k \right)^{-\theta}}. \quad (2.44)$$

This formula is intuitive. The numerator is an index of firm j 's efficiency to sell at k , varying proportionately with the productivity parameter T_j , and inversely with input costs and transportation costs to get from j to k . The formula takes firm j 's efficiency relative to the sum of the efficiency indices across all source locations. In [Eaton and Kortum \(2002\)](#), firms price competitively. [Bernard et al. \(2003\)](#) extend the framework to an oligopoly setting. Under the assumption that demand has constant elasticity, both treatments show that the share of sales at location k , sourced from location j , is given by formula (2.44). Hence, if X^k denotes total industry expenditure at location k , and Y_j^k the sales of firms at j to k , and if Y_j equals total sales at j to all destinations, then

$$Y_j = \sum_{k=1}^S Y_j^k = \sum_{k=1}^S \frac{T_j \left(w_j d_j^k \right)^{-\theta}}{\sum_{s=1}^J T_s \left(w_s d_s^k \right)^{-\theta}} X^k. \quad (2.45)$$

This is a useful equation that links expenditures and sales at each location with the location-level productivity parameters, input prices, and transportation costs. From the formula, we can see that an industry will tend to concentrate at a particular location j if its productivity is high, if input costs are low, and if the costs of transportation to locations with high expenditures are low.²⁹ The second application below uses the same

²⁹ [Anderson and van Wincoop \(2003\)](#) derive a similar equation in an alternative formulation.

Fréchet magic to derive tractable expressions of equilibrium commuting flows between different locations in the same city.

2.4.1.2 Estimation and identification

We now turn to the issue of estimation and identification. To impose more structure on transportation costs, let m_j^k be the distance in miles between locations j and k , and assume the iceberg transportation cost depends only on distance—that is, $d_j^k = f(m_j^k)$, where $f(0) = 1$, and $f'(m) > 0$. Next, define a function $h(m)$ by

$$h(m_j^k) \equiv \left(d_j^k\right)^{-\theta} = f(m_j^k)^{-\theta}. \quad (2.46)$$

We can think of this as a *distance discount*. It equals 1 when the distance is zero and strictly declines as the distance increases, depending on the rate at which the iceberg transportation cost increases, as well as the shape parameter θ of the productivity distribution. Next, define $\gamma_j \equiv T_j w_j^{-\theta}$, a composite of the technology measure T_j , the wage at j , and the shape parameter θ . In a partial equilibrium context, where the wage w_j is fixed and the technology level T_j is exogenous, the composite parameter γ_j can be treated in a structural way now. We discuss alternatives in the discussion of policy below.

Using our definitions of $h(m_j^k)$ and γ_j , we can then rewrite (2.45) as

$$Y_j = \sum_{k=1}^S \frac{\gamma_j h(m_j^k)}{\sum_{s=1}^J \gamma_s h(m_s^k)} X^k, \quad j = 1, \dots, J. \quad (2.47)$$

Suppose for the sake of discussion that the distance discount function $h(\cdot)$ is known for the particular industry under consideration. Suppose we have data $\{Y_j, X^k, m_j^k, \text{ all } j \text{ and } k\}$ —that is, the value of production at each location, absorption at each location, and distance information. The vector of cost efficiencies $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_J)$ is identified from the set of equations given by (2.47). The identification is subject to a rescaling by a positive multiplicative constant, so a normalization is required, e.g., $\gamma_1 = 1$, if $Y_1 > 0$. See Proposition A.1 in the appendix of [Ahlfeldt et al. \(2014\)](#) for a proof that a unique $\boldsymbol{\gamma}$ exists that solves (2.47), again subject to a normalization. The appendix in [Holmes and Stevens \(2014\)](#) describes an iterative procedure to obtain a solution as a fixed point. Think of the γ_j as a location-level fixed effect that is solved for to exactly fit the data. [Redding and Sturm \(2008\)](#) and [Behrens et al. \(2013\)](#) perform similar calculations.

The above consideration takes as given the distance discount $h(m)$. Suppose the discount is unknown *a priori*. In this case, data on the distances that shipments travel are useful. A long tradition in the trade literature examines how trade flows vary with distance; one example is the gravity model considered in [Anderson and van Wincoop \(2003\)](#). Here, we focus on the approach taken in [Holmes and Stevens \(2014\)](#). In the census data used in the study, total shipments originating across all plants at a given location j are observed (this is Y_j). In addition, an estimate of absorption at each destination (i.e., X^k) is also obtained. In addition to these aggregate quantities, the article employs

data from a random sample of individual transactions, for which the origin and destination are provided. Let the distance discount function be parameterized by a vector $\boldsymbol{\eta}$ —that is, we write $h(m, \boldsymbol{\eta})$. The article jointly estimates $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_J)$ and $\boldsymbol{\eta}$ by choosing $(\boldsymbol{\gamma}, \boldsymbol{\eta})$ to maximize the likelihood of the shipment sample, subject to $(\boldsymbol{\gamma}, \boldsymbol{\eta})$, satisfying (2.47) for the given values of Y_j and X^k . If shipments in the data tend to go short distances, the estimated distant discount $h(m, \hat{\boldsymbol{\eta}})$ will tend to drop sharply with distance (examples in the data include industries like ready-mix cement and ice). In cases in which shipments travel long distances, the estimated distance discount will be relatively flat at 1 (an example is medical equipment).

2.4.2 Internal structure of cities

Our discussion is based on the work of Ahlfeldt et al. (2014), who estimate a structural model of the city of Berlin. (See also Duranton and Puga (2015) in this volume for a discussion of the work of Ahlfeldt et al. (2014) that complements ours.) Theories of the internal structure of cities focus on flows of commuters from their place of residence to their place of work, and the spillover benefits from economies of density. The city of Berlin provides a fascinating context because of the way the Berlin Wall blocked such flows. The paper uses data for periods before, during, and after the existence of the Berlin Wall to estimate a rich model that simultaneously takes into account both commuter and spillover flows.

The paper builds on a long tradition in urban economics research on the internal structure of cities, dating back to the literature on the monocentric model of the city. This classic early model is useful for illustrating theoretical points, such as how a change in commuting costs affects land prices. Yet this abstraction, in which land is used for residence and not for production, and where all residents commute to work at a single point, does not correspond to what actual cities look like. Lucas and Rossi-Hansberg (2002) provided an important generalization in which land is used for both residence and production. Yet again, this structure aims at theoretical points, and one abstraction is that a city is a perfect circle with uniform rings. Furthermore, there is no worker heterogeneity, with the implication that all workers living in a given part of the city would commute to the same place for work. Ahlfeldt et al. (2014) estimate a structural model of an actual city, and its approach departs from these various simplifications. Their model explicitly takes into account that land features are not uniform over space and that cities are not circles. It takes into account that individuals are heterogeneous and may vary in their match quality with particular employers, and in match quality with particular places to live. Finally, the model allows for spillovers to arise on the consumption side as well as on the production side.

2.4.2.1 Model development

We provide a brief overview of the modeling setup. Individuals are freely mobile and choose whether or not to live in the city, and if so, where to live and where to work,

from a choice of J discrete locations. Firms are also freely mobile about where to produce, and a given parcel of land can be used for production or residence. Productivity varies across locations, because of the exogenous features of land, as well as endogenously, through the levels of neighboring employment and the resulting spillovers. Specifically, the productivity index A_j at location j is given by

$$A_j = Y_j^\lambda a_j, \quad (2.48)$$

where a_j is the exogenous location quality, and Y_j is aggregated spillovers received by j from all other city locations, defined by

$$Y_j = \sum_{k=1}^J e^{-\delta m_j^k} \tilde{Y}_k, \quad \lambda \geq 0, \delta \geq 0. \quad (2.49)$$

In this expression, $\tilde{Y}_k$ is employment at location k , and m_j^k is the distance between locations i and j . The parameter δ governs how rapidly spillovers decline with distance. The parameter λ determines how the aggregated spillovers convert into productivity gains. Analogously, there is an exogenous consumption amenity level b_j at location j and an endogenous spillover component from neighboring residents, with the same functional form as for the production side, but with different parameters. The last pieces of the model relate to individual choice. Individuals who choose to live in the city obtain match quality draws for every possible combination of where they might live and where they might work. Commuting costs create tension between these two considerations. Besides commuting costs and match quality, individuals need to take into account how wages vary by location in their decision of where to work. In the decision of where to live, they need to take into account housing rents and consumption amenities.

Note that the model is very flexible and general in the way that exogenous productivity a_j is free to vary across locations. Analogously, the exogenous consumption amenity b_j is free to vary. Allowing for this generality is important because if this variation exists and we ignore it, we might mistakenly attribute all the observed concentration of employment or residence to spillovers, when exogenous variations in land quality also play a role.

For technical convenience, analogous to the first application, [Ahlfeldt et al. \(2014\)](#) make use of the Fréchet structure of [Eaton and Kortum \(2002\)](#), regarding the distribution of workplace/residence match qualities. The assumption yields a tractable approach.

2.4.2.2 Estimation and identification

In our first application, the logic behind the identification of location-specific productivities and distance discounting (the parameters given by (γ, η)) is straightforward. The issues are more complex in the [Ahlfeldt et al. \(2014\)](#) model of residential and worker location within a city. We highlight two challenges in particular. First, separating out

natural advantage (given by the exogenous productivity component a_j at each location j) from knowledge spillovers (the elasticity λ listed above) is intrinsically difficult. Suppose we see in the data that at locations with a high density of workers, land rents are high. Is this because locations with high exogenous productivity a_j are attracting a large number of workers and this bids up rents? Or does causation go the other way, such that locations with a high concentration of workers are more productive, which in turn bids up rents? Or does the answer lie somewhere in between?

The second issue is that when there are knowledge spillovers, there is a potential for multiple equilibria to exist at given values of the model's structural parameters. For example, workers might cluster at point A just because everyone else is clustering there (i.e., the cluster is self-fulfilling). Perhaps an alternative equilibrium also exists where workers cluster at some different point B. The possibility of multiplicity has potential implications for estimation and identification as well as for policy analysis.

Ahlfeldt et al. (2014) confront these issues by exploiting the historical context of the Berlin Wall going up and coming down. They treat these events as quasi-experimental variation that can be used to identify the structural parameters of the model. Data were collected at a fine geographic level, 16,000 city blocks, and include the number of residents X_t^j in block j at time t , the number of workers $Y_{j,t}$ employed at j at time t , and the rental price of land $r_{j,t}$ at time t for block j . The wage at location j plays the same role in the Ahlfeldt et al. (2014) model as the productivity variable T_j plays in the industry specialization application, and there is a formula in Ahlfeldt et al. (2014) that is analogous to (2.45). Location-level wages are unobserved and are inferred in a way that is analogous to the way that unobserved location-level productivities were inferred in the regional specialization application.

Let β be a vector that collects all of the various parameters of the model, such as the knowledge spillover elasticity λ and the spatial discount parameter δ that appear in the productivity specification (2.48). Let $a_{j,t}$ and $b_{j,t}$ be the natural advantage parameters for production and consumption at location j at time t , which we write in vector form as $\mathbf{a}_t$ and $\mathbf{b}_t$, with elements for each of the J locations. Let $(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{r}_t)$ be the vector of data that contains the number of residents, number of workers, and the rental rate for each block. Although there may be multiple equilibria, a key result of the paper is that for a fixed parameter vector β and a given data realization $(\mathbf{X}_t, \mathbf{Y}_t, \mathbf{r}_t)$, there exists unique values of $(\mathbf{a}_t, \mathbf{b}_t)$ consistent with equilibrium.³⁰ For intuition, recall the earlier discussion that if in the data we see high concentration and high rents, we can account for these findings by giving all the credit to natural advantage and none to spillovers, or all of the credit to spillovers and none to natural advantage, or something in between. But in the present discussion, when we take the parameter vector β as given, as well as the data, we are *fixing* the credit given to spillovers, and the resulting values $(\mathbf{a}_t, \mathbf{b}_t)$ can be thought of as the residual credit that must be given to natural advantage, in order

³⁰ This is uniqueness, subject to some normalizations.

for the equilibrium conditions to hold. So in terms of estimation, the second issue noted above, about the potential multiplicity of equilibrium, ends up not being a concern.

We now turn to the first challenge, disentangling spillovers and natural advantage. Following the above discussion, for a given set of model parameters and the observed data, the article infers the implied values of natural advantage in production a_j and consumption amenity b_j for each location j . The key identifying assumption is that any changes in these natural advantage variables over time are unrelated to the distance of a location from the Berlin Wall. The article estimates significant levels of spillovers for both production and consumption. Remarkably, the estimates based on what happened between 1936 and 1986, when the Berlin Wall went up, are very similar to the estimates based on 1986 and 2006, when the Berlin Wall went down. The key feature of the data that drives estimates of spillovers is that after the Berlin Wall was erected, land prices collapsed near it. The pattern reversed when the Berlin Wall was taken down. To understand how this works in the model, suppose we shut down knowledge spillovers. The sharp drops in land prices near the Berlin Wall imply that natural advantage must have systematically declined near the Berlin Wall. This is inconsistent with the identifying assumption.

2.4.3 Policy analysis

As emphasized in [Section 2.1](#), a key benefit of the structural approach to empirical work is that prospective policy analysis can be conducted with the estimated model. At the beginning of this section, we mentioned a variety of interesting policy issues that can be addressed with the class of models discussed here. Now we focus on a particular case that is useful for illustrating methodological points. In the model of industry specialization, we evaluate how opening up the domestic industry to foreign competition affects the regional distribution of production. [Holmes and Stevens \(2014\)](#) conduct such an exercise by evaluating the regional impact of imports from China, and here we consider a simpler version of the experiment.

Following our discussion above of the regional specialization model, we begin with our estimates of the vector γ of cost efficiency indices across locations and the parameters η governing distance discounts $h(m, \eta)$. Suppose imports are initially banned. The specific policy change we consider is to allow imports, subject to a quota. Suppose the world market is such that imports will flow in, up to the quota. Suppose the quota is set in such a way that the value of imports will equal 5% of the total domestic market. Assume for simplicity that all imports must go through the same port, which is at some new location $J + 1$, and the distance discount from here to other locations follows the same distance discount estimated in the first stage. Assume that the industry under consideration is relatively small, such that imports do not affect wages. Finally, make Cobb-Douglas assumptions about consumer utility so that relative spending shares on the industry X^k/X^j between any pair of locations k and j do not change.

Putting all of these assumptions together, we see that the policy is equivalent to creating a new location $J + 1$, with its own efficiency index γ_{J+1} and no consumption—that is, $X_{J+1} = 0$ —holding fixed the cost efficiency indices of the other locations $\gamma_j, j \leq J$, and the distance discounts $h(m, \eta)$. For any given value of γ_{J+1} , we can use Equation (2.47), now extended to sum up to $J + 1$, to solve for the sales of each location Y_j^{new} , where “new” means after the policy change. The higher γ_{J+1} , the greater are imports Y_{J+1}^{new} and the lower domestic production at each location $Y_j^{\text{new}}, j \leq J$. We pick γ_{J+1} such that the value of imports Y_{J+1}^{new} is 5% of the domestic market. We then compare Y_j^{new} with Y_j^{old} to examine the regional impact of trade. In general, the effects vary across locations, depending on the role of transportation costs (domestic producers near the port will be hurt more than others), a location’s productivity, and the productivity of a location’s neighbors.

We now have in place an example structural model, for which we laid out the issues of estimation and identification, and have presented an illustrative policy experiment. Next we use the example to address various issues.

First, notice that we were able to conduct this particular experiment without having to unpack the estimated distance function $h(m, \eta)$ into underlying parts. Remember this is a composite of other parameters. We are able to do this because the underlying policy change being considered leaves distance discounting alone. Of course, there are other policy changes, such as infrastructure investment to reduce transportation costs, for which we would need estimates of these deeper structural parameters to conduct policy analysis. [Donaldson \(forthcoming\)](#) needs these deeper structural parameters in his analysis of the productivity effects of the introduction of the railroad network in India. A key step in his analysis is his use of data on how price varies across space to directly infer transportation costs and how these costs changed after the railroad network was introduced.³¹

Second, we left wages unchanged. If the industry being considered accounts for a significant share of a particular location’s employment, then the policy experiment will lead to local wage changes. That is, the cost efficiency parameter $\gamma_j = T_j w_j^{-\theta}$ being held fixed in the exercise now varies. If this is a concern, the analysis must be extended to incorporate a structural model of regional wages. In addition, the shape parameter θ of the productivity distribution needs to be estimated.

Third, we left the productivity parameter T_j unchanged. This is appropriate if productivity reflects natural advantage, but is a concern if knowledge spillovers are potentially important. Suppose, in particular, that the location productivity scaling parameter takes the following form, analogous to that in [Ahlfeldt et al. \(2014\)](#):

$$T_j = a_j N_j^\lambda, \quad (2.50)$$

³¹ For a related analysis, see also [Duranton et al. \(2014\)](#).

where a_j is natural advantage, N_j is industry employment at j , and λ is the knowledge spillover elasticity. So far we have implicitly assumed that $\lambda = 0$, so $T_j = a_j$, but now we consider $\lambda > 0$. In [Eaton and Kortum \(2002\)](#), equilibrium expenditure on inputs at location j is a fraction $\frac{\theta}{1+\theta}$ of revenue, or $w_j N_j = \frac{\theta}{1+\theta} Y_j$. Solving for N_j and substituting (2.50), we can write cost efficiency at j as

$$\gamma_j = T_j w_j^{-\theta} = a_j \left(\frac{\theta Y_j}{w_j} \right)^\lambda w_j^{-\theta}. \quad (2.51)$$

Now suppose we also have data on wages at j . If we take θ and λ as known, following our discussion above, we can solve (2.47) for a unique solution vector $\mathbf{a} = (a_1, a_2, \dots, a_J)$, subject to a normalization. With this setup in place, the analysis can proceed in two ways. The ideal procedure, if feasible, is to go back to the estimation stage to develop a strategy for estimating θ and λ . For example, as in [Ahlfeldt et al. \(2014\)](#), it may be possible to obtain instruments that can be used to construct orthogonality conditions that are satisfied by the vector $\mathbf{a}$ of natural advantages. If estimation of θ and λ is not feasible, then researchers can take a second approach that takes the form of robustness analysis. The estimates under the identifying assumption that $\lambda = 0$ provide the baseline case, and the policy experiment under this assumption is discussed first. Next is a discussion of how results would change if knowledge spillovers are introduced. A variety of estimates of λ can be found in the literature, as discussed in this volume. A value of $\lambda = 0.10$ is generally considered on the high end. Turning to the θ parameter, note that $\frac{\theta}{1+\theta}$ is the variable cost share of revenues. Thus a broad range of θ from 3 to 9 is equivalent to variable cost shares that range from 0.75 to 0.90. This broad range nests values that have been obtained in various applications in the literature (e.g., $\theta = 8.28$ in [Eaton and Kortum, 2002](#)). Now consider re-estimating the model over a grid of θ and λ satisfying $\theta \in [3, 9]$ and $\lambda \in [0, 0.10]$ and resimulating the policy experiment for each case. This provides a range of estimates for the policy effects, with $\lambda = 0$ corresponding to the benchmark case. (In that limit, the choice of θ is irrelevant for the policy experiment.) It may very well be that the baseline results are relatively robust to these alternative assumptions. Transportation cost may be the primary force determining the relative impact of imports across regions (i.e., where those locations closest to ports are affected the most), and knowledge spillovers might be a secondary consideration. If so, the proposed robustness analysis will make this clear. In any case, this discussion highlights how the structural empirical approach yields models that can be built upon and enriched. Rather than speculate about how allowing for agglomeration economies can change an answer, the model can be extended and the answer to the question simulated.

We conclude this discussion of policy experiments by coming back to the issue of multiple equilibria. In the baseline version with $\lambda = 0$, equilibrium is unique. As is well understood in the literature, multiple equilibria may be possible when $\lambda > 0$. In this case,

there is positive feedback, where adding more production lowers costs, increasing the incentive to have still more production, and there are potentially multiple places where an industry might agglomerate. Suppose there is a policy intervention and there are multiple equilibria given the model estimates. Which equilibrium is the relevant one? This issue can be a difficult one, but we can make some observations. First, although multiplicity is possible when $\lambda > 0$, there might be enough curvature (e.g., transportation costs or congestion costs) such that there is a unique equilibrium. If researchers verify uniqueness, this addresses the issue. Second, equilibrium might be unique locally in the vicinity of the baseline case. If the policy intervention is small, a sensible approach may be to focus on the comparative statics of the local equilibrium. Third, it may be possible to estimate the selection process for equilibria, as in [Bajari et al. \(2010a\)](#).

2.4.4 Relation to entry models in the industrial organization literature

When spillovers exist in the models discussed above, interactions are created between decision makers. The study of interactions between decision makers is a general problem in economics. Recently, extensive work has been done on this class of models in the industrial organization literature, focusing on developing partial-solution approaches to study entry by firms into markets, and in particular incorporating dynamics. Here, we connect the discussion above to this literature.

In environments considered in the industrial organization literature, there are often relatively few decision makers, in which case taking into account that entry is discrete may be important. Urban and regional applications often abstract from discreteness in the underlying economic environment, as in the examples above, and this abstraction can be useful when a relatively large number of decision makers are interacting. As research in urban and regional applications takes advantage of new data sets at high levels of geographic resolution, it permits the study of interactions at narrow levels, where there may be relatively few decision makers. In such cases, taking discreteness into account may be useful, and the discussion here illustrates the discrete case. In any case, the partial-solution approaches discussed below can also be scaled up to include cases of large numbers of interacting agents.³² As a starting point for the discussion, a useful step is to review the classic discrete choice model of social interactions in [Brock and Durlauf \(2001\)](#). We can think of this as the approximate state of the literature at the time of publication of the previous handbook (see [Durlauf, 2004](#)). In the model, an agent is making a decision where the agent's payoff depends on the decisions of the other agents. Labeling variables to represent the context of a model of industry agglomeration, suppose that at a given location j , there are I potential entrants indexed by i . Let a_j be a measure of the natural

³² See, for example, [Weintraub et al. \(2008\)](#).

advantage of location j . Let N_j be the total number of firms that enter at location j . Define U_{ij}^E and U_{ij}^N to be firm i 's profit from entering or not entering market j , and suppose profits take the following form:

$$U_{ij}^E = \beta^E + \beta^a a_j + \beta^N N_j + \varepsilon_{ij}^E, \quad (2.52)$$

$$U_{ij}^N = \varepsilon_{ij}^N. \quad (2.53)$$

In this specification, β^a is the weight on natural advantage, and β^N is the weight on firm interactions. The shocks ε_{ij}^E and ε_{ij}^N are independent and identically distributed and are private information observed only by potential entrant i . In a Nash equilibrium, firms will take as given the strategies of the other firms, which specify how their entry decisions will depend on their private shocks. Taking as given these entry strategies by the other firms, let EN_j be the expected count of firm entry perceived by a given firm, conditional on the given firm itself entering. Note $EN_j \geq 1$, because the count includes the firm itself. Substituting expected entry EN_j into the payoff U_{ij}^E , firm i enters if

$$\beta^E + \beta^a a_j + \beta^N EN_j + \varepsilon_{ij}^E \geq \varepsilon_{ij}^N, \quad (2.54)$$

which can be written as a cutoff rule in terms of the difference in shocks,

$$\varepsilon_{ij}^E - \varepsilon_{ij}^N \geq f_{ij}(EN_j) \equiv -(\beta^E + \beta^a a_j + \beta^N EN_j). \quad (2.55)$$

Thus, starting out with a perceived value of expected entry EN_j , we derive the entry rule (2.55), from which we can calculate expected entry. An equilibrium is a fixed point where EN_j maps to itself. As highlighted in [Brock and Durlauf \(2001\)](#), if β^N is positive and large, there can be multiple equilibria. If expected entry is high, then with $\beta^N > 0$, entry is more attractive and high entry is self-fulfilling. If the coefficient on natural advantage β^a is positive, entry will tend to be higher in locations with higher natural advantage.³³

In terms of estimation, [Brock and Durlauf \(2001\)](#) note that if the private shocks are extreme values and if EN_j is observed, then the parameters β^E , β^a , and β^N can be estimated as a standard logit model. Although EN_j may be increasing in a_j , it does so in a nonlinear fashion (through the discrete entry). Since a_j and EN_j are not perfectly collinear, β^a and β^N are separately identified. This is in contrast to the earlier *linear-in-means* formulation in [Manski \(1993\)](#), where it was noted that the analog of EN_j in the model was linear in the analog of a_j , implying that the analogs of β^a and β^N were not separately identified. Researchers are often uncomfortable about relying heavily on functional form assumptions to obtain identification. There is great value in coming up with exclusion restrictions based on the economics of the problem. For example, suppose potential

³³ Note that this monotonicity claim regarding natural advantage a_j ignores complications that may arise with comparative statics when multiple equilibria exist.

entrants vary in productivity ω_i , and suppose the profitability of entry U_{ij}^E above is modified to include an additional term $\beta^\omega \omega_i$ —that is,

$$U_{ij}^E = \beta^E + \beta^\omega \omega_i + \beta^a a_j + \beta^N N_j + \varepsilon_{ij}^E. \quad (2.56)$$

Assume that firm productivities are common knowledge. With $\beta^\omega > 0$, and everything else the same, the higher ω_i , the more likely firm i is to enter. This sets up an exclusion restriction, where a higher value of productivity $\omega_{i'}$ for some other firm i' has no direct effect on firm i 's profitability and affects profitability only indirectly by affecting the likelihood of entry by firm i' .

We now connect the discussion to recent developments in the industrial organization literature. This literature has long been interested in analysis of games with payoff structures such as (2.52), though typically the focus has been on environments in which the interaction parameter β^N is negative—that is, agents are worse off when others enter. For example, if the market is the drugstore market, a firm will be worse off if it has to share the market with more competitors, and in addition the added competition will put downward pressure on prices (Bresnahan and Reiss, 1991). The recent literature has focused on dynamics.³⁴ Going back to the problem as described above, we find dynamics add two elements. First, agents who decide to enter consider not only current profits but also future profits and how future entry will evolve. Second, when agents make entry decisions, in general there may already be incumbent firms in the industry. Although the literature is typically motivated by cases in which $\beta^N < 0$, the technical developments also apply for $\beta^N > 0$.

Let γ_{ijt} be an indicator variable that firm i is an incumbent in location j at time t (i.e., entered previously), and let $\mathbf{y}_t = (\gamma_{1jt}, \gamma_{2jt}, \dots, \gamma_{ijt})$ be the vector listing incumbent status. Analogously, let $\boldsymbol{\omega}$ be the vector of firm productivities. The state of the industry at the beginning of time t at j is $\mathbf{s}_{jt} = (a_j, \boldsymbol{\omega}, \mathbf{y}_t)$ —that is, location natural advantages, firm productivities, and a list of firms that have entered. Let a firm's current period payoff when it participates in market j in period t be given by (2.56). It is straightforward to see how the nested fixed point works here: for a given set of parameters, solve for equilibrium and then vary the parameters to best fit the data according to some metric. However, for computational tractability, the recent literature has focused on two-step approaches, following techniques developed by Hotz and Miller (1993), for discrete choice in labor market applications. The idea is to estimate behavioral relationships in a first stage and then in a second stage back out the parameters that rationalize the behavior.

To explain this, suppose first that the state $\mathbf{s}_{jt} = (a_j, \boldsymbol{\omega}, \mathbf{y}_t)$ is common knowledge for industry participants and is also observed by the econometrician studying the problem (we come back to this below). Moreover, in cases in which there are multiple equilibria, assume the same equilibrium is played conditional on the state $\mathbf{s}_{jt}$ across all the sample

³⁴ See Aguirregabiria and Mira (2010) for a survey.

locations in the data. Given $\mathbf{s}_{jt}$, entry decisions will depend on the realizations of the shocks ε_{ij}^E and ε_{ij}^N for each i and j , and will induce a probability of entry $p_{ij}(\mathbf{s}_{jt})$ for each firm i at j , given $\mathbf{s}_{jt}$. This is a conditional choice probability. Since $\mathbf{s}_{jt}$ is observed by the econometrician, we can obtain an estimate of $\hat{p}_{ij}(\mathbf{s}_{jt})$ from the sample averages. The estimated values $\hat{p}_{ij}(\mathbf{s}_{jt})$ from the first stage summarize an agent's choice behavior. In the second stage, various approaches can recover the structural parameters from the first stage estimates of choice behavior. For the sake of brevity, we consider a simple special case: entry is static (lasts for one period), in which case payoffs look exactly like (2.52). Let $\widehat{E_i N_j}(\mathbf{s}_{jt})$ be an estimate of the expected count of entering firms from the perspective of firm i , given that it enters and given the state. This is constructed as

$$\widehat{E_i N_j}(\mathbf{s}_{jt}) = 1 + \sum_{k \neq i} \hat{p}_{kj}(\mathbf{s}_{jt}). \quad (2.57)$$

If firm i enters, it counts itself in addition to the expected value of all other potential entrants. Now substitute $\widehat{E_i N_j}(\mathbf{s}_{jt})$ for EN_j into (2.56), and the structural parameter vector $\beta = (\beta^E, \beta^\omega, \beta^a, \beta^N)$ can be estimated as a standard logit model.³⁵ The simplicity of the approach is the way in which it takes a potentially complicated model with game-theoretical interactions and boils it down to the estimation of a much more tractable decision-theoretical model. Notice that in the estimation procedure just described, it was not necessary even once to solve for the equilibrium.

Having sketched the approach, we now connect it to our earlier discussion of the work of Ahlfeldt et al. (2014), beginning with the issue of how the potential for multiplicity of equilibria factors into the analysis. In Ahlfeldt et al. (2014), no assumptions about equilibrium selection are made, whereas in the two-step approach, it is necessary to assume that the same equilibrium is played conditional on $\mathbf{s}_{jt}$. Ahlfeldt et al. (2014) provide a full-solution approach. In contrast, the two-step approach is a partial-solution method, and the technical simplicities that it delivers are purchased at the cost of an additional assumption.

Next, recall that Ahlfeldt et al. (2014) are very flexible about allowing for unobserved natural advantage. But ultimately, the paper is able to do this because of the information obtained from the quasi-experimental variation of the Berlin Wall going up and coming back down. The two-step method assumes that the econometrician sees $\mathbf{s}_{jt}$, which is everything except for the private temporary firm-specific shocks ε_{ijt}^E and ε_{ijt}^N . This limitation is a serious one, because the natural expectation is that industry participants have information about locations that an econometrician would not see. Recent work has generalized the two-step approaches to allow for an unobserved, persistent, location-specific quality shock (see Aguirregabiria and Mira, 2007; Arcidiacono and Miller,

³⁵ Bajari et al. (2010b) provide a useful treatment of nonparametric approaches to estimating static models of interactions.

2011; and the discussion in Aguirregabiria and Nevo, 2013). The approach can be viewed as a random effects formulation as opposed to a fixed effect formulation. In particular, permanent location-specific unobserved shocks themselves are not identified, but rather the distribution of the shock is identified. For example, if the pattern in the data is that some locations tend to have persistently low entry levels while other locations have persistently high entry levels, holding fixed the same observable state $\mathbf{s}_{jt}$, this would be rationalized by some dispersion in the random effect.

Two-step approaches have been applied to some topics in urban and regional economics, albeit in only a limited number of cases so far. One example is the work of Suzuki (2013), which uses the approach to examine how land use regulations affect entry and exit in the hotel industry. Another is the work of Bayer et al. (2012), which uses this kind of approach to estimate a model of the demand for housing. In the model, homeowners have preferences over the characteristics of their neighbors and so have to forecast how a neighborhood will evolve. This approach is analogous to a firm making an entry decision in a market and forecasting whether subsequent entry will take place.

An interesting aspect of the two-step approach is the way it provides a bridge between structural estimation and descriptive work. The essence of the first stage is the description of behavior. Yet from this approach, the description of behavior has an interpretation in terms of an equilibrium relationship in a formal model.

2.5. CONCLUSIONS

Structural estimation requires creativity and tenacity; good economic modeling skills; a deep understanding of econometric methods; computational, programming, and data management skills; and an interest in and understanding of public policy. We hope that this survey article will inspire other researchers who are not afraid to work on hard and challenging problems to explore structural estimation approaches in urban economics.

Moving forward, it is not too hard to predict that computer-aided decision making will play a much larger role in the future. Computational capacities, in terms of both software and hardware, will continue to improve. This capacity will provide researchers with the opportunity to develop more powerful algorithms designed to solve complex and challenging problems. By combining the computational power and accuracy of machines with human ingenuity and creativity, we will be able to solve problems that seem completely intractable at this point.

Structural estimation can be viewed as one compelling method for providing quantitative models and algorithms that can be used within a broader framework of decision support systems. In other areas of economics, such as asset pricing and portfolio management, consumer demand analysis, or monetary policy, structurally estimated models are already commonly used to help households, firms, and government agencies make more

informed decisions. The challenge is to develop quantitative models in urban and regional economics that are equally successful. The next generations of urban economists will need to rise to this challenge.

ACKNOWLEDGMENTS

We thank Nate Baum-Snow, Gilles Duranton, Dennis Epple, Vernon Henderson, Andy Postlewaite, and Will Strange for helpful discussions and detailed comments. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Minneapolis, the Federal Reserve Board, or the Federal Reserve System.

REFERENCES

- Aguirregabiria, V., Mira, P., 2007. Sequential estimation of dynamic discrete games. *Econometrica* 75, 1–53.
- Aguirregabiria, V., Mira, P., 2010. Dynamic discrete choice structural models: a survey. *J. Econom.* 156, 38–67.
- Aguirregabiria, V., Nevo, A., 2013. Recent developments in empirical IO: dynamic demand and dynamic games. In: Acemoglu, D., Arellano, M., Deckel, E. (Eds.), *Advances in Economics and Econometrics*. In: Tenth World Congress, vol. 3. Cambridge University Press, Cambridge, pp. 53–122.
- Ahlfeldt, G., Redding, S., Sturm, D., Wolf, N., 2014. The economics of density: evidence from the Berlin Wall. NBER Working paper 20354, July 2014.
- Anderson, J., van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. *Am. Econ. Rev.* 93, 170–192.
- Arcidiacono, P., Miller, R., 2011. Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica* 79, 1823–1867.
- Bajari, P., Kahn, M.E., 2005. Estimating housing demand with an application to explaining racial segregation in cities. *J. Bus. Econ. Stat.* 23, 20–33.
- Bajari, P., Hong, H., Krainer, J., Nekipelov, D., 2010a. Estimating static models of strategic interactions. *J. Bus. Econ. Stat.* 28, 469–482.
- Bajari, P., Hong, H., Ryan, S., 2010b. Identification and estimation of a discrete game of complete information. *Econometrica* 78, 1529–1568.
- Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage premium. *Rev. Econ. Stud.* 79, 88–127.
- Bayer, P., 2001. Exploring differences in the demand for school quality: an empirical analysis of school choice in California, Working paper.
- Bayer, P., Timmins, C., 2005. On the equilibrium properties of locational sorting models. *J. Urban Econ.* 57, 462–477.
- Bayer, P., McMillan, R., Rueben, K., 2004. The causes and consequences of residential segregation: an equilibrium analysis of neighborhood sorting, Working paper.
- Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. *J. Polit. Econ.* 115, 588–638.
- Bayer, P., McMillan, R., Murphy, A., Timmins, C., 2012. A dynamic model of demand for houses and neighborhoods, Working paper.
- Behrens, K., Mion, G., Murata, Y., Sudekum, J., 2013. Spatial frictions. IZA DP Working paper No. 7175.
- Benabou, R., 1996a. Equity and efficiency in human capital investments: the local connection. *Rev. Econ. Stud.* 63, 237–264.
- Benabou, R., 1996b. Heterogeneity, stratification and growth: macroeconomic effects of community structure and school finance. *Am. Econ. Rev.* 86, 584–609.
- Benabou, R., 2002. Tax and education policy in a heterogeneous-agent economy: maximize growth and efficiency? *Econometrica* 70, 481–517.

- Bernard, A., Eaton, J., Jensen, J.B., Kortum, S., 2003. Plants and productivity in international trade. *Am. Econ. Rev.* 93, 1268–1290.
- Berry, S., 1994. Estimating discrete-choice models of product differentiation. *Rand J. Econ.* 25, 242–262.
- Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. *Econometrica* 63, 841–890.
- Berry, S., Linton, O., Pakes, A., 2004. Limit theorems for estimating parameters of differentiated product demand systems. *Rev. Econ. Stud.* 71, 613–654.
- Bewley, T.F., 1981. A critique of Tiebout's theory of local public expenditures. *Econometrica* 49, 713–740.
- Bishop, K., 2011. A dynamic model of location choice and hedonic valuation, *Working paper*.
- Bresnahan, T.F., Reiss, P.C., 1991. Entry and competition in concentrated markets. *J. Polit. Econ.* 99, 977–1009.
- Brock, W., Durlauf, S., 2001. Discrete choice with social interactions. *Rev. Econ. Stud.* 68, 235–260.
- Calabrese, S., Epple, D., Romer, T., Sieg, H., 2006. Local public good provision: voting, peer effects, and mobility. *J. Public Econ.* 90, 959–981.
- Calabrese, S., Epple, D., Romano, R., 2012. Inefficiencies from metropolitan political and fiscal decentralization: failures of Tiebout competition. *Rev. Econ. Stud.* 79, 1081–1111.
- Coate, S., 2011. Property taxation, zoning, and efficiency: a dynamic analysis. NBER Working paper 17145.
- Combes, P., Duranton, G., Gobillon, L., 2011. The identification of agglomeration economies. *J. Econ. Geogr.* 11, 253–266.
- Combes, P., Duranton, G., Gobillon, L., Puga, D., Roux, S., 2012. The productivity advantages of large cities: distinguishing agglomeration from firm selection. *Econometrica* 80, 2543–2594.
- Donaldson, D., forthcoming. Railroads of the Raj: Estimating the impact of transportation infrastructure. *Am. Econ. Rev.*
- Duranton, G., Puga, D., 2015. Urban land use. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam, pp. 467–560.
- Duranton, G., Morrow, P., Turner, M., 2014. Roads and trade: evidence from the US. *Rev. Econ. Stud.* 81 (2), 681–724.
- Durlauf, S., 1996. A theory of persistent income inequality. *J. Econ. Growth* 1, 75–93.
- Durlauf, S., 2004. Neighborhood effects. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, pp. 2173–2242.
- Eaton, J., Kortum, S., 2002. Technology, geography, and trade. *Econometrica* 70, 1741–1779.
- Ellickson, B., 1973. A generalization of the pure theory of public goods. *Am. Econ. Rev.* 63, 417–432.
- Epple, D., Platt, G., 1998. Equilibrium and local redistribution in an urban economy when households differ in both preferences and incomes. *J. Urban Econ.* 43, 23–51.
- Epple, D., Romer, T., 1989. On the flexibility of municipal boundaries. *J. Urban Econ.* 26, 307–319.
- Epple, D., Romer, T., 1991. Mobility and redistribution. *J. Polit. Econ.* 99, 828–858.
- Epple, D., Sieg, H., 1999. Estimating equilibrium models of local jurisdictions. *J. Polit. Econ.* 107, 645–681.
- Epple, D., Filimon, R., Romer, T., 1984. Equilibrium among local jurisdictions: toward an integrated treatment of voting and residential choice. *J. Public Econ.* 24, 281–308.
- Epple, D., Filimon, R., Romer, T., 1993. Existence of voting and housing equilibrium in a system of communities with property taxes. *Reg. Sci. Urban Econ.* 23, 585–610.
- Epple, D., Romer, T., Sieg, H., 2001. Interjurisdictional sorting and majority rule: an empirical analysis. *Econometrica* 69, 1437–1465.
- Epple, D., Gordon, B., Sieg, H., 2010a. Drs. Muth and Mills meet Dr. Tiebout: integrating location-specific amenities into multi-community equilibrium models. *J. Reg. Sci.* 50, 381–400.
- Epple, D., Gordon, B., Sieg, H., 2010b. A new approach to estimating the production function for housing. *Am. Econ. Rev.* 100, 905–924.
- Epple, D., Peress, M., Sieg, H., 2010c. Identification and semiparametric estimation of equilibrium models of local jurisdictions. *Am. Econ. J. Microecon.* 2, 195–220.
- Epple, D., Romano, R., Sieg, H., 2012. The life cycle dynamics within metropolitan communities. *J. Public Econ.* 96, 255–268.
- Epple, D., Jha, A., Sieg, H., 2014. Estimating a game of managing school district capacity as parents vote with their feet, *Working paper*.

- Fernandez, R., Rogerson, R., 1996. Income distribution, communities, and the quality of public education. *Q. J. Econ.* 111, 135–164.
- Fernandez, R., Rogerson, R., 1998. Public education and income distribution: a dynamic quantitative evaluation of education-finance reform. *Am. Econ. Rev.* 88, 813–833.
- Ferreira, F., 2009. You can take it with you: Proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities, *Working paper*.
- Ferreira, M., 2007. Estimating the effects of private school vouchers in multi-district economies. *Am. Econ. Rev.* 97, 789–817.
- Fisher, R., 1935. *Design of Experiments*. Hafner, New York.
- Galliani, S., Murphy, A., Pantano, J., 2012. Estimating neighborhood choice models: lessons from a housing assistance experiment, *Working paper*.
- Geyer, J., Sieg, H., 2013. Estimating an model of excess demand for public housing. *Quant. Econ.* 4, 483–513.
- Glomm, G., Lagunoff, R., 1999. A dynamic Tiebout theory of voluntary vs involuntary provision of public goods. *Rev. Econ. Stud.* 66, 659–677.
- Goodspeed, T., 1989. A reexamination of the use of ability-to-pay taxes by local governments. *J. Public Econ.* 38, 319–342.
- Gould, E., 2007. Cities, workers, and wages: a structural analysis of the urban wage premium. *Rev. Econ. Stud.* 74, 477–506.
- Hansen, L.P., Singleton, K., 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–1286.
- Hastings, J., Kane, T., Staiger, D., 2006. Paternal preferences and school competition: evidence from a public school choice program, *Working paper*.
- Heckman, J., MaCurdy, T., 1980. A life cycle model of female labour supply. *Rev. Econ. Stud.* 47, 47–74.
- Henderson, J.V., Thisse, J.F., 2001. On strategic community development. *J. Polit. Econ.* 109, 546–569.
- Holmes, T.J., 2005. The location of sales offices and the attraction of cities. *J. Polit. Econ.* 113, 551–581.
- Holmes, T., 2011. The diffusion of Wal-Mart and economies of density. *Econometrica* 79, 253–302.
- Holmes, T., Stevens, J., 2014. An alternative theory of the plant size distribution, with geography and intra- and international trade. *J. Polit. Econ.* 122, 369–421.
- Hotz, J., Miller, R., 1993. Conditional choice probabilities and estimation of dynamic models. *Rev. Econ. Stud.* 60, 497–529.
- Judd, K., 1998. *Numerical Methods in Economics*. MIT Press, Cambridge.
- Keane, M., Wolpin, K., 1997. The career decisions of young men. *J. Polit. Econ.* 105, 473–523.
- Kennan, J., Walker, J., 2011. The effect of expected income on individual migration decisions. *Econometrica* 79, 211–251.
- Lucas Jr., R.E., 1976. Econometric policy evaluation: a critique. In: Brunner, K., Meltzer, A. (Eds.), *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, vol 1. American Elsevier, New York, pp. 19–46.
- Lucas Jr., R.E., Rossi-Hansberg, E., 2002. On the internal structure of cities. *Econometrica* 70, 1445–1476.
- Manski, C.F., 1993. Identification of endogenous social effects: the reflection problem. *Rev. Econ. Stud.* 60, 531–542.
- McFadden, D., 1974. The measurement of urban travel demand. *J. Public Econ.* 3, 303–328.
- McFadden, D., 1978. Modelling the choice of residential location. In: Karlqvist, A., Snickars, F., Weibull, J. (Eds.), *Spatial Interaction Theory and Planning Models*. Elsevier North-Holland, Amsterdam, pp. 531–552.
- Murphy, A., 2013. A dynamic model of housing supply, *Working paper*.
- Nechyba, T., 1997. Local property and state income taxes: the role of interjurisdictional competition and collusion. *J. Polit. Econ.* 105, 351–384.
- Nevo, A., 2000. A practitioner's guide to estimation of random-coefficients logit models of demand. *J. Econ. Manag. Strateg.* 9, 513–548.
- Newey, W.K., McFadden, D., 1994. Large sample estimation and hypothesis testing. In: Engle, R.F., McFadden, D.L. (Eds.), *Handbook of Econometrics*, vol. 4. Elsevier, Amsterdam, pp. 2111–2245.

- Neyman, J., 1923. On the application of probability theory to agricultural experiments: essay on principles. *Transl. Stat. Sci.* 5, 465–472.
- Ortalo-Magne, F., Rady, S., 2006. Housing market dynamics: on the contribution of income shocks and credit constraints. *Rev. Econ. Stud.* 73, 459–485.
- Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 1988. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.
- Redding, S., Sturm, D., 2008. The costs of remoteness: evidence from German division and reunification. *Am. Econ. Rev.* 98, 1766–1797.
- Rosenthal, S., Strange, W., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, pp. 2119–2171.
- Rothstein, J., 2006. Good principals or good peers? Parental valuation of school characteristics, Tiebout equilibrium, and the incentive effects of competition among jurisdictions. *Am. Econ. Rev.* 96, 1333–1350.
- Rust, J., 1987. Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher. *Econometrica* 55, 999–1033.
- Rust, J., 1994. Structural estimation of Markov decision processes. In: Engle, R.F., McFadden, D.L. (Eds.), *Handbook of Econometrics*, vol. 4. Elsevier, Amsterdam, pp. 3081–3143.
- Scotchmer, S., 1986. The short-run and long-run benefits of environmental improvement. *Public Econ.* 30, 61–81.
- Sieg, H., Smith, V.K., Banzhaf, S., Walsh, R., 2002. Interjurisdictional housing prices in locational equilibrium. *J. Urban Econ.* 52, 131–153.
- Sieg, H., Smith, V.K., Banzhaf, S., Walsh, R., 2004. Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *Int. Econ. Rev.* 45, 1047–1077.
- Suzuki, J., 2013. Land use regulation as a barrier to entry: evidence from the Texas lodging industry. *Int. Econ. Rev.* 54, 495–523.
- Tiebout, C., 1956. A pure theory of local expenditures. *J. Polit. Econ.* 64, 416–424.
- Todd, P., Wolpin, K., 2006. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *Am. Econ. Rev.* 96, 1384–1417.
- Tra, C., 2010. A discrete choice equilibrium approach to valuing large environmental changes. *J. Public Econ.* 94, 183–196.
- Train, K.E., 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.
- Walsh, R., 2007. Endogenous open space amenities in a locational equilibrium. *J. Urban Econ.* 61, 319–344.
- Weintraub, G., Benkard, C.L., Van Roy, B., 2008. Markov perfect industry dynamics with many firms. *Econometrica* 76, 1375–1411.
- Westhoff, F., 1977. Existence of equilibrium in economies with a local public good. *J. Econ. Theory* 14, 84–112.
- Wu, J., Cho, S., 2003. Estimating households' preferences for environmental amenities using equilibrium models of local jurisdictions. *Scott. J. Polit. Econ.* 50, 189–206.
- Yoon, C., 2012. The decline of the Rust Belt, **Working paper**.

CHAPTER 3

Spatial Methods

Steve Gibbons^{*}, Henry G. Overman^{*}, Eleonora Patacchini[†]

^{*}London School of Economics, London, UK

[†]Cornell University, Ithaca, NY, USA

Contents

3.1. Introduction	116
3.2. Nonrandomness in Spatial Data	120
3.3. Spatial Models	124
3.3.1 Specification of linear spatial models	124
3.3.2 Specifying the interconnections	128
3.3.3 Interpretation	132
3.3.3.1 <i>Spatial versus social interactions</i>	134
3.3.3.2 <i>Pecuniary versus technological externalities</i>	135
3.4. Identification	136
3.4.1 Spatially autocorrelated unobservables, when these are uncorrelated with the observables	136
3.4.1.1 <i>The reflection problem</i>	138
3.4.1.2 <i>Solutions to the reflection problem</i>	140
3.4.2 Spatially autocorrelated unobservables, when these are correlated with the observables	145
3.4.3 Sorting and spatial unobservables	149
3.4.4 Spatial methods and identification	151
3.5. Treatment Effects When Individual Outcomes Are (Spatially) Dependent	152
3.5.1 (Cluster) randomization does not solve the reflection problem	152
3.5.2 Randomization and identification	156
3.6. Conclusions	157
Appendix A: Biases with Omitted Spatial Variables	158
Appendix B: Hypothetical RCT Experiments for Identifying Parameters in the Presence of Interactions	
Within Spatial Clusters	161
References	164

Abstract

This chapter is concerned with methods for analyzing spatial data. After initial discussion of the nature of spatial data, including the concept of randomness, we focus most of our attention on linear regression models that involve interactions between agents across space. The introduction of spatial variables into standard linear regression provides a flexible way of characterizing these interactions, but complicates both interpretation and estimation of parameters of interest. The estimation of these models leads to three fundamental challenges: the “reflection problem,” the presence of omitted variables, and problems caused by sorting. We consider possible solutions to these problems, with a particular focus on restrictions on the nature of interactions. We show that similar assumptions are implicit in the

empirical strategies—fixed effects or spatial differencing—used to address these problems in reduced form estimation. These general lessons carry over to the policy evaluation literature.

Keywords

Spatial analysis, Spatial econometrics, Neighborhood effects, Agglomeration, Weights matrix

JEL Classification Codes

R, C1, C5

3.1. INTRODUCTION

This chapter is concerned with methods for analyzing spatial data. When location is simply a source of additional information on each unit of observation, it adds little to the complexity of analyzing and understanding the causes of spatial phenomena. However, in situations where agents are able to interact, relative locations may play a role in determining the nature of those interactions. In these situations of spatial interdependence, analysis is significantly more complicated and the subject of ongoing epistemological and methodological debate. It is these issues that are the focus of this chapter.

Even when units of observation can be located in some space, it is possible that location is irrelevant for understanding data pertaining to those units. In such circumstances it makes sense to think of the spatial dimension as random—a concept that can be made precise using notions from spatial statistics (Cressie, 1993; Diggle, 2003). In contrast, when location matters, the spatial dimension is nonrandom and our understanding of the data will be increased if we can allow for and explain this nonrandomness. Such nonrandomness is pervasive in areas of interest to urban economics. Why do individuals and firms concentrate geographically in dense (urban) areas? How does concentration affect outcomes and how does this explain why some cities perform better than others? To what extent do firms in particular industrial sectors cluster geographically? Why does this clustering happen and how does it influence outcomes for firms? Is the spatial concentration of poverty within cities a manifestation or a determinant of individual outcomes? Does location determine how individuals, firms, and other organizations, including government, interact and if so, how does this help us understand socioeconomic outcomes?

Answering such questions about nonrandomness is clearly central to increasing our understanding of how urban economies function. Unfortunately, as we explain in detail below, detecting departures from nonrandomness is not always straightforward. Distinguishing between the causes of nonrandom spatial outcomes is exceptionally difficult, because it requires us to distinguish between common influences and interaction effects that might explain the observed nonrandomness. For example, all individuals that live in New York City may be affected by the density of the city, its cost of living, or many other shared environmental factors. As a consequence, their outcomes—such as wages, health,

behavior, and well-being—change together as these factors change. However, this correlation of outcomes across individuals need not imply that these individuals directly influence each other. If, in contrast, individual New Yorkers' behavior is directly influenced by (expectations of) the behavior of other New Yorkers, then the correlation across individuals is the result of social interactions.

Consideration of these issues is further complicated by the fact that the terminology used to talk about these effects is often imprecise and dependent on the disciplinary background. For example, “spatial interactions,” “social interactions,” “neighborhood effects,” “social capital,” “network effects,” and “peer effects” are all terms that are often used synonymously but may have different connotations (Ioannides, 2013). These differences in terminology may also reflect important differences in the theoretical models that underlie empirical specifications. For example, in the network effects literature, the definition of an interaction effect is often based on interdependent objective functions (utility, profit, etc.). If my utility (and choice) is based on yours and vice versa, the equilibrium outcomes observed in the data are a complex function of both utility functions. Common influences do not imply such interdependency. However, social interactions defined more broadly need not involve such direct interdependency in objective functions (Manski, 2000). Social interactions may involve the availability of information, for example, about the value of education, job opportunities, or one's own ability (Banerjee and Besley, 1991). Or they may arise because of the effect that one person's actions have on another owing to the constraints they both face, for example, when one child's misbehavior diverts a teacher's attention from another child, allowing them to misbehave (which is a standard explanation of educational peer effects). In contrast, in the spatial econometrics literature, spatial interactions in outcomes may be posited for individual-level or area-level outcomes with no reference made to any underlying objective function or any other economic microfoundations. Of course, this begs the question whether one could microfound such models without recourse to interdependent objective functions. Many models within the new economic geography tradition show that this is indeed possible. In the Krugman (1991b) core-periphery model, for example, firms are sufficiently small that they ignore their impact on other firms (and hence ignore reactions from those firms), while workers' utility functions depend only on consumption of a continuum of manufacturing sector varieties and an agricultural good (not directly on the utility of other workers). Yet in these models the location of both firms and workers is interdependent in equilibrium.¹ Similarly, in the urban peer effects literature, Benabou (1993) shows how segregation can arise when the skill of neighborhood peers affects the costs of acquiring skills (in schools), and how this in turn can affect the incentives to

¹ Similarly, a range of search models can also be used to provide microfoundations for spatial interactions without the need for interdependent objective functions. See, for example, Patacchini and Zenou (2007) and Zenou (2009).

acquire skills. [Epple and Romano \(2011\)](#) review a range of other theoretical models that explain social interactions without directly interdependent objective functions.

Regardless of the terminology, recent research on spatial econometrics (and the related literature on network effects) has shown that the nature of the interconnection between individuals, firms, or places is crucial when it comes to identifying parameters or causal effects in spatial models that involve interactions. This literature has given us a far better understanding of the kind of data-generating processes where we can, in principle, distinguish between the different causes of nonrandomness and the information that is then needed to do so in practice. In particular, it is important to distinguish between two broad types of interaction structure. On the one hand, there is the context where a group of individuals or firms may influence one another jointly. For example, all firms in a cluster, or individuals in a neighborhood, may jointly impact each other. Estimation in this case would look to determine, for example, whether cluster-level R&D spending determines firm-level R&D spending² or if the local crime rate is relevant to explain the individual propensity to commit crime.³ In this case the interaction scheme is complete because all agents in a given group are connected to all others in the group.

Distinguishing between a common influence and an interaction effect in this setting is particularly challenging, because when one estimates the propensity of a firm or individual to make a decision as a function of the average behavior of its group, a unique type of endogeneity arises. In particular, if outcomes are modeled as a linear function of group outcomes (e.g., R&D), and exogenous individual and group characteristics (e.g., firm age and average firm age), it becomes difficult to distinguish between the influence of the group outcome and other group-level characteristics. Econometrically, problems arise because group-averaged outcomes are perfectly collinear, or nearly collinear, with the group-averaged exogenous variables unless specific types of restrictions are imposed on the structure of interactions, or on other aspects of the specification. Conceptually, the issue is that the average outcome for the group is an aggregation of outcomes or behaviors over other group members, and hence is an aggregation of individual characteristics over other group members. This problem is known as the “reflection problem” ([Manski, 1993](#)). It is an often misunderstood problem, which frequently results in the inappropriate interpretation of neighborhood and peer effects. Specifically, positive significant coefficients on group averages are often misinterpreted as identifying endogenous social interactions even in situations where the full set of exogenous characteristics that determine behavior are not available. This problem is pervasive even in cases when assignment to groups is random as, for example, in [Sacerdote \(2001\)](#).

The alternative to complete interactions occurs in contexts where some, but not all, individuals or firms in a group influence one another: that is, the interaction scheme is

² See, for example, the extensive knowledge production function literature initiated by [Jaffe \(1989\)](#).

³ [Case and Katz \(1991\)](#) provide an early example.

“incomplete.” For example, firm-level R&D may be influenced by interaction with specific peers, rather than a cluster (or industry) as a whole.⁴ If firm A interacts with firm B, firm B interacts with both firm A and firm C but firm C does not interact with firm A, the interaction scheme is not complete. In this case the influence of the group outcome and the influence of other group-level characteristics can, in principle, be separately identified. In a similar vein, individuals may be influenced by only some (rather than all) neighbors when taking decisions. If one can specify the details of such an incomplete interaction scheme, then this avoids the reflection problem. Indeed, this is the “solution” to the identification problem that has traditionally been (implicitly and artificially) imposed in the spatial econometrics literature through the use of standard, *ad hoc* spatial weight matrices (e.g., rook or queen contiguity). We discuss these issues in much more depth below.

Unfortunately, in practice, the number of situations where we have detailed information on the true structure of interactions is limited—especially in terms of common spatial interactions that may be of interest. The problems of distinguishing between different causes become even more pronounced in situations where we do not know all of the relevant individual factors or common influences that explain outcomes, and do not know the structure of interactions or whether the structure of interactions is endogenously determined (i.e., decisions of individual agents determine who is influenced, not just how they are influenced). In these situations, [Gibbons and Overman \(2012\)](#) propose adopting a reduced form approach, focusing on finding credibly exogenous sources of variation to allow the identification of causal processes at work. Again, we discuss these issues further below.

This chapter is organized as follows. We lay out some of the basic intuitions regarding the modeling of spatial data in [Section 3.2](#) and provide more formal consideration in [Section 3.3](#), focusing our attention on the linear regression model with spatial effects. This section also considers the distinction between spatial and social interactions. In [Section 3.4](#) we consider issues relating to identification and estimation with observational data, with a particular focus on how the existence of spatial interactions might complicate the reduced form approach to identification. An alternative to focusing on the reduced form in quasi-experimental settings is to adopt an experimental approach where the researcher uses randomization to provide an exogenous source of variation. Such an approach is particularly associated with the estimation of treatment effects.

⁴ The importance of networks has long been recognized in the literature on research productivity (broadly defined). However, empirical papers have tended to focus on the construction of summary statistics (i.e., social network analysis measures) for use as additional explanatory variables in knowledge production function specifications. See, for example, [Abbasi et al. \(2011\)](#) and [Harhoff et al. \(2013\)](#). A second literature uses shocks to networks as an exogenous source of variation in the composition of peers. See, for example, [Borjas and Doran \(2012\)](#). Only recently has the focus shifted toward network structure as a source of identification, as we discuss further in [Section 3.4](#).

We devote [Section 3.5](#) to the estimation of treatment effects in the presence of spatial interactions. [Section 3.6](#) concludes the chapter.

3.2. NONRANDOMNESS IN SPATIAL DATA

Underlying all spatial data are units of observation that can be located in some space. Locational information provides us with the position of one observation relative to others (distance and direction) and can be recorded in a number of ways. In many examples we will be interested in physical locations, but the methods we discuss can be applied more broadly (e.g., to location within a nonphysical network). [Figure 3.1](#) presents a stylized set of spatial data that allow us to introduce the basic identification problem. Each panel in this figure maps location for two groups of observations. Group membership is identified through the use of different symbols—hollow points to represent membership of group 1, solid points to represent membership of group 2. In the left-hand panel the location of all observations is randomly determined, while in the right-hand panel it is nonrandomly determined (with solid points over represented toward the South and West and hollow points over represented toward the North and East).

The precise meaning of randomness for this kind of spatial data can be formalized using concepts developed for the analysis of spatial point patterns ([Cressie, 1993](#); [Diggle, 2003](#)). Traditionally, that literature has focused on the null hypothesis of complete spatial randomness, which assumes that space is homogeneous, so that points are equally likely to be located anywhere. As argued in [Duranton and Overman \(2005\)](#), this hypothesis is unlikely to be particularly useful in many economic situations where location choices are constrained by a range of factors. To address this problem, those authors propose comparing the distribution of the sample of interest with some reference distribution. In their specific application, the groups of interest are specific industry sectors, while the reference distribution is the location of UK manufacturing as a whole. Comparison to this distribution allows one to test for geographical clustering of specific sectors—in terms of both the extent of clustering and its statistical significance.

For given spatial data, randomness can be uniquely defined (either using the assumption of homogeneous space or relative to some reference distribution) but deviations

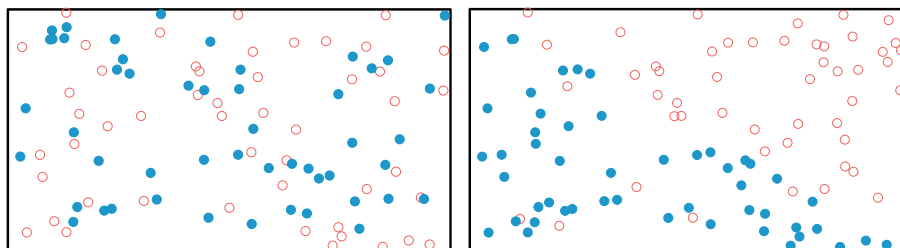


Figure 3.1 Randomness versus nonrandomness.

from randomness can happen along many dimensions. For example, in their study of segregation in the United States, [Massey and Denton \(1987\)](#) characterize racial segregation along five dimensions: evenness, concentration, exposure, clustering, and centralization. In contrast to these multiple causes of nonrandomness, tests for departures from randomness must be based on the calculation of index numbers that characterize the underlying distribution. A given index will have a unique distribution under the null hypothesis, but the power of the test will often depend on the causes of nonrandomness. In many cases, the distribution under the null cannot be derived analytically, leaving tests to rely on bootstrapping to determine appropriate test values. In short, while it may be conceptually simple to define randomness, detecting departures from randomness is more complicated in practice.

Until relatively recently, the mainstream economics literature largely ignored these problems and focused on the use of indices calculated using areal data (e.g., district, region) and constructed to characterize certain features of the data. For example, in the segregation literature, [Cutler et al. \(1999\)](#) use two indices of segregation. The first is a measure of dissimilarity which captures “what share of the black population would need to change areas for the races to be evenly distributed within a city.” The second is a measure of isolation which captures the exposure of blacks to whites. Changes in both these indices over a long time period are then used to characterize the “rise and decline of the American Ghetto.” In the international trade literature, similar indices such as the spatial Gini index and the Krugman specialization/concentration index (which is just two times the dissimilarity index) have been used to describe patterns of specialization and geographical concentration. Again, the focus has usually been on changes over time or on comparisons across geographical areas or industries rather than on the statistical significance of any departure from randomness. [Ellison and Glaeser \(1997\)](#) moved the literature closer to the statistical point pattern literature by worrying about the appropriate definition of randomness (specifically, the extent to which any index of spatial concentration should adjust for industrial concentration). But their criteria for high and moderate spatial concentration relied on the use of arbitrary cutoff points, defined with respect to the observed distribution of index values across industries rather than the underlying distribution of the index conditional on the assumption of randomness. [Combes and Overman \(2004\)](#) provide an overview and assessment of different measures.

Using ideas from the spatial point pattern literature, a number of authors have subsequently developed a new generation of tests for nonrandomness that can be applied to nonaggregated data with detailed location information. All of these tests use information on some moment of the bilateral distribution of distances between points to allow comparison of the sample with the reference distribution. [Duranton and Overman \(2005\)](#) make the case for comparison to be based on the density function for the full set of bilateral distances. In contrast, [Marcon and Puech \(2003\)](#) develop more traditional measures based on the use of cumulative distribution functions (Ripley’s K and L; [Ripley, 1976](#)).

Subsequent contributions to this literature have developed alternative tests which differ in terms of the way in which the moments of the distribution of distances are used to assess for nonrandomness. Some of these alternative tests (e.g., those focusing on distances to the k -nearest neighbors) simplify calculations for large distributions—remembering that the number of bilateral distance calculations increases with the square of the number of sample points. Other authors (e.g., [Klier and McMillen, 2008](#); [Vitali et al., 2009](#); [Ellison et al., 2010](#); [Kosfeld et al., 2011](#)) have suggested approximations or algorithmic improvements for tests based on the complete distribution of bilateral distances that similarly reduce computational complexity. [Scholl and Brenner \(2012\)](#) provide a relatively recent overview of different measures, while [Scholl and Brenner \(2013\)](#) provide discussion of computational issues. Debate still continues as to the “best” method for detecting departures from randomness. Our own view is that in situations where we wish to test for nonrandomness, the choice of the method is a second-order consideration relative to the first-order decision of whether or not to treat space as continuous. If the data allow it, using insights from the spatial point pattern literature and treating space as continuous, rather than discrete, allows for more powerful tests of nonrandomness.

Unfortunately, in many circumstances, researchers have access to only spatial aggregates for units of observations that correspond to areas rather than the individual units of observation. [Duranton and Overman \(2005\)](#) refer to this process of aggregation as moving from “points on a map to units in a box.” Any such discretization and corresponding aggregation implies a loss of information and makes it harder to test for departures from randomness. Still, such areal data are often all that researchers have available to them. In these cases, tests for nonrandomness can be based on the concentration/segregation indices, discussed above, that have traditionally been used in the population and industrial location literature (such as the Herfindahl–Hirschman index, Krugman/dissimilarity index, and Ellison and Glaeser index; see, respectively, [Herfindahl, 1959](#); [Hirschman, 1964](#); [Krugman, 1991a](#); [Ellison and Glaeser, 1997](#)) or on “global indicators of spatial association” developed in the spatial statistics and econometrics literature (such as Moran’s I or Getis–Ord statistics; see, respectively, [Moran, 1950](#); [Getis and Ord, 1992](#)).

Once we have applied one or more of these tests and rejected the null hypothesis of randomness, we may want to find out where within our geographical study area this nonrandomness occurs. For example, once we have established that crime is nonrandom across space in New York, we may want to visualize where in New York the crime hot spots occur. A range of spatial methods exist for doing just that, facilitated today by the integrated data analysis and mapping capabilities of geographical information systems (GIS) and related spatial software. Standard kernel density and spatial interpolation methods can be easily implemented in a modern GIS to visualize these patterns using point pattern data. For more aggregated data “local indicators of spatial association” ([Anselin, 1995](#)) such as the local Moran’s I and Getis–Ord G_i^* statistics (which are simply the spatially disaggregated components of their global counterparts) are also readily

available in standard GIS software to statistically test for and visualize these local spatial departures from randomness (see [Felkner and Townsend, 2011](#), for one example). All these methods are, however, purely descriptive and say nothing about the causes (or consequences) of the departure from randomness. It is these questions which are the main motivation behind the development and application of the spatial methods that are discussed in detail in the remainder of this chapter. Thinking about the possible causes of nonrandom location and the way in which the consequence of nonrandom location feeds back into location decisions gives us some idea about the difficulties that lie ahead. For example, assume that the points in [Figure 3.1](#) represent either firms or workers and the color represents different types of economic activity. There are several ways in which the nonrandom pattern in the right-hand panel in [Figure 3.1](#) can emerge. First, firms may be randomly allocated across space but some characteristic of locations varies across space and influences outcomes. We might think of farmers who are randomly distributed across space, with the type of crops they produce driven by locational differences in underlying soil type and fertility.⁵ Second, location may have no causal effect on outcomes, but outcomes may be correlated across space because heterogeneous individuals or firms are nonrandomly allocated across space. We might think of highly educated workers producing R&D in one area, while less educated workers assemble manufactured goods in another area.⁶ Third, individuals or firms may be randomly allocated across space but they interact, and so a decision by one agent affects outcomes of other agents. We might think of students choosing among different college majors, where the choice of each student influences the choices of their fellow students.⁷ Similarly, in R&D, knowledge might spill over beneficially between nearby scientists, so the decision to undertake research in a specific field, or the registration of patents by inventors, varies systematically across space (as indicated by the color of the dots). Fourth, individuals or firms may be nonrandomly allocated across space and the characteristics of others nearby directly influence individual outcomes. For example, growing up among educated, employed, and successful neighbors might be beneficial in raising children's expectations about their life chances, and this may directly influence their own educational outcomes and through that their employment outcomes.⁸

⁵ See, for example, [Holmes and Lee \(2012\)](#), who attempt to distinguish whether soil characteristics (explanation number 1 in our list) or economies of density (explanation number 3) explain crop choice in North Dakota.

⁶ See, for example, [Ellison and Glaeser \(1997\)](#), who consider the role of “natural advantages” in explaining geographical concentration of industrial activity. Their broad definition of natural advantages allows a role for resources (e.g., coal), factor endowments (e.g., skilled workers), and density to influence geographical concentration. That is, they assess the role of the first, second, and fourth factors (in our list) in determining sector of economic activity.

⁷ See, for example, [Sacerdote \(2001\)](#) and [De Giorgi et al. \(2010\)](#).

⁸ A vast literature on childhood neighborhood effects considers this possibility; for example, [Aaronson \(1998\)](#), [Patacchini and Zenou \(2012\)](#), and [Gibbons et al. \(2013\)](#).

Understanding the causes of nonrandomness requires us to discriminate between these four different causes of nonrandomness in situations where one or more of them may explain departures from randomness. In empirical settings, the situation is further complicated because we may not observe all individual factors that determine outcomes. This makes it even harder to distinguish between different causes of nonrandomness. This adds a further potential explanation for nonrandomness—that individuals appear to be randomly located, in terms of observables, but they are in fact nonrandomly located in terms of unobserved characteristics that determine outcomes. The next section formalizes a number of these issues and considers what information is required to enable us to distinguish between different causes of nonrandomness.

3.3. SPATIAL MODELS

This section sets up a very general framework for linear regression models that involve interactions between agents across space. We show how the standard regression approach can accommodate spatial factors by the addition of “spatial variables.” These allow the outcomes for an individual to be influenced by the choices, outcomes, and characteristics of other individuals who interact with the individual, and by other characteristics of the location of the individual. In practice, these spatial variables are typically constructed as linear combinations of the observations in neighboring locations, aggregated with a sequence of scalar spatial or group weights. Traditionally, the literature has summarized this information in a (spatial) weights matrix (G in the network literature, W in the spatial econometrics literature), constructed on the basis of the definition of reference groups—the set of individuals or firms that may impact other agents’ outcomes. We provide a number of examples below. Both the nature of the reference group and the way in which individual outcomes depend on group membership have fundamental implications for the interpretation, estimation, and identification of spatial models. We deal with questions of interpretation in this section, and also consider the implication for estimation if spatial factors are present, but ignored. The next section then shows how the nature of the reference group, as captured in the structure of the weights matrix, is essential in determining whether the parameters on spatial variables are identified, or can be estimated (and if so, what is the appropriate identification strategy).

3.3.1 Specification of linear spatial models

We start with the standard linear regression model of a variable y relating to some unit of observation i such as a firm, individual, or household (or an areal aggregate of these, e.g., a zip code). For convenience in what follows, we often refer to these units of observation as “individuals.” We suppress the constant term and assume that all variables are in deviations from means, allowing us to write the standard linear regression model as

$$y_i = x_i' \gamma + \varepsilon_i, \quad (3.1)$$

where y_i is some outcome, such as output (for a firm) or income (for an individual), and x_i is a vector of characteristics, such as capital, labor, and material inputs (for a firm), or education, age, gender, etc. (for an individual), which determine outcomes and are observed in the data available. Unobserved characteristics that affect outcomes are represented by ε_i . In what follows we assume that ε_i is random and set aside the potential problems that arise if ε_i is not random and correlated with x_i , since the econometric issues involved in this case are well known and we will not address them here.⁹ This is a completely nonspatial model, in that there is no explicit reference to where individuals are located in space, to any of the characteristics of the space in which they are located, or to any interconnections between individuals. Suppose we have additional information about the geographical locations s of the individuals whose behavior we want to model. This information is what makes data spatial. Variable s_i might be a point in space referenced by coordinates, or a geographical zone, or some other locational identifier (school, position in a network, etc.).

Let us now modify Equation (3.1) by adding new terms that reflect the fact that the individual choice or outcome y_i may be influenced not only by the characteristics of the individual i , but also by the choices, outcomes, and characteristics of other individuals who interact with the individual i and by other characteristics of the location s_i of individual i . Individuals may interact with each other for a number of reasons, but the important point here is that their interaction is based on some relationship in terms of their spatial location s —for example, they are neighbors or belong to some common group. We will say more about how this “neighborliness” or grouping can be defined below. As we have outlined already, spatial patterns arise through two primary channels: (1) the influence of area characteristics on individuals, both in determining the characteristics acquired by individuals, and through the sorting of already heterogenous individuals across space; and (2) the interaction of neighboring individuals with each other. A framework that captures almost anything researchers try to do with linear regressions when investigating the importance of these spatial factors—both how spatial characteristics affect individuals in the economy, and how neighboring individuals affect each other—is based around the following generalization of Equation (3.1):

$$y_i = x_i' \gamma + m_y(y, s)_i \beta + m_x(x, s)_i' \theta + m_z(z, s)_i' \delta + m_v(v, s)_i \lambda + \varepsilon_i. \quad (3.2)$$

Here, as before, y_i is the outcome for an individual at location s_i , and x_i is the vector of characteristics of i . The expressions $m(., s)_i$ are a general representation of “spatial

⁹ A general, textbook-level treatment can be found in Angrist and Pischke (2009). Chapter 1 considers how insights from the experimentalist paradigm advocated by Angrist and Pischke (2009) can be applied to questions of causal inference in urban economics. This chapter complements the chapter by Baum-Snow and Ferreira by specifically considering the complications introduced by spatial or social interactions.

variables,” the interpretation of which we come to in more detail below. These are functions that generate linear, or sometimes nonlinear, aggregations of variables that are spatially connected with location s_i using information on the vector of locations s . We consider four kinds of spatial variables relating to outcomes (y_i), a vector of individual characteristics (x_i), a vector of characteristics (z_i) of other entities or objects (other than individuals i), and a variable that captures all characteristics of either individuals or entities and objects that are unobservable to the econometrician (v_i). We are keeping things very general at this stage, so we allow the form of $m(\cdot, s)_i$ to be different for y , x , z , and v , and indeed for x and z , possibly different for different elements of these vectors, so that each variable could have its own aggregating or averaging function.

The spatial connections between locations, which form the basis for aggregation, can be defined through absolute or relative positions in geographical space, the position within networks, or other methods. In general, these functions $m(\cdot, s)_i$ can be thought of in a number of ways, as forming estimates of the means of the variables or expectations at location s_i , as spatial smoothing functions that estimate how the variables vary over locations s , or as structural representations of the connections between locations s . Depending on the setting, these functions may capture interpersonal effects that are passive or deliberate (which might be distinguished as “externalities” vs. “interactions”). These effects may also occur directly or may instead be mediated through the market (leading, for example, to the distinction between pure/technological externalities and pecuniary externalities).

To give a specific example, the outcome under consideration might be earnings, for individuals, and the aim is to estimate Equation (3.2) on a sample of individuals. If y_i is individual earnings, $m_y(y, s)_i$ allows for the possibility that some spatial aggregation of individual outcomes—for example, the mean earnings for individuals living in the same city—may affect individual earnings. The vector x_i might include individual years of education, so $m_x(x, s)_i$ might be defined to capture the mean years of education in some interconnected group—for example, individuals working in the same city. Vector z_i might include indicators of firm industrial classification in an auxiliary sample of firms, so one component of $m_z(z, s)_i$ could be defined to capture the proportion of firms or the total number of firms in each industry category in i ’s city. Vector z_i might also include average yearly temperature readings from weather stations, such that a second component of $m_z(z, s)_i$ yields mean city temperature. In this example, the share of educated workers (a component of $m_x(x, s)_i$) and the number of firms by sector (a component of $m_z(z, s)_i$) may have a direct effect on earnings or a pecuniary effect (if the share of educated workers is also a measure of labor supply, while the number of firms is also a measure of labor demand).¹⁰ Importantly, Equation (3.2) allows spatial aggregates of the unobservables

¹⁰ This distinction has received some consideration in the literature on human capital externalities (Ciccone and Peri, 2006) but has largely been ignored in the agglomeration literature looking at productivity effects or urban wage premium.

$m_\nu(\nu, s)_i$ to influence γ_i , to allow for the possibility either that individuals interact with each other across space on unobserved dimensions, or that there are spatially correlated shocks from other sources that affect spatially interconnected individuals simultaneously. To continue the example above, ν_i might include individual abilities that are not represented in x , or unobserved productive advantages of the places s in which individuals are located, but which are not represented by variables in z . Again, the spatial aggregate $m_\nu(\nu, s)_i$ might then be defined as the mean of these unobserved factors. It is, of course, possible to add a time dimension to this specification, for estimation on a panel or repeated cross sections of individuals, but for now we focus on the cross-sectional case only.

For a set of observations on variables at locations s_j , the “spatial” variables $m_\nu(\nu, s)_i$ are typically linear combinations of the observations in neighboring locations, aggregated with a sequence of scalar spatial or group weights $g_{ik}(s_i, s_j)$ that depend on the distance (or some other measure of the degree of interconnection) between observations at the corresponding locations s_i and s_j . Let us define

$$m_x(x, s_i) = \sum_{j=1}^M g_{ij}(s_i, s_j) \cdot x_j = G_{xi}x, \quad (3.3)$$

where G_{xi} is a $1 \times M$ row vector of the set of weights relating to location s_i , and x is an $M \times 1$ column vector of x for locations $s_1, s_2, \dots, s_M$. Sometimes it is more convenient to work with matrix notation for all observations i , where G is an $N \times M$ matrix, so

$$m_x(x, s) = G_x x, \quad (3.4)$$

and similarly for z , γ , and ν . Note that in cases where spatial variables are created by aggregating over the N individuals for whom Equation (3.2) is to be estimated, $N = M$. With use of Equation (3.4) and similar expressions for γ , x , and ν , Equation (3.2) becomes

$$\gamma = X\gamma + G_\gamma\gamma\beta + G_x X\theta + G_z Z\delta + G_\nu \nu\lambda + \epsilon. \quad (3.5)$$

This notation is favored in the spatial econometrics literature, where the weights matrix is usually designated using W instead of G , assumed common across variables (so $W_\gamma = W_x = W_z = W_\nu$), and $W\gamma$, WX , WZ , and $W\nu$ are called “spatial lags.” Restrictions on Equation (3.5) yield a typology of spatial econometrics models—for example, the spatially autoregressive (SAR) model ($\delta = 0, \lambda = 0, \theta = 0$), the spatially lagged x model¹¹ ($\beta = 0, \lambda = 0$), the spatial Durbin model ($\lambda = 0$), and the spatial error model ($\beta = 0, \delta = 0$). In what follows, we use the notation G in preference to W , because W has become associated with a set of spatial weights which specify *ad hoc* connections between

¹¹ The distinction between Z and X is often irrelevant in much applied spatial econometrics research, which usually works with aggregated spatial data units. In this case the data for individuals (x) and for other spatial entities (z) have already implicitly been through a first stage of aggregation. Hence, the standard terminology refers simply to the spatially lagged x model without distinguishing between x and z .

neighboring places, and with a spatial econometrics literature that seeks to distinguish between competing models through statistical testing of model fit. Instead, we wish to focus attention on the fact that the nature of interactions within social and spatial groups is central to theoretical interpretation, identification, and estimation.

In contrast, the social interactions literature favors an alternative notation, where Equations (3.2) and (3.5) are typically written out in terms of expected values of the variables in the groups to which i belongs. Here, the expected values are taken to imply the mean characteristics (observed or unobserved) of the group, or expectations about behaviors or characteristics which are unobserved by individuals or not yet realized. The structural specification analogous to Equation (3.2) in the social interactions literature is thus

$$y_i = x_i' \gamma + E(y|G_i) \beta + E(x|G_i)' \theta + E(z|G_i)' \delta + E(v|G_i)' \lambda + \varepsilon_i. \quad (3.6)$$

In practice, in empirical implementations, the expectations are replaced by empirical counterparts with the estimates $\hat{E}(y|G_i) = G_y \gamma$, $\hat{E}(x|G_i) = G_x x$, and $\hat{E}(z|G_i) = G_z z$ so the spatial models and social interactions models are for the most part isomorphous. [Manski \(1993\)](#) introduced a useful and popular typology of interaction terms in this kind of specification. In this typology, β represents “endogenous” effects, whereby individuals’ behavior, outcome, or choices respond to the anticipated behavior outcome or choices of the other members in their reference group. In contrast, θ represents “contextual” or “exogenous” interactions in which individuals respond to observable exogenous or predetermined characteristics of their group (e.g., age and gender). Manski refers to λ as “correlated” effects, in which peer-group-specific unobservable factors affect both individual and peer behavior. For example, children in a school class may be exposed to common factors such as having unobservably good teachers, which can lead to correlation between individuals and peers which look like interactions, but are not. Of course, some of these peer-group-specific factors may also be observable (e.g., teacher qualifications or salaries), and the effects of these observable characteristics are captured in our notation by δ .

3.3.2 Specifying the interconnections

We now turn to the various ways that are used in the literature to define reference groups—the set of agents that impact other agents’ outcomes. Both the nature of the reference group and the way in which individual outcomes depend on group membership have fundamental implications for the interpretation, estimation, and identification of spatial models.

The most basic structure for G , and one that is implicitly used in many regression applications that are not ostensibly “spatial,” is a block grouping structure. Assume that there are N individuals (or firms, households, areas, etc.; although we continue to focus on individuals for ease of exposition) divided into $k = 1, \dots, K$ groups, each

with n_k members, $i = 1, \dots, n_k$, $\sum_{k=1}^K n_k = N$. The interaction scheme can be represented by a matrix $G = \{g_{ij}\}$ whose generic element g_{ij} would be 1 if i is connected to j (i.e., interacts with j) and 0 otherwise. Usually, such matrices are row normalized, such that premultiplying an $N \times 1$ vector x by the $N \times N$ matrix G generates an $N \times 1$ vector of spatial averages.¹² For example, consider seven individuals, from each of two neighborhoods: $k = 1, 2$. Individuals $i = \{1, 2, 3\}$ belong to neighborhood $k = 1$ and individuals $i = \{4, 5, 6, 7\}$ belong to neighborhood $k = 2$. The associated G matrix is shown below:

$$G = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 3 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 5 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 6 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 7 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}, \quad GG = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 3 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 5 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 6 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 7 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}. \quad (3.7)$$

Notice that in this example, the weights are set to $1/n_k$, where n_k is the number of neighbors in group k , to achieve row normalization. More importantly, this matrix has two important properties. First, it is block diagonal, and transitive such that the neighbors of i 's neighbors are simply i 's neighbors. Second, it is symmetric-idempotent, and as a result $GG = G$. This feature will be both useful for interpretation and harmful to estimation. The interpretation is clear: all individuals from 1 to 3 and from 4 to 7 are in a given neighborhood and therefore the spatial influence is constrained to that neighborhood. Indeed, in this case, the values that populate the matrix indicate both group membership and the extent of the influence of any one individual on other individuals. This will not be the case with other specifications of G .

A simple modification that is commonly used in practice is to exclude i from being his or her own neighbor, by putting zeros on the diagonal. This maintains the transitive property, although the matrix is no longer idempotent, for example,

¹² We discuss averaging versus aggregating in more detail below.

$$G = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 3 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 5 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ 6 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ 7 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \end{bmatrix}, \quad GG = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 2 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 3 & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & \frac{1}{3} & \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \\ 5 & 0 & 0 & 0 & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} & \frac{2}{9} \\ 6 & 0 & 0 & 0 & \frac{2}{9} & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} \\ 7 & 0 & 0 & 0 & \frac{2}{9} & \frac{2}{9} & \frac{2}{9} & \frac{1}{3} \end{bmatrix}. \quad (3.8)$$

A simple structure for G that breaks both the transitivity property and the idempotent property could be based on the two nearest neighbors, where 1 is nearest to 2 and 7, 2 is nearest to 1 and 3, 3 is nearest to 2 and 4, 4 is nearest to 3 and 5, 5 is nearest to 4 and 6, and 6 is nearest to 5 and 1. The associated G matrix is shown below, and it is clear in this case that $GG \neq G$ —that is, the neighbors of i 's neighbors are not simply i 's neighbors:

$$G = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} \\ 2 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 \\ 3 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 \\ 4 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 5 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ 6 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 7 & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix}, \quad GG = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{2}{9} \\ 2 & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} \\ 3 & \frac{1}{9} & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & 0 & 0 \\ 4 & 0 & \frac{1}{9} & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} & 0 \\ 5 & 0 & 0 & \frac{1}{9} & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} & \frac{1}{9} \\ 6 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{2}{9} & \frac{1}{3} & \frac{2}{9} \\ 7 & \frac{2}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{2}{9} & \frac{1}{3} \end{bmatrix}. \quad (3.9)$$

Similar matrices would summarize the pattern of influence in a situation where individuals are asked to name their two closest friends.¹³ Of course, the number of neighbors need not be the same for all i . Allowing for varying numbers of bordering neighbors, this

¹³ See, for example, the National Longitudinal Study of Adolescent Health, which asks adolescents in grades 7–12 to name up to five male and five female friends. Fryer and Torelli (2010), Calvó-Armengol et al. (2009), Weinberg (2007), and Ioannides (2013) provide other examples.

form of the G matrix gives a contiguity matrix that is commonly used in the spatial econometrics literature for regressions involving areas (districts, regions, etc., rather than individuals) in which the weights are constructed to indicate whether areas share a border. The previous example would correspond to the contiguity matrix for seven areas located sequentially around a circle, with area 1 contiguous to areas 2 and 7, area 2 contiguous to areas 1 and 3, etc.

As should be clear from these three examples, different specifications of G provide a fairly flexible way of constructing spatially weighted variables. A nonexhaustive list of other common structures includes constructing G on the basis of

- “buffers” based on the choice of a fixed distance threshold within which interaction occurs;
- queen or rook contiguity (for geographies with two or higher dimensions), the distinction between the two being whether to regard areas touching at a vertex as contiguous or only those sharing a common border;
- inverse distance weighting;
- connectivity measures along some network.

Observe that the matrix G could be symmetric or asymmetric, depending on the nature of the interactions. It is symmetric in case of bilateral influences between any two units, and—in the case of row normalization—when each unit has the same number of neighbors. It will be asymmetric if interactions are assumed to flow one way, or if units have different numbers of neighbors. The appropriate definition will, of course, depend on the specific application. Note also that the spatial grouping or weights matrix can be defined so that it generates either spatial averages or spatial aggregates of neighboring observations. To produce averages, the G matrix must be row normalized as in the examples above, so that the weights in any row sum to 1. That is, for the spatial weights corresponding to an observation at location s , the weighting vector is

$$G_i = 1 / \sum_{j=1}^M g_{ij}(s_i, s_j) \times [g_{i1}(s_i, s_1) \quad g_{i2}(s_i, s_2) \quad \dots \quad g_{iN}(s_i, s_N)],$$

while for aggregation, the weighting vector is simply

$$G_i = [g_{i1}(s_i, s_1) \quad g_{i2}(s_i, s_2) \quad \dots \quad g_{iN}(s_i, s_N)].$$

The distinction between these two operations could be important, since aggregation adds up the effects of neighboring individuals, firms, or places, thus taking into account the number of these within the appropriate group as specified by the weighting structure. In contrast, averaging takes out any influence from the number of individuals, firms, or places that are close by. Which of these schemes is appropriate is essentially a theoretical consideration. Averaging has been the standard approach in most fields, including those on neighbor and peer effects (Epple and Romano, 2011). Aggregating is more appropriate,

and is usually applied, in work on agglomeration, or transport accessibility where the focus is on economic mass or “market potential” (Graham, 2007; Melo et al., 2009), although the literature on human capital externalities in cities has generally favored averaging (see Chapter 5). In cases where there is no guidance from economic considerations, it may be possible to use statistical tests to choose between the different specifications. In regression specifications such as (3.2) it is in principle straightforward to test whether to use aggregation or averaging, since both versions are nested within the expression $n_{ki}m_x(x, s)_i'\theta_1 + m_x(x, s)_i'\theta_2 + n_{ki}\theta_3$, in which n_{ki} is the group size for person i , $m_x(x, s)_i$ is a row-normalized (averaging) aggregator, and $n_{ki}m_x(x, s)_i$ is the interaction of the two, which gives non-row-normalized (aggregating) specification. Including all these terms in a regression specification and testing for restrictions on the parameters would provide one way to distinguish these cases statistically, with $\theta_2 = \theta_3 = 0$, $\theta_1 \neq 0$ implying aggregation, and $\theta_1 = 0$, $\theta_2 \neq 0$, $\theta_3 \neq 0$ implying that separate mean and group size effects are more relevant. There may, of course, be practical collinearity problems when implementing such a test. Liu et al. (2014) provide another test procedure to discriminate between the local-average and local-aggregate models with network data.

Another potentially important consideration is whether or not the number of individuals in the groups over which variables are averaged increases as the sample size increases (“infill” asymptotics). The number of cases over which the averages are constructed increases with sample size for inverse distance weighting or fixed distance buffer groups, and may also do so with block diagonal structures (e.g., if the block specifies different cities, and the cases are individuals). In contrast, this is not necessarily the case with contiguity matrices based on a fixed geographical structure of areas (unless sample size is increased by adding more observations of the same areas over time), or with a fixed number of nearest neighbors or friends. Sample size increases in this case require obtaining more groups (“increasing domain” asymptotics). This issue is important because it affects the way the variance of the spatial means $m_x(x, s)_i$, $m_v(x, s)_i$ behaves as the sample size increases, which will naturally matter when we come to consider questions of identification and estimation of these spatial models.

3.3.3 Interpretation

A vast range of empirical studies on urban, regional, and neighborhood questions, plus research on peer groups and other social interactions, have been based on some version of Equation (3.2). Usually in such studies, the primary focus is on estimating one or more elements of δ or θ , the effect of spatially aggregated observed characteristics for individuals (x_i) or other entities (z_i) on individual outcomes y_i ; or sometimes on estimating β , the effect of neighboring individual outcomes (y_i) on the outcome of an individual entity.

For example, in a typical study of neighborhood effects on the education of children, y would be a child’s educational attainment, $G_y y$ (using matrix notation) would be the

mean of the attainment of neighboring children, x could include child prior achievement, age, gender, and family background, G_{xx} might include the mean of these characteristics among neighboring children, and G_{zz} might include attributes of the child's home location (average local school quality, number of libraries, or average distance to nearest schools). Potentially unobserved factors in $G_{\nu\nu}$ include the quality of teaching in the local school, motivation and aspirations of neighbors, other local resources that facilitate education, etc. This literature is discussed in [Chapter 9](#). To take a second example, studies of agglomeration effects on firm productivity typically specify y_i as firm output, restrict the coefficient on $G_y y$, $\beta = 0$, and define G_{xx} as a measure of employment density based on aggregating neighboring firm employment or G_{zz} as a measure of market potential based on aggregating population or income in an auxiliary population sample or census. Firm characteristics such as capital, labor, and material inputs appear in x . Unobservables in $G_{\nu\nu}$ probably include climate, terrain, and other local productive advantages. Depending on whether the specification was in terms of G_{xx} or G_{zz} , the coefficient θ or δ would then be interpreted as an estimate of the impact of agglomeration economies on total factor productivity. [Chapter 5](#) provides a summary of this literature.

The aim of researchers employing a specification such as Equation (3.2) for these kinds of applications is usually to estimate the “causal” relationship between changes in one or more of the right-hand-side variables and changes in y_i . A good definition of causality is the subject of much debate, and there are a number of interpretations.¹⁴ One definition of a causal estimate is the expected change in y in response to an exogenous manipulation of some particular right-hand-side variable, including any indirect effects that operate through other determinants of y that may also be influenced by the exogenous manipulation of the right-hand-side variable in question. Another definition is the expected change in y for a change in x , with all other factors being held constant. We do not worry too much about these definitions here, except to note that neither looks particularly satisfactory in terms of understanding the parameter β on $G_y y$. Since $G_y y$ is an aggregate of the dependent variable, there is no sense in which it can be directly, exogenously manipulated within the population or sample to which Equation (3.2) relates. Nor can it be changed while holding other factors constant, since if other factors are constant, then y is constant and so is $G_y y$. To return to the education example, it is impossible to think of a hypothetical experiment that would directly manipulate average neighborhood educational outcomes. Instead, one would have to manipulate some other determinant of educational outcomes (e.g., teacher quality in G_{zz} , or neighborhood composition G_{xx} or the unobserved determinants of $G_{\nu\nu}$) that in turn change average educational outcomes. But in this case this implies a change in

¹⁴ See, for example, the “Con out of Economics” symposium in the *Journal of Economic Perspectives*, 24 (2) (spring 2010). See also [Heckman \(2005\)](#).

G_{zz} , G_{xx} or $G_{\nu\nu}$, and G_{yy} . As we shall see below, there are structures of G for which we could think of (3.2) applying to one subgroup of the population, while we causally manipulate G_{yy} by changing G_{zz} or G_{xx} for some other subgroup of the population to which they are connected. We return to this issue in [Section 3.5](#). Given these conceptual problems, an alternative is to approach Equation (3.2) as a structural, law-like relationship that determines the process generating y , with the goal of estimating the parameters characterizing this process, setting aside questions over the causal interpretation of β . In this case, the specification to be estimated will need to be derived from some underlying theoretical model. [Chapter 2](#) provides further discussion.

3.3.3.1 *Spatial versus social interactions*

A particular class of the spatial models described above, which adopt a structural interpretation of the parameter β on G_{yy} , are so-called social interactions models. Social interactions models, as a class, are concerned with modeling these interactions between agents at the microlevel. More specifically, social interactions models are concerned with estimating the parameters that describe the way individuals behave given what they can observe about the group to which they belong, and especially how they expect other individuals in their group to behave. These models and their behavioral foundations have been the focus of much recent attention in the research literature, and are discussed in greater detail in [Chapter 9](#). They provide two crucial insights in the context of the spatial methods considered here. First, as a result of this research, considerable progress has been made in our understanding of the importance of the structure of G in achieving identification of the class of models that involve endogenous interactions in outcomes G_{yy} . We discuss this in the next section. Second, and perhaps less widely recognized, is that the social interactions literature clarifies the circumstances in which the structural equation for y will involve terms in G_{yy} .

In fact, there is a sense in which these social interaction models in which individuals make simultaneous decisions about some action are the only class of models for which the structural equation for y will involve terms in G_{yy} . To see this, note that in any situation where there is no direct interaction in decisions, we should be able to explain the outcome for individual i as a function of own characteristics and group characteristics without needing to know G_{yy} . A concrete example may help clarify this. Imagine a situation where an individual is deciding on the price at which he or she will sell his or her house. We might think that one piece of information the individual will use to set prices is the price of any neighboring houses that have been sold recently. In such situations, it may be convenient to model individual house prices as a function of neighborhood house prices G_{yy} . But this cannot be the structural form, because the timing of sales means that the prices for earlier houses are not determined by the future sales prices of neighboring houses (ignoring any expectation effects that may influence the demand for housing). With information on both prices and the timing of sales, the appropriate structural form

involves no term in $G_y\gamma$ because the sales prices of neighboring houses are predetermined from the point of view of any individual price and should thus be treated as an element of X .¹⁵ In contrast, the structural equation for γ will involve $G_y\gamma$ in situations of social interaction where decisions are simultaneous. For example, a teenager's decision to start smoking may be dependent on the simultaneous decisions of his or her friends ($G_y\gamma$)—which implies a joint decision based on what each expects the other to do—although even here, an individual's decision to start smoking may be more affected by what that individual observe his or her friends already doing (in which case timing matters and $G_y\gamma$ does not enter the structural form for γ).¹⁶

Another way of putting this is that the scope for including spatial lags in γ is more limited than would seem to be implied by the applied spatial econometrics literature. Indeed, in that literature, terms in $G_y\gamma$ are often included without any consideration of whether decisions that determine γ are truly simultaneous. In some circumstances, this assumption may be justified. For example, in the tax competition literature, local tax rates are a function of neighboring government tax rates if governments simultaneously set taxes in response to (expectations of) taxes in contiguous neighboring jurisdictions. More generally, however, many spatial models simply assume that any interaction (between individuals in neighborhoods or schools, between neighboring or otherwise interconnected firms, between inventors and other agents of innovation, between neighboring governments and other institutions, etc.) can be used to justify the inclusion of terms in $G_y\gamma$.

3.3.3.2 *Pecuniary versus technological externalities*

Another important distinction, but one that has received relatively little attention in the literature, is whether spatial interactions arise as a result of pecuniary or technological externalities. As we discussed above, in the general spatial model terms in G_y , GX , and GZ can capture interactions that either occur directly or are mediated through the market (i.e., may capture either technological or pecuniary externalities, respectively). We have provided several examples where either may arise. For example, models in the new economic geography tradition can motivate empirical specifications that model employment in area i as a function of employment in nearby areas G_y . As we explained in [Section 3.1](#), in these models firms are sufficiently small that they ignore their impact on other firms (and hence ignore reactions from those firms), while workers' utility functions depend only on

¹⁵ For an empirical example, see [Eerola and Lyytikäinen \(2012\)](#), who use the partial release of public information on past house sales to examine the impact of information on past transactions on current house prices. [Ioannides and Zabel \(2008\)](#), [Kiel and Zabel \(2008\)](#), and [Ioannides \(2013\)](#) provide a more general discussion of neighborhood effects on housing demand and the use of neighborhood information in hedonic regressions.

¹⁶ See, for example, [Krauth \(2005\)](#) and [Nakajima \(2007\)](#). [Simons-Morton and Farhat \(2010\)](#) provide a review of the literature on peer group influences on adolescent smoking.

consumption of a continuum of manufacturing sector varieties and an agricultural good (not directly on the utility of other workers). Given that, at least in the general spatial form, these two kinds of externalities are observationally equivalent, it is likely that theory will need to provide additional structure if applied work is going to distinguish between these different sources of interaction. [Chapter 2](#) provides further discussion.

3.4. IDENTIFICATION

All researchers working with spatial data have to confront fundamental challenges that render the identification and estimation of Equation (3.2) a difficult empirical exercise. These challenges are (a) the so-called reflection problem, (b) the presence of correlated unobservables or common shocks, and (c) sorting—that is, the presence of omitted variables which are correlated with location decisions and outcomes. Problem (a) occurs when the aim is to estimate β (i.e., the effect of group outcomes or behavior on individual outcomes) as distinct from θ (i.e., the effect of group characteristics), while problems (b) and (c) may arise regardless of whether we are estimating models with or without endogenous interactions. We consider these problems in turn and discuss the solutions proposed in the existing literature.

3.4.1 Spatially autocorrelated unobservables, when these are uncorrelated with the observables

Even in the simplest setting where we know the structure of group membership and the individual and group variables that determine outcomes, the reflection problem can prevent the estimation of all coefficients of interest. The problem arises when the aim is to separately estimate β (the effect of group outcomes or behavior on individual outcomes) and θ (the effect of group characteristics) in situations where there are unobservable factors that also vary at the group level. The presence of these variables means that estimation must rely on recovering the structural parameters from parameters on the exogenous variables in the reduced form. This is usually not possible without imposing further restrictions.

To focus on this specific issue, let us initially assume that group membership is exogenous and that these unobservables are uncorrelated with the observable characteristics. This spatial autocorrelation in unobservables could occur because individuals are interacting on unobserved dimensions. For example, in a model of neighborhood effects on school grades, individual effort (unobserved by the researcher) may influence other individuals' effort within the neighborhood, even before the outcomes of that effort—school grades (y)—are observed. Or it could occur because the group members are exposed to similar unobservables. For example, in a model of the effect of cluster employment on firm employment, different clusters could be subjected to area shocks that are not directly related to the performance of the cluster. Both these processes show up as autocorrelated unobservables, so are observationally equivalent from the researcher's perspective.

As mentioned above, [Manski \(1993\)](#) refers to these as “correlated effects,” the presence of group-specific unobservable factors, uncorrelated with individual observables, but affecting both individual and group behavior. Spatial econometricians refer to models containing these spatially autocorrelated unobservables as spatial error models. Applied economists in many other fields generally refer to these as “common shocks” to capture the idea that individuals in spatial or peer groups are subject to unobserved influences in common. These group-specific differences in unobservables are almost inevitable in situations where estimation is based on observational survey, census, or administrative data, and there is no explicit manipulation of the data by experimentation or policy. In situations where we are not interested in the estimation of β , the presence of these unobservable factors that are uncorrelated with x and z requires no more than adjustment to standard errors. Standard approaches to correcting the standard errors in the case of intragroup correlation and groupwise heteroscedasticity can be applied in this case ([Cameron and Miller, 2015](#)). However, these methods require discrete spatial groups, with no intergroup correlation, and can seem *ad hoc* in settings where space is best thought of as continuous. [Conley \(1999\)](#) provides analogous methods for continuous space. For a deeper discussion of these issues, see [Barrios et al. \(2012\)](#). Alternatively, researchers could resort to Monte Carlo methods in which the null distribution is simulated by random assignment across space, an approach that is common in spatial statistics.¹⁷ Unfortunately, in models involving $G_y\gamma$ the implications are more serious.

For models involving $G_y\gamma$, the presence of unobserved effects, even if uncorrelated with the included variables, leads to a basic estimation problem because the ordinary least squares (OLS) estimate of β —the endogenous effect or SAR parameter—is biased and inconsistent. The intuition behind this is simply that the model is a simultaneous equation model. For any individual i , group outcomes $G_y\gamma$ are partly determined by the outcome for individual i . Therefore, group outcomes for individual i , $G_y\gamma$, are explicitly correlated with individual i 's own unobservables. In other words, the spatial lag term contains the dependent variable for “neighbors” (i.e., members of the same group), which in turn contains the spatial lag for their neighbors, and so on, leading to a nonzero correlation between the spatial lag $G_y\gamma$ and the error terms—that is,¹⁸

$$p\lim_{n \rightarrow \infty} = n^{-1}(G_y\gamma'\epsilon) = 0. \quad (3.10)$$

¹⁷ Tests for spatial autocorrelation in the residuals from a regression analysis can also be helpful in establishing whether such corrections to the standard errors are justified. These tests can be based on Moran's I or other statistics that measure spatial autocorrelation, as outlined in [Section 3.2](#).

¹⁸ More technically, the pure SAR model $y = G_y\gamma\beta + \epsilon$ has the following reduced form: $y = (I - G_y\beta)^{-1}\epsilon$. Hence, $G_y\gamma = G_y(I - G_y\beta)^{-1}\epsilon$. Let us define $S = G_y(I - G_y\beta)^{-1}$, then $E(G_y\gamma'\epsilon) = E(\epsilon'^{-1}G_y'\epsilon) = E(\text{tr}(S\epsilon'\epsilon)) = \text{tr}(S)E(\epsilon'\epsilon) \neq 0$. There is no reason to believe that $\text{tr}(S) = 0$.

As a consequence, OLS estimates of parameters in a specification such as Equation (3.5) are inherently biased, unless $\beta = 0$. This is a mechanical endogeneity problem generated by the two-way feedback between individuals in a spatial setting. Much spatial econometrics, since [Anselin \(1988\)](#), is concerned specifically with this problem and adopts maximum likelihood methods or instrumental variables estimators (in the case where there are exogenous variables in the model).¹⁹ While this basic estimation problem is pervasive, solutions to it are well understood. The biases that arise in situations where $G_y\gamma$ determines γ but is omitted from the estimating equation are also well understood and are discussed in Appendix A. The much more substantive problem concerns the question of whether the underlying parameters are identified (or, equivalently, whether valid instruments are available). It is to this issue that we now turn.

3.4.1.1 The reflection problem

To focus on this specific issue, let us define these unobservables as $u = G_v\nu\lambda + \varepsilon$. We assume these are uncorrelated with the observable characteristics x and z —that is, there is no sorting and no omitted spatial variables (we return to this problem in [Section 3.4.3](#)). Using this definition of u , we can write Equation (3.5) as

$$\gamma = X\gamma + G_y\gamma\beta + G_xX\theta + G_zZ\delta + u. \quad (3.11)$$

Premultiplying by $G_y\gamma$ gives

$$G_y\gamma = G_yX\gamma + G_yG_y\gamma\beta + G_yG_xX\theta + G_yG_zZ\delta + G_yu. \quad (3.12)$$

Now, the spatial aggregate or average γ , $G_y\gamma$ is explicitly correlated with u by virtue of the model structure, even if $E[u|X, Z] = 0$. Evidently then $E[u|G_y\gamma] \neq 0$, and least squares estimates of Equation (3.11) are biased. Given this dependence of the spatial average γ on the remaining spatially averaged unobservables (the common unobserved interactions/shocks/correlated effects), methods for estimating β in Equation (3.11) must rely on being able to recover the parameters β , θ , and δ from parameters on the exogenous observables X and Z in the reduced form. The reduced form is obtained by substituting out $G_y\gamma$ in Equation (3.11) to obtain an expression that contains only the exogenous variables and their spatial lags. Unfortunately, in general, it is not easy to recover these parameters from the reduced form without imposing further restrictions.

The fundamental issue which makes it difficult to recover the parameters in Equation (3.11) from its reduced form is that, in this linear specification, the spatially averaged outcomes $G_y\gamma$ are likely to be perfectly collinear with the spatially averaged

¹⁹ See [Lee \(2004\)](#) for details of the maximum likelihood approach and [Kelejian and Prucha \(1998, 1999, 2004, 2010\)](#) for details of the instrumental variables approach. A basic review of the estimation methods for linear spatial models can be found in [Anselin \(1988\)](#).

exogenous variables $G_x X$ and $G_x Z$, except in so far as $G_y y$ is determined by the spatial unobservables u . This holds unless specific types of restrictions are imposed on the structure of G , or on other aspects of the specification, as we discuss in detail below. In other words, $m_y(y, s)_i$ is an aggregation of outcomes or behaviors over “neighbors” (i.e., members of the relevant group) at location s_i , and hence is an aggregation of $m_x(x, s)_i$, $m_z(z, s)_i$ (and u) over neighbors at s_i .

This is easiest to see if we choose the very simple mean-creating, block diagonal, idempotent, and transitive grouping structure as in Equation (3.7), and define a common $G = G_y = G_x = G_z$. In this case,

$$y = X\gamma + G\gamma\beta + GX\theta + GZ\delta + u, \quad (3.13)$$

$$\begin{aligned} G\gamma &= GX\gamma + G\gamma\beta + GX\theta + GZ\delta + Gu \\ &= GX(\gamma + \theta)/(1 - \beta) + GZ\delta/(1 - \beta) + Gu/(1 - \beta). \end{aligned} \quad (3.14)$$

Plugging the expression for $G\gamma$ in Equation (3.14) into the expression for y yields a reduced form:

$$y = X\gamma/(1 - \beta) + GX(\gamma\beta + \theta)/(1 - \beta) + GZ\delta/(1 - \beta) + u + Gu\beta/(1 - \beta), \quad (3.15)$$

$$y = X\tilde{\gamma} + GX\tilde{\theta} + GZ\tilde{\delta} + \tilde{u}. \quad (3.16)$$

The parameters β , θ , and δ cannot be separately identified from the composite parameters $\tilde{\theta} = (\gamma\beta + \theta)/(1 - \beta)$ and $\tilde{\delta} = \delta/(1 - \beta)$ in this reduced form. This is the [Manski \(1993\)](#) “reflection problem,” which Manski originally discussed in the context of social interactions, where we are trying to infer whether individual behavior is influenced by the average behavior of the group to which the individual belongs. Although our exposition above assumes an idempotent G matrix, the problem is not limited to only that case. For example, the problem still arises if, as is common practice in spatial econometrics, we exclude the influence of an individual i on itself in defining G —that is, we set the diagonals to zero to render G nonidempotent as in Equation (3.8). To see this, define $G^\star$ and G as zero-diagonal and non-zero-diagonal matrices for the same grouping structure, with equal-size groups with M members. It follows that

$$G^\star = \frac{M}{M-1}G - \frac{1}{M-1}I.$$

It is evident from this that there is no additional information in $G^\star$ that could be used for identification, since it only differs from G in subtracting the contribution made to each group by individual i . To see this more formally, define $a = \frac{M}{M-1}$ and $b = \frac{1}{M-1}$. Now, using the zero-diagonal grouping matrix in Equation (3.13) and disregarding $G_z z$, for which the concept of zero diagonals is irrelevant since the z come from entities other than the individuals under investigation,

$$\begin{aligned}
\gamma &= X\gamma + G^*\gamma\beta + G^*X\theta + u \\
&= X\gamma + G\gamma\beta b + GX\theta b - a\gamma\beta - aX\theta + u \\
&= G\gamma\beta b + X(\gamma - a\theta)/(1 + a\beta) + GX\theta b/(1 + a\beta) + u/(1 + a\beta).
\end{aligned} \tag{3.17}$$

Evidently, comparing Equation (3.17) with Equation (3.13), we see there is no gain from using zero diagonals in terms of identification, when group sizes are equal, because we have no additional exogenous variables. A similar argument holds when group sizes are large, because $M \xrightarrow{\lim} \infty$ $a = 1$ and $M \xrightarrow{\lim} \infty$ $b = 0$, so $M \xrightarrow{\lim} \infty$ $G^* = G$. The reflection problem carries through in general to any case where $G\gamma$, GX , GZ forms the averages or expectations of γ , X , and Z conditional on the groups defined by G .²⁰

To summarize, to be able to estimate an equation such as (3.5) or (3.6), the researcher must be able to observe differences between the spatial means defined by $G\gamma$, G_xX , G_zZ in the data, otherwise there is insufficient variation to allow estimation. But if group-specific differences lead to variation in $G\gamma$, G_xX , G_zZ , then they almost certainly lead to differences between groups in terms of unobservables. In large groups of individuals (e.g., census data from cities), these differences can arise only because there is nonrandom sorting of individuals across space. In smaller groups (e.g., samples based on friendship networks), the process of assignment to these groups must also be nonrandom, or else the groups must be sufficiently small that the researcher can make an estimation from the random sampling variation in the group means. Of course, if the researcher is conducting an experiment or is investigating the consequences of a specific policy intervention, then that researcher may have much greater control over assignment of individuals to groups and manipulation of the variables of interest, G_xX and G_zZ . We return to discuss these issues in Section 3.5. But for observational data, the reflection problem is very likely to occur unless we are able to impose further restrictions.

3.4.1.2 Solutions to the reflection problem

There are a number of possible solutions to the identification challenges arising from the reflection problem.

First, since the issue originates in the fact that individual outcomes are linear in group-mean outcomes, and group-mean outcomes are, in turn, linear in group-mean characteristics, the use of nonlinear functional forms provides one parametric solution

²⁰ In cases where the group size is small and varies across groups, it is technically possible to identify the parameters in Equation (3.13), with a zero-diagonal block diagonal matrix, as discussed in, for example, Lee (2007) and Bramoullé et al. (2009). This identification comes from the fact that the neighborhood or peer effect for individuals in a given group is a weighted average of the simple mean in the group (from which we have shown that β is not identified) and their own contribution to the mean. These weights vary with group size. The relationship between the simple mean generated by G and the mean generated by G^* is, for a given individual, $G_i^*\gamma = \frac{M_k}{M_k-1}G_i\gamma - \frac{\gamma_i}{M_k-1}$. Technically, identification can come from the weights $\frac{M_k}{M_k-1}$. This is clearly a tenuous source of identification, particularly if there are separate group size impacts (i.e., direct effects) of M_k on the outcome. In addition, in practice, problems may arise because as the group sizes become similar, $\text{Var}(M_k) \rightarrow 0$, and as the group sizes become large, $\frac{M_k}{M_k-1} \rightarrow 1$ and $\frac{1}{M_k-1} \rightarrow 0$.

(e.g., Brock and Durlauf, 2001). For instance, if an outcome is binary (e.g., either to smoke or not to smoke) and thus the probability of smoking is nonlinear in individual characteristics, then identification could come from the assumed functional form of the relationship between covariates and the probability of smoking. However, these kinds of structural assumptions clearly assume that the theoretical structure is known a priori. Further discussion can be found in Chapter 9 and Ioannides (2013). Empirical examples can be found in Sirakaya (2006), Soetevant and Kooreman (2007), Li and Lee (2009), Krauth (2005), and Nakajima (2007).

A second strategy would be to impose restrictions on the parameters on the basis of theoretical reasoning. Obviously, as discussed above, setting $\beta = 0$ and assuming away endogenous effects would be one solution, but would not be very helpful if the aim is to estimate β or we are interested in a structural estimate of γ . Restrictions on some or all of the coefficients on group-means GX are another possibility. That is, if there is some x_r that affects outcomes whose group-mean does not affect outcomes, then the group-average can be used as an instrument for $G\gamma$ in Equation (3.13). These assumptions are quite difficult to defend, and the exclusion restrictions on θ can appear arbitrary. Goux and Maurin (2007), for example, experiment with using neighbors' age as an instrument for neighbors' educational achievement in their study of neighborhood effects in France, but recognize that neighbors' age may have direct effects. Gaviria and Raphael (2001) simply assume away all contextual effects from GX completely.

The third strategy builds on our discussion of the interaction matrix G in Section 3.3.2. It relies on imposing a specific structure for the interaction matrix G that is not block diagonal or transitive, and has the property that $GG \neq G$. This approach to identification has long been proposed in the spatial econometrics literature (Kelejian and Prucha, 1998). Recently, this same approach has been the focus of a number of papers dealing with the identification and estimation of peer effects with network data (e.g., Bramoullé et al., 2009; Calvó-Armengol et al., 2009; Lee et al., 2010; Lin, 2010; Liu and Lee, 2010; Liu et al., 2012).

In the general spatial model in Equation (3.11), if G is characterized by a known nonoverlapping group structure, such that $G_\gamma G_\gamma \neq G_\gamma$, $G_\gamma G_x \neq G_x$, or $G_\gamma G_z \neq G_z$, then the parameters β , θ , and δ can be separately identified. More explicitly, suppose $G_\gamma = G_x = G_z = G$, but $GG \neq G$. As before we can get an expression for $G\gamma$ by multiplying through by G :

$$\gamma = X\gamma + G\gamma\beta + GX\theta + GZ\delta + u, \quad (3.18)$$

$$\begin{aligned} G\gamma &= GX\gamma + G\gamma\beta + GX\theta + GZ\delta + Gu \\ &= GX(\gamma + \theta)/(1 - \beta) + GZ\delta/(1 - \beta) + Gu/(1 - \beta). \end{aligned} \quad (3.19)$$

Now, however, when we plug $G\gamma$ back into the estimating equation, the fact that $GG \neq G$ means we end up with additional terms in G^2X , G^2Z , and $G^2\gamma$ (using the notation that $GG = G^2$). Repeated substitution for $G\gamma$ gives the reduced form of Equation (3.11) as

$$\begin{aligned} \gamma = & X\gamma + GX(\gamma\beta + \theta) + G^2X(\gamma\beta^2 + \theta\beta) + G^3X(\gamma\beta^3 + \theta\beta^2) \\ & + \dots + GZ\delta + G^2Z\delta\beta + G^3Z\delta\beta^2 + \dots + u + Gu\beta + G^2u\beta^2 + \dots \end{aligned} \quad (3.20)$$

In this case, in comparison with Equation (3.15), there are additional exogenous variables which are the spatially double-lagged and spatially multiply lagged observables G^2X , $G^3X, \dots$ and G^2Z , $G^3Z, \dots$ which affect γ only via their influence on $G_y\gamma$. There are at least as many reduced form parameters as structural parameters, so technically, the structural parameters are identified. For example, the ratio of the coefficients on the corresponding elements of the vectors GZ and G^2Z provides an estimate of β . That estimate, combined with the estimate of γ (the coefficient on X) can then be used to back out θ from the coefficient on GX . Alternatively, we could use terms in G^2X , $G^3X, \dots$ and G^2Z , $G^3Z, \dots$ as an instrument directly for $G_y\gamma$ using two-stage least squares. The intuition behind this result is simple: when the interaction structure is incomplete, we can find “neighbors of my neighbors” whose behavior influences me only via the influence that they have on my neighbor. The characteristics of these second-degree neighbors are thus correlated with my neighbors’ behavior, but have no direct influence on my behavior, satisfying the relevance and excludability criterion for a valid instrument.

In principle, these results are widely applicable, because in many real-world contexts, an individual or firm may not necessarily be influenced by all the others in a given group. For example, firms in an industry may not be in contact with all the others in the industry, but may be in contact only with those firms from which they buy inputs. Or a child may not be affected by all children in its school, but may be affected only by those children with whom that child is friends on Facebook. These cases are examples of an incomplete network—that is, everybody is not connected with everybody else. Rather, each individual has its own group of contacts, which differ from individual to individual. When this occurs, $GG \neq G$, and this solves the reflection problem as just discussed. The network structure provides a good context to summarize the intuition for the formal result. Consider a simple network with three individuals A, B, and C as illustrated in Figure 3.2.

A and B play piano together and B and C swim together, but A and C have never met. Then, the only way C could influence A’s behavior is through B. The characteristics of C are thus a good instrument for the effect of the behavior of B on A because they certainly influence the behavior of B but they do not influence directly the behavior of A.

To identify network effects, one needs only one such intransitivity; however, in most real-world networks, there are a very large number of them.

While in principle this solution to the reflection problem might apply in a large number of situations, its application in many spatial settings is problematic. The identification

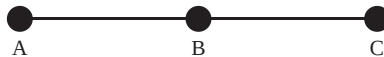


Figure 3.2 A simple network.

strategy relies on having detailed and accurate data on the interactions between agents (i.e., one needs to know exactly who interacts with whom). In particular, it hinges upon nonlinearities in group membership (i.e., on the presence of intransitive triads). If links are incorrectly specified, then the exclusion restrictions are violated. Going back to our example in Figure 3.2, if C in fact knows A but we assume that she does not, then identification fails. In the network literature, restrictions on the interaction scheme are often imposed on the basis of data that specifically seek to identify relevant linkages (Bramoullé et al., 2009; Calvó-Armengol et al., 2009; Lee et al., 2010; Lin, 2010; Liu and Lee, 2010; Liu et al., 2012) or are explicitly derived from theory.

In contrast, in the spatial econometrics literature, the requirement that $GG \neq G$ has been largely met through the use of *ad hoc* spatial weight matrices pulled from a pick-list of popular forms—for example, constructed on the basis of rook or queen contiguity, or inverse distance weighting, which are non-block diagonal and nonidempotent as discussed in Section 3.3.2. In our view, while $GG \neq G$ provides a solution to the reflection problem, any such restrictions require careful justification on the basis of institutions, policy, or theory, or (as in the network literature) need to be imposed on the basis of data that specifically seek to identify relevant linkages. This is something which is very hard to achieve when simply imposing many of the popular spatial weight matrices.

Unfortunately, identification fails if these restrictions (whether carefully justified, based on data, or imposed *ad hoc*) are invalid. The network literature suggests that the problems of missing data (on nodes, but not on links) may be less severe. Helmers and Patnam (2014), Liu et al. (2012), and Liu et al. (2013) present Monte Carlo evidence on the bias of the estimator when misspecification of the social network structure is due to data for individuals missing at random because of sampling (but where all links are observed). Liu et al. (2013) develop a nonlinear estimator designed to address sampling issues over networks. The common finding seems to be that random sampling with known network structure induces a consistent downward bias in the estimates at all sample sizes and at all spatial parameter values. That is to say, as in more standard settings, nonsystematic measurement error causes attenuation bias on the parameters of interest. This implies that, in the presence of a known network structure but random measurement error for nodes, estimated coefficients are likely to provide a lower bound for the importance of social interactions. There is little chance, however, that random measurement errors are inducing us to detect the presence of peer effects when they are not existent (see Conley and Molinari, 2007; Kelejian and Prucha, 2007 for studies showing the robustness of variance-covariance estimators to location misspecification). In other words, if G is known and the only source of measurement error is random missing data for specific nodes, point estimates of peer effects are likely to be higher and standard errors remain roughly unchanged. Note, however, that these results do not provide much reassurance in situations where missing data are nonrandom or where there are errors on the interaction structure (e.g., due to the endogeneity of the interaction structure, missing

links in the network, or the fact that the restriction $GG \neq G$ has been arbitrarily imposed by choosing one of the popular spatial weight matrices).

Even when G is known and the network is incomplete, so that G^2X , G^3X , G^2Z , G^3Z (and so on) provide valid instruments, the weakness of the instruments may prove a serious threat to identification and estimation.²¹ This weak instruments problem arises if the instruments G^2X , G^3X , G^2Z , G^3Z (and so on) are highly correlated with the explanatory variables GX and GZ , so that, conditional on GX and GZ , there is little variation in the instruments. Therefore, while identification is technically possible, there may be little variation in the instruments to allow estimation. This is potentially a serious problem when G represents spatial connections between neighboring agents or places, when G is row normalized so that it creates the means of the neighbors (as G is commonly specified), and where there is strong spatial autocorrelation in X and Z (usually the case empirically). In this case Gx , for example, estimates the mean of a variable x at each location on the basis of the values of x at neighboring locations, G^2x estimates the means at each location on the basis of the means of the means of x at each location, and so on. So, Gx , G^2x , and G^3x are all just estimates of the mean of x at each location using different weighting schemes. Indeed, this use of neighbors to estimate location-specific means underpins nonparametric kernel regression methods, and spatial interpolation methods in GIS applications. In practice, in cases where the groups formed by G are small (e.g., three nearest neighbors, or contiguous districts), there may be enough sampling variation in these means to ensure that Gx , G^2x , G^3x , and higher-order spatial lags are not perfectly collinear, so estimation may be possible. The problem is, however, potentially especially serious in the situations, noted at the end of [Section 3.3](#), where the numbers of observations in a group becomes very large. The means estimated by Gx , G^2x , and G^3x converge to the population mean of x at each location as the group size goes to infinity, implying the spatial lags are all perfectly collinear and so identification fails.²²

This weak instruments problem is potentially less pervasive in peer group network applications with individual data (see [Chapter 9](#)) when the information on social connections is rich and if individuals make diverse and idiosyncratic choices about their friends. In this case, unlike the spatial setting with spatial autocorrelation, the characteristics of an individual's friends provide little or no information about the individual's own characteristics. However, in cases where peer groups are formed by strongly assortative or

²¹ As discussed in [Bound et al. \(1995\)](#), weak instruments lead to a number of problems. The two-stage least squares estimator with weak instruments is biased for small samples. Any inconsistency from a small violation of the exclusion restriction is magnified by weak instruments. Finally, estimated standard errors may be too small. [Stock et al. \(2002\)](#) propose a first-stage F test that can be used to guide instrument choice when there are concerns about weak instruments.

²² For example, the mean of a variable x among the 1000 nearest neighbors of an individual will not be very different from the mean among the 1000 nearest neighbors of that individual's nearest neighbor, so Gx , G^2x , G^3x , and so on will be almost perfectly collinear.

disassortative matching processes, the weak instruments issue may still create a potential threat to estimation and identification.²³

We have considered three possible solutions to the reflection problem—the use of functional form, the imposition of exclusion restrictions, and the use of an incomplete interactions matrix such that $GG \neq G$. The last of these, in particular, has received considerable attention in the recent social interactions literature focusing on the identification and estimation of peer effects with network data. These methods may be applicable in a broader set of spatial settings. However, any such restrictions require careful justification on the basis of institutions, policy, or theory, or need to be imposed on the basis of data that specifically seek to identify relevant linkages. While these issues have received careful consideration in both the networks literature and the theoretical spatial econometrics literature, much applied work continues to rely on *ad hoc* restrictions implicitly imposed through the choice of popular spatial weight matrices.

3.4.2 Spatially autocorrelated unobservables, when these are correlated with the observables

So far we have set aside the possibility, explicit in Equation (3.2) or (3.5), that there are spatial or group-specific unobservables, $m_\nu(\nu, s)_i$ or $G_\nu\nu$ using the matrix form, which are correlated with the explanatory variables. The second challenge arises once we drop this assumption and allow for the possibility that unobservables $u = G_\nu\nu\lambda + \varepsilon$ are correlated with the observable characteristics x and z . In many situations observable individual, location, and neighbor characteristics x , $G_x x$, and $G_z z$ are very likely related to the unobservable location and neighbor characteristics $G_\nu\nu$. We can identify two mechanisms. First, group membership is exogenous and the correlation arises because of spatially omitted variables that are correlated for individuals in the same group. These omitted variables may directly affect y , or they may determine x or z and hence indirectly affect y . Second, group membership is endogenous and the correlation arises because of the sorting of individuals with different characteristics x into locations with different $G_\nu\nu$. For example, in the agglomeration literature the link between urban wages and urban education may arise because cities that offer high returns to education have unobserved characteristics that encourage individuals to acquire more schooling (as in the literature on human capital externalities, reviewed in [Moretti, 2004](#)), or highly educated workers may move into cities that offer high returns to their education (as in the urban wage premium literature; e.g., [Combes et al., 2008](#)). In either case, if the factors that determine city-specific returns to education are not all observable, x and spatial aggregates of x (i.e., $G_x x$) or variables that are included in $G_z z$ are correlated with $G_\nu\nu$.

²³ [Lee and Liu \(2010\)](#) propose a generalized method of moments with additional instruments to try to circumvent the weak instrument problem.

It is important to note that while the urban economics literature has traditionally recognized these two mechanisms through which $G_x x$ and $G_z z$ may be correlated with $G_v \nu$, it has tended to treat these symmetrically. However, in most cases “sorting” is better thought of as the situation where group membership is endogenous. That is, the correlation between $G_x x$ or $G_z z$ and $G_v \nu$ arises because G_x , G_z , and G_v are endogenous. In this subsection, we set aside this possibility to consider the situation where group membership is exogenous (although not necessarily fixed over time) and correlation arises because of spatially omitted variables that are correlated for individuals in the same group.

Suppose that the aim is to estimate a specification without endogenous interactions, either because endogenous interactions are being ruled out, or because this is viewed as the reduced form of a model with endogenous specifications. Restricting our attention to spatial interactions that can be represented by a set of spatial weight matrices implies

$$y = X\gamma + G_x X\theta + G_z Z\delta + G_v \nu\lambda + \varepsilon. \quad (3.21)$$

Standard nonexperimental approaches to estimating Equation (3.21) all involve, in some way, transforming the estimating equation in a way that “partials” out $G_v \nu$ so that it no longer enters the estimating equation. For example, an increasingly common way to partial out $G_v \nu$ is to apply “spatial differencing,” which transforms all variables by subtracting some appropriately constructed spatial mean (Holmes, 1998). Assume, for the moment, that we know G_v , then spatial differencing is equivalent to premultiplying Equation (3.21) by a transformation matrix $[I - G_v]$ to give (where ζ is another random error term)

$$y - G_v y = (X - G_v X)\gamma + (G_v - G_v G_x)X\theta + (G_z - G_v G_z)Z\delta + (G_v - G_v G_v)\nu\lambda + \zeta. \quad (3.22)$$

If $\text{plim}(G_v - G_v G_v)\nu = 0$, this transformation eliminates spatial unobservables $G_v \nu$, allowing consistent estimation of Equation (3.22) by OLS. Clearly, from the above, this condition will hold when we know G_v and where G_v has an idempotent structure (e.g., block group structures similar to the example in Equation (3.7)), in which case $G_v - G_v G_v = 0$, so

$$y - G_v y = (X - G_v X)\gamma + (G_v - G_v G_x)X\theta + (G_z - G_v G_z)Z\delta + \zeta. \quad (3.23)$$

This is just a standard fixed effects estimator, in which variables have been differenced from some group mean (where the groups are defined by G_v) or where the regression includes a set of dummy variables for the groups defined by G_v .

Indeed, if we have panel data providing multiple observations for individuals over time and define G_v to have a block group structure for each individual, this is just the standard fixed effects estimator. The transformation matrix $[I - G_v]$ eliminates the individual-level mean and allows us to consistently estimate Equation (3.21) providing that group-level characteristics are correlated only with time-invariant individual-level unobservables. Individual-level time-varying shocks will still lead to inconsistent estimates if they are correlated with group-level characteristics. This is the approach adopted

in the standard mincerian wage regression approach to estimating city-level productivity or wage differences (Combes et al., 2008; Di Addario and Patacchini, 2008; Mion and Naticchioni, 2009; De la Roca and Puga, 2014; Gibbons et al., 2014; and many others). In that literature, the identifying assumption is that city location (i.e., group membership) can be correlated with time-invariant individual characteristics (such as ability), but not with time-varying shocks (e.g., to an individual's income).

Just as with the standard individual fixed effects approach, there are evidently further limitations to the application of spatial differencing. Suppose in the absence of any other information, we simply assume that the spatial weighting/grouping functions $m(.,s)$ are the same for all variables—that is, $G_x = G_z = G_v = G$. In this case, Equation (3.23) reduces to

$$\gamma - G\gamma = (X - GX)\gamma + \zeta. \quad (3.24)$$

Note that spatial differencing removes both $GX\theta$ and $GZ\delta$, so while the parameters γ on X are identified, the parameters on the spatial variables GX or GZ are not. This is, of course, just the standard problem that the parameters on variables that are collinear with group fixed effects cannot be estimated. Clearly, if one is willing to assume that the structure of connections in terms of unobservables G_v is different from the ones in terms of observables (G_x and G_z), then demeaning the variables using the spatial means of G_v would not eliminate GX and GZ and allow estimation of θ and δ .²⁴ However, imposing a different structure of connections for the observables and unobservables is a strong assumption. This discussion illustrates a crucial point: even in the most basic strategy for eliminating spatial unobservables, researchers are making fairly strong assumptions about the structure of the implied interconnections between observations, and the structure of the (implicit) G matrices that link different observations together on observable and unobservable dimensions.

There are cases where this assumption may serve as a reasonable approximation. For example, a study of neighborhood effects on labor market outcomes might be prepared to assume that the observable variables of interest—for example, neighborhood unemployment rates—are linked at the neighborhood level (defined by G_x), but that unobservable labor market demand factors (G_v) operate at a large labor market level. A good research design should ground this identifying assumption on sound theoretical reasoning or on supporting evidence (e.g., about institutional arrangements).

One increasingly popular approach in spatial settings, “boundary-discontinuity” design (which is a particular spatial case of regression discontinuity design), provides an explicit justification for having a distinct set of weights for observables and unobservables. In this setup, the researcher cites institutional and policy-related rules as a justification for assuming that the spatial connections between places in terms of the

²⁴ Estimation of γ does not require this assumption as shown above.

characteristics of interest are very different from those that affect unobservables ν . This difference may arise because, for example, administrative boundaries create discontinuities in the way $G_z Z$ varies over space but (so it is assumed) do not create discontinuities in the way $G_\nu \nu$ varies over space. Typical applications include studies of the effects of school quality on house prices (Black 1999), the effect of local taxes on firm employment (Duranton et al., 2011), and the evaluation of area-based initiatives (Mayer et al., 2012; Einio and Overman, 2014). This boundary-discontinuity design amounts to defining G_ν to be a block diagonal matrix, in which pairs of places that share the same nearest boundary and are close to the boundary (e.g., within some distance threshold) are assigned equal nonzero (row-normalized) weights. G_z , on the other hand, is structured such that a row for an individual i , located at s_i , assigns nonzero weights to places on the same side of the administrative boundary, and zero weights (or much smaller weights) to places in different administrative districts to location s_i . Restricting G_ν in this way implicitly assumes that observations close to an administrative boundary share the same spatial unobservables, but that area-level determinants are at work at the administrative district or sub-administrative district level. The main threat to identification in this boundary-discontinuity regression discontinuity design is that this assumption may not hold. For example, individuals may sort across the boundary in response to cross-boundary differences in $G_z Z$, so unobserved individual characteristics will differ across the boundary, leading to a change in $G_\nu \nu$ across the boundary. Again, note that it is the assumptions on the structure of $G_\nu \nu$ that have failed in this example.

There are also extensions to the spatial differencing/fixed effects idea in which G is not idempotent, but $\text{plim}[G_\nu G_\nu] = \text{plim}[G_\nu]$. This would be true for any case in which G_ν forms an estimate of the mean of ν at each location s , because $E[E[\nu|s]|s] = E[\nu|s]$. This is the case if each row of G , $g(s)$ is structured such that it comprises a sequence of weights $[g_{i1} \ g_{i2} \ g_{i3} \ \dots]$ which decline with the distance of locations 1,2,3,... from location s , and sum to 1, which yields a standard kernel weighting structure. Applications of this approach are given in Gibbons and Machin (2003) and Gibbons (2004). However, the basic problem remains that the spatial weights used to aggregate spatial variables of interest $G_x X\theta$ and $G_z Z\delta$ must be different from the spatial weights used in the transformation to sweep out the unobservables ν .

As with the reflection problem, if $G_\nu = G_x = G_z = G_\nu = G$ is known and the network is incomplete, then $G^2 X$, $G^3 X$, $G^2 Z$, $G^3 Z$,... continue to provide valid instruments for G_ν , although not for G_x or G_z . That is, an incomplete structure for G can solve the reflection problem and allow estimation of the coefficient on endogenous effects ($G_\nu \gamma$) in the presence of peer-group-specific effects that are correlated with observables. But this cannot provide us with an estimate of the coefficients on either G_x or G_z . More generally, the other way to think about these spatial models with sorting and correlated spatial shocks is in terms of the class of general problems where x and z may be correlated with the error term and to look for ways of instrumenting using variables that are

exogenous but correlated with the included variables. This approach requires theoretical reasoning about appropriate instruments. However, even then, the instruments must be orthogonal to the spatial unobservables, so it is often necessary to apply instrumental variables combined with spatial-differencing-based methods (see, e.g., [Duranton et al., 2011](#)).

In a nutshell, when group membership is exogenous and there are unobservable variables that are correlated with observables, our ability to estimate coefficients of interest depends on the structure of the spatial interactions. If we are willing to assume that the interconnections between individuals on these unobserved dimensions are best described by a matrix of interconnections G_ν that is symmetric and idempotent, then these unobservables can be partialled out using standard differencing/fixed effects methods. If we wish to estimate the coefficients on the spatial explanatory variables $G_x X$, $G_z Z$, we must further assume that the interconnections between individuals that form the group-level or spatial averages of the explanatory variables (i.e., G_x and G_z) must be different from G_ν . If this assumption holds, the spatial differencing/fixed effects design eliminates the spatially correlated unobservables, but does not eliminate the spatial explanatory variables. Neither of these assumptions is sufficient to allow the estimation of $G_y \gamma$. If we wish to estimate the coefficient on $G_y \gamma$, then we must assume a known incomplete interaction matrix. This solves the reflection problem and allows the estimation of the coefficient on $G_y \gamma$ but not on $G_x X$ or $G_z Z$ (in either the structural or the reduced form).

Note that the issues and solutions discussed in this section are essentially the same as those for standard omitted variables, but where the correlation between unobservables and observables arises through channels that may not be immediately obvious without thinking about the spatial relationships at work. A subtler consequence of omitted spatial variables is the so-called modifiable areal unit problem (see, e.g., [Openshaw, 1983](#); [Wong, 2009](#); [Briant et al., 2010](#)) in which estimates of parameters can change as the spatial aggregation of the units of analysis changes. We say more about this issue in Appendix A.

3.4.3 Sorting and spatial unobservables

In the previous section we considered the possibility, explicit in Equation (3.2) or Equation (3.5), that there are spatial or group-specific unobservables, $m_\nu(\nu, s)_i$ or $G_\nu \nu$ using the matrix form, which are correlated with the explanatory variables. Our discussion there assumed that group membership was exogenous. In this section we allow for the possibility that group membership is endogenous so that the correlation between $G_x x$ and $G_z z$ with $u = G_\nu \nu + \varepsilon$ stems from individual-level decisions about group membership. As discussed above, while the urban economics literature has traditionally recognized these two mechanisms through which $G_x x$ and $G_z z$ may be correlated with $G_\nu \nu$, it has tended to treat these symmetrically. However, when group membership is endogenous, the correlation between $G_x x$ or $G_z z$ and $G_\nu \nu$ arises because G_x , G_z , and G_ν are endogenous.

If the individual-level variables that affect location also affect outcomes, then a fixed effects approach can do little to alleviate this problem as the individual-level unobservables would not be eliminated when subtracting a group-mean. To return to the urban wage premium example, including individual-level and city-level fixed effects does not consistently identify the urban wage premium if unobserved shocks (e.g., a change in labor market circumstances) affect both wages and location.

In much of the urban economics literature, the response to this problem has been to suggest that this is the best that can be achieved in the absence of random allocation across locations (we consider this further in the next section). An alternative is to impose more structure on the location problem. Ioannides and Zabel (2008), for example, use factors influencing neighborhood choice as instruments for neighbors' housing structure demand when estimating neighborhood effects in housing structure demand. The literature on equilibrium sorting models and hedonics may lead to further theoretical insights into identification of neighborhood effects when the researcher is prepared to impose more structure on the neighborhood choice process (Kuminoff et al., 2013).

Various estimation techniques have recently been developed in the econometrics of network literature to address the issue of endogenous group membership. These have not yet been applied in spatial settings although they may be helpful (particularly for researchers taking a more structured approach). There are three main methodological approaches. In the first approach, parametric modeling assumptions and Bayesian inferential methods are employed to integrate a network formation model with the model of behavior over the formed networks. The selection equation is based on individual decisions and considers all the possible couple-specific correlations between unobservables. This is a computationally intense method where the network formation and the outcome equation are estimated jointly (Goldsmith-Pinkham and Imbens, 2013; Hsieh and Lee, 2013; Mele, 2013; Del Bello et al., 2014; Patacchini and Rainone, 2014). The alternative approach is the frequentist approach, where a selection equation based on individual decisions is added as a first step prior to modeling outcome decisions. An individual-level selection correction term is then added in the outcome equation. The properties of the estimators are analytically derived. Observe that, while the idea is similar to a Heckman-type estimation, inference is more difficult because of the complex cross-sectional interaction scheme. This approach is considered in Liu et al. (2012). Finally, another strategy is to deal with possible network endogeneity by using a group-level selection correction term. The group-level selection correction term can be treated as a group fixed effect or can be estimated directly. Estimation can follow a parametric approach as in Lee (1983) or a semiparametric approach as in Dahl (2002). This method is considered in Horrace et al. (2013).

In the peer groups/social interactions literature that employs the network structure as a source for identification, network or "component" fixed effects can sometimes be used to control for sorting into self-contained networks or subsets of the networks (Bramoullé

et al., 2009; Calvó-Armengol et al., 2009; Lee et al., 2010; Lin, 2010; Liu and Lee, 2010). For example, children whose parents have a low level of education or whose level of education is worse than average in unmeasured ways are more likely to sort into groups with low human capital peers. If the variables that drive this process of selection are not fully observable, potential correlations between (unobserved) group-specific factors and the target regressors are major sources of bias. The richness of social network data (where we observe individuals over networks) provides a possible way out through the use of network fixed effects, for groups of individuals who are connected together, assuming individuals fall into naturally disconnected subgroups, or some cutoff in terms of connectivity can be used for partitioning into subgroups. Network fixed effects are a potential remedy for selection bias that originates from the possible sorting of individuals with similar unobserved characteristics into a network. The underlying assumption is that such unobserved characteristics are common to the individuals within each network partition.²⁵ This may be a reasonable assumption where the networks are quite small—for example, a network of school students. When networks contain instead a large number of agents who are not necessarily drawn together by anything much in common—for example, a network of LinkedIn connections—this is no longer a viable strategy as it is not reasonable to think that the unobserved factors are variables which are common to all members. As another example, networks of transactions in the housing market that involve a large number of properties may contain different types of unobservables for different properties, even though all the properties belong to the same network of buyers and sellers. In this case, the use of network fixed effects would not eliminate endogeneity problems. A similar context is provided by trading networks with financial data. Also in this case, when the number of transactions is high, the use of network fixed effect is not a valid strategy, although network topology can still contain valuable information (see Cohen-Cole et al., 2014). Obviously, it must also be feasible to partition individuals into mutually exclusive sets of individuals (or units) who are not directly or indirectly related in the network in order to define the fixed effects, so this is not a solution in networks where all individuals are indirectly related to each other.

3.4.4 Spatial methods and identification

To summarize, all researchers working with spatial data face fundamental identification and estimation challenges. Spatial methods can provide a partial solution to these challenges. Restrictions on functional form, on the exogenous variables that directly determine outcomes, and on the nature of interactions may solve the reflection problem and allow identification of interaction effects. But identification fails if these restrictions

²⁵ Testable implications of this assumption can be verified using the recent approach proposed by Goldsmith-Pinkham and Imbens (2013). Patacchini and Venanzoni (2014) apply this approach to an urban topic.

are invalid. Further challenges to identification arise if there are omitted variables that are correlated with observables. These challenges arise when estimating models with or without endogenous interactions. Standard solutions to these problems (e.g., fixed effects, spatial differencing) imply restrictions on the nature of spatial interactions. Reformulating these approaches within a spatial econometrics framework makes these restrictions explicit. If the omitted variables problem arises because of sorting across space (i.e., location is endogenous), this raises further identification problems. Again, reformulating sorting within the spatial econometrics framework, specifically as giving rise to an endogenous interaction matrix, helps clarify these issues. The network literature and the spatial econometrics literature suggest some solutions to the sorting problem although all of these require further assumptions and restrictions on the model that determines location. In situations where researchers are unwilling to impose these restrictions, it is often suggested that the use of standard spatial methods (e.g., fixed effects or spatial differencing) provides the best estimates that we can hope for in the absence of random allocation across locations. Unfortunately, recent literature questions the extent to which even random allocation may help. It is to this question that we now turn.

3.5. TREATMENT EFFECTS WHEN INDIVIDUAL OUTCOMES ARE (SPATIALLY) DEPENDENT

In this section, we recast the discussion so far in terms of the framework used in the policy evaluation literature, where the aim is to estimate the treatment (causal) effect of some policy intervention.²⁶ We consider the extent to which explicit experiments—for example, randomized controlled trials (RCTs)—can be designed to overcome the basic identification problems discussed above. Doing so helps reinforce the intuition provided above by considering the issues within a different conceptual framework, as well as providing a link to the evaluation literature that applies RCTs in settings where spatial or network dependence may be important.

3.5.1 (Cluster) randomization does not solve the reflection problem

As discussed above, the reflection problem can prevent estimation of β (the effect of neighbor outcomes or behavior on individual outcomes) separately from θ (the effect of neighbor characteristics) in situations where there are unobservable factors that also vary at the group level. Unfortunately as this section shows, without the imposition of further restrictions, randomization does not generally solve the reflection problem.

²⁶ A burgeoning literature considers the application of treatment effect analysis to economic problems. Early surveys include those of Angrist and Krueger (1999) and Heckman et al. (1999), while Lee (2005) provides a book-level treatment. Angrist and Pischke (2011), among a number of others, provide further discussion.

To think this through, consider the design of an experiment that would identify the parameters from a standard linear (spatial) interactions model where outcome y is determined by both individual characteristics and the outcome, observed and unobserved characteristics of some reference group (for simplicity we disregard Z or assume it is subsumed in X , and we suppress the constant):

$$y = X\gamma + G_y\gamma\beta + G_xX\theta + u. \quad (3.25)$$

If each individual is a member of at most one reference group (i.e., G is block diagonal), then an RCT could use the existing reference groups (summarized by G) as the basis for the random allocation of treatment. That is, the group, rather than the individuals, can be randomized into treatment. This is the approach taken by cluster randomized trials, which have seen widespread application in the public health literature (see, e.g., [Campbell et al., 2004](#)). Note that, although G may be endogenously determined, randomization of groups into treatment ensures that u is uncorrelated with treatment status (at least when there are a large number of available groups). We can model treatment as changing some element of x_i for all members of treated groups while holding everything else constant. Given that there is complete interaction within each group (and assuming G is row normalized), $G_y\gamma$ and G_xX form the sample mean within each group. Thus, treatment affects individuals directly through x_i , and indirectly via both $G_y\gamma$ and G_xX . As highlighted by [Manski \(2013\)](#), and discussed further below, these assumptions imply restrictions on the treatment response functions (which characterize the way in which outcomes change with treatment) that are not trivial.

Suppose we have just two groups, group 0 and group 1, with random assignment of treatment to all members of group 1 rather than to members of group 0. We have

$$\text{Treatment group: } E[y|1] = E[x|1](\gamma + \theta)/(1 - \beta) + E[u|1]/(1 - \beta), \quad (3.26)$$

$$\text{Control group: } E[y|0] = E[x|0](\gamma + \theta)/(1 - \beta) + E[u|0]/(1 - \beta), \quad (3.27)$$

where random assignment implies $E[y|1] - E[y|0] = 0$, given that $E[x|1] - E[x|0] = 0$, $E[u|1] - E[u|0] = 0$. Now we expose all members of the treatment group to some known treatment, by changing some element of x_i for all members of the treatment group (group 1) while holding everything else constant, to give $E[x|1] - E[x|0] = x^*$. This gives the reduced form, causal effect of the treatment:

$$\begin{aligned} E[y|1] - E[y|0] &= (E[x|1] - E[x|0])(\gamma + \theta)/(1 - \beta) \\ &= x^*(\gamma + \theta)/(1 - \beta). \end{aligned} \quad (3.28)$$

For many policy evaluation purposes this is sufficient, but it is clear that cluster randomization does not solve the reflection problem and allow the separate estimation of γ , θ , and $(1 - \beta)$. With control over within-cluster assignment to treatment it is possible to go further (under the assumptions imposed so far) and separately identify the direct effect of the intervention (γ) from the effects due to social interactions. We show

an example in Appendix B. Note, however, that control over group membership when individuals are members of only one group (i.e., G is block diagonal) does not provide a solution to the reflection problem or allow us to separately identify θ or $(1 - \beta)$.

In addition, note that applying cluster randomization to existing reference groups raises issues with respect to inference when (a) group membership is endogenous, or (b) there are omitted group-specific variables that affect outcomes. Both situations imply that the characteristics of individuals are correlated with the characteristics of others in their group. This within-group correlation in terms of either observable or unobservable characteristics (often referred to as intraclass correlation) reduces the effective sample size in a way that depends on both the size of the within-group correlation and the average group size relative to the total sample size. When within-group correlation equals 1 (so that individuals are identical within groups in terms of characteristics which determine y), the effective sample size is equal to the number of groups. When within-group correlation in the characteristics that determine y is 0, the effective sample size is equal to the total number of individuals in the two groups. For intermediate situations, basing inference only on the number of groups will result in standard errors that are too large, while using the total number of individuals will result in standard errors that are too small. Using conservative standard errors (based on group size) will exacerbate concerns over power (i.e., the probability of correctly rejecting the null hypothesis of no treatment effect when the null is false) in situations where the number of groups is small and the within-group correlation is large.

In situations where the researcher has control over group membership, random assignment of individuals to treatment and control groups, rather than random assignment of treatment to all members of existing groups, helps address these concerns over inference. This is because individual-level randomization reduces this within-group correlation in terms of both observable and unobservable characteristics, given that group membership is no longer endogenously determined. It also ensures that u is uncorrelated with treatment status in situations where unobservable characteristics are correlated within groups (as will usually be the case when group membership is endogenous). However, even if we randomly allocate individuals to treatment and control groups, if we want these individuals in the treated group to interact, then they have to be colocated somewhere and if they are colocated, then they will be subject to place-specific unobservables. Therefore, even this form of randomization does not completely eliminate the problems for inference induced by treating people in groups.

In practice, it is perhaps difficult to think of situations where we would have such strong control over both group membership and treatment assignment within groups. But thinking about the appropriate RCT helps clarify intuition about the kind of quasi-random variation needed to achieve identification of the direct effect γ separately from the effects of interaction between agents. Conditional on the assumption about the

treatment response function,²⁷ an RCT with control over both group membership and individual assignment into treatment allows us to eliminate biases due to selection on unobservables into the two groups, and to estimate the reduced form effect of changes in x and group average $\bar{x}$. The quasi-experimental methods for causal analysis on non-experimental data discussed in [Chapter 1](#) are therefore perfectly applicable to this problem providing they can use two sources of quasi-random variation: the first to determine assignment into treatment, the second to determine assignment into the reference group. Note, however, that simple treatment/control randomization does not solve the “reflection” problem of separate identification of β and θ , so clearly methods based on quasi-random variation will also fail in this respect.

Is there an experiment that separately identifies β and θ ? As before, we must impose more structure on the problem to achieve identification. It should be clear from [Section 3.4](#) that an appropriate identification strategy must rely on overlapping but incomplete network structures (i.e., a nonidempotent G matrix with intransitive network relationships). Appendix B provides an example of a simply hypothetical experiment that fulfills these criteria.

As can be seen, the requirements for a successful RCT to identify the separate causal parameters in the general spatial model of Equation (1) are rather stringent. Two key components are required: (a) randomization into different groups; (b) a known and enforceable “incomplete” network structure that defines the permissible interactions between agents in these groups. Even then there are evidently problems when trying to design such a hypothetical experiment to answer questions that are specifically spatial, such as questions about neighborhood effects or geographical spillovers. For example, in the hypothetical experiment discussed in Appendix B, individuals are assigned into a control group and three treatment groups (groups 1–3). The crucial restriction for identification is that individuals in group 1 are connected to individuals in group 2 and individuals in group 2 are connected to individuals in group 3, but individuals in groups 1 and 3 are not connected. If the connections are spatial, then ensuring compliance is not so straightforward, since group 1 must overlap with group 2 in space and group 2 must overlap with group 3 in space, so it is very hard to ensure that group 3 does not overlap with group 1 in geographical space. Given the difficulties of designing a hypothetical experiment to recover these parameters, it becomes clear that recovering them from observational data when there is no explicit randomization and/or the true network structure of G is unknown is going to be difficult.

The situation is further complicated once we relax the assumption on the treatment response function that we have imposed so far (i.e., that treatment affects individuals directly through x_i , and indirectly via both $G_y\gamma$ and G_xX). As emphasized by [Manski \(2013\)](#), once we allow for the possibility of social interaction, it is hard to maintain

²⁷ That is, that treatment affects individuals directly through x_i , and indirectly via both $G_y\gamma$ and G_xX .

the assumption that individual outcomes only vary with own treatment, and not with treatment of other members of the population. That is, the stable unit treatment value assumption (Rubin, 1978) that underpins much of the treatment effects literature is unlikely to hold. As Manski (2013) makes clear, the stable unit treatment value assumption, or “individualistic treatment response” assumption (as he calls it) is quite restrictive in situations that allow for social interaction. Indeed, in the examples above, we dropped this assumption to allow the treatment effect to depend on both the individual treatment and the average level of treatment in the group (as captured by $G_{j\gamma}$ and $G_x X$). Manski (2013) defines this as a functional interaction response (the interaction occurs only through some function of the distribution of treatments across the groups—in this case the mean). Relaxing this assumption would give us what Manski calls distributional interactions (where individual treatment response depends on the distribution of treatments across others in the group but not on the size of the group or the identity of those treated). A further relaxation gives anonymous interactions (the outcome of person j is invariant with respect to permutations of the treatments received by other members of his group, but the size of the group could matter). Progressively weaker assumptions on the treatment response function make identification more difficult. The situation is further complicated if we allow reinforcing or opposing interactions (two examples of “semimonotone treatment response functions”). Treatment could also influence group structure if, for example, treatment is observable and individuals sort on the basis of treatment. In short, even in situations where G is known and structured such that $GG \neq G$, further assumptions on the nature of the treatment response function are required to identify treatment effects of interest. The literature that considers these issues is in its infancy.

3.5.2 Randomization and identification

It is increasingly common for the applied urban economics literature to suggest that the application of spatial methods (e.g., fixed effects, spatial differencing) represents the “best we can do” in the absence of explicit randomization. While this may be true, this section showed that randomization itself may be insufficient to solve fundamental identification problems, especially where the aim is to identify endogenous neighborhood effects or spillovers of the SAR variety in spatial econometrics. Even in situations where the researcher has control over group structure and treatment, identification of β (the effect of neighbor outcomes or behavior on individual outcomes) separately from θ (the effect of neighbor characteristics) is not straightforward. Uncertainty about treatment response (i.e., the appropriate functional form) or the endogeneity of group membership (especially to treatment) further complicates the problem, as well as providing an additional set of challenges to researchers interested in identifying reduced form treatment effects. The nascent literature considering this latter issue is yet to receive widespread

consideration in the applied treatment effects literature. However, this emerging literature makes it clear that much applied work relies on restrictions on the treatment response function, in particular the individual treatment response assumption, which may not hold in practice. Dealing with these issues is one of the key challenges facing those who wish to develop and apply the treatment effects approach in spatial settings.

3.6. CONCLUSIONS

This chapter has been concerned with methods for analyzing spatial data. After initial discussion of the nature of spatial data and measuring and testing for departures from randomness, we focused most of our attention on linear regression models that involve interactions between agents across space. The introduction of spatial variables—functions that generate (usually linear) aggregations of variables that are spatially connected with a specific location using information on all locations—into standard linear regression provides a flexible way of characterizing these interactions. The introduction of these spatial variables complicates both interpretation and estimation of model parameters of interest. This raises the question of whether one could ignore these spatial variables and still correctly determine the impact of some specific variable x on some outcome y ? As is usually the case, however, model misspecification—in this case ignoring interactions between individuals when they are relevant—means that OLS results may be misleading. In some circumstances—for example, when we are interested in the impact of some policy intervention x on some outcome y —the OLS bias may not be problematic. In other cases, this bias will be a problem. This is one reason to consider how to estimate models which allow for spatial interactions. A second, more substantive, reason is that the spatial interactions themselves may be objects of interest.

Once we switch focus to the estimation of models including spatial variables, we face three fundamental challenges which are particularly important in the spatial setting: the so-called reflection problem, the presence of omitted variables that imply correlated effects (or common shocks), and problems caused by sorting.

In most settings using observational data, the reflection problem is very likely to occur unless we are able to impose further restrictions. We consider three possible solutions involving restrictions on the functional form, (exclusion) restrictions on the exogenous variables that directly determine outcomes, and restrictions on the nature of interactions. This last solution has been widely applied in the spatial econometrics literature through the use of *ad hoc* spatial weight matrices that assume interactions are incomplete, so have the property that $GG \neq G$. This strategy has been more recently applied in the social interaction literature, which exploits the architecture of network contacts to construct valid instrumental variables for the endogenous effect (i.e., by using the characteristics of indirect friends). However, in our view, these restrictions require careful justification on the basis of institutions, policy, or theory (or need to be imposed on the basis of data

that identify relevant linkages). These issues have received careful consideration in the networks and theoretical spatial econometrics literature, but much applied work continues to rely on *ad hoc* restrictions imposed through the choice of popular spatial weight matrices. Unfortunately, identification fails if these restrictions (whether carefully justified or imposed *ad hoc*) are invalid.

For some, especially those working within the experimentalist paradigm, the information requirements associated with these techniques are sufficiently profound that they may favor estimation of the reduced form with a specific focus on addressing problems created by sorting and omitted spatial variables. However, as we have shown, similar assumptions on the structure of G are implicit in the frequently applied empirical strategies—fixed effects or spatial differencing—used to address these problems. Our discussion above makes these assumptions explicit, which suggests that there may be an argument for greater use of the general spatial form in structuring applied microeconomic studies. Unfortunately, when the source of the omitted variables is due to endogenous sorting, it is very difficult to make progress without imposing further assumptions on the process that determines location. We show that these general lessons carry over to the policy evaluation literature, where the aim is to estimate the causal effect of some policy intervention. In particular, the requirements for a successful RCT to identify the separate causal parameters in the general spatial model are stringent. The difficulties inherent in designing the hypothetical experiment serve to emphasize the challenges for studies using observational data as well as pointing out the limits of RCTs in addressing these problems.

If there is one overarching message to emerge from this chapter, it is that while the use of spatial statistics and econometrics techniques to answer relevant questions in urban economics is certainly a promising avenue of research, the use of these techniques cannot be mechanical. As we discussed in this chapter, there are a variety of challenges and various possible solutions. Ultimately, the choice of the most appropriate model, identification, and estimation strategy depends on the mechanism underlying the presence of spatial effects and cannot be based only on statistical considerations.

APPENDIX A: BIASES WITH OMITTED SPATIAL VARIABLES

Even when estimation of spatial or social interactions is not the main goal, omission of salient spatial variables and variables capturing social interactions can obviously have important consequences for the estimates of other parameters. This is just a standard omitted variables problem. In the main text, we show that interactions between individuals may stem from the effects of (1) group-level individual characteristics, (2) group-level characteristics of other entities or objects, or (3) the outcomes for other individuals in the reference group. Omitting any of these sources of interaction leads to biases on the estimates of the effects of the other variables, although the importance of these biases in practice depends to some extent on the intended purpose of the estimation.

Suppose interactions really occur only through group-level characteristics—that is, contextual effects—so Equation (3.5) becomes (using matrix notation)

$$\gamma = X\gamma + G_x X\theta + \varepsilon.$$

Now suppose we try to estimate γ using a (misspecified) standard regression model in which individual outcomes depend only on own characteristics:

$$\gamma = X\gamma + \varepsilon. \quad (\text{A.1})$$

There is now a standard omitted variables bias due to omission of $G_x X\theta$, given that $G_x X$ is correlated with X by construction. The bias in the OLS estimate of γ is increasing in the importance of neighbors' or peers' characteristics in determining individual outcomes, θ :

$$\hat{\gamma}_{\text{OLS}} = \gamma + (X'X)^{-1} X' G_x X\theta. \quad (\text{A.2})$$

An analogous argument holds for omission of external attributes of the group $G_z Z$, when the correct specification is

$$\gamma = X\gamma + G_z Z\delta + \varepsilon,$$

although clearly the magnitude of the bias will depend on the extent to which $G_z Z$ and X are correlated.

Suppose instead that interactions genuinely occur as a result of individuals' responses to other individuals' outcomes—that is, endogenous effects—so Equation (3.5) becomes

$$\gamma = X\gamma + G_y \gamma \beta + \varepsilon.$$

If we mistakenly estimate γ using Equation (A.1), the OLS estimator is

$$\begin{aligned} \hat{\gamma}_{\text{OLS}} &= \gamma + (X'X)^{-1} X' G_y \gamma \beta \\ &= \gamma + (X'X)^{-1} X' G_y X \gamma \beta + (X'X)^{-1} X' G_y^2 \gamma \beta^2 \\ &= \gamma + (X'X)^{-1} X' G_y X \gamma \beta + (X'X)^{-1} X' G_y^2 X \gamma \beta^2 \\ &\quad + (X'X)^{-1} X' G_y^3 X \gamma \beta^3 + \dots \end{aligned} \quad (\text{A.3})$$

by repeated substitution, implying an infinite polynomial series of bias terms. OLS will be biased if $\beta > 0$. The bias goes to infinity when β approaches 1 (where the estimator is not defined) and it goes to 0 as β goes to 0. The intuitive reason for this bias is simply that the effect of X operating through γ is amplified through feedback between neighbors or peers, with the effect of X on one individual having an effect on its neighbor, and vice versa. In the case where G_y is a simple symmetric block diagonal, mean-creating matrix such as Equation (3.7), this bias expression simplifies to

$$\hat{\gamma}_{\text{OLS}} = \gamma + (X'X)^{-1} X' G_y X \gamma \beta / (1 - \beta). \quad (\text{A.4})$$

Finally, let us consider the case where interactions occur in terms of both group-level characteristics and outcomes—that is, the real relationship is

$$\gamma = X\gamma + G_y\gamma\beta + G_xX\theta + \varepsilon.$$

If we estimate γ using model (A.1)—that is, omitting both endogenous effects, $G_y\gamma$, and contextual effects, G_xX —the OLS estimator is

$$\begin{aligned}\hat{\gamma}_{OLS} &= \gamma + (X'X)^{-1}X'G_xX\theta + (X'X)^{-1}X'G_y\gamma\beta \\ &= \gamma + (X'X)^{-1}X'G_xX\theta + (X'X)^{-1}X'G_yX\gamma\beta \\ &\quad + (X'X)^{-1}X'G_yG_xX\theta\beta + (X'X)^{-1}X'G_y^2\gamma\beta^2 \\ &= \gamma + (X'X)^{-1}X'G_xX\theta + (X'X)^{-1}X'G_yX\gamma\beta \\ &\quad + (X'X)^{-1}X'G_yG_xX\theta\beta + (X'X)^{-1}X'G_y^2X\gamma\beta^2 \\ &\quad + (X'X)^{-1}X'G_y^2G_xX\theta\beta^2 + \dots,\end{aligned}\tag{A.5}$$

and again if $G_y = G_x = G$ is a simple block diagonal mean-creating idempotent matrix, this simplifies to

$$\hat{\gamma}_{OLS} = \gamma + (X'X)^{-1}X'GX(\gamma\beta + \theta)/(1 - \beta).\tag{A.6}$$

If we disregard the pathological case where $\beta\gamma = -\delta$, OLS will be biased, with the bias depending on both β and θ . The bias goes to infinity when β goes to 1 or θ goes to infinity and it goes to 0 if both β and θ go to 0. Again the bias is intuitive and includes effects due to omitted contextual interactions working through θ and the individual impacts γ , both amplified by the feedback effect between neighbors β .

Of course, for a policy maker interested in the effect of some treatment X , this “biased” parameter is exactly what that policy maker is interested in: the reduced form effect of the policy, taking into account the amplifying effects of the spatial interactions between agents—both in the sense that individuals are affected by their own treatment γ and the treatment of their neighbors δ , and because there is feedback via the outcomes that the treatments induced (the multiplicative factor $1/(1 - \beta)$). Whether this estimate should be considered the “causal” effect of treatment depends on the definition of causality as discussed in the main text, although in the usual interpretation in the program effects literature this biased parameter is indeed a causal parameter. Regardless, this reduced form interpretation of the OLS coefficient is the fundamental reason why researchers interested in policy treatment effects may care more about other threats to identification than about carefully delineating the various types of spatial or social interaction. We discussed these issues further in [Section 3.5](#).

In some situations, where researchers are interested in trying to understand the structure of spatial and social interactions out of curiosity, rather than for any instrumental policy purpose, this reduced form interpretation is not very helpful. A researcher may be interested specifically in the identification of the structural parameter γ , or the interaction terms θ and β may be of substantive interest. If simply disregarding the interaction effects is not an attractive option, the researcher needs to adopt methods for estimation

which allow for the inclusion of these interactions, although as we have shown in [Section 3.4](#), identification of these parameters is not easy.

Omitting spatial variables can also lead to a lot of confusion, because it gives rise to the problem usually called the modifiable areal unit problem (see, e.g., [Openshaw, 1983](#); [Wong, 2009](#); [Briant et al., 2010](#)). This refers to the empirical observation that estimates of parameters can change substantially as the researcher changes the level of spatial aggregation of the data on which the analysis is conducted (moving, for example, from individual microdata, to districts to regions, or even abstract regular geometric aggregations as shown in [Briant et al., 2010](#)). The reasons for this problem in regression applications are clear from the above discussion, in that changing the level of aggregation changes the relative weights of the individual effects γ and the effects arising from spatial interactions (or other spatial variables). For example, suppose the underlying relationship at the individual level is

$$y = X\gamma + G_x X\theta + \varepsilon$$

as in the first example above, and we estimate a regression of y on X using individual data, omitting the spatial variable $G_x X$. Then as shown above, the OLS estimate is $\hat{\gamma}_{OLS} = \gamma + (X'X)^{-1}X'G_x X\theta$. This is a weighted average of γ and θ which depends on the sample covariance between $G_x X$ and X and the sample variance of X . As we perform aggregation up from the individual level to higher geographical levels of aggregation, the weight on θ increases, until, if we perform estimation at the level of aggregation defined by G_x —that is, we estimate $G_x y = G_x X\gamma + G_x X\theta + \varepsilon$ —we obtain $\hat{\gamma}_{OLS} = \gamma + \theta$. Similar issues arise if the omitted variable is not $G_x X$, but is any other spatial variable that is correlated with X .

APPENDIX B: HYPOTHETICAL RCT EXPERIMENTS FOR IDENTIFYING PARAMETERS IN THE PRESENCE OF INTERACTIONS WITHIN SPATIAL CLUSTERS

In [Section 3.5](#) we noted that standard clustered RCT designs can identify only a composite parameter characterizing a combination of the direct effects of an intervention plus the social multiplier effects from contextual and endogenous interactions between treated individuals in spatial clusters. However, we noted that experiments could potentially be designed to recover some or all of these parameters. Here, we provide some simple examples, which we hope further elucidate the more general problems of identifying the parameters in models with spatial and social interaction.

The standard clustered RCT experiment described around Equation (3.26) allowed us to estimate the overall effect of a policy intervention x^* in the presence of interactions within the randomly treated spatial clusters: $E[y|1] - E[y|0] = x^*(\gamma + \theta)/(1 - \beta)$.

Suppose now, rather than randomly treating some clusters (treatment) and not others (control), we have control over the share of individuals who are randomly treated within each cluster. We use s to denote the share of individuals who are treated within a cluster, such that for those individuals $E[x|1] - E[x|0] = x^*$, but for the cluster we have $E[x|s] = x^*s$.

From this experiment we could estimate the means of the outcomes for the treated individuals in each cluster, the nontreated individuals in each cluster, and the mean outcome in each cluster, which would vary with the share s treated.²⁸ Mean outcome in cluster is:

$$\begin{aligned} E[y|s] &= \beta E[y|s] + x^*s(\gamma + \theta) \\ &= x^*s(\gamma + \theta)/(1 - \beta). \end{aligned} \quad (\text{B.1})$$

Individual treated directly in cluster with share s treated

$$\begin{aligned} E[y|1, s] &= \beta E[y|s] + x^*(\gamma + s\theta) \\ &= x^*s[\beta(\gamma + \theta)/(1 - \beta) + \theta] + \gamma x^*. \end{aligned} \quad (\text{B.2})$$

Individual not treated directly, in cluster with share s treated

$$\begin{aligned} E[y|0, s] &= \beta E[y|s] + x^*s\theta \\ &= x^*s[\beta(\gamma + \theta)/(1 - \beta) + \theta]. \end{aligned} \quad (\text{B.3})$$

And subtracting the mean for those not treated from the mean of those treated recovers the direct effect of the treatment:

$$E[y|1, s] - E[y|0, s] = x^*\gamma. \quad (\text{B.4})$$

Hence, with two or more clusters available, with different shares treated, we can identify γ and a composite parameter representing the strength of social interactions $\beta(\gamma + \theta)/(1 - \beta) + \theta$. However, this still does not provide a solution to the reflection problem and allow the separate estimation of θ and $(1 - \beta)$.²⁹

Attempting to separately identify the endogenous interactions β is more complex, and requires that the experimental structure mimics the intransitive network grouping structure discussed as a prerequisite for identification in [Section 3.4](#). The idea is to create some groups of individuals who are treated directly, some groups of individuals who are treated indirectly through interaction with the individuals treated directly (endogenous and contextual effects), and some individuals who are treated only indirectly through interaction with others who are treated only indirectly (endogenous effects).

We create four groups of individuals (groups 0, 1, 2, and 3), in which group 0 is a control group. Individuals are randomly assigned to equal-size groups 1, 2, and 3 in triads

²⁸ Here we are assuming the standard linear in means expression for individual outcomes as in (3.6).

²⁹ We could also use group assignment to identify γ and $\theta/(1 - \beta)$ by completely isolating some agents. For isolated agents, the difference in expected outcomes between treated and untreated individuals is $E[y|1] - E[y|0] = (E[x|1] - E[x|0])\gamma = x^*\gamma$, which provides estimates of the direct effect γ .

in which an individual in group 1 interacts with an individual in group 2 and this individual in group 2 also interacts with an individual in group 3, but the individual in group 1 does not interact with an individual in group 3. Also, for simplicity of notation, we assume that individuals in a given group cannot interact with other individuals in that group. Again, we set aside practical considerations about how this system of interactions might be enforced. Agents are randomized across all three groups, so $E[y|j] - E[y|k] = E[x|j] - E[x|k] = E[u|j] - E[u|k] = 0$ for all j and k . Group 1 is subject to an intervention x^*

For a simple example of only two agents in each group, the structure of the G matrix is, by design,

$$G = \begin{bmatrix} & a & b & c & d & e & f & g & h \\ a & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ d & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ e & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 \\ f & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 \\ g & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ h & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix},$$

where a and b belong to group 0, c and d belong to group 1, e and f belong to group 2, and g and h belong to group 3. Clearly $GG \neq G$, so we could simply apply the results from [Section 3.4](#). Once again, however, we think it is instructive to work through this specific example within the case-control RCT paradigm to further develop understanding of how identification is achieved and what this tells us about how difficult this might be in nonexperimental settings.

Following the standard structure of linear interactions and using the notation $DE[x_i|j] = E[x_i|j] - E[x_i|0]$ and so on (i.e., differences from control group means), we find the expressions for individuals in each group are as follows:

$$E[y|0] = E[x|0]\gamma + E[u|0], \quad (\text{B.5})$$

$$E[y|1] = E[y|2]\beta + E[x|1]\gamma + E[u|1], \quad (\text{B.6})$$

$$E[y|2] = (E[y|1] + E[y|3])\beta/2 + (E[x|1] + E[x|3])\theta/2 + E[x|2]\gamma + E[u|2], \quad (\text{B.7})$$

$$E[y|3] = E[y|2]\beta + E[x|2]\theta + E[x|3]\gamma + E[u|3]. \quad (\text{B.8})$$

With randomization and intervention in group 1,

$$DE[y|1] = DE[y|2]\beta + x^*\gamma, \quad (\text{B.9})$$

$$DE[y|2] = (DE[y|1] + DE[y|3])\beta/2 + x^*\theta/2, \quad (\text{B.10})$$

$$DE[y|3] = DE[y|2]\beta. \quad (\text{B.11})$$

We get the reduced form for $DE[y|2]$ by substituting $DE[y|1]$ and $DE[y|3]$ in Equation (B.10):

$$\begin{aligned} DE[y|2] &= DE[y|2]\beta^2 + x^*(\gamma\beta + \theta)/2 \\ &= x(\gamma\beta + \theta)/2(1 - \beta^2) \\ &= x^*\pi, \end{aligned} \tag{B.12}$$

where π is the composite parameter $(\gamma\beta/2 + \theta)/2(1 - \beta^2)$

Since $DE[y|3] = x^*\pi\beta$ and $DE[y_i|2] = x^*\pi\beta = DE[y|3]/DE[y|2]$. In other words, an estimate of the endogenous interaction coefficient β could be obtained from this experiment by taking the difference between means outcomes of group 3 and group 0, and dividing by the difference in means between group 2 and group 0. This is equivalent to an instrumental variables estimate, using the intervention x^* as an instrument for $DE[y|2]$ in the regression of $DE[y|3]$ on $DE[y|2]$ (with obvious parallels to the way identification is achieved in the network literature as described in Section 3.4).

REFERENCES

- Aaronson, D., 1998. Using sibling data to estimate the impact of neighborhoods on children's educational outcomes. *J. Hum. Resour.* 33 (4), 915–946.
- Abbasi, A., Altmann, J., Hossain, L., 2011. Identifying the effects of co-authorship networks on the performance of scholars: a correlation and regression analysis of performance measures and social network analysis measures. *J. Informetr.* 5 (4), 594–607.
- Angrist, J., Krueger, A., 1999. Empirical strategies in labor economics. In: Ashenfelter, A., Card, D. (Eds.), *Handbook of Labor Economics* 3A. North-Holland, Amsterdam.
- Angrist, J., Pischke, J.S., 2009. *Mostly harmless econometrics*. Princeton University Press, Princeton.
- Angrist, J., Pischke, J.S., 2011. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J. Econ. Perspect.* 24, 3–30.
- Anselin, L., 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.
- Anselin, L., 1995. Local indicators of spatial association. *Geogr. Anal.* 27 (2), 93–115.
- Banerjee, A., Besley, T., 1991. *Peer Group Externalities and Learning Incentives: A Theory of Nerd Behavior*. Princeton University, Mimeo.
- Barrios, T., Diamond, R., Imbens, G.W., Kolesar, M., 2012. Clustering, spatial correlations, and randomization inference. *J. Am. Stat. Assoc.* 107 (498), 578–591.
- Benabou, R., 1993. Workings of a city: location, education, and production quarterly. *J. Econ.* 108, 619–652.
- Black, S.E., 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 577–599.
- Borjas, G., Doran, K., 2012. The collapse of the Soviet Union and the productivity of American mathematicians. *Q. J. Econ.* 127 (3), 1143–1203.
- Bound, J., Jaeger, D., Baker, R., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *J. Am. Stat. Assoc.* 90 (430), 443–450.
- Bramoullé, Y., Djebbari, H., Fortin, B., 2009. Identification of peer effects through social networks. *J. Econom.* 150, 41–55.
- Briant, A., Combes, P.P., Lafourcade, M., 2010. Dots to boxes: do the size and shape of spatial units jeopardize economic geography estimations? *J. Urban Econ.* 67 (3), 287–302.
- Brock, W.A., Durlauf, S.N., 2001. Interactions-based models. In: Heckman, J.J., Leamer, E.E. (Eds.), *Handbook of Econometrics*, first ed., vol. 5. Elsevier, pp. 3297–3380 (Chapter 54).

- Calvó-Armengol, A., Patacchini, E., Zenou, Y., 2009. Peer effects and social networks in education. *Rev. Econ. Stud.* 76, 1239–1267.
- Cameron, A.C., Miller, D.L., 2015. A practitioner's guide to cluster-robust inference. *J. Hum. Resour.* forthcoming.
- Campbell, M.K., Elbourne, D.R., Altman, D.G., 2004. CONSORT statement: extension to cluster randomised trials. *BMJ* 328, 702.
- Case, A., Katz, L., 1991. The company you keep: the effects of family and neighborhood on disadvantaged youths. National Bureau of Economic Research, Inc, NBER Working papers 3705.
- Ciccone, A., Peri, G., 2006. Identifying human-capital externalities: theory with applications. *Rev. Econ. Stud.* 73 (2), 381–412, Oxford University Press.
- Cohen-Cole, E., Kirilenko, A., Patacchini, E., 2014. Trading networks and liquidity provision. *J. Financ. Econ.* 113 (2), 235–251.
- Combes, P.P., Overman, H.G., 2004. The spatial distribution of economic activities in the European Union. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics. Cities and Geography*, vol. 4. Elsevier, Amsterdam.
- Combes, P.P., Duranton, G., Gobillon, L., 2008. Spatial wage disparities: sorting matters!. *J. Urban Econ.* 63 (2), 723–742.
- Conley, T.G., 1999. GMM estimation with cross sectional dependence. *J. Econom.* 92 (1), 1–45, Elsevier.
- Conley, T.G., Molinari, F., 2007. Spatial correlation robust inference with errors in location or distance. *J. Econom.* 140, 76–96.
- Cressie, N.A.C., 1993. *Statistics for Spatial Data*. John Wiley, New York.
- Cutler, D.M., Glaeser, E.L., Vigdor, J.L., 1999. The rise and decline of the American Ghetto. *J. Polit. Econ.* 107 (3), 455–506.
- Dahl, G.B., 2002. Mobility and the returns to education: testing a Roy model with multiple markets. *Econometrica* 70, 2367–2420.
- De Giorgi, G., Pellizzari, M., Redaelli, S., 2010. Identification of social interactions through partially overlapping peer groups. *Am. Econ. J. Appl. Econ.* 2 (2), 241–275.
- De la Roca, J., Puga, D., 2014. Learning by working in big cities. CEMFI.
- Del Bello, C., Patacchini, E., Zenou, Y., 2014. Peer effects: social or geographical distance? Working paper.
- Di Addario, S., Patacchini, E., 2008. Wages and the city. Evidence from Italy. *Labour Econ.* 15 (5), 1040–1061.
- Diggle, P.J., 2003. *Statistical Analysis of Spatial Point Patterns*. Oxford University Press, New York.
- Duranton, G., Overman, H.G., 2005. Testing for localisation using micro geographic data. *Rev. Econ. Stud.* 72, 1077–1106.
- Duranton, G., Gobillon, L., Overman, H.G., 2011. Assessing the effects of local taxation using microgeographic data. *Econ. J.* 121, 1017–1046.
- Eerola, E., Lyytikäinen, T., 2012. On the role of public price information in housing markets. Government Institute for Economic Research, VATT Working papers 30/2012.
- Einio, E., Overman, H.G., 2014. The effects of spatially targeted enterprise initiatives: evidence from UK LEGI. LSE.
- Ellison, G., Glaeser, E.L., 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. *J. Polit. Econ.* 105 (5), 889–927, University of Chicago Press.
- Ellison, G., Glaeser, E.L., Kerr, W., 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. *Am. Econ. Rev.* 100, 1195–1213.
- Epple, D., Romano, R.E., 2011. Peer effects in education: a survey of the theory and evidence. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), *Handbook of Social Economics*, vol. 1B. Elsevier, Amsterdam (Chapter 20).
- Felkner, J.S., Townsend, R.M., 2011. The geographic concentration of enterprise in developing countries. *Q. J. Econ.* 126 (4), 2005–2061.
- Fryer, R., Torelli, P., 2010. An empirical analysis of 'Acting White'. *J. Public Econ.* 94 (5–6), 380–396.
- Gaviria, A., Raphael, S., 2001. School-based peer effects and juvenile behavior. *Rev. Econ. Stat.* 83 (2), 257–268, MIT Press.
- Getis, A., Ord, J.K., 1992. The analysis of spatial association by use of distance statistics. *Geogr. Anal.* 24, 189–206.
- Gibbons, S., 2004. The costs of urban property crime. *Econ. J.* 114 (498), F441–F463.

- Gibbons, S., Machin, S., 2003. Valuing English primary schools. *J. Urban Econ.* 53 (2), 197–219.
- Gibbons, S., Overman, H.G., 2012. Mostly pointless spatial econometrics. *J. Reg. Sci.* 52 (2), 172–191.
- Gibbons, S., Silva, O., Weinhardt, F., 2013. Everybody needs good neighbours? Evidence from students' outcomes in England. *Econ. J.* 123 (571), 831–874.
- Gibbons, S., Overman, H.G., Pelkonen, P., 2014. Area disparities in Britain: understanding the contribution of people versus place through variance decompositions. *Oxf. Bull. Econ. Stat.* 76 (5), 745–763.
- Goldsmith-Pinkham, P., Imbens, G.W., 2013. Social networks and the identification of peer effects. *J. Bus. Econ. Stat.* 31, 253–264.
- Goux, D., Maurin, E., 2007. Close neighbours matter: neighbourhood effects on early performance at school. *Econ. J.* 117 (523), 1193–1215, Royal Economic Society.
- Graham, D.J., 2007. Agglomeration, productivity and transport investment. *J. Transp. Econ. Policy* 41 (3), 317–343.
- Harhoff, D., Hiebel, M., Hoisl, K., 2013. The impact of network structure and network behavior on inventor productivity. Munich Center for Innovation and Entrepreneurship Research (MCIER). Max Planck Institute.
- Heckman, J., 2005. The scientific model of causality. *Sociol. Method.* 35 (1), 1–97.
- Heckman, J., Lalonde, R., Smith, J., 1999. The economics and econometrics of active labour market programs. In: Ashenfelter, A., Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3A, North-Holland, Amsterdam.
- Helmerts, C., Patnam, M., 2014. Does the rotten child spoil his companion? Spatial peer effects among children in rural India. *Quant. Econ.* 5 (1), 67–121.
- Herfindahl, O.C., 1959. *Copper Costs and Prices: 1870–1957*. The John Hopkins Press, Baltimore, MD.
- Hirschman, A.O., 1964. The paternity of an index. *Am. Econ. Rev.* 54 (5), 761.
- Holmes, T., 1998. The effect of state policies on the location of manufacturing: evidence from state borders. *J. Polit. Econ.* 106, 667–705.
- Holmes, T.J., Lee, S., 2012. Economies of density versus natural advantage: crop choice on the back forty. *Rev. Econ. Stat.* 94 (1), 1–19, MIT Press.
- Horrace, C.W., Liu, X., Patacchini, E., 2013. Endogenous network production function with selectivity. Syracuse University, **Working paper**.
- Hsieh, C.S., Lee, L.F., 2013. A social interaction model with endogenous friendship formation and selectivity. Ohio State University, **Working paper**.
- Ioannides, Y., 2013. *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton University Press, Amsterdam.
- Ioannides, Y., Zabel, J., 2008. Interactions, neighbourhood selection and housing demand. *J. Urban Econ.* 63, 229–252.
- Jaffe, A., 1989. Real effects of academic research. *Am. Econ. Rev.* 79 (5), 957–970.
- Kelejian, H.H., Prucha, I.R., 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbance. *J. Real Estate Financ. Econ.* 17, 99–121.
- Kelejian, H.H., Prucha, I.R., 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. *Int. Econ. Rev.* 40, 509–533.
- Kelejian, H.H., Prucha, I.R., 2004. Estimation of simultaneous systems of spatially interrelated cross sectional equations. *J. Econom.* 118, 27–50.
- Kelejian, H., Prucha, I.R., 2007. HAC estimation in a spatial framework. *J. Econom.* 140, 131–154.
- Kelejian, H.H., Prucha, I.R., 2010. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *J. Econom.* 157, 53–67.
- Kiel, K., Zabel, J., 2008. Location, location, location: the 3L approach to house price determination. *J. Hous. Econ.* 17, 175–190.
- Klier, T., McMillen, D.P., 2008. Evolving agglomeration in the U.S. auto supplier industry. *J. Reg. Sci.* 48 (1), 245–267.
- Kosfeld, R., Eckey, H.-F., Lauridsen, J., 2011. Spatial point pattern analysis and industry concentration. *Ann. Reg. Sci.* 47, 311–328.
- Krauth, B., 2005. Peer effects and selection effects on smoking among Canadian youth. *Can. J. Econ.* 38 (3), 414–433.

- Krugman, P., 1991a. *Geography and Trade*. MIT Press, Cambridge, MA.
- Krugman, P., 1991b. Increasing returns and economic geography. *J. Polit. Econ.* 99 (3), 483–499.
- Kuminoff, N., Kerry Smith, V., Timmins, C., 2013. The new economics of equilibrium sorting and policy evaluation using housing markets. *J. Econ. Lit.* 51 (4), 1007–1062.
- Lee, L.-F., 1983. Generalized econometric models with selectivity. *Econometrica* 51, 507–512.
- Lee, L.-F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial econometric models. *Econometrica* 72, 1899–1926.
- Lee, M.-J., 2005. *Micro-Econometrics for Policy, Program and Treatment Effects*. Oxford University Press, Oxford.
- Lee, L.-F., 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J. Econom.* 140, 333–374.
- Lee, L.-F., Liu, X., 2010. Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances. *Econ. Theory* 26, 187–230.
- Lee, L.-F., Liu, X., Lin, X., 2010. Specification and estimation of social interaction models with network structures. *Econom. J.* 13, 145–176.
- Li, J., Lee, L., 2009. Binary choice under social interactions: an empirical study with and without subjective data on expectations. *J. Appl. Econ.* 24, 257–281.
- Lin, X., 2010. Identifying peer effects in student academic achievement by a spatial autoregressive model with group unobservables. *J. Urban Econ.* 28, 825–860.
- Liu, X., Lee, L.-F., 2010. GMM estimation of social interaction models with centrality. *J. Econom.* 159, 99–115.
- Liu, X., Patacchini, E., Zenou, Y., Lee, L.-F., 2012. Criminal networks: who is the key player? CEPR Discussion Paper No. 8772.
- Liu, X., Patacchini, E., Rainone, E., 2013. The allocation of time in sleep: a social network model with sampled data. CEPR Discussion Paper No. 9752.
- Liu, X., Patacchini, E., Zenou, Y., 2014. Endogenous peer effects: local aggregate or local average? *J. Econ. Behav. Organ.* 103, 39–59.
- Manski, C.F., 1993. Identification of endogenous effects: the reflection problem. *Rev. Econ. Stud.* 60, 531–542, 84, 600–616.
- Manski, C.F., 2000. Economic analysis of social interactions. *J. Econ. Perspect.* 14 (3), 115–136.
- Manski, C.F., 2013. Identification of treatment response with social interactions. *Econom. J.* 16 (1), S1–S23.
- Marcon, E., Puech, F., 2003. Evaluating the geographic concentration of industries using distance-based methods. *J. Econ. Geogr.* 4 (3), 409–428.
- Massey, D.S., Denton, N.A., 1987. Trends in the residential segregation of Blacks, Hispanics, and Asians: 1970–1980. *Am. Sociol. Rev.* 94, 802–825.
- Mayer, T., Mayneris, F., Py, L., 2012. The impact of urban enterprise zones on establishments location decisions: evidence from French ZFUs. PSE.
- Mele, A., 2013. Approximate variational inference for a model of social interactions. Working papers 13–16, NET Institute.
- Melo, P.C., Graham, D.J., Noland, R.B., 2009. A meta-analysis of estimates of urban agglomeration economies. *Reg. Sci. Urban Econ.* 39, 332–342.
- Mion, G., Naticchioni, P., 2009. The spatial sorting and matching of skills and firms. *Can. J. Econ.* 42, 28–55 [Revue canadienne d'économique].
- Moran, P.A.P., 1950. Notes on continuous stochastic phenomena. *Biometrika* 37 (1), 17–23.
- Moretti, E., 2004. Human capital externalities in cities. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics. Cities and Geography*, vol. 4. Elsevier, Amsterdam.
- Nakajima, R., 2007. Measuring peer effects on youth smoking behaviour. *Rev. Econ. Stud.* 74, 897–935.
- Openshaw, S., 1983. *The Modifiable Areal Unit Problem*. Geo Books, Norwich.
- Patacchini, E., Rainone, E., 2014. The word on banking—social ties, trust, and the adoption of financial products. EIEF Discussion Paper No. 1404.
- Patacchini, E., Venanzoni, G., 2014. Peer effects in the demand for housing quality. *J. Urban Econ.* 83, 6–17.
- Patacchini, E., Zenou, Y., 2007. Spatial dependence in local unemployment rates. *J. Econ. Geogr.* 7, 169–191.

- Patacchini, E., Zenou, Y., 2012. Neighborhood effects and parental involvement in the intergenerational transmission of education. *J. Reg. Sci.* 51 (5), 987–1013.
- Ripley, B.D., 1976. The second-order analysis of stationary point processes. *J. Appl. Probab.* 13, 255–266.
- Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6 (1), 34–58.
- Sacerdote, B., 2001. Peer effects with random assignment: results for Dartmouth roommates. *Q. J. Econ.* 116, 681–704.
- Scholl, T., Brenner, T., 2012. Detecting spatial clustering using a firm-level cluster index. *Working papers on Innovation and Space* 02.12: 1–29.
- Scholl, T., Brenner, T., 2013. Optimizing distance-based methods for big data analysis. Philipps-Universität Marburg, *Working papers on Innovation and Space*.
- Simons-Morton, B., Farhat, T., 2010. Recent findings on peer group influences on adolescent smoking. *J. Prim. Prev.* 31 (4), 191–208.
- Sirakaya, S., 2006. Recidivism and social interactions. *J. Am. Stat. Assoc.* 101 (475), 863–875.
- Soetevant, A., Kooreman, P., 2007. A discrete choice model with social interactions: with an application to high school teen behaviour. *J. Appl. Econ.* 22, 599–624.
- Stock, J., Wright, J., Yogo, M., 2002. A survey of weak instruments and weak identification in generalized method of moments. *J. Bus. Econ. Stat.* 20 (4), 518–529.
- Vitali, S., Mauro, N., Fagiolo, G., 2009. Spatial localization in manufacturing: a cross-country analysis. *LEM Working paper Series* 4, 1–37.
- Weinberg, R., 2007. Social interactions with endogenous associations. NBER Working paper No. 13038.
- Wong, D., 2009. The modifiable areal unit problem (MAUP). In: Fotheringham, A.S., Rogerson, P. (Eds.), *The SAGE Handbook of Spatial Analysis*. Sage Publications Ltd, London, pp. 105–124.
- Zenou, Y., 2009. *Urban Labour Markets*. Cambridge University Press, Cambridge.

SECTION II

Agglomeration and Urban Spatial Structure

This page intentionally left blank

CHAPTER 4

Agglomeration Theory with Heterogeneous Agents

Kristian Behrens^{*,†,‡,§}, Frédéric Robert-Nicoud^{§,¶,||}

^{*}Department of Economics, Université du Québec à Montréal, Montréal, QC, Canada

[†]National Research University, Higher School of Economics, Moscow, Russia

[‡]CIRPÉE, Université du Québec à Montréal, Montréal, QC, Canada

[§]CEPR, London, UK

[¶]Geneva School of Economics and Management, Université de Genève, Genève, Switzerland

^{||}SERC, The London School of Economics and Political Science, London, UK

Contents

4.1. Introduction	172
4.2. Four Causes and Two Moments: A Glimpse at the Data	175
4.2.1 Locational fundamentals	175
4.2.2 Agglomeration economies	176
4.2.3 Sorting of heterogeneous agents	178
4.2.4 Selection effects	181
4.2.5 Inequality and city size	184
4.2.6 City size distribution	184
4.2.7 Assembling the pieces	184
4.3. Agglomeration	187
4.3.1 Main ingredients	187
4.3.2 Canonical model	188
4.3.2.1 <i>Equilibrium, optimum, and maximum city sizes</i>	188
4.3.2.2 <i>Size distribution of cities</i>	193
4.3.2.3 <i>Inside the “black boxes”: extensions and interpretations</i>	197
4.3.3 The composition of cities: industries, functions, and skills	201
4.3.3.1 <i>Industry composition</i>	202
4.3.3.2 <i>Functional composition</i>	206
4.3.3.3 <i>Skill composition</i>	210
4.4. Sorting and Selection	211
4.4.1 Sorting	212
4.4.1.1 <i>A simple model</i>	212
4.4.1.2 <i>Spatial equilibrium with a discrete set of cities</i>	213
4.4.1.3 <i>Spatial equilibrium with a continuum of cities</i>	217
4.4.1.4 <i>Implications for city sizes</i>	219
4.4.1.5 <i>Some limitations and extensions</i>	220
4.4.1.6 <i>Sorting when distributions matter (a prelude to selection)</i>	222
4.4.2 Selection	226
4.4.2.1 <i>A simple model</i>	227
4.4.2.2 <i>CES illustration</i>	229

4.4.2.3	<i>Beyond the CES</i>	230
4.4.2.4	<i>Selection and sorting</i>	231
4.4.2.5	<i>Empirical implications and results</i>	232
4.5.	Inequality	234
4.5.1	Sorting and urban inequality	235
4.5.2	Agglomeration and urban inequality	236
4.5.3	Selection and urban inequality	237
4.6.	Conclusions	239
	Acknowledgments	240
	References	241

Abstract

This chapter surveys recent developments in agglomeration theory within a unifying framework. We highlight how locational fundamentals, agglomeration economies, the spatial sorting of heterogeneous agents, and selection effects affect the size, productivity, composition, and inequality of cities, as well as their size distribution in the urban system.

Keywords

Agglomeration, Heterogeneous agents, Selection, Sorting, Inequality, City size distribution

JEL Classification Codes

R12, D31

4.1. INTRODUCTION

Cities differ in many ways. A myriad of small towns coexist with medium-sized cities and a few urban giants. Some cities have a diversified economic base, whereas others are specialized by industry or by the functions they perform. A few large cities attract the brightest minds, while many small ones can barely retain their residents. Most importantly, however, cities differ in productivity: large cities produce more output per capita than small cities do. This *urban productivity premium* may occur because of locational fundamentals, because of agglomeration economies, because more talented individuals sort into large cities, or because large cities select the most productive entrepreneurs and firms. The literature from [Marshall \(1890\)](#) on has devoted most of its attention to agglomeration economies, whereby a high density of firms and workers generates positive externalities to other firms and workers. It has done so almost exclusively within a representative agent framework. That framework has proved extremely useful for analyzing many different microeconomic foundations for the urban productivity premium. It is, however, ill-suited to study empirically relevant patterns such as the over representation of highly

educated workers and highly productive firms in large cities. It has also, by definition, very little to say on distributional outcomes in cities.

Individual-level and firm-level data have revealed that the broad macro relationships among urban aggregates reflect substantial heterogeneity at the micro level. Theorists have started to build models to address these issues and to provide microeconomic foundations explaining this heterogeneity in a systematic manner. This chapter provides a *unifying framework of urban systems* to study recent developments in agglomeration theory. To this end, we extend the canonical model developed by [Henderson \(1974\)](#) along several dimensions, in particular to heterogeneous agents.¹ Doing so allows us to analyze urban macro outcomes in the light of microheterogeneity, and to better understand the patterns substantiated by the data. We also show how this framework can be used to study under-researched issues and how it allows us to uncover some caveats applying to extant theoretical work. One such caveat is that sorting and selection are intrinsically linked, and that assumptions which seem reasonable in partial equilibrium are inconsistent with the general equilibrium logic of an urban systems model.

This chapter is organized as follows. [Section 4.2](#) uses a cross section of US cities to document the following set of stylized facts that we aim to make sense of within our framework:

- Fact 1 (size and fundamentals): the population size and density of a city are positively correlated with the quality of its fundamentals.
- Fact 2 (urban premiums): the unconditional elasticity of mean earnings and city size is about 8%, and the unconditional elasticity of median housing rents and city size is about 9%.
- Fact 3 (sorting): the share of workers with at least a college degree increases with city size.
- Fact 4 (selection): the share of self-employed is negatively correlated with urban density and with net entry rates of new firms, so selection effects may be at work.
- Fact 5 (inequality): the Gini coefficient of urban earnings is positively correlated with city size and the urban productivity premium increases with the education level.
- Fact 6 (Zipf's law): the size distribution of US places follows closely a log-normal distribution and that of US metropolitan statistical areas (MSAs) follows closely a power law (aka Zipf's law).

The rest of this chapter is devoted to theory. [Section 4.3](#) sets the stage by introducing the canonical model of urban systems with homogeneous agents. We extend it to allow for

¹ Worker and firm heterogeneity has also sparked new theories in other fields. See, for example, the reviews by [Grossman \(2013\)](#) and [Melitz and Redding \(2014\)](#) of international trade theories with heterogeneous workers and heterogeneous firms, respectively.

heterogeneous fundamentals across locations and show how the equilibrium patterns that emerge are consistent with facts 1 (size and fundamentals), 2 (urban premiums), and, under some assumptions, 6 (Zipf's law). We also show how cities differ in their industrial and functional specialization. [Section 4.4](#) introduces heterogeneous agents and shows how the model with sorting replicates facts 2 (urban premiums), 3 (sorting), and 6 (Zipf's law). The latter result is particularly striking since it arises in a static model and relies solely on the sorting of heterogeneous agents across cities. We also show under what conditions the model with heterogeneous agents allows for selection effects, as in fact 4 (selection), what their citywide implications are, and how they are linked to sorting. [Section 4.5](#) builds on the previous developments to establish fact 5 (inequality). We show how worker heterogeneity, sorting, and selection interact with agglomeration economies to deliver a positive equilibrium relationship between city size and urban inequality. This exercise also reveals that few general results are known, and much work remains to be done in this area.

Before proceeding, we stress that our framework is purely static. As such, it is ill-equipped to study important fluctuations in the fate of cities such as New York, which has gone through periods of stagnation and decline before emerging, or more recently Detroit and Pittsburgh. Housing stocks and urban infrastructure depreciate only slowly, so housing prices and housing rents swing much more than city populations do ([Henderson and Venables, 2009](#)). The chapter by [Desmet and Henderson \(2015\)](#) in this handbook provides a more systematic treatment of the dynamic aspects and evolution of urban systems.

We further stress that the content of this chapter reflects the difficult and idiosyncratic choices that we made in the process of writing it. We have opted to study a selective set of topics in depth rather than cast a wide but shallow net. We have, for instance, limited ourselves to urban models and largely omitted “regional science” and “new economic geography” contributions. Focusing on the macro aspects and on heterogeneity, we view this chapter as a natural complement to the chapter by [Duranton and Puga \(2004\)](#) on the microfoundations for urban agglomeration economies in volume 4 of this handbook series. Where [Duranton and Puga \(2004\)](#) take city sizes mostly as given to study the microeconomic mechanisms that give rise to agglomeration economies, we take the existence of these citywide increasing returns for granted. Instead, we consider the urban system and allow for worker and firm mobility across cities to study how agglomeration economies, urban costs, heterogeneous locational fundamentals, heterogeneous workers and firms, and selection effects interact to shape the size, composition, productivity, and inequality of cities. In that respect, we build upon and extend many aspects of urban systems that have been analyzed before without paying much attention to micro level heterogeneity (see [Abdel-Rahman and Anas, 2004](#) for a survey).

4.2. FOUR CAUSES AND TWO MOMENTS: A GLIMPSE AT THE DATA

To set the stage and organize our thoughts, we first highlight a number of key stylized facts.² We keep this section brief on purpose and paint only the big picture related to the four fundamental causes that affect the first two moments of the income, productivity, and size distributions of cities. We report more detailed results from empirical studies as we go along.

The *four fundamental causes* that we focus on to explain the sizes of cities, their composition, and the associated productivity gains are (a) locational fundamentals, (b) agglomeration economies, (c) the spatial sorting of heterogeneous agents, and (d) selection effects. These four causes influence—either individually or jointly—the spatial distribution of economic activity and the *first moments* of the productivity and wage distributions within and across cities. They also affect—especially jointly—the *second moments* of those distributions. The latter effect, which is important from a normative perspective, has received little attention until now.

4.2.1 Locational fundamentals

Locations are heterogeneous. They differ in endowments (natural resources, constructible area, soil quality, etc.), in accessibility (presence of infrastructures, access to navigable rivers and natural harbors, relative location in the urban system, etc.), and in many other first- and second-nature characteristics (climate, consumption and production amenities,

² *Data sources:* The “places” data come from the “Incorporated Places and Minor Civil Divisions Datasets: Subcounty Resident Population Estimates: April 1, 2010 to July 1, 2012” file from the US Census Bureau (SUB-EST2012.csv). It contains 81,631 places. For the big cities, we use 2010 Census and 2010 American Community Survey 5-year estimates (US Census Bureau) data for 363 continental US MSAs. The 2010 data on urban clusters come from the Census Gazetteer file (Gaz_ua_national.txt). We aggregate up urban clusters at the metropolitan and micropolitan statistical area level using the “2010 Urban Area to Metropolitan and Micropolitan Statistical Area (CBSA) Relationship File” (ua_cbsa_rel_10.txt). From the relationship file, we compute MSA density for the 363 continental MSAs (excluding Alaska, Hawaii, and Puerto Rico). We also compute “cluster density” at the MSA level by keeping only the urban areas within an MSA and by excluding MSA parts that are not classified as urban areas (variable ua = 99999). This yields two density measures per MSA: overall density, D , and cluster density, b . We further have the total MSA population and “cluster” population. We also compute an “urban cluster” density measure in the spirit of [Wheeler \(2004\)](#), where the cluster density of an MSA is given by the population-weighted average density of the individual urban clusters in the MSA. The “MSA geological features” variable is constructed using the same US Geological Survey data as in [Rosenthal and Strange \(2008b\)](#): seismic hazard, landslide hazard, and sedimentary bedrock. For illustrative purposes, we take the logarithm of the sum of the three measures. The data on firm births, firm deaths, and the number of small firms come from the County Business Patterns (files msa_totals_emplchange_2009–2010.xls and msa_naicssector_2010.xls) of the US Census Bureau. The data on natural amenities come from the US Department of Agriculture (file natamenf_1.xls). Lastly, the data on state-level venture capital come from the National Venture Capital Association (file RegionalAggregateData42010FINAL.xls).

geological and climatic hazards, etc.). We regroup all these factors under the common header of *locational fundamentals*. The distinctive characteristics of locational fundamentals are that they are exogenous to our static economic analysis and that they can either attract population and economic activity (positive fundamentals such as a mild climate) or repulse them (negative fundamentals such as exposure to natural hazards). The left panel in Figure 4.1 illustrates the statistical relationship between a particular type of (positive) amenities and the size of US MSAs. The MSA amenity score—constructed by the US Department of Agriculture—draws on six underlying factors: mean January temperature; mean January hours of sunlight; mean July temperature; mean July relative humidity; the percentage of water surface; and a topography index.³ Higher values of the score are associated with locations that display better amenities—for example, sunny places with a mild climate, both of which are valued by residents.

As can be seen from the left panel in Figure 4.1, locations well endowed with (positive) amenities are, on average, larger. As can be seen from the right panel in Figure 4.1, locations with worse geological features (higher seismic or landslide hazard, and a larger share of sedimentary bedrock) are, on average, smaller after partialling out the effect of amenities.⁴

While empirical work on city sizes and productivity suggests that locational fundamentals may explain about one-fifth of the observed geographical concentration (Ellison and Glaeser, 1999), theory has largely ignored them. Locational fundamentals do, however, interact with other agglomeration mechanisms to shape economic outcomes. They pin down city locations and explain why those locations and city sizes are fairly resilient to large shocks or technological change (Davis and Weinstein, 2002; Bleakley and Lin, 2012). As we show later, they may also serve to explain the size distribution of cities.

4.2.2 Agglomeration economies

Interactions within and between industries give rise to various sorts of complementarities and indivisibilities. We regroup all those mechanisms under the common header

³ Higher mean January temperature and more hours of sunlight are positive amenities, whereas higher mean July temperature and greater relative humidity are disamenities. The topography index takes higher values for more difficult terrain (ranging from 1 for flat plains to 21 for high mountains) and thus reflects, on the one hand, the scarcity of land (Saiz, 2010). On the other hand, steeper terrain may offer positive amenities such as unobstructed views. Lastly, a larger water surface is a consumption amenity but a land supply restriction. Its effect on population size is *a priori* unclear.

⁴ The right panel in Figure 4.1 shows that worse geological features are positively associated with population size when one does not control for amenities. The reason is that certain amenities (e.g., temperature) are valued more highly than certain disamenities (e.g., seismic risk). This is especially true for California and the US West Coast, which generate a strong positive correlation between seismic and landslide hazards and climate variables.

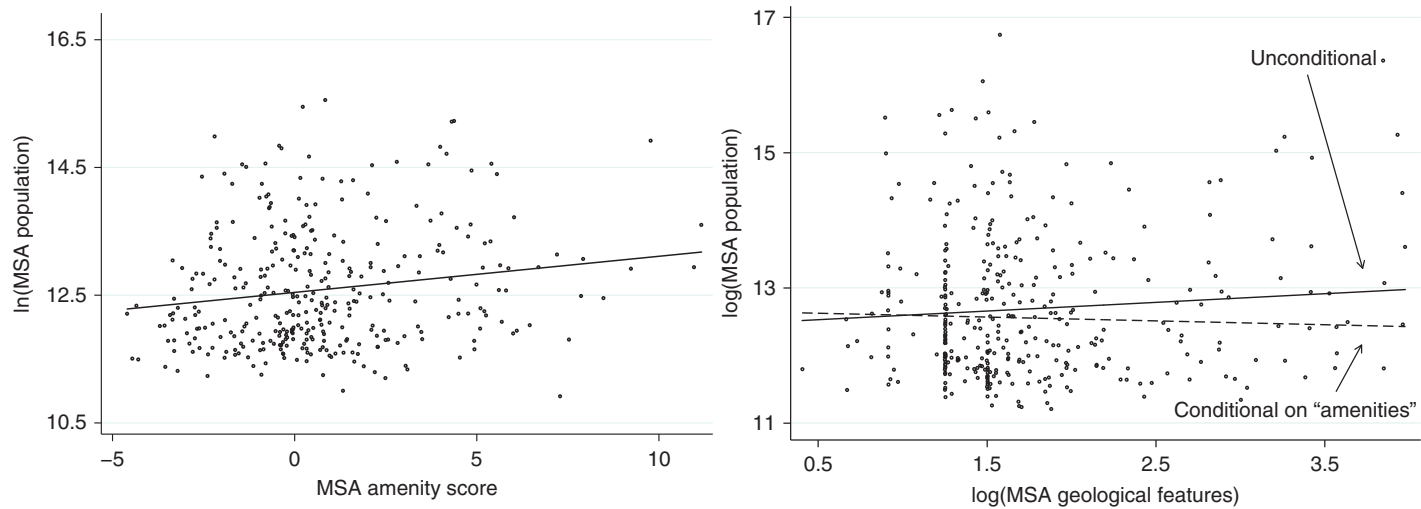


Figure 4.1 *Fundamentals*. MSA population, climatic amenities, and geological disamenities. *Notes:* Authors' calculations based on US Census Bureau, US Department of Agriculture, and US Geological Survey data for 343 and 340 MSAs in 2010 and 2007. See footnote 2 for details. The "MSA geological features" is the product of landslide, seismic hazard, and the share of sedimentary bedrock. The slope in the left panel is 0.057 (standard error 0.019). The unconditional slope in the right panel is 0.059 (standard error 0.053), and the conditional slope is -0.025 (standard error 0.047).

agglomeration economies. These include matching, sharing, and learning externalities (Duranton and Puga, 2004) that can operate either within an industry (localization economies) or across industries (urbanization economies). Labor market pooling, input-output linkages, and knowledge spillovers are the most frequently invoked Marshallian mechanisms that justify the existence of citywide increasing returns to scale.

The left panel in Figure 4.2 illustrates the presence of agglomeration economies for our cross section of US MSAs. The unconditional size elasticity of mean household income with respect to urban population is 0.081 and statistically significant at 1%. This estimate falls within the range usually found in the literature: the estimated elasticity of income or productivity with respect to population (or population density) is between 2% and 10%, depending on the method and the data used (Rosenthal and Strange, 2004; Melo et al., 2009). The right panel in Figure 4.2 depicts the corresponding urban costs (“congestion” for short), with the median gross rent in the MSA as a proxy. The estimated elasticity of urban costs with respect to urban population is 0.088 in our sample and is statistically significant at 1%. Observe that the two estimates are very close: the difference of 0.007 is statistically indistinguishable from zero.⁵ Though the measurement of the urban congestion elasticity has attracted much less attention than that of agglomeration economies in the literature, so that it is too early to speak about a consensual range for estimates, recent studies suggest that the gap between urban congestion and agglomeration elasticities is positive yet tiny (Combes et al., 2014). We show later that this has important implications for the spatial equilibrium and the size distribution of cities.

4.2.3 Sorting of heterogeneous agents

Though cross-city differences in size, productivity, and urban costs may be the most visible ones, cities also differ greatly in their composition. Most basically, cities differ in their industrial structure: diversified and specialized cities coexist, with no city being a simple replica of the national economy (Helsley and Strange, 2014). Cities may differ both horizontally, in terms of the *set* of industries they host, and vertically, in terms of the *functions* they perform (Duranton and Puga, 2005). Cities also differ fundamentally in their human capital, the set of workers and skills they attract, and the “quality” of their entrepreneurs and firms. These relationships are illustrated in Figure 4.3, which shows that the share of the highly skilled in an MSA is strongly associated with the MSA’s size (left panel) and density (right panel). We group under the common header *sorting* all mechanisms that imply that heterogeneous workers, firms, and industries make heterogeneous location choices.

⁵ The estimated standard deviation of the difference is 0.011, with a *t* statistic of 0.63 and a *p* value of 0.53.

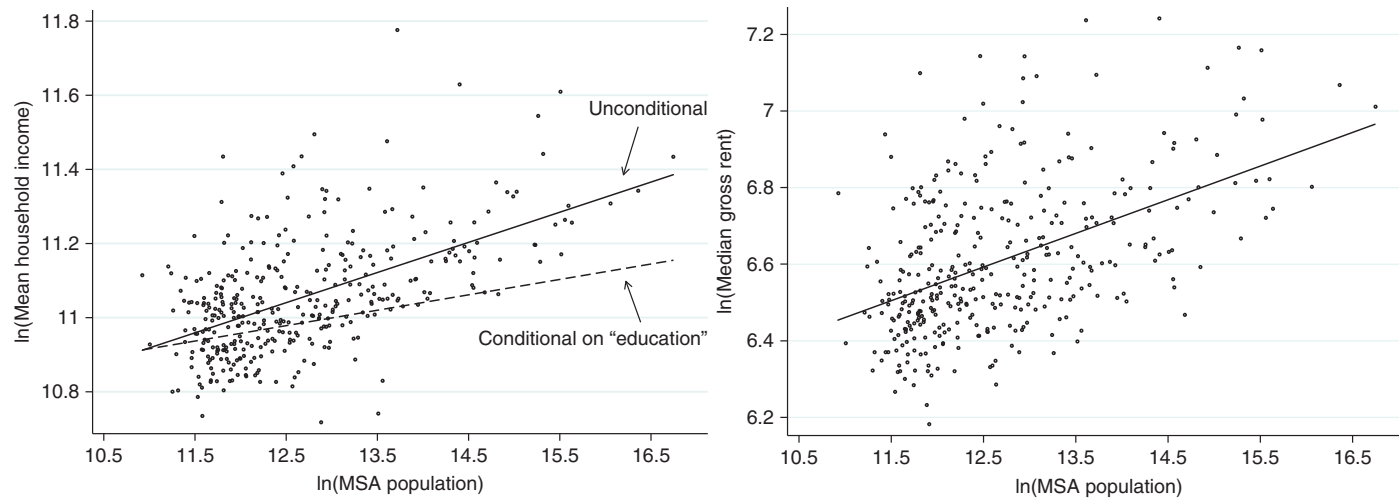


Figure 4.2 *Agglomeration*. MSA population, mean household income, and median rent. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The unconditional slope in the left panel is 0.081 (standard error 0.006), and the conditional slope is 0.042 (standard error 0.005). The slope in the right panel is 0.088 (standard error 0.008).*

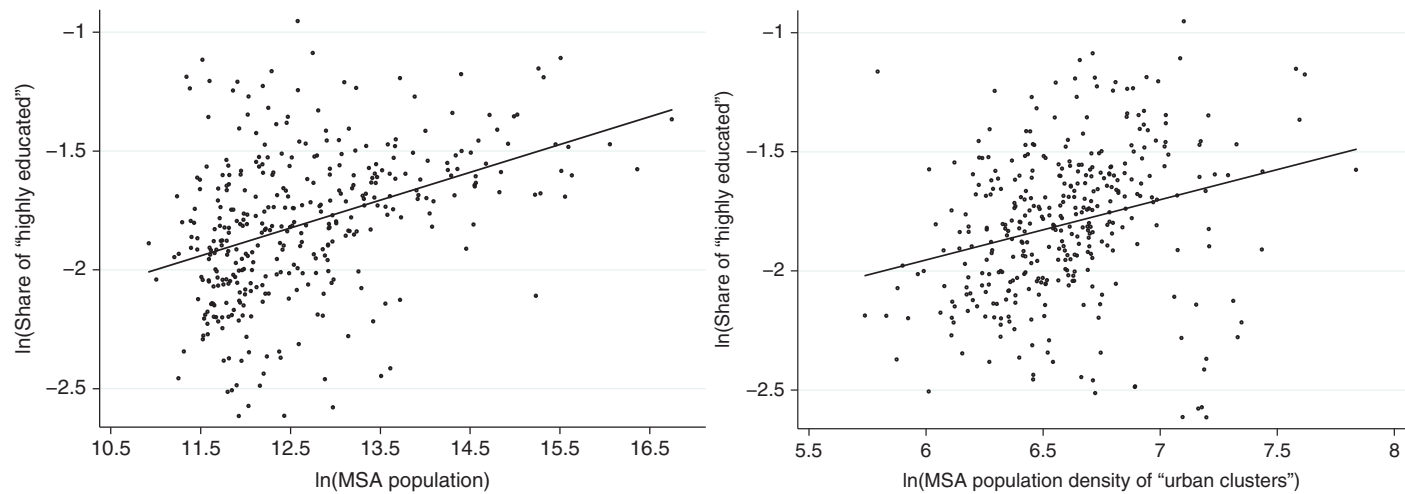


Figure 4.3 *Sorting. MSA population, cluster density, and share of “highly educated” workers. Notes: Authors’ calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The slope in the left panel is 0.117 (standard error 0.014). The slope in the right panel is 0.253 (standard error 0.048).*

The consensus in the recent literature is that sorting is a robust feature of the data and that differences in worker “quality” across cities explain up to 40–50% of the measured size–productivity relationship (Combes et al., 2008). This is illustrated in the left panel in Figure 4.2, where the size elasticity of wages falls from 0.081 to 0.049 once the share of “highly skilled” is introduced as a control.⁶ Although there are some sectoral differences in the strength of sorting, depending on regional density and specialization (Matano and Naticchioni, 2012), sorting is essentially a broad-based phenomenon that cuts across industries: about 80% of the skill differences in larger cities occur within industries, with only 20% accounted for by differences in industrial composition (Hendricks, 2011).

4.2.4 Selection effects

The size, density, industrial composition, and human capital of cities affect entrepreneurial incentives and the relative profitability of different occupations. Creating a firm and running a business also entails risks that depend, among other factors, on city characteristics. Although larger cities provide certain advantages for the creation of new firms (Duranton and Puga, 2001), they also host more numerous and better competitors, thereby reducing the chances of success for budding entrepreneurs and nascent firms. They also increase wages, thus changing the returns of salaried work relative to self-employment and entrepreneurship. We group under the common header *selection* all mechanisms that influence agents’ occupational choices and the choice of firms and entrepreneurs to operate in the market.

Figure 4.4 illustrates selection into entrepreneurship across US MSAs. Although there is no generally agreed upon measure of “entrepreneurship,” we use the share of self-employed in the MSA, or the average firm size, or the net entry rate (firm births minus firm deaths over total number of firms), which are standard proxies in the literature (Glaeser and Kerr, 2009).⁷ As can be seen from the left panel in Figure 4.4, there is no clear relationship between MSA size and the share of self-employed in the United States. However, Table 4.1 shows that there is a negative and significant relationship

⁶ How to conceive of “skills” or “talent” is a difficult empirical question. There is a crucial distinction to be made between horizontal skills and vertical talent (education), as emphasized by Bacolod et al. (2009a,b, 2010). That distinction is important for empirical work or for microfoundations of urban agglomeration economies, but less so for our purpose of dealing with cities from a macro perspective. We henceforth use the terms “skills,” “talent,” and “education” interchangeably and mostly conceive of skills, talent, or education as being vertical in nature.

⁷ Glaeser and Kerr (2009, pp. 624–627) measure entrepreneurship by “new entry of stand-alone plants.” They focus on “manufacturing entrepreneurship” only, whereas our data contain all firms. They note that their “entry metric has a 0.36 and 0.66 correlation with self-employment rates in the year 2000 at the city and state levels, respectively. Correlation with average firm size is higher at -0.59 to -0.80 .” Table 4.1 shows that our correlations have the same sign, though the correlation with average size is lower.

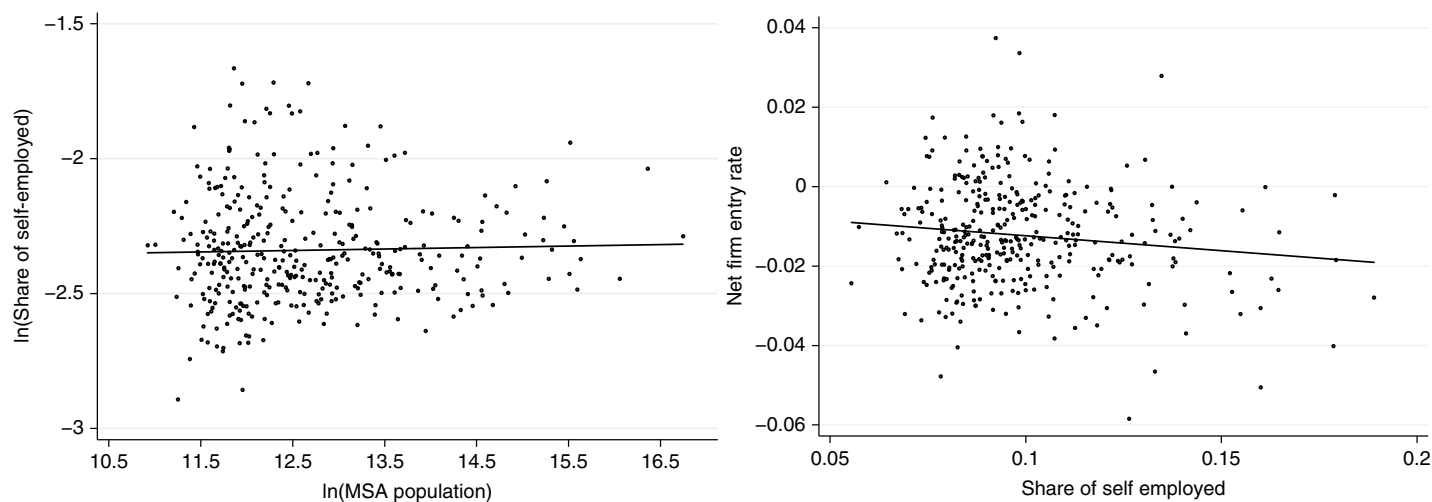


Figure 4.4 *Selection. MSA population, share of self-employed, and net entry rates. Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The slope in the left panel is 0.005 (standard error 0.010). The slope in the right panel is -0.075 (standard error 0.031).*

Table 4.1 Correlations between alternative measures of “entrepreneurship” and MSA size
“Entrepreneurship” measures

Variables	Self-employed (share)	log (Average firm employment)	Entry rate	log (MSA population)
log (MSA population)	0.0062	0.3502*	0.5501*	—
log (MSA density)	−0.1308*	0.3359*	0.2482*	0.6382*
log (Average firm employment)	−0.7018*	—	−0.1394*	0.3502*
Exit rate	0.3979*	−0.2019*	0.7520*	0.5079*
Entry rate	0.3498*	−0.1394*	—	0.5501*
Net entry rate	−0.1258*	0.1144*	0.2119*	−0.0231
Churning	0.4010*	−0.1826*	0.9193*	0.5664*
Venture capital deals (number per capita)	0.1417*	−0.1396*	−0.0197	0.1514*
Venture capital invest (\$ per capita)	0.0791	−0.1028	0.0314	0.1403*
Venture capital invest (\$ per deal)	0.1298*	−0.1366*	0.1139	0.0871
Share of highly educated	0.2006*	0.0104	0.2414*	0.4010*

See footnote 2 for information on the data used. The three venture capital variables are constructed at the state level only (using state-level population for per capita measures). Multistate MSA values are averaged across states. We indicate by asterisks correlations that are significant at the 5% level.

between MSA density and the share of self-employed.⁸ Furthermore, as can be seen from the right panel of Figure 4.4 and from the last column of Table 4.1, the net entry rate for firms is lower in larger MSAs. Also, larger cities or cities with more self-employment have smaller average firm sizes, and the latter two characteristics are positively associated with firm churning and different measures of venture capital investment.⁹

The right panel in Figure 4.4 and some correlations in Table 4.1 are suggestive of the possible existence of “selection effects.” For example, firm (churning) turnover is substantially higher in bigger cities. We will show that the existence and direction of selection effects with respect to market size or density is theoretically ambiguous: whether more or fewer firms survive or whether the share of entrepreneurs increases or decreases strongly depends on modeling choices. This finding may explain why the current empirical evidence is inconclusive.

⁸ The estimated density elasticity from a simple ordinary least squares regression is -0.032 and statistically significant at 1%.

⁹ A word of caution is in order. The venture capital data are available only at the state level, and per capita figures are relative to state population. Hence, we cannot account for within-state variation in venture capital across MSAs.

4.2.5 Inequality and city size

The size and density of cities are correlated with their composition, with the occupational choices of their residents, and with the success probabilities of businesses. They are also correlated with inequality in economic outcomes. That larger cities are more unequal places is a robust feature of the data (Glaeser et al., 2010; Baum-Snow and Pavan, 2014). This is illustrated in Figure 4.5.

The left panel depicts the relationship between MSA size and inequality as measured by the Gini coefficient of income. The human capital composition of cities has a sizable effect on inequality: the size elasticity of the Gini coefficient falls from 0.011 to 0.008 once education (as measured by the share of college graduates) is controlled for. Size, however, also matters for inequality beyond the sorting of the most educated agents to the largest cities. One of the reasons is that agglomeration interacts with human capital sorting and with selection to “dilate” the income distribution (Combes et al., 2012; Baum-Snow and Pavan, 2014). As can be seen from the right panel in Figure 4.5, the size elasticity of income increases across the income distribution, thus suggesting that agglomeration economies disproportionately accrue to the top of the earnings or productivity distribution of workers and firms.

4.2.6 City size distribution

The spatial distribution of population exhibits strong empirical regularities in many countries of the world. Figure 4.6 illustrates these strong patterns for the US data. Two aspects are worth mentioning. First, as can be seen from the left panel in Figure 4.6, the distribution of populated places in the United States is well approximated by a log-normal distribution (Eeckhout, 2004). As is well known, the upper tail of that distribution is difficult to distinguish from a Pareto distribution. Hence, the size distribution of the largest cities in the urban system approximately follows a power law. That this is indeed a good approximation can be seen from the right panel in Figure 4.6: the size distribution of large US cities follows Zipf’s law—that is, it follows a Pareto distribution with a unitary shape parameter (Gabaix and Ioannides, 2004; Gabaix, 1999).¹⁰

4.2.7 Assembling the pieces

The foregoing empirical relationships point toward the key ingredients that agglomeration models focusing on citywide outcomes should contain. While prior work has essentially focused on those ingredients individually, we argue that looking at them jointly is important, especially if distributional issues are of concern. To

¹⁰ Rozenfeld et al. (2011) have shown that even the distribution of US “places” follows Zipf’s law when places are constructed as geographically connected areas from satellite data. This finding suggests that the distribution is sensitive to the way space is (or is not) partitioned when constructing “places,” which is reminiscent of the classic “modifiable areal unit problem” that plagues spatial analysis at large.

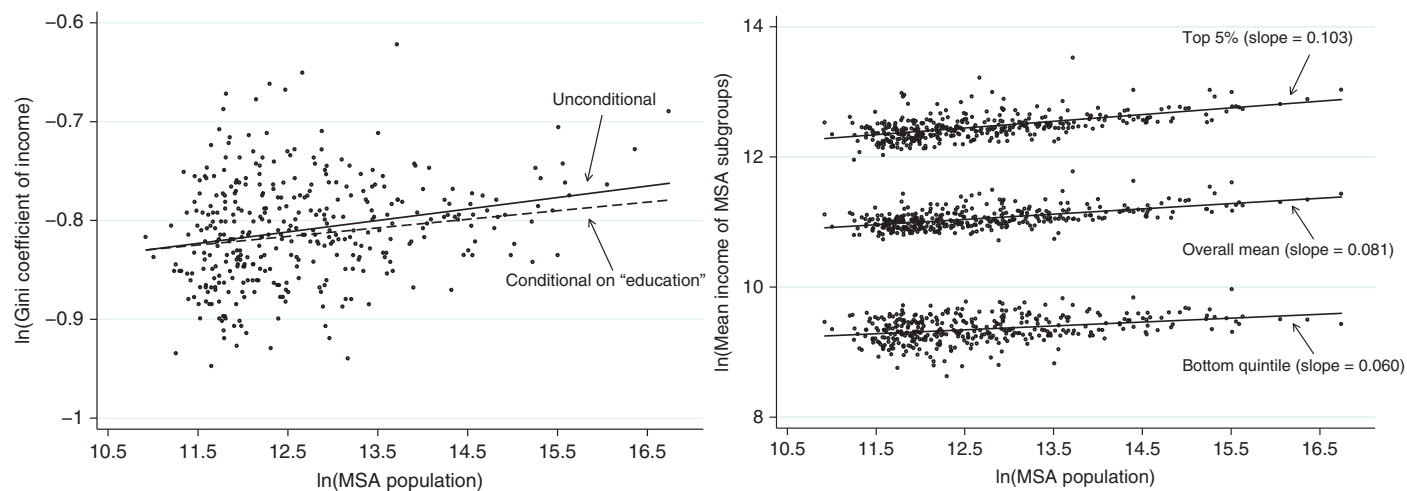


Figure 4.5 *Inequality. MSA population, Gini coefficient, and mean incomes by groups. Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. See footnote 2 for details. The unconditional slope in the left panel is 0.012 (standard error 0.003), and the conditional slope is 0.009 (standard error 0.002). The slopes in the right panel are provided in the figure, and they are all significant at 1%.*

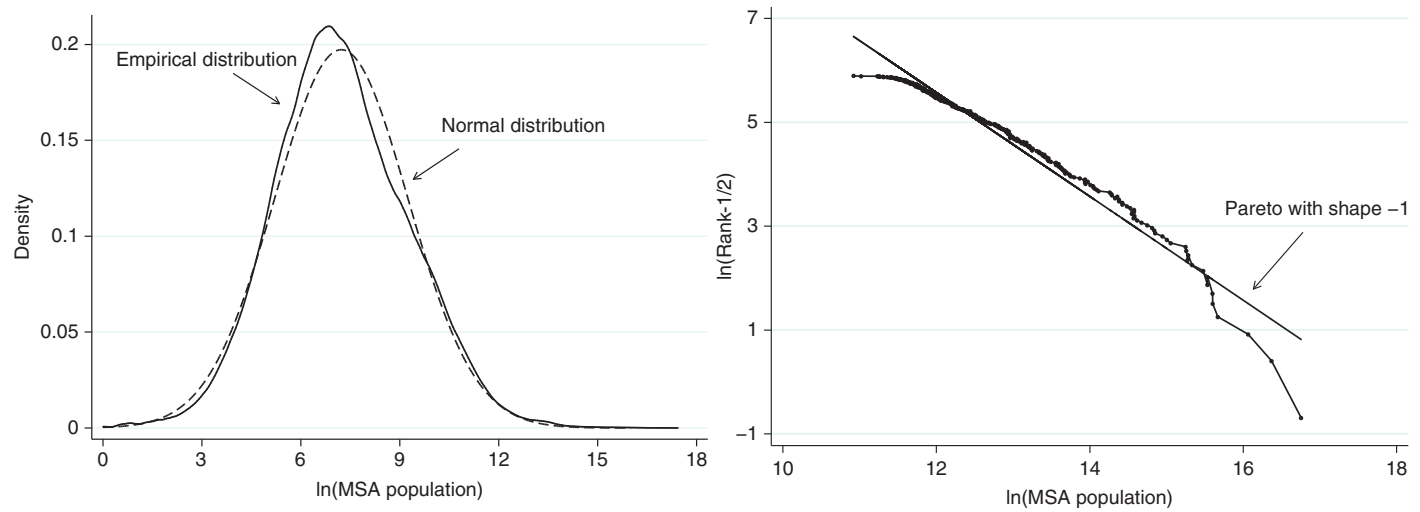


Figure 4.6 *Size distribution. Size distribution of places and the rank-size rule of cities. Notes: Authors' calculations based on US Census Bureau data for 81,631 places in 2010 (left panel) and 363 MSAs in 2010 (right panel). See footnote 2 for details. The estimated slope coefficient in the right panel is -0.922 (standard error 0.009). We subtract $1/2$ from the rank as in [Gabaix and Ibragimov \(2011\)](#).*

understand how the four causes (heterogeneous fundamentals, agglomeration economies, and the sorting and selection of heterogeneous agents) interact to shape the two moments (average and dispersion) of the productivity and income distributions, consider the following simple example. Assume that more talented individuals, or individuals with better cognitive skills, gain more from being located in larger cities (Bacolod et al., 2009a). The reasons may be that larger cities are places of intense knowledge exchange, that better cognitive skills allow individuals to absorb and process more information, that information is more valuable in bigger markets, or any combination of these. The complementarity between agglomeration economies—knowledge spillovers in our example—and agents’ talent leads to the sorting of more able agents into larger cities. Then, more talented agents make those cities more productive. They also make them places where it is more difficult to succeed in the market—as in the lyrics of Scorsese’s eponymous movie “New York, New York, if I can make it there, I’ll make it anywhere.” Selection effects and increasing urban costs in larger cities then discourage less able agents from going there in the first place, or “fail” some of them who are already there. Those who do not fail, however, reap the benefits of larger urban size. Thus, the interactions between sorting, selection, and agglomeration economies shape the wage distribution and exacerbate income inequality across cities of different sizes. They also largely contribute to shaping the equilibrium size distribution of cities.

4.3. AGGLOMERATION

We start by laying out the framework upon which we build throughout this chapter. That framework is flexible enough to encompass most aspects linked to the size, composition, and productivity of cities. It can also accommodate the qualitative relationships in the data we have highlighted, and it lends itself quite naturally to empirical investigation. We are not interested in the precise microeconomic mechanisms that give rise to citywide increasing returns; we henceforth simply assume their existence. Doing so greatly eases the exposition and the quest for a unified framework. We enrich the canonical model as we go along and as required by the different aspects of the theory. Whereas we remain general when dealing with agglomeration economies throughout this chapter, we impose more structure on the model when analyzing sorting, selection, and inequality. We first look at agglomeration theory when agents are homogeneous in order to introduce notation and establish a (well-known) benchmark.

4.3.1 Main ingredients

The basic ingredients and notation of our theoretical framework are the following. First, there is set $\mathcal{C}$ of sites. Without loss of generality, one site hosts at most one city. We index cities—and the sites at which they are developed—by c and we denote by C their

endogenously determined number, or mass. Second, there is a (large) number I of perfectly competitive industries, indexed by i . Each industry produces a homogeneous final consumption good. For simplicity, we stick to the canonical model of [Henderson \(1974\)](#) and we abstract from intercity trade costs for final goods. We later also introduce non-traded goods specific to some cities.¹¹ Production of each good requires labor and capital, both of which are freely mobile across cities. Workers are hired locally and paid city-specific wages, whereas capital is owned globally and fetches the same price everywhere. We assume that total output, Y_{ic} , of industry i in city c is given by

$$Y_{ic} = \mathbb{A}_{ic} \mathbb{L}_{ic} K_{ic}^{1-\theta_i} L_{ic}^{\theta_i}, \quad (4.1)$$

where $\mathbb{A}_{ic}$ is an industry- and city-specific productivity shifter, which we refer to as “total factor productivity” (TFP); K_{ic} and L_{ic} denote the capital and labor inputs, respectively, with economy-wide labor share $0 < \theta_i \leq 1$; and $\mathbb{L}_{ic}$ is an agglomeration effect external to firms in industry i and city c .

Since final goods industries are perfectly competitive, firms in those industries choose labor and capital inputs in Equation (4.1) taking the TFP term, $\mathbb{A}_{ic}$, and the agglomeration effect, $\mathbb{L}_{ic}$, as given. In what follows, bold capitals denote aggregates that are external to individual economic agents. For now, think of them as black boxes that contain standard agglomeration mechanisms (see [Duranton and Puga, 2004](#) and [Puga, 2010](#) for surveys on the microfoundations of urban agglomeration economies). We later open those boxes to look at their microeconomic contents, especially in connection with the composition of cities and the sorting and selection of heterogeneous agents.

4.3.2 Canonical model

To set the stage, we build a simple model of a system of cities in the spirit of the canonical model of [Henderson \(1974\)](#). In that canonical model, agglomeration and the size distribution of cities are driven by some external agglomeration effect and the unexplained distribution of TFP across sites. We assume for now that there is no heterogeneity across agents, but locational fundamentals are heterogeneous.

4.3.2.1 Equilibrium, optimum, and maximum city sizes

Consider an economy with a single industry and labor as the sole primary input ($I = 1$ and $\theta_i = 1$). The economy is endowed with $\bar{L}$ homogeneous workers who distribute themselves across cities. City formation is endogenous. All cities produce the same homogeneous final good, which is freely tradeable and used as the numeraire. Each city has an exogenous TFP $\mathbb{A}_c > 0$. These city-specific TFP terms are the locational

¹¹ A wide range of nontraded consumer goods in larger cities are clearly a force pushing toward agglomeration. In recent years, the literature has moved away from the view whereby cities are exclusively places of production to conceive of “consumer cities” as places of consumption of local amenities, goods, and services ([Glaeser et al., 2001](#); [Lee, 2010](#); [Couture, 2014](#)).

fundamentals linked to the sites at which the cities are developed. In a nutshell, $\mathbb{A}_c$ captures the comparative advantage of site c to develop a city: sites with a high TFP are particularly amenable to hosting a city. Without loss of generality, we index cities in decreasing order of their TFP: $\mathbb{A}_1 \geq \mathbb{A}_2 \geq \dots \geq \mathbb{A}_C$.

For cities to arise in equilibrium, we further assume that production exhibits increasing returns to scale at the city level. From (4.1), aggregate output Y_c is such that

$$Y_c = \mathbb{A}_c \mathbb{L}_c L_c. \quad (4.2)$$

Perfect competition in the labor market and zero profits yield a citywide wage that increases with city size: $w_c = \mathbb{A}_c \mathbb{L}_c$. The simplest specification for the external effect $\mathbb{L}_c$ is that it is governed by city size only: $\mathbb{L}_c = L_c^\epsilon$. We refer to $\epsilon \geq 0$, a mnemonic for “external,” as the elasticity of agglomeration economies with respect to urban population. Many microeconomic foundations involving matching, sharing, or learning externalities give rise to such a reduced-form external effect (Duranton and Puga, 2004). Workers spend their wage net of urban costs on the numeraire good. We assume that per capita urban costs are given by L_c^γ , where the parameter γ is the congestion elasticity with respect to urban size. This can easily be microfounded with a monocentric city model in which γ is the elasticity of the commuting cost with respect to commuting distance (Fujita, 1989). We could also consider that urban costs are site specific and given by $\mathbb{B}_c L_c^\gamma$. If sites differ both in productivity $\mathbb{A}_c$ and in urban costs $\mathbb{B}_c$, most of our results go through by redefining the net advantage of site c as $\mathbb{A}_c/\mathbb{B}_c$. We henceforth impose $\mathbb{B}_c = 1$ for all c for simplicity. Assuming linear preferences for consumers, the utility level associated with living in city c is

$$u_c(L_c) = \mathbb{A}_c L_c^\epsilon - L_c^\gamma. \quad (4.3)$$

Throughout this chapter, we focus our attention on either of two types of allocation, depending on the topic under study. We characterize the allocation that prevails with welfare-maximizing local governments when studying the composition of cities in Section 4.3.3. We follow this normative approach for the sake of simplicity. In all other cases, we characterize an equilibrium allocation. We also impose the “full-employment condition”

$$\sum_{c \in \mathcal{C}} L_c \leq \bar{L}. \quad (4.4)$$

When agents are homogeneous and absent any friction to labor mobility, a *spatial equilibrium* requires that there exists some *common* equilibrium utility level $u^* \geq 0$ such that

$$\forall c \in \mathcal{C}: (u_c - u^*) L_c = 0, \quad u_c \leq u^*, \quad (4.5)$$

and (4.4) holds. That is to say, all nonempty sites command the same utility level at equilibrium. The spatial equilibrium is “the single most important concept in regional and

urban economics . . . the bedrock on which everything else in the field stands” (Glaeser, 2008, p. 4). We will see later that this concept needs to be modified in a fundamental way when agents are heterogeneous. We maintain the free-mobility assumption throughout the chapter unless otherwise specified. The utility level (4.3) and the indifference conditions (4.5) can be expressed as follows:

$$u_c = \mathbb{A}_c L_c^\epsilon \left(1 - \frac{L_c^{\gamma-\epsilon}}{\mathbb{A}_c} \right) = u^*, \quad (4.6)$$

which can be solved for the equilibrium city size L_c^* as a function of u^* . This equilibrium is stable only if the marginal utility decreases with city size for all cities with a positive equilibrium population, which requires that

$$\frac{\partial u_c}{\partial L_c} = \epsilon \mathbb{A}_c L_c^{\epsilon-1} \left(1 - \frac{\gamma L_c^{\gamma-\epsilon}}{\epsilon \mathbb{A}_c} \right) < 0 \quad (4.7)$$

holds at the equilibrium city size L_c^* . It is easy to show from Equations (4.6) and (4.7) that a stable equilibrium necessarily requires $\gamma > \epsilon$ —that is, urban costs rise faster than urban productivity as the urban population grows. In that case, city sizes are bounded so that not everybody ends up living in a single megacity. We henceforth impose this parameter restriction. Empirically, $\gamma - \epsilon$ seems to be small, and this has important theoretical implications as shown later.

There exist many decentralized equilibria that simultaneously satisfy the full-employment condition (4.4), the indifference condition (4.6), and the stability condition (4.7). The existence of increasing returns to city size for low levels of urban size is the source of potential coordination failures in the absence of large agents able to coordinate the creation of new cities, such as governments and land developers.¹² The precise equilibrium that will be selected—both in terms of sites and in terms of city sizes—is undetermined, but it is *a priori* constrained by the distribution of the $\mathbb{A}_c$ terms, by the number of sites at which cities can be developed, and by the total population of the economy. Figure 4.7 illustrates a decentralized equilibrium with three cities with different underlying TFPs, $\mathbb{A}_1 > \mathbb{A}_2 > \mathbb{A}_3$. This equilibrium satisfies (4.4), (4.6), and (4.7) and yields utility u^* to all urban dwellers in the urban system. Other equilibria may be possible, with fewer or more cities (leading to, respectively, higher and lower equilibrium utility). To

¹² The problem of coordination failure stems from the fact that the utility of a single agent starting a new city is zero, so there is no incentive to do so. Henderson and Venables (2009) develop a dynamic model in which forward-looking builders supply nonmalleable housing and infrastructure, which are sunk investments. In such a setting, either private builders or local governments can solve the coordination problem, and the equilibrium city growth path of the economy becomes unique. Since we do not consider dynamic settings and we focus on static equilibria, we require “static” mechanisms that can solve the coordination problem. Heterogeneity of sites and agents will prove useful here. In particular, heterogeneous agents and sorting along talent across cities may serve as an equilibrium refinement (see Section 4.4). Also, adding a housing market as in Lee and Li (2013) allows one to pin down city sizes.

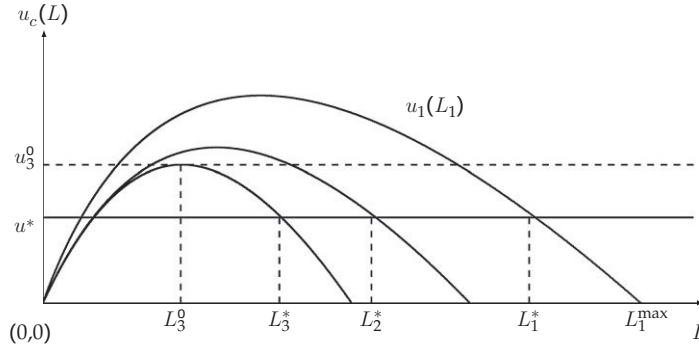


Figure 4.7 City sizes with heterogeneous $\mathbb{A}_c$ terms.

solve the equilibrium selection problem, the literature has often relied on the existence of large-scale, competitive land developers. When sites are homogeneous, the equilibrium with land developers is both unique and (generally) efficient, arguably two desirable properties (see [Henderson, 1988](#), and [Desmet and Henderson, 2015](#); see also [Becker and Henderson 2000b](#), on the political economy of city formation). When sites are heterogeneous, any decentralized equilibrium (absent transfers across sites) will generally be inefficient though the equilibrium with land developer may be efficient. Providing a full characterization of such an equilibrium is beyond the scope of this chapter.¹³ Equilibria feature cities that are larger than the size that a utility-maximizing local government

¹³ In [Behrens and Robert-Nicoud \(2014a\)](#), we show that the socially optimal allocation of people across cities and the (unique) equilibrium allocation with perfectly competitive land developers coincide and display the following features: (a) only the most productive sites are developed and more productive sites host larger cities; (b) (gross) equilibrium utility increases with $\mathbb{A}_c$ and equilibrium utility net of equilibrium transfers to competitive land developers is equalized across cities and is weakly smaller than u_C^0 , where u_C^0 is the maximum utility that can be achieved at the least productive populated urban site (thus all developers owning inframarginal sites make pure profits); (c) the socially optimal size of any city c is strictly lower than $L_c^{\max}$; and (d) the socially optimal size of any city c is strictly larger than the size chosen by local governments L_c^0 for all cities but the smallest, for which the two may coincide. If $C \subseteq \mathbb{R}$ and if $\mathbb{A}(c)$ is a continuous variable, then $u^* \leq u_C^0$ and $L_C^* \geq L_C^0$. Note that the allocation associated with *local governments that can exclude people* (implementing zoning restrictions, greenbelt policies, or city boundaries) and that maximize the welfare of their current residents violates the indifference condition (4.6) of the standard definition of the urban equilibrium because

$$u(L_c^0) = \frac{\gamma - \epsilon}{\epsilon} \left(\frac{\epsilon}{\gamma} \mathbb{A}_c \right)^{\frac{\gamma}{\gamma - \epsilon}}$$

increases with $\mathbb{A}_c$. That is, residents of high-amenity places are more fortunate than others because their local authorities do not internalize the adverse effects of restricting the size of their community on others. This raises interesting public policy and political economy questions—for example, whether high-amenity places should implement tax and subsidy schemes to attract certain types of people and to expand beyond the size L_c^0 chosen in the absence of transfers. [Albouy and Seegert \(2012\)](#) make several of the same points and analyze under what conditions the market may deliver too many and too small cities when land is heterogeneous and when there are cross-city externalities due to land ownership and federal taxes.

would choose. From a national perspective, some cities may be oversized and some undersized when sites are heterogeneous.¹⁴ In order to characterize common properties of decentralized equilibria, we first derive bounds on feasible city sizes. Let $L_c^{\max}$ denote the maximum size of a city, which is determined by the utility that can be secured by not residing in a city and which we normalize to zero for convenience. Hence, plugging $u^* = 0$ into (4.6) and solving for L_c yields

$$L_c^{\max} = \mathbb{A}_c^{\frac{1}{\gamma-\epsilon}}. \quad (4.8)$$

Let L_c^o denote the size that would be implemented by a *local* government in city c that can restrict entry but cannot price discriminate between current and potential residents, and that maximizes the welfare of its residents. This provides a lower bound to equilibrium city sizes by (4.7) and $\gamma > \epsilon$. Maximizing (4.3) with respect to L_c and solving for L_c^o yields

$$L_c^o = \left(\frac{\epsilon}{\gamma} \mathbb{A}_c \right)^{\frac{1}{\gamma-\epsilon}}. \quad (4.9)$$

Equations (4.8) and (4.9) establish that the lower and upper bounds of city sizes are both proportional to $\mathbb{A}_c^{1/(\gamma-\epsilon)}$. At any spatial equilibrium, the utility level u^* is in $[0, u_C^o]$, where u_C^o is the maximum utility that can be achieved in the city with the smallest $\mathbb{A}_c$ (in the decentralized equilibrium with three cities illustrated in Figure 4.7, u_C^o is u_3^o). Cities are oversized in any equilibrium such that $u^* < u_C^o$ because individuals do not take into account the negative impact they impose on other urban dwellers at the margin when making their location decisions. This coordination failure is especially important when thinking about the efficiency of industrial coagglomeration (Helsley and Strange, 2014), as we discuss in Section 4.3.3.1.

What can the foregoing results for the bounds of equilibrium city sizes teach us about the equilibrium city size distribution? Rearranging (4.6) yields

$$L_c^* = \left(\mathbb{A}_c - \frac{u^*}{L_c^{*\epsilon}} \right)^{\frac{1}{\gamma-\epsilon}}. \quad (4.10)$$

Equation (4.10) shows that L_c^* is smaller than but gets closer to $\mathbb{A}_c^{1/(\gamma-\epsilon)}$ when L_c^* becomes large (to see this, observe that $\lim_{L_c^* \rightarrow \infty} u^*/L_c^{*\epsilon} = 0$). Therefore, the *upper tail* of the equilibrium city size distribution L_c^* inherits the properties of the TFP distribution in the same way as L_c^o and $L_c^{\max}$ do. In other words, the distribution of $\mathbb{A}_c$ is crucial for determining the distribution of equilibrium sizes of large cities. We trace out implications of that property in the next section.

¹⁴ The optimal allocation requires one to equalize the net marginal benefits across all occupied sites. Henderson (1988) derives several results with heterogeneous sites, some of them heuristically. See also Vermeulen (2011), Albouy and Seegert (2012), and Albouy et al. (2015).

We can summarize the properties of the canonical model, characterized by Equations (4.7)–(4.10), as follows:

Proposition 4.1 (equilibrium size). *Let $\gamma > \epsilon > 0$ and assume that the utility level enjoyed outside cities is zero. Then any stable equilibrium features city sizes $L_c^* \in [L_c^o, L_c^{\max}]$ and a utility level $u^* \in [0, u_C^o]$. Equilibrium city sizes are larger than the sizes chosen by local governments and both L_c^o and $L_c^{\max}$ are proportional to $\mathbb{A}_c$. Finally, in equilibrium the upper tail of the size distribution of cities follows the distribution of the TFP parameters $\mathbb{A}_c$.*

Four comments are in order. First, although all agents are free to live in cities, some agents may opt out of the urban system. This may occur when the outside option of not living in cities is large and/or when the number of potential sites for cities is small compared with the population. Second, not all sites need to develop cities. Since both L_c^o and $L_c^{\max}$ increase with $\mathbb{A}_c$, this is more likely to occur for any given number of sites if locational fundamentals are good, since L_c^* is bounded by two terms that both increase with $\mathbb{A}_c$.¹⁵ Third, the empirical link between city size and $\mathbb{A}_c$ (with an index of natural amenities or with geological features as a proxy) is borne out in the data, as illustrated by the two panels in Figure 4.1. Regressing the logarithm of the population on the MSA amenity score yields a positive size elasticity of 0.057, statistically significant at the 1% level. Lastly, we argued in Section 4.2.2 that $\gamma - \epsilon$ is small in the data. From Proposition 4.1 and from Equation (4.10), we thus obtain that small differences in the underlying $\mathbb{A}_c$ terms can map into large equilibrium size differences between cities. In other words, we may observe cities of vastly different sizes even in a world where locational fundamentals do not differ much across sites.

4.3.2.2 Size distribution of cities

One well-known striking regularity in the size distribution of cities is that it is roughly log-normal, with an upper tail that is statistically indistinguishable from a Pareto distribution with unitary shape parameter: Zipf's law holds for (large) cities (Gabaix, 1999; Eeckhout, 2004; Gabaix and Ioannides, 2004).¹⁶ Figure 4.6 depicts those two properties.

¹⁵ It is reasonable to assume that sites are populated in decreasing order of productivity. Bleakley and Lin (2012, p. 589) show that “locational fundamentals” are good predictors of which sites develop cities. Focusing on “breaks” in navigable transportation routes (portage sites; or hubs in Behrens, 2007), they find that the “footprint of portage is evident today [since] in the south-eastern United States, an urban area of some size is found nearly every place a river crosses the fall line.” Those sites are very likely places to develop cities. One should keep in mind, however, that with sequential occupation of sites in the presence of taste heterogeneity, path dependence is an issue (Arthur, 1994). In other words, the most productive places need not be developed first, and depending on the sequence of site occupation, there is generally a large number of equilibrium development paths.

¹⁶ The log-normal and the Pareto distributions theoretically have very different tails, but those are arguably hard to distinguish empirically. The fundamental reason is that, by definition, we have to be “far” in the tail, and any estimate there is quite imprecise owing to small sample size (especially for cities, since there are only very few very large ones).

The canonical model has been criticized for not being able to deliver empirically plausible city size distributions other than if ad hoc assumptions are made on the distribution of $\mathbb{A}_c$. Recent progress has been made, however, and the model can generate such distributions on the basis of fairly weak assumptions on the heterogeneity of sites.¹⁷ Proposition 4.1 reveals that the size distribution of cities inherits the properties of the distribution of $\mathbb{A}_c$, at least in the upper tail of that distribution. In particular, if $\mathbb{A}_c$ follows a power law (or a log-normal distribution), then L_c also follows a power law (or a log-normal distribution) in the upper tail. The question then is why $\mathbb{A}_c$ should follow such a specific distribution. Lee and Li (2013) have shown that if $\mathbb{A}_c$ consists of the product of a large number of underlying factors a_{fc} (where $f = 1, 2, \dots, F$ indexes the factors) that are randomly distributed and not “too strongly correlated,” then the size distribution of cities converges to a log-normal distribution and is generally consistent with Zipf’s law in its upper tail. Formally, this result is the static counterpart of random growth theory that has been widely used to generate city size distributions in a dynamic setting (Gabaix, 1999; Eeckhout, 2004; Duranton, 2006; Rossi-Hansberg and Wright, 2007). Here, the random shocks (the factors) are stacked in the cross section instead of occurring through time. The factors can be viewed broadly as including consumption amenities, production amenities, and elements linked to the land supply in each location. Basically, they may subsume all characteristics that are positively associated with the desirability of a location. Each factor can also depend on city size—that is, it can be subject to agglomeration economies as captured by $a_{fc} L_c^{\epsilon_f}$. Let

$$\mathbb{A}_c \equiv \prod_f a_{fc} \quad \text{and} \quad \mathbb{L}_c \equiv \prod_f L_c^{\epsilon_f} \quad (4.11)$$

and assume that production is given by (4.2). Let $\epsilon \equiv \sum_f \epsilon_f$ subsume the agglomeration effects generated by all the underlying factors. Consistent with the canonical model, we assume that congestion economies dominate agglomeration economies at the margin—that is, $\gamma > \epsilon$. Plugging $\mathbb{A}_c$ and $\mathbb{L}_c$ into (4.8), and assuming that the outside option leads to a utility of zero so that $u^* = 0$, we find the equilibrium city size is $L_c^* = \mathbb{A}_c^{1/(\gamma-\epsilon)}$. Letting $a_{fc} \equiv \ln a_{fc}$ and taking the logarithm, we then can rewrite this as

$$\ln L_c^* = \frac{1}{\gamma - \epsilon} \left(\sum_{f=1}^F \hat{\alpha}_{fc} + \sum_{f=1}^F \bar{\alpha}_{fc} \right), \quad (4.12)$$

where we denote by $\hat{\alpha}_{fc} = \ln a_{fc} - \ln \bar{a}_{fc}$ the demeaned log factor, and where $\bar{a}_{fc}$ is the geometric mean of the a_{fc} terms. As shown by Lee and Li (2013), one can then apply a particular variant of the central-limit theorem to the sum of centered random variables $\sum_{f=1}^F \hat{\alpha}_{fc}$ in (4.12) to show that the city size distribution converges asymptotically to a

¹⁷ As shown in Section 4.4.1, there are other mechanisms that may serve the same purpose when heterogeneous agents sort across cities. Hsu (2012) proposes yet another explanation, based on differences in fixed costs across industries and central place theory, to generate Zipf’s law.

log-normal distribution $\ln \mathcal{N}\left(\frac{1}{\gamma-\epsilon} \sum_{j=1}^J \bar{\alpha}_{fc}, \frac{\sigma^2_F}{(\gamma-\epsilon)^2}\right)$, where σ^2 is the limit of the variance of the partial sums.¹⁸

As with any asymptotic result, the question arises as to how close one needs to get to the limit for the approximation to be reasonably good. Lee and Li (2013) use Monte Carlo simulations with randomly generated factors to show that (a) the size distribution of cities converges quickly to a log-normal distribution, and (b) Zipf's law holds in the upper tail of the distribution even when the number of factors is small and when they are quite highly correlated. One potential issue is, however, that the random factors do not correspond to anything we can observe in the real world. To gauge how accurate the foregoing results are when we consider “real factors” and not simulated ones, we rely on US Department of Agriculture county-level amenity data to approximate the a_{fc} terms. We use the same six factors as for the amenity score in Section 4.2.1 to construct the corresponding $\mathbb{A}_c$ terms.¹⁹

The distribution of the $\mathbb{A}_c$ terms is depicted in the left panel in Figure 4.8, which contrasts it with a normal distribution with the same mean and standard deviation. As can be seen, even a number of observable factors as small as six may deliver a log-normal distribution.²⁰ However, even if the distribution of factors is log-normal, they should be *strongly and positively associated with city size* for the theory to have significant explanatory power. In words, large values of $\mathbb{A}_c$ should map into large cities. As can be seen from the right panel in Figure 4.8, although there is a positive and statistically significant association between locational fundamentals and city sizes, that relationship is very fuzzy. The linear correlation for our 363 MSAs of the logarithm of the population and the amenity terms is only 0.147, whereas the Spearman rank correlation is 0.142. In words, only about 2.2% of the size distribution of MSAs in the United States is explained by the factors underlying our $\mathbb{A}_c$ terms, even if the latter are log-normally distributed.²¹

¹⁸ As shown by expression (4.12), a key requirement for the result to hold is that the functional forms are all multiplicatively separable. The ubiquitous Cobb–Douglas and constant elasticity of substitution (CES) specifications satisfy this requirement.

¹⁹ The factors are mean January temperature, mean January hours of sunlight, the inverse of mean July temperature, the inverse of mean July relative humidity, the percentage of water surface, and the inverse of the topography index. We take the logarithm of each factor, center the values, and sum them up to generate a county-specific value. We then aggregate these county-specific values by MSA, weighting each county by its land-surface share in the MSA. This yields MSA-specific factors $\mathbb{A}_c$ which map into an MSA size distribution.

²⁰ Using either the Shapiro–Wilk, the Shapiro–Francia, or the skewness and kurtosis tests for normality, we cannot reject at the 5% level (and almost at the 10% level) the null hypothesis that the distribution of our MSA amenity factors is log-normal.

²¹ This may be because we focus on only a small range of consumption amenities, but those at least do not seem to matter that much. This finding is similar to the that of Behrens et al. (2013), who use a structural model to solve for the logit choice probabilities that sustain the observed city size distribution. Regressing those choice probabilities on natural amenities delivers a small positive coefficient, but which does not explain much of the city size distribution either.

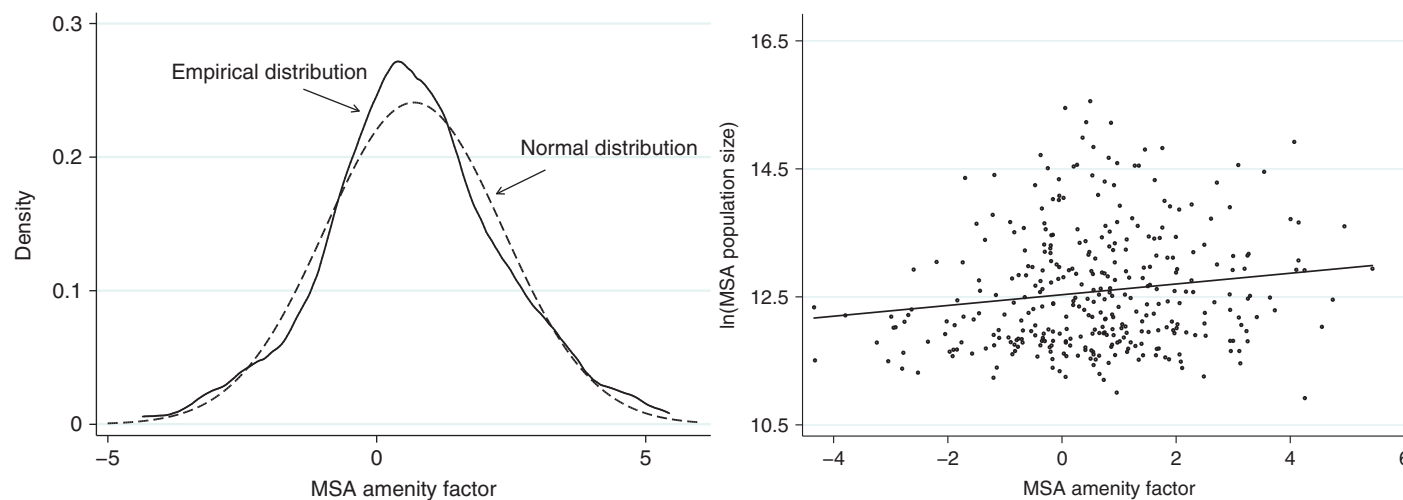


Figure 4.8 Log-normal distribution of MSA amenity factors $\mathbb{A}_c$, and factors-city size plot. *Notes: Authors' calculations based on US Census Bureau data for 363 MSAs in 2010. The MSA amenity factors are constructed using US Department of Agriculture amenity data. See footnotes 2 and 19 for details. The estimated slope coefficient in the right panel is 0.083 (standard error 0.031).*

Log-normality of $\mathbb{A}_c$ does not by itself guarantee that the resulting distribution matches closely with the ranking of city sizes, which thus breaks the theoretical link between the distribution of amenities and the distribution of city sizes. This finding also suggests that, as stated in [Section 4.2.1](#), locational fundamentals are no longer a major determinant of observed city size distributions in modern economies. We thus have to find alternative explanations for the size distribution of cities, a point we come back to in [Section 4.4.1.4](#).

4.3.2.3 Inside the “black boxes”: extensions and interpretations

We now use the canonical model to interpret prior work in relation to its key parameters ϵ , γ , and $\mathbb{A}_c$. To this end, we take a look inside the “black boxes” of the model.

Inside ϵ

The literature on agglomeration economies, as surveyed in [Duranton and Puga \(2004\)](#) and [Puga \(2010\)](#), provides microeconomic foundations for ϵ . For instance, if agglomeration economies arise as a result of input sharing, where Y_c is a CES aggregate of differentiated intermediate inputs produced under increasing returns to scale (as in [Ethier, 1982](#)), using local labor only, then $\epsilon = 1/(\sigma - 1)$, where $\sigma > 1$ is the elasticity of substitution between any pair of inputs. If, instead, production of Y_c requires the completion of an exogenous set of tasks and urban dwellers allocate their time between learning, which raises their effective amount of productive labor with an elasticity of $\theta \in (0, 1)$, and producing (as in [Becker and Murphy, 1992](#); [Becker and Henderson, 2000a](#)), then larger cities allow for a finer division of labor and this gives rise to citywide increasing returns, with $\epsilon = \theta$.²² The same result is obtained in a model where workers have to allocate a unit of time across tasks, and where learning-by-doing increases productivity for a task with an elasticity of θ . What is remarkable in all these models is that, despite having very different underlying microeconomic mechanisms, they generate a reduced-form citywide production function given by (4.2), where only the structural interpretation of ϵ changes. The empirical literature on the estimation of agglomeration economies, surveyed by [Rosenthal and Strange \(2004\)](#) and [Melo et al. \(2009\)](#), estimates this parameter to be in the range from 0.02 to 0.1 for a variety of countries and using a variety of econometric techniques. The consensus among urban economists nowadays is that the “true” value of ϵ is closer to the lower bound, especially when unobserved heterogeneity is controlled for using individual data and when different endogeneity concerns are properly addressed (see the chapter by [Combes and Gobillon, 2015](#) in this handbook).

²² Agglomeration economies may stem from investment in either vertical talent or horizontal skill ([Kim, 1989](#)). Larger markets favor investment in horizontal skills (which are useful in specific occupations) instead of vertical talent (which is useful in any occupation) because of better matching in thicker markets.

Inside γ

The literature on the microeconomic foundations of urban costs, γ , is much sparser than the literature on the microeconomic foundations of agglomeration economies. In theory, γ equals the elasticity of the cost per unit distance of commuting to the central business district in the one-dimensional Alonso–Muth–Mills model (see also Fujita and Ogawa, 1982; Lucas and Rossi-Hansberg, 2002). It also equals the elasticity of utility with respect to housing consumption in the Helpman (1998) model with an exogenous housing stock. The empirical literature on the estimation of γ is scarcer still: we are aware of only Combes et al. (2014). This is puzzling since the relative magnitude of urban costs, γ , and of agglomeration economies, ϵ , is important for understanding a variety of positive and normative properties of the spatial equilibrium. Thus, precise estimates of *both elasticities* are fundamental. The simplest models with linear cities and linear commuting costs suggest a very large estimate of $\gamma = 1$. This is clearly much too large compared with the few available estimates, which are also close to 2%.

Inside $\mathbb{A}_c$

The TFP parameters $\mathbb{A}_c$ are related to the industrial or functional composition of cities, the quality of their sites, and their commuting infrastructure. We have seen that heterogeneity in site-specific underlying factors may generate Zipf's law. However, just as the random growth version of Zipf's law, that theory has nothing to say about the microeconomic contents of the $\mathbb{A}_c$ terms. Heterogeneity in sites may stem from many underlying characteristics: production and consumption amenities, endowments, natural resources, and *locational advantage* in terms of transportation access to markets. This issue has received some attention in the new economic geography literature, but multiregion models are complex and thus have been analyzed only sparsely. The reason is that with multiple cities or regions, the relative position matters for access to demand (a positive effect) and exposure to competition (a negative effect). The urban literature has largely ignored costly trade between cities: trade costs are usually either zero or infinite, just as in classical trade theory.

Behrens et al. (2009) extend the “home market effect” model of Krugman (1980) to many locations. There is a mobile increasing returns to scale sector that produces differentiated varieties of a good that can be traded across space at some cost, and there is an immobile constant returns to scale sector that produces some freely traded good. The latter sector differs exogenously by productivity across sites, with productivity $1/z_c$ at site c . Sites also differ in their relative advantage for the mobile sector as compared with the outside sector: $a_c = (1/m_d)/(1/z_d)$. Finally, locations differ in access to each other: transportation costs across all sites are of the iceberg type and are represented by some $C \times C$ matrix Φ , where the element $\phi_{c,c'}$ is the freeness of trade between sites c and c' . Specifically, $\phi_{c,c'} \in [0, 1]$, with $\phi_{c,c'} = 0$ when trade between sites c and c' is prohibitively costly and $\phi_{c,c'} = 1$ when bilateral trade is costless. Behrens et al. (2009)

show that the equilibrium per capita output of site c is given by $y_c = \mathbb{A}_c$, with $\mathbb{A}_c \equiv A_c(\Phi, \{a_c\}_{c \in C}, 1/z_c)$. Per capita output increases with the site's productivity, which is a complex combination of its own productivity parameters ($1/z_c$ and a_c) and some spatially weighted combination of the productivity parameters of all other sites, and interacts with the spatial transportation cost structure of the economy. Intuitively, sites that offer better access to markets—that are closer to more productive markets, where incomes are higher—have a locational advantage in terms of access to consumers. However, those markets are also exposed to more competition from more numerous and more productive competitors, which may partly offset that locational advantage. The spatial allocation of firms across sites, and the resulting productivity distribution, crucially depends on the equilibrium trade-off between these two forces.²³

Another model that can be cast into our canonical mold is that of [Desmet and Rossi-Hansberg \(2013\)](#). In their model, per capita output of the homogeneous numeraire good in city c is given by

$$y_c = A_c \mathbb{L}_c k_c^{1-\theta} h_c^\theta, \quad (4.13)$$

where k_c and h_c are per capita capital and hours worked, respectively, A_c is a city-specific productivity shifter, and $\mathbb{L}_c = L_c^\epsilon$ is the agglomeration externality. Observe that Equation (4.13) is identical to our expression (4.1), except for the endogenous labor-leisure choice: consumers are endowed with one unit of time that can be used for work, h_c , or leisure, $1 - h_c$. They have preferences $v_c = \ln u_c + \psi \ln(1 - h_c) + a_c$ that are log-linear in consumption of the numeraire, u_c (which is, as before, income net of urban costs), leisure, and consumption amenities a_c .

In each city c of size L_c , a local government levies a tax τ_c on total labor income $L_c w_c h_c$ to finance infrastructure that is used for commuting. A consumer's consumption of the numeraire good is thus given by $u_c = w_c h_c (1 - \tau_c) - R_c$, where R_c is the per capita urban costs (commuting plus land rents) borne by a resident of city c . Assuming that cities are monocentric, and choosing appropriate units of measurement, we obtain per capita urban costs $R_c = L_c^\gamma$.

Consumers choose labor and leisure time to maximize utility and producers choose labor and capital inputs to minimize costs. Using the optimal choice of inputs, as well as the expression for urban costs R_c , we obtain per capita consumption and production as follows:

$$u_c = \theta(1 - \tau_c) y_c - L_c^\gamma \quad \text{and} \quad y_c = \kappa A_c^{\frac{1}{\theta}} L_c^{\frac{\epsilon}{\theta}} h_c,$$

²³ The same holds in the model of [Behrens et al. \(2013\)](#). In that model, cross-city differences in market access are subsumed by the selection cutoff for heterogeneous firms. We deal more extensively with selection effects in [Section 4.4.2](#).

where $\kappa > 0$ is a bundle of parameters. Desmet and Rossi-Hansberg (2013) show that $h_c \equiv h_c(\tau_c, A_c, L_c)$ is a monotonically increasing function of L_c : agents work more in bigger cities (Rosenthal and Strange, 2008a). Thus $u_c = \mathbb{A}_c h_c(\tau_c, A_c, L_c) L_c^{\epsilon/\theta} - L_c^\gamma$, where $\mathbb{A}_c \equiv \mathbb{A}_c(\tau_c, A_c) = \kappa \theta (1 - \tau_c) A_c^{1/\theta}$. If utility were linear in consumption and labor supply were fixed (as we have assumed so far), we would obtain an equilibrium relationship that is structurally identical to Equation (4.3). The cross-city heterogeneity in taxes, τ_c , and productivity parameters, A_c , serves to shift up or down the equilibrium city sizes via the TFP term $\mathbb{A}_c$.²⁴ However, labor supply is variable and utility depends on income, leisure, and consumption amenities. Hence, the spatial equilibrium condition requiring the equalization of utility is slightly more complex and is given by

$$\ln [\mathbb{A}_c h_c(\tau_c, A_c, L_c) L_c^{\epsilon/\theta} - L_c^\gamma] + \psi \ln [1 - h_c(\tau_c, A_c, L_c)] + a_c = u^*, \quad (4.14)$$

for some u^* that is determined in general equilibrium by the mobility of agents. The equilibrium allocation of homogeneous agents across cities depends on the cross-city distribution of three elements: (a) local taxes, τ_c , also referred to as “labor wedges”; (b) exogenous productivity differences, A_c ; and (c) differences in exogenous consumption amenities, a_c . Quite naturally, the equilibrium city size L_c^* increases with A_c and a_c , and decreases with τ_c .

The key contribution of Desmet and Rossi-Hansberg (2013) is to apply their spatial general equilibrium model (4.14) in a structural way to the data.²⁵ To this end, they first estimate the productivity shifters A_c and the labor wedges τ_c from their structural equations, and infer the amenities a_c such that—conditional on the labor wedges and productivity shifters—the model replicates the observed distribution of city sizes for 192 US cities in 2005–2008. They then evaluate the correlation between the implied a_c and a variety of quality-of-life measures usually used in the literature. Having thus calibrated the model, they finally perform an “urban accounting” exercise. The objective is to quantify the respective contribution of the different wedges—labor τ_c , productivity

²⁴ The full model of Desmet and Rossi-Hansberg (2013) is more complicated since they also make taxes endogenous. To pin them down, they assume that the local government must provide a quantity of infrastructure proportional to the product of wages and total commuting costs in the city, scaled by some city-specific government inefficiency g_c . Assuming that the government budget is balanced then requires that $\tau_c \propto g_c L_c^\gamma$ —that is, big cities with inefficient governments have higher tax rates.

²⁵ For more information on the use of structural methods in urban economics, see the chapters by Holmes and Sieg (2014) in this volume of the handbook. Behrens et al. (2013) perform a similar analysis in a very different setting. They use a multicity general equilibrium model that builds on the monopolistic competition framework developed by Behrens and Murata (2007). In that framework, heterogeneous firms produce differentiated varieties of a consumption good that can be traded at some cost across all cities. The key objective of Behrens et al. (2013) is to quantify how trade frictions and commuting costs affect individual city sizes, the size distribution of cities, and aggregate productivity. They find that the city size distribution is fairly stable with respect to trade frictions and commuting costs.

A_c , and amenities a_c —to city sizes, to welfare, and to the city size distribution. This is achieved by simulating counterfactual changes when one of the three channels— τ_c , a_c , or A_c —is shut down—that is, what happens if “we eliminate differences in a particular characteristic by setting its value to the population weighted average”? (Desmet and Rossi-Hansberg, 2013, p. 2312). They obtain large population reallocations but small welfare effects.²⁶ In words, the movement of agents across cities in response to possibly large shocks yields only fairly small welfare gains (see also Behrens et al. 2014a). These results are quite robust to the inclusion of consumption and production externalities in the US data. By contrast, applying their model to Chinese data, Desmet and Rossi-Hansberg (2013) obtain fewer population movements but larger welfare effects.

4.3.3 The composition of cities: industries, functions, and skills

Until now, cities differ only in terms of exogenous fundamentals. That cities also differ in their industrial structure is probably the most obvious difference that meets the eye. Cities differ further in many other dimensions, especially in the functions they perform and in whom inhabits them. In this section, we cover recent studies that look at the interactions between agglomeration economies and the industrial, functional, and skill composition of cities. Abdel-Rahman and Anas (2004) and Duranton and Puga (2000) offer comprehensive treatments of the earlier literature, and many of the results we derive on industry composition belong to it. With respect to industry composition, the production mix of large cities is more diversified than that of small ones (Henderson, 1997; Helsley and Strange, 2014). Also, large and small cities do not specialize in the same sectors, and their industrial composition can change rapidly as there is substantial churning of industries (Duranton, 2007).²⁷ Regarding functional composition, large firms increasingly slice up the value chain and outsource tasks to independent suppliers. Cities of different sizes specialize in different tasks or functions along the value chain, with larger cities attracting the headquarters and small cities hosting production and routine tasks (Duranton and Puga, 2005; Henderson and Ono, 2008). Finally, cities differ in terms of their skill composition. Large cities attract a larger fraction of highly skilled workers than small cities do (Combes et al., 2008; Hendricks, 2011).

²⁶ Behrens et al. (2013) reach the opposite conclusion in a model with heterogeneous agents. Shutting down trade frictions and urban frictions, they find that population reallocations are rather small, but that welfare and productivity gains may be substantial. As pointed out by Behrens et al. (2013), the rather small welfare effects in their model are driven by their assumption of homogeneous agents.

²⁷ Smaller cities usually produce a subset of the goods produced in larger cities. See the “number-average size rule” put forward in the empirical work of Mori et al. (2008).

4.3.3.1 Industry composition

We modify Equation (4.1) as follows. Consider an economy with I different industries. Let p_i denote the price of good i , which is freely traded, and let Y_i denote physical quantities. Then the value of output of industry i in city c is

$$p_i Y_{ic} = p_i \mathbb{J}_c \mathbb{U}_c \mathbb{L}_{ic} \mathbb{A}_{ic} L_{ic}, \quad (4.15)$$

where $\mathbb{L}_{ic}$ now captures the extent of *localization economies* (namely, to what extent local employment in a given industry contributes to scale economies external to individual firms belonging to that industry), $\mathbb{U}_c$ captures the extent of *urbanization economies* (namely, to what extent local employment, whatever its industry allocation, contributes to external scale economies), and $\mathbb{J}_c$ captures the external effects of industry diversity, following [Jacobs \(1969\)](#). In (4.15), we have made the assumption that urbanization and Jacobs externalities affect all sectors in the same way; this is for simplicity and to avoid a proliferation of cases.

An equilibrium in this model requires that (a) workers of any city c earn the same nominal wage in all active industries in that city—that is, $w_c \geq p_i \mathbb{J}_c \mathbb{U}_c \mathbb{L}_{ic} \mathbb{A}_{ic}$ with equality for all i such that $L_{ic} > 0$ —and (b) that they achieve the same utility in all populated cities—that is, $u_c = w_c - L_c^\gamma = u^*$ for some u^* , if $L_c > 0$. The simplest functional forms consistent with localization economies and urbanization economies are $\mathbb{L}_{ic} = L_{ic}^\nu$ and $\mathbb{U}_c = L_c^\epsilon$, respectively. A simple functional form for Jacobs externalities that enables us to encompass several cases studied by the literature is given by

$$\mathbb{J}_c = \left[\sum_{i=1}^I \left(\frac{L_{ic}}{L_c} \right)^\rho \right]^{\frac{1}{\rho}}, \quad (4.16)$$

where $\rho < 1$ is a parameter governing the complementarity among the different industries: ρ is negative when employment levels in various industries are strongly complementary, positive when they are substitute, and tends to unity when variety does not matter (since $\lim_{\rho \rightarrow 1} \mathbb{J}_c = 1$).²⁸ In (4.16), diversification across industries brings external benefits to urban labor productivity. To see this, note that $\mathbb{J}_c \in \{0, 1\}$ if c is fully specialized in some industry, and $\mathbb{J}_c = I^{-1 + (1/\rho)}$ when all industries are equally represented.²⁹ In the latter case, $\mathbb{J}_c > 1$ (diversification raises urban productivity) because $\rho < 1$. Observe also that (4.16) is homogeneous of degree zero by construction so that it is a pure measure of the industrial diversity of cities (size effects are subsumed in $\mathbb{U}_c$ and $\mathbb{L}_{ic}$).

Specialization

Consider first the model of [Fujita and Thisse \(2013, Chapter 4\)](#). In this case, Jacobs and urbanization economies are absent ($\rho = 1$ and $\nu = 0$) and there are no exogenous

²⁸ See [Helsley and Strange \(2011\)](#) for recent microeconomic foundations to Jacobs externalities.

²⁹ If $L_{ic} = L_c$ for some i , then $\mathbb{J}_c = 0$ if $\rho \leq 0$ and $\mathbb{J}_c = 1$ if $\rho > 0$.

differences across sites ($\mathbb{A}_{ic} = \mathbb{A}_i$, for all c). Output of any industry is freely traded among all cities. Thus, there is no benefit in bringing two or more different industries to the same city (Henderson, 1974). A simple proof of this is by contradiction. Assume that an arbitrary city of size L_c is hosting at least two different industries. The per capita urban cost is L_c^γ . Per capita gross income of workers in industry i is equal to $\mathbb{A}_i L_{ic}^\epsilon$. The fact that there is more than one industry in city c implies $L_{ic} < L_c$. Consider next another city c' specialized in industry i , with employment $L_{c'} = L_{ic'} = L_{ic}$. Then, per capita income of workers in industry i net of urban costs is equal to $\mathbb{A}_i L_{ic'}^\epsilon - L_{c'}^\gamma$, which is strictly larger than $\mathbb{A}_i L_{ic}^\epsilon - L_c^\gamma$ because $L_{ic'} = L_{ic}$ and $L_{ic} < L_c$. Hence, a competitive land developer could profitably enter and create a specialized city c' and attract the workers of industry i who are located in city c . No diversified city exists in equilibrium. The unique spatial equilibrium of this model of urban systems has cities specialized by industry, and their (optimal) sizes depend only on the industry in which they specialize. We can therefore label cities by their industry subscripts only and write

Proposition 4.2 (industrial specialization). *Assume that $\rho = 1$, $\nu = 0$, and $\mathbb{A}_{ic} = \mathbb{A}_i$ for all i and all c . Then all cities are specialized by industry at the unique spatial equilibrium with competitive land developers, and their size is optimal:*

$$L_i = \left(p_i \frac{\epsilon}{\gamma} \mathbb{A}_i \right)^{\frac{1}{\gamma - \epsilon}}. \quad (4.17)$$

The proof of the first part (specialization) is given in the text above. The second part follows from the fact that competitive land developers create cities that offer the largest possible equilibrium utility to agents, which, given specialization, yields the same result as in the foregoing section where we considered a single industry. Note that the distribution of $L_c^{\gamma - \epsilon}$ need no longer follow the distribution of $\mathbb{A}_c$ in a multi-industry environment; (endogenous) prices in (4.17) may break the link between the two that Proposition 4.1 emphasizes. Note that cities are fully specialized and yet their size distribution approximately follows Zipf's law in the random growth model of Rossi-Hansberg and Wright (2007).

Industry assignment

The literature on the assignment of industries, occupations, and/or skills to cities dates back to Henderson (1974, 1988). Ongoing work by Davis and Dingel (2014) does this in a multidimensional environment using the tools of assignment theory (Sattinger, 1993; Costinot, 2009).³⁰ Here, we are interested in the assignment of industries to urban sites. In order to connect tightly with the framework we have developed so far, we assume that

³⁰ See also Holmes and Stevens (2014) for an application to the spatial patterns of plant-size distributions, and Redding (2012) for an application to regional inequality and welfare.

industries are distinct in their degree of localization economies, now given by ϵ_i . Furthermore, the suitability of each site for an industry may differ, and there is a large finite set $\mathcal{C} = \{1, 2, \dots, C\}$ of sites. We maintain $\nu = 0$ and $\rho = 1$. We denote by $\mathbb{A}_{ic}$ the site-specific TFP shifter for industry i . Assume that all goods can be traded at no cost, so nominal wage net of urban cost provides a measure of utility. We further assume that all goods are essential—that is, they must be produced in some city. There are local city governments that create cities in order to maximize utility of their residents. Agents are mobile between sectors within each city. We disregard integer constraints and assume that all cities are fully specialized (this is literally true if $\mathcal{C}$ is a continuum).

We solve the problem in three steps. First, we solve for the city size chosen by each local government c conditional on industry i . As shown by [Proposition 4.2](#), if cities are fully specialized then the size chosen by the local government of a city developed at site c and specialized in industry i is given by (4.17). It offers utility

$$u_{ic} = \left(\frac{\gamma}{\epsilon_i} - 1 \right) \left(p_i \frac{\epsilon_i}{\gamma} \mathbb{A}_{ic} \right)^{\frac{\gamma}{\gamma - \epsilon_i}} \quad (4.18)$$

to its residents. Second, local governments choose to specialize their city in the industry that yields the highest utility—namely, they solve $\max_i u_{ic}$. Cities thus specialize according to their comparative advantage. The nature of this comparative advantage is a mixture of Ricardian technology and external scale economies. To see the first part of this statement, let us get rid of differences in external scale economies and temporarily impose $\epsilon_i = \epsilon$ for all i . Consider two cities, c and d . City c specializes in the production of good i and city d specializes in the production of good j if the following chain of comparative advantage holds:

$$\frac{A_{cj}}{A_{ci}} < \frac{p_i}{p_j} < \frac{A_{dj}}{A_{di}}.$$

This is the well-known chain of Ricardian comparative advantage, as was to be shown.

It is not possible to write such an expression for the more interesting case $\epsilon_i \neq \epsilon_j$. The solution here is to tackle the problem as an assignment problem where we match industries to cities following the method developed by [Costinot \(2009\)](#). This is our third and final step. Taking logarithms and differentiating (4.18), one can easily verify that

$$\frac{\partial^2 \ln u_{ic}}{\partial \epsilon_i \partial \mathbb{A}_{ic}} = \frac{\gamma}{(\gamma - \epsilon_i)^2} \frac{1}{\mathbb{A}_{ic}} > 0;$$

that is, utility is log-supermodular in industry-site characteristics $\mathbb{A}_{ic}$ and agglomeration economies ϵ_i . The outcome is then an allocation with positive assortative matching (PAM) between industries and cities. The quality of urban sites and the strength of agglomeration economies are complements: high- $\mathbb{A}_{ic}$ cities specialize in the production of high- ϵ_i goods.

The results above crucially hinge on the complementarity between industries and sites, the presence of local governments (which can exclude migrants from joining a city), and the absence of Jacobs externalities. When agents are free to migrate across cities, and in the presence of cross-industry externalities, [Helsley and Strange \(2014\)](#) show that inefficient coagglomeration of industries generally takes place. Migration is a very weak disciplining device for efficiency. Specialized cities are generally too big, whereas coagglomerated cities are generally too big and do not contain the right mix of industries.³¹ Part of the problem with multiple industries and cross-industry externalities stems from the fact that *distributions matter*—that is, the optimal location of one industry is conditional on the distribution of industries across cities. In that case, (log)-supermodularity may fail to hold, which can lead to many patterns that do not display regular assignments of industries to sites. A similar issue arises in the context of the sorting of heterogeneous workers that we study in [Section 4.4](#).

Urban sectoral specialization fully accounts for city size differences in this model. However, that cities are fully specialized is counterfactual, and so industry specialization cannot be the main ingredient of a reasonable static explanation for Zipf's law (fact 6). The model would at least need to be combined with a “random growth component” in the spirit of [Lee and Li \(2013\)](#), as discussed in [Section 4.3.2.2](#), or some self-selection constraints of heterogeneous workers in the presence of sorting, as discussed in [Section 4.4.1.4](#). Alternatively, we can consider under what conditions cities end up with a diversified industrial structure in equilibrium.

Diversification

In general, the optimal industry composition of urban employment depends on the tension between foregone localization economies and higher urban costs, on the one hand, and the Jacobian benefits of diversity—or citywide “economies of scope” to use the terminology of [Abdel-Rahman and Anas \(2004\)](#)—on the other hand.³² To see this, assume that all industries are symmetric and all sites are homogeneous ($\mathbb{A}_{ic} = \mathbb{A} > 0$, for all c and all i). Then the optimal allocation implies $p_i = p$ for all i . Without further loss of generality, we choose units so that $p\mathbb{A} = 1$. Consider two cities of equal size L . City c is fully specialized ($L_{ic} = L$ for some i , and $L_{jc} = 0$, for all $j \neq i$) and city c' is fully diversified ($L_{ic'} = L/I$ for all i). Urban costs are the same in both cities under our working

³¹ The result regarding the inefficiency of coagglomeration has important implications for empirical research. Indeed, empirical work on agglomeration economies increasingly looks at coagglomeration patterns ([Ellison et al., 2010](#)) to tease out the relative contribution of the different Marshallian mechanisms for agglomeration. The underlying identifying assumption is that the observed coagglomeration is “efficient” so that nominal factor returns fully reflect the presence and strength of agglomeration economies. As shown by [Helsley and Strange \(2014\)](#), this will unfortunately not be the case.

³² See also [Abdel-Rahman and Fujita \(1993\)](#). By assuming free trade among cities, we omit another potential reason for the diversification of cities: to save on transportation costs ([Abdel-Rahman, 1996](#)).

assumption. The nominal wage in city c is equal to $w_c = L^{\epsilon+\nu}$, whereas the nominal wage in city c' is equal to $w_{c'} = L^{\epsilon+\nu} I^{-\epsilon} I^{-1+1/\rho}$ by inserting $\mathbb{J}_{c'} = I^{-1+1/\rho}$ and $L_{ic'} = L/I$ into (4.15). It immediately follows that $w_{c'} > w_c$ if and only if $1 + \epsilon < 1/\rho$ —that is, the optimal city is diversified if the benefits from diversification, $1/\rho$, are large relative to the scope of localization economies, ϵ . Since $\epsilon > 0$, the foregoing case arises only if $\rho < 1$ —that is, if there is complementarity among sectors.³³

4.3.3.2 Functional composition

The slicing up of the value chain across space (offshoring) and beyond firm boundaries (outsourcing) also has implications for the composition of cities (Ota and Fujita, 1993; Rossi-Hansberg et al., 2009). Duranton and Puga (2005) and Henderson and Ono (2008) report that cities are increasingly specialized by function, whereas Rossi-Hansberg et al. (2009) report a similar pattern *within* cities: urban centers specialize in complex tasks and the suburbs specialize in the routine (back office) tasks.

In this subsection, we are interested in the location of the various activities of firms and no longer in the industrial composition of cities. We thus start by considering a single, representative industry. We briefly turn to the multi-industry case at the end of this subsection.

Representative industry

We follow Duranton and Puga (2005) and Ota and Fujita (1993) and consider the location decisions of a firm regarding its various tasks in light of the proximity-localization trade-off. These authors adopt a technological view of the firm in which the costs of coordinating a firm's headquarter and production facilities increase with the geographical distance separating them. Henderson and Ono (2008) report empirical evidence that is consistent with this view. We encapsulate these models into our framework as follows. Each firm conducts headquarter and manufacturing activities, and each activity benefits from its own localization economies. That is to say, the proximity of the headquarters of other firms enhances the productivity of the headquarters of a typical firm, and the proximity of the manufacturing plants of other firms enhances the productivity of its own manufacturing plant. There are two types of tasks, M (for “manufacturing”) and H (for “headquarter”), each being specific to one type of activity. All workers in the economy are equally able to perform either task. Let the subscripts v and f pertain to vertically integrated and to functionally specialized cities, respectively. The output of the representative firm of a typical industry is equal to

³³ The assumption $\rho > 1$ is the opposite to the assumption made by Jane Jacobs and is consistent with Sartre's view that “Hell is other people”—namely, diversity lowers the productivity of everybody. In this case, $\mathbb{J}_c = I^{-1+1/\rho} < 1$ if c is fully diversified and $\mathbb{J}_c = 1$ if c is fully specialized. Clearly, urban labor productivity is higher in the former case than in the latter case. This force comes in addition to urban congestion forces and, therefore, also leads to specialized cities.

$$Y_v = \mathbb{A}(\mathbb{M}\mathbb{M})^\lambda(\mathbb{H}\mathbb{H})^{1-\lambda} \quad (4.19)$$

if this firm locates its headquarter and manufacturing tasks in the same city (i.e., this city is vertically integrated), and $Y_f = Y_v/\tau$ if it locates these units in two distinct cities (i.e., cities are vertically disintegrated). In expression (4.19), $0 < \lambda < 1$ is the share of manufacturing labor in production, M and H are manufacturing and headquarter employment of the representative firm, $\mathbb{M}$ and $\mathbb{H}$ denote localization economies specific to each type of task, and $\tau > 1$ is a Samuelson “iceberg” cost of coordinating remote headquarter and manufacturing activities. As before, the simplest specification for localization economies is $\mathbb{M} = M^\epsilon$ and $\mathbb{H} = H^\nu$, where ϵ and ν are the size elasticities of agglomeration economies specific to plants and to headquarters, respectively. To stress the main insights of the model in the simplest possible way, we impose symmetry between tasks by assuming $\nu = \epsilon$ and $\lambda = 1/2$.³⁴ Let $h \equiv H/(H + M)$ denote the share of workers performing headquarter tasks in production, and let $L \equiv H + M$ denote the size of the workforce. The model being symmetric in H and M , we can anticipate that the optimal allocation is symmetric too. We may write per capita (average) utility as

$$u(\mathbb{I}_v) = \tau^{\mathbb{I}_v-1} \mathbb{A}[(1-h)h]^{\frac{1+\epsilon}{2}} L^\epsilon - \mathbb{I}_v L^\gamma - (1-\mathbb{I}_v) L^\gamma [(1-h)^{1+\gamma} + h^{1+\gamma}], \quad (4.20)$$

where $\mathbb{I}_v = 1$ if firms are spatially vertically integrated and $\mathbb{I}_v = 0$ if headquarter and manufacturing activities are located in distinct, functionally specialized cities. The key trade-off between proximity (due to $\tau > 1$) and local congestion (due to $h^{1+\gamma} + (1-h)^{1+\gamma} < 1$) is clearly apparent in (4.20).

Consider first the case of a *vertically integrated* city—namely, a city that contains vertically integrated firms only ($\mathbb{I}_v = 1$). The optimal size and composition of that city are

$$L_v = \left(\frac{\epsilon}{\gamma} \frac{\mathbb{A}}{2^{1+\epsilon}} \right)^{\frac{1}{\gamma-\epsilon}} \quad \text{and} \quad h_v = \frac{1}{2}, \quad (4.21)$$

respectively. Observe that the expression characterizing the optimal integrated city size in (4.21) is structurally identical to (4.9) in the canonical model.

Turning to the case $\mathbb{I}_v = 0$ of *functional cities*—namely, of cities that specialize fully in either headquarter or manufacturing activities—we again have $h_f = 1/2$, so the optimal headquarter-city and manufacturing-city sizes are given by

³⁴ In practice, agglomeration effects are stronger for high-end services (Combes et al., 2008; Davis and Henderson, 2008; Dekle and Eaton, 1999). Note that $\nu > \epsilon$ would imply that service cities are larger than manufacturing cities, in line with the evidence. It can also explain part of the painful adjustment of many former manufacturing powerhouses such as Detroit and Sheffield. We thank Gilles Duranton for pointing this out to us.

$$H_f = M_f = \left(\frac{\epsilon \mathbb{A}}{\gamma 2\tau} \right)^{\frac{1}{\gamma-\epsilon}}. \quad (4.22)$$

We next compare the normative properties of the allocations in (4.21) and (4.22) by plugging the relevant values into the expressions for $u(\mathbb{I}_v)$ in (4.20). In both cases, congestion costs are equal to a fraction ϵ/γ of output at the optimal allocations. Both output and congestion costs are lower in the allocation with functional cities than in the allocation with vertically integrated cities. Which of the two dominates depends on the parameters of the model. Specifically, average utility (consumption of the numeraire good Y) with vertically integrated cities and cities specialized by function is given by

$$u_v \equiv u(1) = \frac{\gamma-\epsilon}{\epsilon} \left(\frac{\epsilon \mathbb{A}}{\gamma 2^{1+\epsilon}} \right)^{\frac{\gamma}{\gamma-\epsilon}} \quad \text{and} \quad u_f \equiv u(0) = \frac{\gamma-\epsilon}{\epsilon} \left(\frac{\epsilon \mathbb{A}}{\gamma 2\tau} \right)^{\frac{\gamma}{\gamma-\epsilon}}, \quad (4.23)$$

respectively. The following results then directly follow by inspection of (4.21), (4.22), and (4.23):

Proposition 4.3 (functional specialization). *Functional cities are larger than vertically integrated cities and yield higher utility if and only if coordination costs are low enough and/or localization economies are strong enough:*

$$u_f > u_v \quad \text{and} \quad H_f = M_f > L_v \quad \text{if and only if} \quad 1 \leq \tau < \tau_{vf} \equiv 2^\epsilon. \quad (4.24)$$

When coordination costs are low, the output forgone by coordinating manufacturing activities from a remote headquarters is low. If we keep in mind that the congestion cost is a constant proportion of output, it then follows that the size of functional cities, and the per capita consumption of the numeraire good, decreases with the coordination costs. Strong agglomeration economies by function magnify the level of output lost or saved relative to the allocation with vertically integrated cities.

Duranton and Puga (2005) insist on the time-series implication of Proposition 4.3 (see also the chapter by Desmet and Henderson, 2015 in this volume): cities increasingly specialize by function as coordination costs fall over time owing to technical changes in communication technologies. We can also stress the following cross-sectional implication of Proposition 4.3 when industries differ in the scope of agglomeration economies: given τ , an industry with little scope for localization economies (a low ϵ) is more likely to be vertically integrated and to form vertically integrated cities than an industry with a higher ϵ .

Functional composition with several industries

We encapsulate (4.15) and (4.16) into (4.19) in order to study the determinants of the localization of headquarter and manufacturing services of *different* industries in the presence of urbanization and Jacobs externalities. Specifically, consider I symmetric industries with production functions

$$Y_i(\mathbb{I}_v) = \tau^{\mathbb{I}_v - 1} \mathbb{A} (\mathbb{M} M_i)^{\frac{1}{2}} (\mathbb{H} H_i)^{\frac{1}{2}}, \text{ where } \mathbb{M} = \left(\sum_{j=1}^I M_j^\rho \right)^{\frac{\epsilon}{\rho}} \text{ and } \mathbb{H} = \left(\sum_{j=1}^I H_j^\rho \right)^{\frac{\epsilon}{\rho}}.$$

We make two observations about this specification. First, the model is symmetric across industries and production factors. We readily anticipate that any optimal allocation will be symmetric in these variables too. Second, this specification assumes away localization economies. Urbanization economies operate if $\epsilon > 0$ and so do Jacobs economies if $\rho < 1$. Assuming these inequalities hold implies that all industries will be represented in all optimal cities. Then the only relevant question is whether the planner creates vertically integrated cities or functionally specialized cities.

Assume that preferences are symmetric in all goods, so $p_i = p$ for all i . Let $p \equiv 1$ by choice of the numeraire. Output in a vertically integrated city of size L is given by

$$Y_v \equiv \sum_{i=1}^I Y_i(1) = I \mathbb{A} \left[I \left(\frac{L}{2I} \right)^\rho \right]^{\frac{\epsilon}{\rho}} \frac{L}{2I} = \mathbb{A} I^{(\frac{1}{\rho}-1)\epsilon} \left(\frac{L}{2} \right)^{1+\epsilon},$$

where the first equality makes use of the symmetry of the model (and of $M_i = H_i = L/(2I)$ for all i in particular), and the second equality simplifies the expressions. Maximizing per capita output net of urban costs $u = Y/L - L^\gamma$ with respect to L and solving for L yields

$$L_v = \left(\frac{\epsilon \mathbb{A} I^{(\frac{1}{\rho}-1)\epsilon}}{\gamma 2^{1+\epsilon}} \right)^{\frac{1}{\gamma-\epsilon}},$$

which is identical to (4.21) for $I = 1$. We turn now to the joint output of a pair of functional cities (a manufacturing and a headquarter city). Let $M = H = L/2$ denote the (common) size of these cities. Then the joint output is given by

$$Y_f \equiv \sum_{i=1}^I Y_i(0) = \frac{\mathbb{A}}{\tau} I^{(\frac{1}{\rho}-1)\epsilon} \left(\frac{L}{2} \right)^{1+\epsilon}.$$

Maximizing per capita output net of urban costs $u = Y/L - 2(L/2)^\gamma$ with respect to L and solving for $L/2$ yields

$$M_f = H_f = \left(\frac{\epsilon \mathbb{A} I^{(\frac{1}{\rho}-1)\epsilon}}{\gamma 2\tau} \right)^{\frac{1}{\gamma-\epsilon}},$$

which is again identical to (4.22) for $I = 1$. The per capita utility levels u_v and u_f evaluated at the optimal city sizes are proportional to the expressions in (4.23), namely,

$$u_v \equiv u(1) = \frac{\gamma - \epsilon}{\epsilon} \left(\frac{\epsilon \mathbb{A} I^{(\frac{1}{\rho}-1)\epsilon}}{\gamma 2^{1+\epsilon}} \right)^{\frac{\gamma}{\gamma-\epsilon}} \text{ and } u_f \equiv u(0) = \frac{\gamma - \epsilon}{\epsilon} \left(\frac{\epsilon \mathbb{A} I^{(\frac{1}{\rho}-1)\epsilon}}{\gamma 2\tau} \right)^{\frac{\gamma}{\gamma-\epsilon}}.$$

It then immediately follows that the conditions in (4.24) hold in the current setting too. We conclude that cities specialize by function if and only if coordination costs are low enough and/or if urbanization economies are strong enough.

Nursery cities and the life cycle of products

Our framework is also useful to link the life cycle of products to the location of tasks along the value chain. [Duranton and Puga \(2001\)](#) provide evidence from France and the United States that firms locate their innovation activities in *large and diverse* “nursery cities” and afterward relocate the production tasks to smaller manufacturing cities specialized by industry. The reason is that firms face uncertainty and need to discover their optimal production process in the early stages of the product life cycle and afterward want to exploit localization economies in production once they have discovered and mastered the optimal mass production process.

[Duranton and Puga \(2001\)](#) propose a dynamic model with microeconomic foundations that accounts for these facts. It is, however, possible to distill the spirit of their approach using our static framework. The development phase of a product consists of trials and errors and the local experiences of all industries are useful to any other industry: everybody learns from the errors and successes of everyone else.³⁵ Thus, at the innovation stage urbanization and Jacobs economies dominate, while localization economies are relatively unimportant. In the context of Equations (4.15) and (4.16), the presence of urbanization and Jacobs economies at the development stage implies $\nu^I > 0$ (*size matters*) and $\rho^I < 1$ (*diversity matters*), where the superscript I stands for “innovation.” Conversely, localization economies prevail for manufacturing tasks, implying $\epsilon^M > 0$, while urbanization and Jacobs externalities are relatively unimportant at the production stage: $\nu^M = 0$ and $\rho^M = 1$, where the superscript M stands for “manufacturing.”

4.3.3.3 Skill composition

[Hendricks \(2011\)](#) reports that large US cities are relatively skill abundant and that 80% of the skill abundance of a city is unrelated to its industry composition. Put differently, all industries are more skill intensive in large cities than in small cities. Furthermore, the urban premium of skilled workers is unrelated to the industry that employs them, which is suggestive of the existence of human capital externalities that operate broadly across industries in the city (see [Moretti, 2004](#) for a survey of the empirical evidence).

To see how our framework can make sense of these patterns, assume that there are two types of labor in the economy, unskilled workers and skilled workers. Let L_c denote

³⁵ Using a model where the success or failure of firms shapes the beliefs of entrants as to how suitable a region is for production, [Ossa \(2013\)](#) shows that agglomeration may take place even when there are no external effects in production. Large cities may in part be large because they signal to potential entrants that they provide an environment amenable to the successful development of new products.

the size of a city, and h_c denote its fraction of skilled workers. Assume that the per capita output of a representative industry net of urban costs is given by

$$u_c = \mathbb{A}_c [\mathbb{L}_c h_c^\rho + (1 - h_c)^\rho]^{\frac{1}{\rho}} - L_c^\gamma,$$

where $\rho < 1$ and $\mathbb{L}_c = L_c^\epsilon$. This expression assumes skill-biased scale effects, whereas local production amenities $\mathbb{A}_c$ are Hicks neutral as before. Maximizing per capita output net of urban costs with respect to the composition and the size of an arbitrary city yields

$$L_c = \left(\frac{h_c}{1 - h_c} \right)^{\frac{1-\rho}{\epsilon}} \quad \text{and} \quad L_c^{\gamma-\epsilon} = \frac{\epsilon \mathbb{A}_c}{\gamma \rho} h_c^\rho (1 - h_c)^{-\frac{(1-\rho)^2}{\rho}}, \quad (4.25)$$

respectively. City size, L_c , and city skill abundance, h_c , are positively correlated by the first expression in (4.25), and both increase with local amenities $\mathbb{A}_c$ under some regularity condition.³⁶ This generates the positive correlation between skill abundance and city size uncovered by [Hendricks \(2011\)](#).

While the foregoing mechanism relies on the heterogeneity in the TFP terms, $\mathbb{A}_c$, and skill-biased scale effects to generate the positive correlation between size and skills, we now show that the sorting of heterogeneous individuals across cities generates the same relationship without imposing such assumptions.

4.4. SORTING AND SELECTION

Our objective in this section is to propose a framework of sorting of heterogeneous agents *across* cities and selection of heterogeneous agents *within* cities. In what follows, we refer to *sorting* as the heterogeneous location choices of heterogeneous workers or firms. We refer to *selection* as either an occupational choice (workers) or a market-entry choice (firms). Our framework is simple enough to highlight the key issues and problems associated with those questions and to encompass recent models that look at them in greater detail. We also highlight two fundamental difficulties that plague sorting and selection models: the general equilibrium feedbacks that arise in cities and the choice of functional forms. In sorting models, general equilibrium feedbacks preclude in many cases supermodularity, thus making the problem of assignment of heterogeneous agents to cities a fairly complicated one. In selection models, selection effects can go in general

³⁶ Using both expressions to eliminate L_c yields the following implicit equation for h_c as a function of $\mathbb{A}_c$ and of the other parameters of the model:

$$\frac{h_c^{(1-\rho)\frac{\gamma}{\epsilon}-1}}{(1-h_c)^{(1-\rho)(\frac{\gamma}{\epsilon}-\frac{1}{\rho})}} = \mathbb{A}_c \frac{\epsilon}{\rho\gamma}.$$

If $\frac{\gamma}{\epsilon} > \min\{\frac{1}{1-\rho}, \frac{1}{\rho}\}$ then h_c increases with $\mathbb{A}_c$.

either way, thereby precluding clear comparative static results in the absence of specific functional forms. Although several tricks have been used in the literature to cope with both issues, we argue that any analysis of sorting across cities and selection within cities is complicated and unlikely to yield very robust theoretical results. It is here that interactions between theory and empirical analysis become important to select (no pun intended) the “correct” models.

4.4.1 Sorting

We first analyze sorting and show that it is closely related to selection in general equilibrium. This will serve as a basis for the analysis of selection in the next subsection.

4.4.1.1 A simple model

We develop a simple reduced-form extension of the canonical model of [Henderson \(1974\)](#) in which individuals are endowed with heterogeneous ability. Within that model, we then derive (a) a spatial equilibrium with sorting, (b) limiting results when the size elasticity of agglomeration economies, ϵ , and the size elasticity of urban costs, γ , are small, as vindicated by the data, and (c) limiting results on the city size distribution when γ/ϵ is close to 1. We then show how our model encompasses or relates to recent models in the literature that have investigated either the sorting of workers ([Behrens et al., 2014a](#); [Davis and Dingel, 2013](#); [Eeckhout et al., 2014](#)) or the sorting of firms ([Baldwin and Okubo, 2006](#); [Forslid and Okubo, 2014](#); [Gaubert, 2014](#); [Nocke, 2006](#)) across locations. Let $t \in [\underline{t}, \bar{t}]$ denote some individual characteristic that is distributed with probability distribution function $f(\cdot)$ and cumulative distribution function $F(\cdot)$ in the population. For short, we refer to t as “talent.” More able workers have higher values of t . As in the canonical urban model, workers are free to move to the city of their choice. We assume that total population is fixed at $\bar{L}$. The number C of cities, as well as their sizes L_c , are as before endogenously determined by workers’ location choices. Yet, the *talent composition* of each city is now endogenous and determined by the location choices of heterogeneous individuals. Each worker chooses one city in equilibrium, so $\bar{L} = \sum_c L_c$.

We assume that a worker with talent t supplies t^a efficiency units of labor, with $a > 0$. Labor in city c is used to produce a freely traded homogeneous final consumption good under the constant returns to scale technology (4.2). We ignore site heterogeneity by letting $\mathbb{A}_c = \mathbb{A}$ for all c . Hence, $w_c = \mathbb{A}L_c$ is the wage per efficiency unit of labor. Assuming that agglomeration economies depend solely on city size and are given by $\mathbb{L}_c \equiv L_c^\epsilon$, and that preferences are linear, the utility of a type t agent in city c is given by

$$u_c(t) = \mathbb{A}L_c^\epsilon t^a - L_c^\gamma. \quad (4.26)$$

Note the complementarity between talent and agglomeration economies in (4.26): a larger city size L_c disproportionately benefits the most talented agents. This is the basic force pushing toward the sorting of more talented agents into larger cities, and it

constitutes the “micro-level equivalent” of (4.25) in the previous section. Observe that there are no direct interactions between the talents of agents: the sorting of one type into a location does not depend on the other types present in that location. This assumption, used for example in Gaubert (2014) in the context of the spatial sorting of firms, is restrictive yet simplifies the analysis greatly.³⁷ When the payoff to locating in a city depends on the *composition* of that city—which is itself based on the choices of all other agents—things become more complicated. We return to this point in Section 4.4.1.6.

Using (4.26), one can readily verify that the single-crossing property

$$\frac{\partial^2 u_c}{\partial t \partial L_c}(t) > 0 \quad (4.27)$$

holds. Hence, utility is *supermodular* in talent and city size, which implies that there will be PAM in equilibrium (Sattinger, 1993). In a nutshell, agents will sort themselves across cities according to their talent. As can be anticipated from (4.26) and (4.27), not all types of agents will choose the same city in equilibrium. The reason is that urban costs are not type specific, unlike urban premia. Hence, only the more talented agents are able to pay the higher urban costs of larger cities, because they earn more, whereas the less talented agents choose to live in smaller cities, where urban costs are also lower.³⁸

4.4.1.2 Spatial equilibrium with a discrete set of cities

Let $\mathcal{C} = \{1, 2, \dots, C\}$ be an exogenously determined set of cities. Because of PAM in (4.27), we know that agents of similar talent will end up locating in similar cities. Hence, we can look at equilibria that induce a partition of talent across cities. Denote by t_c the talent thresholds that pin down the marginal agent who is indifferent between two consecutive cities c and $c + 1$. By definition of those thresholds, it must be that

³⁷ Gaubert (2014) uses a setting similar to ours yet focuses on the sorting of heterogeneous firms. In her model, trade is costless, which implies that the spatial distribution of firms across cities has no impact on the industry price index. Thus, the location choices of firms are driven by city sizes, and not by the composition of cities in terms of the productivity of the firms they host or the overall spatial distribution of the industry.

³⁸ PAM need not hold in sorting models, especially in general equilibrium. For example, in Mori and Turrini (2005), who build on the work of Krugman (1991), more skilled agents are *less sensitive to market size* because they can more easily absorb the extra costs incurred for trading their good across regions. When trade costs are high enough, this effect may imply that there is a (rather counterfactual) negative relationship between market size and sorting along skills: the more skilled may actually concentrate in the smaller region. Wrede (2013) extends the work of Mori and Turrini (2005) to include housing à la Helpman (1998) and by dropping communication costs. His model is then close to ours and predicts that there is sorting along talent across regions, with the more talented region being larger and commanding higher wages and housing prices. Venables (2011) develops a model of imperfect information in which the most talented workers signal their ability by living in large, expensive cities.

$$\mathbb{A} L_c^\epsilon t_c^a - L_c^\gamma = \mathbb{A} L_{c+1}^\epsilon t_c^a - L_{c+1}^\gamma, \quad \text{so} \quad t_c^a = \frac{1}{\mathbb{A}} \frac{1 - \left(\frac{L_c}{L_{c+1}}\right)^\gamma}{1 - \left(\frac{L_c}{L_{c+1}}\right)^\epsilon} L_{c+1}^{\gamma-\epsilon}. \quad (4.28)$$

As in the canonical model in [Section 4.3.2](#), expressions (4.28) provide only bounds on the distribution of talent and the corresponding city sizes that can be sustained as equilibria. Any equilibrium must exhibit a partition of talent and a monotonic increase in city sizes associated with higher talent because of PAM. Without any coordinating device such as local developers or local governments, a large number of equilibria can be potentially sustained under sorting.

For expositional purposes, let us assume $\epsilon, \gamma \rightarrow 0$ and $\gamma/\epsilon \rightarrow 1$. In words, we assume that the size elasticity of agglomeration economies, ϵ , and the size elasticity of urban costs, γ , are both “small” and of similar magnitude. Although it is debatable what “small” means in numerical terms, the empirical partial correlations of $\hat{\epsilon} = 0.081$ and $\hat{\gamma} = 0.088$ in our data (see [Section 4.2](#)) imply that $\hat{\gamma}/\hat{\epsilon} = 1.068$, which is close to 1, and that the gap $\hat{\gamma} - \hat{\epsilon} = 0.007$ is small and statistically indistinguishable from zero. Recent estimates of γ and ϵ using microdata and a proper identification strategy find even smaller values and a tiny gap $\gamma - \epsilon$ between them ([Combes et al., 2008, 2014](#)). Using the foregoing limit for the ratio on the left-hand side of (4.28), relationship (4.28) can be rewritten as follows:

$$t_c^a \approx \frac{1}{\mathbb{A}} L_{c+1}^{\gamma-\epsilon} \lim_{\epsilon, \gamma \rightarrow 0} \frac{1 - \left(\frac{L_c}{L_{c+1}}\right)^\gamma}{1 - \left(\frac{L_c}{L_{c+1}}\right)^\epsilon} = \frac{1}{\mathbb{A}} \frac{\gamma}{\epsilon} L_{c+1}^{\gamma-\epsilon}. \quad (4.29)$$

Taking ratios, we can express condition (4.29) in c and $c-1$ as follows:

$$\left(\frac{t_c}{t_{c-1}}\right)^a = \left(\frac{L_{c+1}}{L_c}\right)^{\gamma-\epsilon} \Rightarrow L_{c+1} = L_c \left(\frac{t_c}{t_{c-1}}\right)^{\gamma-\epsilon} > L_c, \quad (4.30)$$

where the last inequality comes from $\gamma > \epsilon$ and $t_c > t_{c-1}$. Under our approximation, city size can be directly expressed as a function of the talent of its least talented resident:

$$L_c = L(t_c) = \left(\frac{\epsilon}{\gamma} \mathbb{A} t_c^a\right)^{\frac{1}{\gamma-\epsilon}}. \quad (4.31)$$

Clearly, equilibrium city sizes increase with the talent threshold: more talented cities, with a larger t_c , are bigger in equilibrium.³⁹ Recalling that available estimates of $\gamma - \epsilon$

³⁹ This holds for any partition of talents across cities. Even when there are multiple equilibria, every equilibrium is such that an upward shift of any threshold is accompanied by an increase in city sizes. Clearly, (4.31) depends strongly on the limits. Yet, when the city size distribution has a sufficiently fat upper tail, L_c/L_{c+1} rapidly becomes small, and thus (4.28) implies that $t_c^a \approx L_{c+1}^{\gamma-\epsilon}/\mathbb{A}$. The qualitative implications of (4.31) then approximately carry over to that case.

are a fraction of a percentage point, we find the elasticity $1/(\gamma - \epsilon)$ in the expression above is extremely large: small cross-city differences in talent translate into huge differences in city sizes. More talented cities also have a higher average productivity. Let

$$\bar{t}_c \equiv \left(\int_{t_c}^{t_{c+1}} t^a dF_c(t) \right)^{\frac{1}{a}} \quad (4.32)$$

denote the city's average talent, where $F_c(\cdot)$ is the city-specific talent distribution. We then have $\gamma_c = \mathbb{A}_c L_c^\epsilon$, where $\mathbb{A}_c \equiv \mathbb{A} \bar{t}_c^a$ is the city-specific TFP term, which depends on site characteristics $\mathbb{A}$ —common to all sites in the simple model—and the sites' endogenously determined composition in terms of human capital, $\bar{t}_c$. Hence, productivity gains depend on agglomeration economies in a classical sense (via L_c^ϵ) and via a human capital composition effect (via $\bar{t}_c^a$). The latter accounts for about 40–50% of the observed differences in wages between cities of different sizes (Combes et al., 2008). Turning to utility, from (4.26) we have

$$u_c(t) = \left(\frac{\epsilon}{\gamma} \mathbb{A}_c^a \right)^{\frac{\gamma}{\gamma-\epsilon}} \left[\frac{\gamma}{\epsilon} \left(\frac{t}{\bar{t}_c} \right)^a - 1 \right], \quad \text{so} \quad \bar{u}_c = \gamma_c - L_c^\gamma = \left(\frac{\epsilon}{\gamma} \mathbb{A}_c^a \right)^{\frac{\gamma}{\gamma-\epsilon}} \left[\frac{\gamma}{\epsilon} \left(\frac{\bar{t}_c}{\bar{t}_c} \right)^a - 1 \right].$$

The utility in the first expression is increasing in own talent and ambiguous in the city's minimum talent t_c . On the one hand, a more talented city means more effective units of labor and thus higher productivity *ceteris paribus*, and this benefits all urban dwellers and especially the more talented; see Moretti (2004) for a comprehensive review of the literature on human capital externalities in cities. On the other hand, talented cities are bigger by (4.31) and congestion costs larger, which hurts all urban dwellers equally. The second expression reveals that in the limiting case where $\bar{t}_c/t_c$ is approximately constant across cities (as in Behrens et al. 2014a), average utility is convex in t_c : more talented agents are able to leverage their talent by forming larger cities. We have thus established the following result:

Proposition 4.4 (sorting and city size). *In the simple sorting model, equilibrium city size, L_c , and per capita output, γ_c , are increasing functions of the average talent, $\bar{t}_c$, of the agents located in the city. The equilibrium utility of an agent t located in city c is increasing in own talent t and ambiguous in t_c .*

Figure 4.9 illustrates the sorting of agents across three cities. Agents with the lowest talent pick cities of type 1, which are small. Agents with intermediate talent pick cities of type 2, which are larger. Agents with the highest talent pick cities of type 3, which are larger still. As shown before, the equilibrium relationship between talent and utility—and between talent and city size—is convex. More talented agents gain the most from being in large cities, and large cities must be “sufficiently larger” to discourage less talented agents from going there.

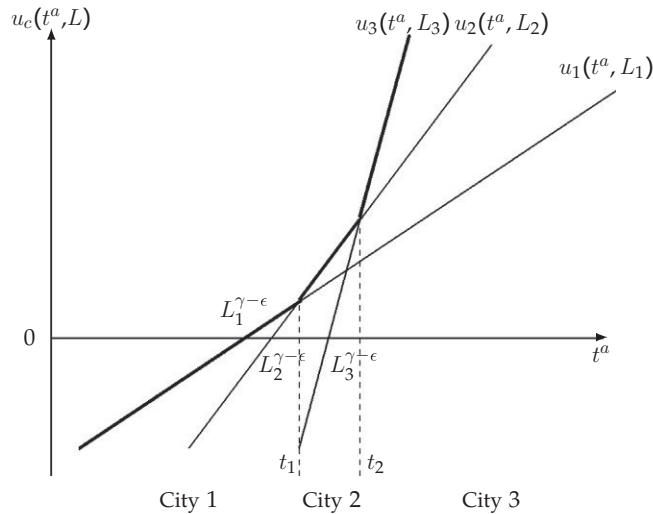


Figure 4.9 Sorting of heterogeneous agents across three cities.

Three remarks are in order. First, the least talented agent pins down the city size that makes that agent indifferent. Any increase in the size of the city would lead the agent to deviate to a smaller city in order to save on urban costs. In each city, more talented individuals naturally receive higher utility. Second, and as a direct consequence of the previous point, the standard condition for a spatial equilibrium in the absence of mobility frictions—namely, the equalization of utility across all locations—breaks down *since no type is generically represented in all cities*. Except for the marginal types who are indifferent between exactly two cities, all agents are strictly better off in the city of their choice.⁴⁰ In words, the ubiquitous condition of equal utility across all populated places naturally ceases to hold in a world where agents differ by type and where different types opt for different locations. The formulation of the spatial equilibrium in (4.6)—“the field’s central theoretical tool” (Glaeser and Gottlieb, 2009, p. 984)—must be modified. This has fundamental theoretical and empirical implications.⁴¹ Lastly, the positive correlation between “talent” and city size is strongly borne out in the data, as can be seen from the left panel in Figure 4.3. Sorting matters!

⁴⁰ Much of the literature has recently moved away from the idea of a simple spatial equilibrium without frictions or heterogeneity and with equalization of utilities across locations. Behrens et al. (2013), Diamond (2013), Gaubert (2014), and Kline and Moretti (2014) all relax this condition either by introducing mobility frictions explicitly or by assuming that agents have locational taste differences. The latter has been previously applied to new economic geography models by, for example, Murata (2003) and Tabuchi et al. (2002) in order to obtain equilibria that vary smoothly with the parameters of the models.

⁴¹ For instance, regressing individual earnings on a measure of citywide average human capital leads to biased results in the presence of self-selection of agents across locations (this bias is positive if agents with similar abilities make similar choices because the error term is positively correlated with $\tilde{t}^a$).

In the foregoing, we looked at “discrete cities,”—that is, cities that span some talent range $[t_c, t_{c+1}]$. Discrete cities induce a discrete partition of the talent space. Though this is empirically relevant because cities host agents of multiple talents, the downside is that the model is quite hard to work with since there is a continuum of equilibria. To solve the model implies specifying a partition, solving for relative city sizes, and choosing a scale for absolute city sizes (by specifying the outside option). Depending on the choice of partition and scale, a multitude of equilibria may be sustained. Part of the problem comes from the fact that we assign a predetermined city structure to agents and then check the equilibrium conditions. Alternatively, we may consider a setting without any predetermined structure in which agents can form any type of city in terms of size and composition.

4.4.1.3 Spatial equilibrium with a continuum of cities

Assume next that agents can choose cities optimally in the sense that they decide—conditional on their talent—which city size they prefer to live in. Formally, an agent with talent t maximizes his or her utility with respect to city size—that is, the agent picks one city size from the menu of all possible city sizes. Here, we assume that the set of cities $\mathcal{C} = [0, C]$ is a continuum. All cities can potentially be formed and the mass (number) of cities C is an endogenous variable. This is essentially the model developed by [Behrens et al. \(2014a\)](#). The first-order condition of that problem is given by⁴²

$$\max_{L_c} u_c(t) \Rightarrow \mathbb{A}\epsilon L_c^{\epsilon-1} t^a - \gamma L_c^{\gamma-1} = 0, \quad (4.33)$$

which yields the preferred city size of agents with talent t :

$$L_c(t) = \left(\frac{\epsilon}{\gamma} \mathbb{A} t^a \right)^{\frac{1}{\gamma-\epsilon}}. \quad (4.34)$$

It is easily verified that the second-order condition holds at the equilibrium city sizes.

Five comments are in order. First, comparing Equations (4.31) and (4.34) reveals that they have the same structure. The difference is that (4.31) applies to the marginal agent, whereas (4.34) applies to any agent. The equilibrium with a large number of discrete cities approaches the one where agents can sort across a continuum of cities.

⁴² It is here that the assumption that the city composition does not matter becomes important. In general, the problem of an agent would involve two dimensions: the choice of a city size, and the choice of a city composition. The latter makes matters complicated. [Behrens et al. \(2014a\)](#) simplify the problem by focusing on “talent-homogeneous” cities—that is, cities which host only one type of talent. In that case, solving for $L_c(t)$ involves solving a differential equation. In our simple model, the talent composition does not matter, so size is the only choice variable and cities will trivially be “talent homogeneous,” as shown by (4.34).

The intuition is that in the continuous model, all agents are almost indifferent between cities of similar sizes. Yet, every agent has his or her own preferred size, depending on his or her talent.

Second, (4.34) gives a relationship that uniquely maps talents into city size: two different agents would optimally choose to not live in a city of the same size. This significantly narrows down the composition of cities in terms of talents: cities are talent homogeneous, and PAM implies that more talented agents choose to live in larger cities. We trace out the implications of this for the city size distribution in the next subsection. Since every agent picks his or her preferred city, this is a stable equilibrium in the sense that no one can profitably deviate. There are potentially many equilibria with a partition of talent across cities (see the discrete setting in the previous subsection), but in that case not all agents live in a city of the size they would prefer had they the choice of city size. How such an equilibrium, where agents can form the number of cities they wish and each agent chooses to live in a city with his or her preferred size, is actually implemented in the static model is an open question.

Third, having talent heterogeneity and a continuum of cities convexifies the problem of allocating agents to cities. We can think about this convexification as follows. In the discrete case, the utility of type t in city c is $u_c(t) = \mathbb{A}L_c^\epsilon(t^a - t_c^a\epsilon/\gamma)$, which is a linear function of t^a (recall that L_c depends only on the marginal type t_c). A change in L_c in city c will change the talent composition of that city (see Figure 4.9), yet can be sustained as an equilibrium if the change in L_c is not too large: city sizes are not uniquely determined. In the continuous case, the utility of type t in a city of optimal size is $u_c(t) = \mathbb{A}L_c^\epsilon t^a(1 - \epsilon/\gamma) = (\epsilon/\gamma)^{\epsilon/(\gamma-\epsilon)}(\mathbb{A}t^a)^{\gamma/(\gamma-\epsilon)}(1 - \epsilon/\gamma)$, which is a strictly convex function of t^a . The convexification stems from the fact that an increase in talent raises utility more than linearly as city size changes with the talent of its representative urban dweller. Contrary to the discrete case, the size–talent relationship is uniquely determined. Intuitively, a city cannot grow larger or smaller than (4.34) because of the existence of arbitrarily similar cities in terms of size and talent to which agents could deviate to get higher utility.

Fourth, per capita output in a type t city is given by $\gamma_c = \mathbb{A}L_c^\epsilon t^a$. If we take logarithms, this becomes either

$$\ln \gamma_c = \kappa_1 + \epsilon \ln L_c + a \ln t_c \quad (4.35)$$

or

$$\ln \gamma_c = \kappa_2 + \gamma \ln L_c, \quad (4.36)$$

where (4.36) is obtained by making use of (4.34). Hence, a log–log regression of productivity γ_c on size L_c yields either the elasticity of agglomeration economies in (4.35), where sorting is controlled for, or the elasticity of urban costs in (4.36), where sorting is not controlled for.

Last, taking logarithms of (4.34), we obtain $\ln t_c = \kappa + \frac{\gamma-\epsilon}{a} \ln L_c$, where κ is some constant term. When $\gamma - \epsilon$ is small, the elasticity of talent with respect to city size is small: the size elasticity of “education” with respect to city size is 0.117 in our US data (see the left panel in Figure 4.3). The fact that large cities are only slightly more “talented”—as measured by educational attainment of the city population—is the mirror image of the property that small differences in education have to be offset by large differences in city sizes. Thus, a small elasticity of talent with respect to city size is in no way indicative that sorting is unimportant, as some authors have sometimes argued.

4.4.1.4 Implications for city sizes

As shown before, the sorting of heterogeneous individuals across cities gives rise to cities of different equilibrium sizes. What does the theory imply for the *size distribution* of cities? We now use the model with a continuum of cities to show that the implications for that distribution are striking. Observe first that the “number” of agents of talent t in the population is given by $\bar{L}f(t)$. As shown before, agents of talent t prefer cities of size $L(t)$ as given by (4.34). Assume that $n(t)$ of such cities form. Since all agents choose a city in equilibrium, it must be the case that $\bar{L}f(t) = n(t)L(t)$ or, equivalently,

$$n(t) = \frac{\bar{L}f(t)}{L(t)}. \quad (4.37)$$

Let C denote the total mass of cities in the economy. The cumulative distribution $N(\cdot)$ of cities is then given by

$$N(\tau) = \frac{\bar{L}}{C} \int_0^\tau \frac{f(t)}{L(t)} dt.$$

Using the relationship between talent and size (4.34), we have

$$\frac{f(t)}{L(t)} = \frac{f\left(\xi L(t)^{\frac{\gamma-\epsilon}{a}}\right)}{L(t)} \quad \text{and} \quad dL = \frac{a}{\xi(\gamma-\epsilon)} L(t)^{1-\frac{\gamma-\epsilon}{a}} dt,$$

where $\xi \equiv \left(\frac{\epsilon}{\gamma} \mathbb{A}\right)^{-\frac{1}{a}}$ is a positive bundle of parameters. With use of the distribution of talent and the change in variable from talent to city size, the density and the cumulative distribution of city sizes are given by

$$n(L) = \frac{\bar{L}\eta\xi}{C} f(\xi L^\eta) L^{\eta-2} \quad \text{and} \quad N(L) = \frac{\bar{L}\eta\xi}{C} \int_0^\ell f(\xi \ell^\eta) \ell^{\eta-2} d\ell, \quad (4.38)$$

with $\eta \equiv \frac{\gamma-\epsilon}{a}$. The first-order approximation of (4.38) around $\eta = 0$ is given by

$$n(L) = \kappa L^{-2}, \quad (4.39)$$

where $\kappa \equiv \frac{\bar{L}\eta\xi}{C}f(\xi) > 0$ is a positive constant (recall that η remains positive). Using this expression and the full-employment condition, $\bar{L} = \int_{L(\underline{t})}^{L(\bar{t})} n(L)LdL$, and solving for the equilibrium mass of cities yields

$$C = \eta\xi f(\xi) [\ln L(\bar{t}) - \ln L(\underline{t})] \bar{L};$$

that is, the number of cities is proportional to the size of the population. The urban system displays constant returns to scale in equilibrium. Thus, by inspection of Equation (4.39), we can show (Behrens et al., 2014a).

Proposition 4.5 (Zipf's law). *Assume that agents sort across cities according to (4.34). Then the size distribution of cities follows a Pareto distribution with shape parameter -1 in the limit $\eta \equiv \frac{\gamma-\epsilon}{a} \rightarrow 0$.*

The right panel in Figure 4.6 illustrates that relationship. That Zipf's law holds in this model is remarkable because it *does not depend on the underlying distribution of talent in the population*. In other words, when $\gamma - \epsilon$ is small—as seems to be the case in the data—the city size distribution in the model converges to Zipf's law irrespective of the underlying talent distribution.⁴³ Crucial for obtaining this result are two relatively reasonable requirements. First, the “number” of cities—more precisely the mass of cities—associated with each level of talent is endogenously determined. Second, city sizes are also endogenously determined and agents can sort themselves across cities of their preferred type. Since agents of any type t have a preferred city size that is a continuous function of their talent, taking that talent to a sufficiently large power implies that the resulting city size distribution is of the Zipf type.

Random growth models also (approximately) generate Zipf's law in the steady state if Gibrat's law holds. The latter has been challenged lately on empirical grounds (see Michaels et al., 2012). Desmet and Rappaport (2013) show that Gibrat's law appears to settle once the distribution is of the Zipf type (and not the other way round). The model in this subsection displays one possible mechanism to generate Zipf's law, like the models in Hsu (2012) and Lee and Li (2013).⁴⁴ One distinct advantage of our model is that it generates Zipf's law for plausible values of the parameters irrespective of the underlying distribution of talent (which we do not observe).

4.4.1.5 Some limitations and extensions

The model developed in Section 4.4.1.1 has the virtue of simplicity. The flip side is that it naturally has a number of shortcomings. Firstly, like almost any model in the literature

⁴³ Behrens et al. (2014a) show that convergence to Zipf's law is very fast as η gets smaller. For empirically plausible values of η , the simulated city size distribution is indistinguishable from a Pareto distribution with unitary shape parameter.

⁴⁴ Hsu (2012) also generates Zipf's law using a static framework. The mechanism, based on central place theory and fixed costs, is however very different from the other two models reviewed here.

(e.g., [Mori and Turrini, 2005](#); [Nocke, 2006](#); [Baldwin and Okubo, 2006](#); [Okubo et al., 2010](#)), it predicts *strict sorting* along a single dimension. Yet, it is well known that there is a significant overlap of productivities in cities. Larger cities host, on average, more able agents, yet there is nothing close to a clear partition along firm productivity and individual education across cities in the data ([Combes et al., 2012](#); [Eeckhout et al., 2014](#); [Forslid and Okubo, 2014](#)). For example, although the correlation between the share of highly skilled workers and city size in the United States is statistically very significant (see the left panel in [Figure 4.3](#)), the associated R^2 in the log–log regression is only 0.161.⁴⁵

Our simple model with a continuum of cities can easily be extended in the spirit of [Behrens et al. \(2014a\)](#) to allow for incomplete sorting along productivity. The idea is to have a two-stage process, where agents sort on an *ex ante* signal (their talent), but where *ex post* productivity is uncertain. Assume that after choosing a city c , each agent gets hit by a random productivity shock $s \in [0, \bar{s}_c]$, with cumulative distribution function $G_c(\cdot)$. We can think about s as being luck or “serendipity”—the agent is in the right place at the right time. The efficiency units of labor the agent can supply depend on the agent’s talent t and the shock s in a multiplicative way: $\varphi \equiv s \times t$. Denote by $\Phi_c(\cdot)$ the distribution of *productivity* in city c . Clearly, even two cities with similar yet different talent compositions will end up having largely overlapping productivity distributions. We then have the following expected wage in city c with average talent $\bar{t}_c$ defined in (4.32):

$$\mathbb{E}w_c(t) = \mathbb{A}L_c^\epsilon \int_0^{\bar{t}_c \bar{s}_c} \varphi^a d\Phi_c(\varphi) = \underbrace{\mathbb{A} \left(\int_0^{\bar{s}_c} s^a dG_c(s) \right)}_{=\mathbb{A}_c(\mathbb{A}, \bar{t}_c, G_c(\cdot))} \bar{t}_c^a L_c^\epsilon.$$

Clearly, the TFP term $\mathbb{A}_c$ is city specific and a function of sorting and of a city-specific distribution of shocks, and there is a nondegenerate distribution of wages and productivities in all cities. The distribution of productivity of cities endowed with highly talented individuals stochastically dominates the distribution of less talented cities.⁴⁶

Another way to generate incomplete sorting is to assume that agents choose locations on the basis of a random component in their objective function, as in [Behrens et al. \(2013\)](#) or [Gaubert \(2014\)](#). The idea is that the location choices of consumers and firms have a deterministic component (profit or indirect utility) as well as a probabilistic component. Under standard assumptions on the distribution of the probabilistic component—if it

⁴⁵ Sorting by skills in the United States increased between 1980 and 2000. [Diamond \(2013\)](#) studies its consequences for welfare inequality.

⁴⁶ It may be reasonable to assume that the shocks may be, on average, better in larger cities as the result of various insurance mechanisms, better opportunities, etc. This is an additional force pushing toward sorting through the TFP terms: more talented agents will go to places with better shocks since they stand to gain more from good shocks and to lose less from bad shocks.

follows a type I extreme value distribution—location choice *probabilities* are then of the logit form and allow for incomplete sorting across locations: observationally identical agents need not make the same location decisions. More talented agents will, on average, pick larger cities, but the distribution of types is fuzzy across cities. The same result can be achieved by including a deterministic type-independent “attachment to home” component as in [Wrede \(2013\)](#).

Finally, the foregoing models predict PAM: larger cities host, on average, more talented individuals, and the productivity distribution in larger cities first-order stochastically dominates that in smaller cities. However, some recent empirical evidence documents that the right *and* the left tails for the productivity distributions of French workers ([Combes et al., 2012](#)), US workers ([Eeckhout et al., 2014](#)), and Japanese firms ([Forslid and Okubo, 2014](#)) are both fatter in larger cities. In other words, larger markets seem to attract both the most and the least productive workers and firms. Large cities are thus more unequal since they host a disproportionate share of both highly productive and poorly productive agents. While the empirical evidence on two-way sorting is certainly intriguing and points to the existence of some nontrivial complementarities, existing models of two-way sorting still fall short of providing either theoretically plausible or empirically testable mechanisms.⁴⁷ The over representation of the left tail of skills in larger cities could be due to many things, including more generous welfare policies, complementarities between skilled and unskilled workers (e.g., rich households employing unskilled workers for housekeeping and child care activities), greater availability of public housing, effects of migrants, or the presence of public transportation as pointed out by [Glaeser et al. \(2008\)](#). As we argue in the next section, complex general equilibrium effects in the presence of selection effects can generate supermodularity for the upper tail and submodularity for the lower tail of the skill distribution. While the jury is not yet in as to what may drive two-way sorting, we believe that more work is needed in that direction.

4.4.1.6 Sorting when distributions matter (a prelude to selection)

In the simple model in [Section 4.4.1.1](#), individuals make location choices by looking at the sizes and average talent of cities only: a more talented city is a city endowed with more efficiency units of labor per capita. *Per se*, there are no benefits or drawbacks associated with living in a talented city. Yet, there are a number of reasons to believe that the talent composition of a city directly matters for these choices in subtler ways. On the one hand,

⁴⁷ Whether or not the patterns in the data are due to “two-way sorting” or “sorting and selection” is a priori unclear, as we will emphasize in the next section. There may be one-way sorting—larger markets attract more able agents—but selection afterward fails a certain share of them. Those agents end up as low-productivity ones, a pattern that we see in the data.

locating in a city with more talented entrepreneurs may provide a number of upsides, such as access to cheaper intermediates or higher wages for workers. It may also allow more productive interactions among workers, who learn from each other, especially when the quality of learning depends on the talent of the other agents (Davis and Dingel, 2013). Locating in a place with many talented people may, on the other hand, also have its downsides. Most notably, it toughens up competition since any agent has to compete against more numerous and more talented rivals. Whatever the net effect of the pros and cons, it should be clear that, in general, the location decision of any agent is at least partly based on where other agents go—that is, sorting is endogenous to the whole distribution of talent across cities. Sorting when the whole distribution of talent matters is formalized in both Behrens et al. (2014a) and Davis and Dingel (2013). Behrens et al. (2014a) consider that agents sort across cities on the basis of their talent. As in Section 4.4.1.5, productivity φ is the product of “talent” and “luck.” Agents who are productive enough—their productivity exceeds some endogenous city-specific *selection cutoff* $\underline{\varphi}_c$ —become entrepreneurs and produce local intermediates that are assembled at the city level by some competitive final sector using a CES aggregator. They earn profits $\pi_c(\varphi)$. The remaining agents become workers and supply φ^a units of efficient labor, as in our simple model, and earn $w_c \varphi^a \leq \pi_c(\varphi)$. In that context, wages and per capita output in city c are, respectively, given by

$$w_c = \frac{1}{1+\epsilon} \left(\int_{\underline{\varphi}_c}^{\infty} \varphi^{\frac{1}{\epsilon}} d\Phi_c(\varphi) \right)^{\epsilon} L_c^{\epsilon} \quad \text{and} \quad y_c = \underbrace{\left(\int_{\underline{\varphi}_c}^{\infty} \varphi^{\frac{1}{\epsilon}} d\Phi_c(\varphi) \right)^{\epsilon} \left(\int_0^{\underline{\varphi}_c} \varphi^a d\Phi_c(\varphi) \right)}_{=\mathbb{A}_c(\underline{\varphi}_c, \Phi_c)} L_c^{\epsilon}, \quad (4.40)$$

where $\Phi_c(\cdot)$ is the city-specific productivity distribution. Observe that the TFP term $\mathbb{A}_c$ is endogenous and depends on sorting (via the productivity distribution Φ_c) and selection (via the cutoff $\underline{\varphi}_c$). The same holds true for wages. This affects the location decisions of heterogeneous agents in nontrivial ways. In the model of Behrens et al. (2014a), the random shocks s occur after a city has been chosen. Individuals' location decisions are thus based on the expected utility that an agent with talent t obtains in all cities. For some arbitrary city c , this expected utility is given by

$$\mathbb{E}u_c(t) = \int_0^{\bar{s}_c} \max\{\pi_c(st), w_c(st)^a\} dG_c(s) - L_c^{\gamma}.$$

It should be clear from the foregoing expression that a simple single-crossing property $\frac{\partial^2 \mathbb{E}u_c}{\partial t \partial L_c}(t) > 0$ need not generally hold. The reason is that both the selection cutoff $\underline{\varphi}_c$ and the whole productivity distribution $\Phi_c(\cdot)$ depend on the city size L_c in general equilibrium. As shown in Section 4.4.2, it is generally not possible to assess whether larger

markets have tougher selection ($\partial \underline{q}_c / \partial L_c > 0$) or not. Thus, it is also *a priori* not possible to make clear statements about sorting: PAM does not hold in general.

Another way in which the talent composition of a city may matter for sorting is when there are learning externalities. Consider the following simplified variant of the model of [Davis and Dingel \(2013\)](#). There are two types of workers. The first type produces non-tradable goods under constant returns to scale and no externalities. The second type produces some costlessly traded good. Productivity in that sector is subject to learning externalities. Each worker has t units of efficient labor, which can be used either for work or for learning from others. In equilibrium, workers with $t \geq \underline{t}_c$ engage in the production of traded goods in city c , whereas the others produce nontraded goods. In other words, the model features occupational selection. Let $\beta \in (0, 1)$ denote the share of time a worker devotes to learning (this is a choice variable). The output of a type t worker in city c employed in the traded sector is given by⁴⁸

$$y_c(t) = (\beta t)^{\alpha_c} [(1 - \beta) t \mathbb{L}_c]^{1 - \alpha_c}, \quad (4.41)$$

where the first part is the output from allocating time to work, and where the second part is the productivity-enhancing effect of learning. Here, $\alpha_c \in (1/3, 1/2)$ is a city-specific parameter that subsumes how important learning is for an agent's productivity. Expression (4.41) reveals the basic force pushing toward ability sorting: more talented agents benefit more from larger learning externalities.

Maximizing (4.41) with respect to β yields $\beta^* = \frac{\alpha_c}{1 - 2\alpha_c}$, which increases with α_c and is independent of talent.⁴⁹ The learning externality, $\mathbb{L}_c$, depends on the time that all agents in the city allocate to that activity (a scale effect), and to the average talent of agents in the city (a composition effect). Let us assume that

$$\mathbb{L}_c = \mathbb{L}_c^\epsilon \cdot \bar{t}_c, \quad \text{where} \quad \mathbb{L}_c = L_c \int_{t \geq \underline{t}_c} (1 - \beta_c) dF_c(t) \quad \text{and} \quad \bar{t}_c = \frac{1}{1 - F_c(\underline{t}_c)} \int_{t \geq \underline{t}_c} t dF_c(t) \quad (4.42)$$

are the scale and the composition effects, respectively. The former effect can be computed as $\mathbb{L}_c = L_c \frac{1 - 3\alpha_c}{1 - 2\alpha_c} [1 - F_c(\underline{t}_c)]$ and implies that there is greater potential for spillovers when more agents engage in learning. The second effect implies that the quality of learning increases with the average talent of those who are engaged in learning. Both depend on the selection of agents, as captured by the selection threshold $\underline{t}_c$.

Substituting β^* and expressions (4.42) into (4.41), we obtain the average productivity in city c :

⁴⁸ This specification rules out the “no learning” equilibria that arise in [Davis and Dingel \(2013\)](#). Those equilibria are of no special interest.

⁴⁹ Although it may seem reasonable to consider that more talented workers stand to gain more from learning as in [Davis and Dingel \(2013\)](#) and should thus choose higher β values in equilibrium, our assumption simplifies the model while still conveying its key insights.

$$\gamma_c = \underbrace{\kappa_c \bar{t}_c^{2-\alpha_c} [1 - F_c(\underline{t}_c)]^{\epsilon(1-\alpha_c)+1}}_{=\mathbb{A}_c(\underline{t}_c, F_c)} L_c^{\epsilon(1-\alpha_c)}, \quad (4.43)$$

where κ_c is a term that depends on α_c , β , and ϵ . The TFP term $\mathbb{A}_c$ again depends on the endogenous allocation of talents across cities, $F_c(\cdot)$, and selection into occupations within cities (as captured by $\underline{t}_c$). In general, the threshold is itself a function of city size and the distribution of talent across cities. In a nutshell, $\underline{t}_c$, $F_c(\cdot)$, and L_c are simultaneously determined at the city level, and the locational equilibrium condition, whereby each agent picks his or her preferred location, must hold. Note the similarity between (4.40) and (4.43). Both models predict that sorting and selection interact to determine the productivity advantage of cities. We return to this point below.

Although the sorting of workers across cities has attracted the most attention, a growing literature looks at the sorting of firms (see, e.g., Baldwin and Okubo, 2006; Forslid and Okubo, 2014; Nocke, 2006; Okubo et al., 2010). In a subnational context, we can think about the sorting of firms in the same way as we think about the sorting of entrepreneurs since it is fair to say that most firms move with the people running them.⁵⁰ Gaubert (2014) assumes that a firm's realized productivity is given by $\psi(t, L_c)$, where t is the firm's intrinsic productivity. The latter interacts, via ψ , with agglomeration economies with city size L_c as a proxy. With use of a simple single-sector variant of Gaubert's multi-industry CES model, the profit of a firm with productivity t is given by

$$\pi_c(t) = \mathbb{A}_c \mathbb{P}_c^{\sigma-1} \left(\frac{\psi(t, L_c)}{w_c} \right)^{\sigma-1}, \quad (4.44)$$

where $\mathbb{A}_c$ is a city-specific TFP shifter, $\mathbb{P}_c$ is the city-specific CES price aggregator, w_c is the city-specific wage, and $\sigma > 1$ is the demand elasticity. As can be seen from (4.44), the firm-level productivity t interacts with city size L_c both directly, via the reduced-form function ψ , and indirectly via the citywide variables $\mathbb{A}_c$, $\mathbb{P}_c$, and w_c . Taking logarithms of (4.44) and differentiating, and noting that none of the citywide variables $\mathbb{A}_c$, $\mathbb{P}_c$, and w_c depend on a firm's individual t , we see that the profit function is *log-supermodular* in t and L_c if and only if ψ is log-supermodular:

⁵⁰ Empirical evidence suggests that the bulk of the spatial differences in wages is due to the sorting of workers (Combes et al., 2008), with only a minor role for the sorting of firms by size and productivity (Mion and Naticchioni, 2009). Furthermore, it is difficult to talk about the sorting of firms since, for example, less than 5% of firms relocate in France over a 4-year period (Duranton and Puga, 2001). Figures for other countries are fairly similar, and most moves are short distance moves within the same metro area. Entry and exit dynamics thus drive observed patterns, and those are largely due to selection effects.

$$\frac{\partial^2 \ln \pi_c(t)}{\partial L_c \partial t} > 0 \Leftrightarrow \frac{\partial^2 \ln \psi(t, L_c)}{\partial L_c \partial t} > 0.$$

In words, the profit function inherits the log-supermodularity of the reduced-form productivity function ψ , which then implies that more productive firms sort into larger cities.

Four comments are in order. First, this sorting result generically holds only if profits are log-linear functions of citywide aggregates and ψ . The latter is the case with CES preferences. Relaxing CES preferences implies that individual profit is generically not multiplicatively separable in ψ and L_c ; in that case, log-supermodularity of ψ is neither necessary nor sufficient to generate log-supermodularity of π . Second, log-linearity of profits implies that only the direct interactions between t and L_c matter for the sorting of firms. If we relax the (relatively strong) assumption of log-supermodularity of ψ , the model by [Gaubert \(2014\)](#) would also be a model of sorting where the (endogenous) productivity distribution of cities influences location choices in a nontrivial way. As such, it would be extremely hard to solve as we argue in the next subsection. Third, with proper microeconomic foundations for sorting and selection (more on this below), it is not clear at all that ψ is log-supermodular in t and L_c in equilibrium. Fourth, in general equilibrium, the indirect interactions of city size via $\mathbb{P}_c$ and w_c with the individual t may suffice to induce sorting. For example, in the model with an inelastic housing stock as in [Helpman \(1998\)](#), $w(L_c)$ is an increasing function of L_c to compensate mobile workers for higher housing costs. This has opposite effects on profits (higher costs reduce profits, but there are citywide income effects) which may make larger cities more profitable for more productive agents and thereby induce sorting. How these general equilibrium effects influence occupational choice and interact with sorting is the focus of the next subsection.

4.4.2 Selection

We now touch upon an issue that has rightly started attracting attention in recent years: selection. Before proceeding, it is useful to clarify the terminology. We can think of two types of selection: *survival selection* and *occupational selection*. Survival selection refers to a *stochastic selection* of the Hopenhayn–Melitz type where entrants have to pay some sunk entry cost, then discover their productivity, and finally decide whether or not to stay in the market ([Hopenhayn, 1992](#); [Melitz, 2003](#); [Melitz and Ottaviano, 2008](#); [Zhelobodko et al., 2012](#)). Occupational selection refers to a *deterministic selection* where agents decide whether to run firms or to be workers, depending on their talent ([Lucas, 1978](#)).⁵¹ For

⁵¹ In a spatial context, the former has been investigated by [Ottaviano \(2012\)](#), [Behrens et al. \(2014b\)](#), and [Behrens and Robert-Nicoud \(2014b\)](#). The latter has been analyzed by [Davis and Dingel \(2013\)](#), [Behrens et al. \(2014a\)](#), and [Behrens et al. \(2014c\)](#).

simplicity, we deal only with occupational selection in what follows.⁵² The selection cut-off t_c for talent in city c then determines how agents are split among different occupational groups (firms or entrepreneurs vs. workers).

Our aim is not to provide a full-fledged model of selection, but rather to distill some key insights. Our emphasis is on the interactions between selection, sorting, and agglomeration. We show in this section that selection and sorting are causally linked, observationally equivalent, and, therefore empirically very difficult to disentangle (Combes et al., 2012). We also show that the impact of market size on selection is generally ambiguous in economic models—that is, it is unclear whether larger markets have more or fewer firms (entrepreneurs) and whether market size is associated with a procompetitive effect. This result is largely due to the general equilibrium interactions between selection, sorting, and agglomeration.

4.4.2.1 A simple model

While sorting can be studied under fairly general assumptions, studying selection requires imposing more structure on the model. More precisely, we need a model in which the *relative position* of an agent—as compared with the other agents in the market—matters. Models of imperfect competition with heterogeneous agents usually satisfy that requirement. Selection can thus be conveniently studied in general equilibrium models of monopolistic competition with heterogeneity, where the payoff to one agent depends on various characteristics such as market size, the skill composition of the market, and the number of competitors. Developing a full model is beyond the scope of this chapter, but a simple reduced-form version will allow us to highlight the key issues at hand.

Consider a set of heterogeneous producers (entrepreneurs) who produce differentiated varieties of some nontraded consumption good or service in city c . We denote by $F_c(\cdot)$ the cumulative distribution of talent in city c , with support $[\underline{t}_c, \bar{t}_c]$. To make our point clearly, we take that distribution, and especially $\bar{t}_c$, as *given* here—that is, we ignore sorting across cities. The reason is that sorting and selection are difficult to analyze jointly. We discuss the difficulties of allowing for an endogenous talent distribution $F_c(\cdot)$, as well as the interaction of that distribution with selection, later in this section.

Workers earn w_c per efficiency unit of labor, and workers with talent t supply t^a efficiency units. We assume that entrepreneurial productivity increases with talent. We further assume that talented individuals have a comparative advantage in becoming entrepreneurs (this requires entrepreneurial earnings to increase with t at a rate higher than a), so the more talented agents (with $t > t_c$) operate firms as entrepreneurs in

⁵² See Melitz and Redding (2014) for a recent review of survival selection in international trade. Mrázová and Neary (2012) provide additional details on selection effects in models with heterogeneous firms.

equilibrium. We refer to t_c as the *occupational selection cutoff* (or cutoff, for short). An entrepreneur with talent t hires $1/t$ efficiency units of labor to produce a unit of output. Entrepreneurs maximize profits, which we assume are given by

$$\pi_c(t) = \left(p_c(t) - \frac{w_c}{L_c^\epsilon t} \right) L_c x_c(t), \quad (4.45)$$

where $p_c(t)$ is the price of the variety sold by the entrepreneurs, L_c^ϵ is a reduced-form agglomeration externality, and $L_c x_c(t)$ is the total demand faced by the entrepreneur in city c , $x_c(t)$ being the per capita demand.⁵³ Observe from expression (4.45) the complementarity between entrepreneurial talent, t , and the agglomeration externality, L_c^ϵ . As argued before, this is a basic force pushing toward sorting along skills into larger cities. However, in the presence of selection, things are more complicated since profits depend in a nontrivial way on market size in general equilibrium. As shown in the next section, the complementarity is also a basic force that dilates the income distribution of entrepreneurs and, therefore, leads to larger income inequality in bigger cities.

Maximizing profits (4.45) with respect to prices yields the standard condition

$$p_c(t) = \frac{\mathcal{E}_{x,p}}{\mathcal{E}_{x,p} - 1} \frac{w_c}{L_c^\epsilon t}, \quad (4.46)$$

where $\mathcal{E}_{x,p} = 1/r(x_c(t))$ is the price elasticity of per capita demand $x_c(t)$, which can be expressed using the “relative love for variety” (RLV), $r(\cdot)$ (Zhelobodko et al., 2012).⁵⁴ The profit of an agent who produces a variety with talent $t \geq t_c$ located in a city of size L_c is then given by

$$\pi_c(t) = \underbrace{\frac{r(x_c(t))}{1 - r(x_c(t))}}_{=\mu(t, t_c, L_c)} \frac{w_c}{t} L_c^{1-\epsilon} x_t, \quad (4.47)$$

where $\mu(t, t_c, L_c)$ denotes the *profit margin* of a type t agent in a city with cutoff t_c and size L_c .

The set of entrepreneurs who produce differentiated varieties is endogenously determined by the cutoff t_c . More formally, agents self-select into occupations (entrepreneurs

⁵³ For simplicity, we assume that aggregate demand $X_c(t) = L_c x_c(t)$. This will hold true in quasi-linear settings or when preferences are such that aggregate demand depends on some summary statistic (a “generalized Lagrange multiplier”). The latter property amounts to imposing some form of quasi separability on the inverse of the subutility function as in Behrens and Murata (2007).

⁵⁴ In additively separable models, where utility is given by $U = \int u(x_i) dF_i(t)$, we have $\mathcal{E}_{x,p} = 1/r(x_t)$, where $r(x) = -xu''(x)/u'(x) \in (0, 1)$. Condition (4.46) links the firms’ markups solely to the properties of the subutility function u (via the RLV). The way that market size affects selection crucially depends on the properties of $r(\cdot)$ and, therefore, on the properties of preferences. Note that $r(\cdot)$ is a function of individual consumption x_t and that it will, in general, be neither a constant nor a monotonic function.

vs. workers) on the basis of the maximum income they can secure. The *selection condition* that pins down the marginal entrepreneur is as follows:

$$\pi_c(t_c) - w_c t_c^a L_c^\xi = 0, \quad (4.48)$$

where L_c^ξ is an agglomeration externality that makes workers more productive (increases their effective labor). In words, the marginal entrepreneur earns profits equal to the wage he or she could secure as a worker, whereas all agents with talent t such that $\pi_c(t) > w_c t^a L_c^\xi$ choose to become entrepreneurs and the others become workers.

The key questions to be addressed are the following. What is the impact of city size L_c on the occupational structure via t_c , and how does the talent composition of the city, $F_c(\cdot)$, and various agglomeration externalities, interact with selection? We look at the distribution of incomes within and across groups in the next section.

4.4.2.2 CES illustration

To keep things simple, let us start with the well-known case of CES preferences: $u(x) = x^\rho$. In that case $r(x_c(t)) = 1 - \rho$ is constant and independent of individual consumption (and thus of city size). Aggregate CES demand can be expressed as $L_c x_c(t) = L_c [\mathbb{A}_c / p_c(t)]^{1/(1-\rho)}$, where $\mathbb{A}_c$ is some city-specific market aggregate that depends on the distribution of income in the city but that is taken as given by each entrepreneur. From (4.46), we have constant markup pricing: $p_c(t) = w_c / (\rho L_c^\epsilon t)$.

Plugging $x_c(t)$ and $p_c(t)$ into profits yields

$$\pi_c(t) = \rho^{\frac{\rho}{1-\rho}} (1 - \rho) L_c^{1 + \epsilon \frac{\rho}{1-\rho}} \mathbb{A}_c^{\frac{1}{1-\rho}} \left(\frac{w_c}{t} \right)^{\frac{\rho}{\rho-1}}.$$

The occupational selection condition $\pi_c(t_c) = w_c t_c^a L_c^\xi$ can then be written as

$$L_c^{1 + \epsilon \frac{\rho}{1-\rho} - \xi} \left(\frac{\mathbb{A}_c}{w_c} \right)^{\frac{1}{1-\rho}} = t_c^{a - \frac{\rho}{1-\rho} \frac{\rho}{\rho-1}} \frac{1}{1 - \rho}. \quad (4.49)$$

In general equilibrium, the term $\mathbb{A}_c / w_c$ is pinned down by the citywide market clearing condition. Consider the labor market clearing condition: agents who do not become entrepreneurs are workers who will be hired by the entrepreneurs. That condition is given by

$$\int_{t_c}^{t_c} t^a L_c^\xi dF_c(t) = \int_{t_c}^{\bar{t}_c} \frac{L_c x_c(t)}{L_c^\epsilon t} dF_c(t). \quad (4.50)$$

Inserting the expression $L_c x_c(t) = L_c (\mathbb{A}_c / p_c(t))^{1/(1-\rho)}$ and simplifying, we obtain the relationship

$$\begin{aligned}
& \underbrace{L_c^{1+\epsilon \frac{\rho}{1-\rho} - \xi \left(\frac{\mathbb{A}_c}{w_c}\right)^{\frac{1}{1-\rho}}}}_{\text{ZPC}} \rho^{\frac{1}{1-\rho}} \int_{t_c}^{\bar{t}_c} t^{\frac{\rho}{1-\rho}} dF_c(t) = \int_{\underline{t}_c}^{t_c} t^a dF_c(t) \\
& \Rightarrow t_c^{a - \frac{\rho}{1-\rho} \frac{\rho}{1-\rho}} \rho^{\frac{\rho}{1-\rho}} \int_{t_c}^{\bar{t}_c} t^{\frac{\rho}{1-\rho}} dF_c(t) = \int_{\underline{t}_c}^{t_c} t^a dF_c(t),
\end{aligned}$$

where we have replaced ZPC by the selection condition (4.49). As can be seen, the last condition depends only on the selection cutoff t_c . Hence, conditional on the distribution of skills—as captured by the distribution $F_c(\cdot)$ and the support $[\underline{t}_c, \bar{t}_c]$ —the selection cutoff t_c is independent of city size, although profits are increasing as the direct effect of L_c . The reason is that $\mathbb{A}_c/w_c$ is endogenously determined in the citywide general equilibrium. Any increase in L_c triggers an inverse fall in $\mathbb{A}_c/w_c$, so profits and workers' wages increase in the same proportion in equilibrium. Consequently, city size L_c has no bearing on selection when preferences are of the CES type. Two cities with different sizes but identical skill composition have the same selection cutoff and the same share of entrepreneurs. These findings seem to be in line with the empirical results obtained by Combes et al. (2012) and with the observation that the share of self-employed (a proxy for “entrepreneurship”) is independent of city size in the United States (see the left panel in Figure 4.4). Observe though that there is still an effect of sorting on selection: a city c with a better underlying skill distribution than a city c' —for example, because $F_c(\cdot)$ first-order stochastically dominates $F_{c'}(\cdot)$ —has a larger t_c in equilibrium.

There are two main take-away messages from the foregoing analysis. First, selection effects are inherently a general equilibrium phenomenon. Since large cities (especially MSAs) can be viewed as large economic systems, taking into account general equilibrium effects strikes us as being important. Disregarding those effects may lead to erroneous assessments as to the impacts of market size and talent composition on economic outcomes. Larger cities may be tougher markets, but they are also bigger and richer markets. Taking into account income effects and resource constraints is an important part of the analysis. Second, sorting induces selection. Once sorting has been controlled for, there may or may not be an additional effect of market size on selection. In other words, larger markets may or may not have “tougher selection” (conditional on sorting). The absence of selection effects due to market size in the above example is an artifice of the CES structure where markups are constant (Zhelobodko et al., 2012; Behrens et al., 2014a,c). Yet, selection is still influenced by the talent composition of the city. General equilibrium effects matter.

4.4.2.3 Beyond the CES

The CES structure is arguably an extremely special one. Unfortunately, little is known about selection with more general preferences and demands. What is known is that the selection cutoff t_c usually depends on L_c in general equilibrium, essentially since markups

are variable and a function of L_c . Two models where market size matters for the selection of heterogeneous producers are those of [Ottaviano \(2012\)](#) and [Behrens and Robert-Nicoud \(2014b\)](#). They build on the [Melitz and Ottaviano \(2008\)](#) quadratic preferences model to study the relationship between market size and selection in a new economic geography and in a monocentric city setting, respectively. However, sorting along skills is absent in those models. The same holds true for the models building on constant absolute risk aversion preferences ([Behrens et al., 2013, 2014b](#)). We are not aware of any model displaying between-city sorting in the presence of nontrivial selection effects.

[Behrens et al. \(2014c\)](#) use general additive preferences in a quasi-linear setting to show that larger markets may have either tougher selection (fewer entrepreneurs) or weaker selection (more entrepreneurs), depending crucially on the properties of preferences.⁵⁵ In specifications that many consider as being the normal case (e.g., [Vives, 2001](#)), demands become less elastic with consumption levels, so larger cities have tougher selection and fewer entrepreneurs.⁵⁶ We suspect that models where larger markets put downward pressure on prices and markups may yield additional effects of selection on sorting. However, to the best of our knowledge, little progress has been made in that direction to date.

4.4.2.4 Selection and sorting

How do selection and sorting interact? In the foregoing, we developed a simple example that shows that sorting induces selection, even when market size does not matter directly. Clearly, selection also has an impact on sorting by changing the payoff structure for agents. The basic question for sorting is always whether larger markets are more profitable places for more talented entrepreneurs. From (4.47), the single-crossing condition can be expressed as follows (recall that we hold the distribution of talent $F_c(\cdot)$ in the city fixed):

$$\begin{aligned} \frac{\partial^2 \pi_c(t)}{\partial L_c \partial t} &= (1 - \epsilon) L_c^{-\epsilon} \left(\frac{\partial x}{\partial t} \mu + \frac{\partial \mu}{\partial t} x \right) + L_c^{1-\epsilon} \left(\frac{\partial^2 \mu}{\partial t \partial L_c} x + \frac{\partial \mu}{\partial t} \frac{\partial x}{\partial L_c} + \frac{\partial^2 x}{\partial t \partial L_c} \mu + \frac{\partial x}{\partial t} \frac{\partial \mu}{\partial L_c} \right) \\ &\quad + \frac{\partial t_c}{\partial L_c} L_c^{1-\epsilon} \left(\frac{\partial^2 \mu}{\partial t \partial t_c} x + \frac{\partial^2 x}{\partial t \partial t_c} \mu + \frac{\partial \mu}{\partial t} \frac{\partial x}{\partial t_c} + \frac{\partial x}{\partial t} \frac{\partial \mu}{\partial t_c} \right). \end{aligned}$$

The first term on the right-hand-side above is the “profit margin effect,” which depends on how markups and output change with productivity. First, more productive firms sell larger quantities ($\partial x / \partial t > 0$; [Zhelobodko et al., 2012](#)). Second, the effect of productivity on profit margins ($\partial \mu / \partial t$) is generally ambiguous and depends on whether the RLV $r(\cdot)$ is

⁵⁵ The impact of a change in city size L_c on the selection cutoff t_c —and thus on the share of entrepreneurs and the range of varieties—can go either way, depending on the scale elasticity of $u(\cdot)$ and its RLV.

⁵⁶ This class of preferences includes the quasi-linear quadratic model of [Melitz and Ottaviano \(2008\)](#), [Ottaviano \(2012\)](#), and [Behrens and Robert-Nicoud \(2014b\)](#), as well as the constant absolute risk aversion specification of [Behrens and Murata \(2007\)](#) and [Behrens et al. \(2013, 2014b\)](#).

an increasing or decreasing function of productivity. In the CES case, the first term is unambiguously positive, but this is not a general result.

The second term captures the interactions between talent and size that influence the entrepreneur's profits. This term cannot be unambiguously signed either. Whereas the terms $\partial x/\partial t$ and $\partial x/\partial L_c$ are generally positive and negative, respectively, the other terms cannot be signed a priori. For example, per unit profit may increase or decrease with market size and with productivity under reasonable specifications for preferences.

The last term, which we call the *selection effect* ($\partial t_c/\partial L_c$), is also ambiguous. The basic selection term $\partial t_c/\partial L_c$ cannot be signed in general, as we have argued above. The reason is that it depends on many features of the model, in particular on preferences.

To summarize, even in simple models of selection with heterogeneous agents, little can be said a priori on how agents sort across cities in general equilibrium. The main reason for this negative result is that sorting induces selection (via $F_c(\cdot)$ and L_c), and that selection changes the payoffs to running firms. Depending on whether those payoffs rise or fall with city size for more talented agents, we may or may not observe PAM sorting across cities. Supermodularity may fail to hold, and analyzing sorting in the absence of supermodularity is a difficult problem. Many equilibria involving nontrivial patterns of sorting may in principle be sustained.

4.4.2.5 Empirical implications and results

Distinguishing between sorting and selection has a strong conceptual basis: it is location choice versus occupation (either as a choice or as an outcome). Distinguishing between the two is hard empirically. The key difficulties are illustrated in Figure 4.10. The arrows labeled (a) in Figure 4.10 show that there is a causal relationship from the talent composition to the size of a city: tougher cities repel agents. *Ceteris paribus*, people rather want to be “first in the village rather than second in Rome.” We refer to this as

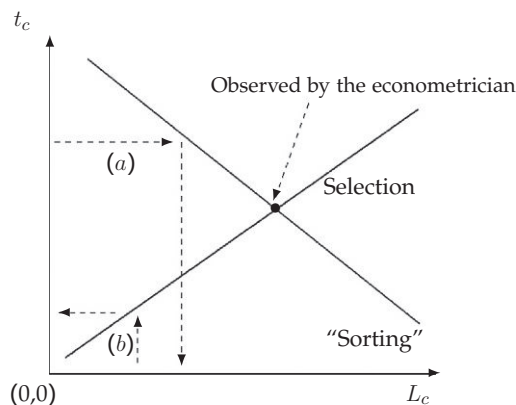


Figure 4.10 Interactions between sorting and selection.

sorting. The arrows labeled (b) in Figure 4.10 show that there is also a causal relationship in the opposite direction, from city size to talent: the talent composition of a city changes with its size. We refer to this as *selection*. The econometrician observes the equilibrium tuples (t_c, L_c) across the urban system. To identify selection, it is necessary to have exogenous shifts in sorting and vice versa. This is difficult, since sorting is itself endogenous. In the end, distinguishing sorting from selection *ex post* is very difficult since both are observationally equivalent and imply that the productivity composition varies systematically across markets.⁵⁷

The empirical evidence on selection effects to date is mixed. This may be a reflection of their theoretical ambiguity, or of their intrinsic relationship with sorting effects. Di Addario and Vuri (2010) find that the share of entrepreneurs increases with population and employment density in Italian provinces. However, once individual characteristics and education are controlled for, the share of entrepreneurs decreases with market size. The probability of young Italian college graduates being entrepreneurs 3 years after graduation decreases by 2–3 percentage points when the population density of a province doubles. About one-third of this “selection effect” seems to be explained by increased competition among entrepreneurs within industries. However, conditional on survival, successful entrepreneurs in dense provinces reap the benefits of agglomeration: their income elasticity with respect to city size is about 2–3%. Sato et al. (2012) find similar results for Japanese cities. Using survey data, they document that the *ex ante* share of individuals who desire to become entrepreneurs is higher in larger and denser cities: a 10% increase in density increases the share of prospective entrepreneurs by about 1%. It, however, reduces it *ex post* by more than that, so the observed rate of entrepreneurship is lower in denser Japanese cities.

To summarize, the empirical evidence suggests that larger markets have more prospective entrepreneurs (more entrants), but only a smaller share of those entrants survive (tougher selection).⁵⁸ Those who do survive in larger markets perform, however, significantly better, implying that denser markets will also be more unequal. Additional evidence for positive selection effects in larger markets in the United States is provided by Syverson (2004, 2007) and by Campbell and Hopenhayn (2005). By contrast, Combes et al. (2012) find no evidence for selection effects—defined as the left truncation of the productivity distribution of firms—when comparing large and small French cities. This finding relies on the identifying assumption that the underlying (unobserved) productivity distributions are the same in small and large cities, and the results are consistent with the CES model.

⁵⁷ Okubo et al. (2010) refer to the “spatial selection” of heterogeneous agents when talking about “sorting.” That terminology clearly reveals how intrinsically linked sorting and selection really are.

⁵⁸ The theoretical predictions of the model of Behrens and Robert-Nicoud (2014b) are consistent with this finding.

4.5. INEQUALITY

Heterogeneous agents face heterogeneous outcomes. Hence, it is natural to study issues related to the second moments of the distributions of outcomes. Specifically, one may ask if larger cities are more unequal places than small towns? What mechanisms drive the dispersion of income in large cities? And how does inequality depend on sorting and selection?

We have seen in the previous sections how the size (agglomeration economies) and composition (selection and sorting) of cities influence occupational choices and individual earnings. They thus naturally influence the distribution of earnings within cities. [Figure 4.5](#) reports that large cities are more unequal than smaller ones and suggests that this effect is the joint outcome of composition and size effects (left panel) and an urban premium that varies across the wage distribution (right panel). Indeed, the partial correlation between city size and city Gini coefficient is positive, whether we control for the talent composition of cities (using the share of college graduates as a proxy) or not, and it is larger when we control for it (dashed line) than when we do not (solid line).

Studying the causes and effects of urban inequality is important for at least two reasons. First, earning and wealth inequality seems to be on the rise in many countries ([Piketty, 2014](#)), and understanding this rise at the country level requires at least a partial understanding of the positive relationship between city size and earnings inequality. Indeed, [Baum-Snow and Pavan \(2014\)](#) report that at least a quarter of the overall increase in earnings inequality in the United States over the period 1979–2007 is explained by the relatively high growth of earnings inequality in large urban areas.⁵⁹ Second, earnings inequality at the local level matters per se: people perceive inequality more strongly when they see it at close range, and cities are not only the locus where inequality materializes, but they are also hosts to mechanisms (sorting and selection) that contribute to changes in that inequality. As such, focusing on cities is of primary interest when designing policies that aim at reducing inequality and its adverse social effects. This is a complex issue because ambitious redistributive policies at the local level may lead to outflow of wealthy taxpayers and an inflow of poor households, a phenomenon that is thought to have contributed to the financial crisis that hit New York City in the 1970s.

Let $y(t, L_c, F_c)$ denote the earnings of an individual with talent t who lives in city c of population size L_c and talent composition F_c . It immediately follows that the earnings distribution in any city inherits some properties of its talent distribution, and also that its size and its composition both affect its shape. In this section, we consider two modifications of (4.27) to study how the composition and the size of cities are related to urban inequality as measured by the Gini coefficient of city earnings. We start with sorting.

⁵⁹ The measure of earnings inequality in [Baum-Snow and Pavan \(2014\)](#) is the variance of the logarithm of hourly wages.

4.5.1 Sorting and urban inequality

Consider first the following slightly generalized version of (4.26):

$$\gamma(t, L_c, F_c) = \mathbb{A}_c t^a L_c^\epsilon, \quad (4.51)$$

where $\mathbb{A}_c$ is the usual TFP shifter and F_c is the talent composition of c . To fix ideas, assume that the distribution of talent F_c is city specific and log-normal with⁶⁰

$$\ln t \sim \mathcal{N}(\mu_{tc}, \sigma_{tc}^2). \quad (4.52)$$

Assumptions (4.51) and (4.52) together imply that earnings γ in city c are also log-normally distributed and the Gini coefficient is a function of the standard deviation of the logarithm of earnings in city c only (Aitchison and Brown, 1963):

$$\text{Gini}(L_c, F_c) = 2\Phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right) - 1, \quad (4.53)$$

where $\Phi(\cdot)$ is the cumulative of the normal distribution and $\sigma_{yc} = a\sigma_{tc}$ is the standard deviation of the logarithm of earnings. It immediately follows from $\Phi'(\cdot) > 0$ and the definition of σ_{yc} that earnings inequality increases with talent inequality (a composition effect)—namely,

$$\frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{tc}} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{yc}} \frac{\partial \sigma_{yc}}{\partial \sigma_{tc}} = a\sqrt{2}\phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right) > 0, \quad (4.54)$$

where $\phi(\cdot)$ is the density of the normal distribution, and the second equality follows from the definition of σ_{yc} . Observe that city size has no direct effect on the Gini coefficient of earnings.⁶¹ This is because agglomeration economies benefit all talents in the same proportion in (4.51).

We know from the previous section that sorting and selection effects imply that the composition of large cities differs systematically from the composition of smaller ones. That is to say, L_c and F_c are jointly determined in general equilibrium. We may thus write

$$\frac{d\text{Gini}(L_c, F_c)}{dL_c} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{tc}} \frac{d\sigma_{tc}}{dL_c},$$

where the partial derivative is from (4.54). This simple framework is consistent with the positive partial correlation between the urban Gini coefficient and city size in the left panel in Figure 4.5 if and only if $d\sigma_{tc}/dL_c > 0$. If urban talent heterogeneity increases with city size, as in Combes et al. (2012) and Eeckhout et al. (2014), or if large cities

⁶⁰ This convenient assumption allows us to parameterize the whole distribution of talents with only two parameters, μ_{tc} and σ_{tc} , which simplifies the analysis below.

⁶¹ Note that urban size has a positive effect on the variance of earnings, $\text{var}_{yc} = \exp(2\mu_{yc} + \sigma_{yc}^2) [\exp(\sigma_{yc}^2) - 1]$, where $\mu_{yc} = \mu_{tc} + \ln \mathbb{A}_c + \epsilon \ln L_c$.

attract a disproportionate share of talented workers (so the variance of talents increases with city size), then this inequality holds. Glaeser et al. (2009) report that differences in the skill distribution across US MSAs explain one-third of the variation in Gini coefficients. Variations in the returns to skill may explain up to half of the cross-city variation in income inequality according to the same authors. We turn to this explanation next.

4.5.2 Agglomeration and urban inequality

Agglomeration economies affect all talents to the same degree in the previous subsection. This is counterfactual. Using individual data, Wheeler (2001) and Baum-Snow and Pavan (2012) estimate that the skill premium and the returns to experience of US workers increase with city size.⁶² A theoretical framework that delivers a positive relationship between city size and the returns to productivity is provided in Davis and Dingel (2013) and Behrens and Robert-Nicoud (2014b). We return to the latter in some detail in Section 4.5.3. To the best of our knowledge, the assignment mechanism similar to Rosen's 1981 "superstar effect" of the former—with markets suitably reinterpreted as urban markets—and the procompetitive effects that skew market shares toward the most productive agents of the latter are the only mechanisms to deliver this theoretical prediction.

To account for this, we now modify (4.26) as follows:

$$\gamma(t, L_c, F_c) = \mathbb{A}_c L_c^{a + \epsilon t}, \quad \text{where } t \sim \mathcal{N}(\mu_t, \sigma_t). \quad (4.55)$$

These expression differ from (4.51) and (4.52) in two ways. First, γ is log-supermodular in size and talent in (4.55) but it is only supermodular in (4.51): "simple" supermodularity is not enough to drive complementarity between individual talent and city size. Second, talent is normally distributed and we assume that the composition of talent is constant across cities—that is, $F_c = F$ for all c .

As before, our combination of functional forms for earnings and the distribution of talent implies that the distribution of earnings is log-normal and that the city Gini coefficient is given by (4.53). The novelty is that the standard deviation of the logarithm of earnings increases with city size, which is consistent with the empirical finding of Baum-Snow and Pavan (2014):

$$\sigma_{y_c} = \sigma_t \epsilon \ln L_c. \quad (4.56)$$

Combining (4.53) and (4.56) implies that urban inequality increases with city size:

⁶² See also Baum-Snow and Pavan (2014) for evidence consistent with this mechanism. These authors also report that the positive relationship between urban inequality and city size strengthened between 1979 and 2007, explaining a large fraction of the rise in within-group inequality in the United States.

$$\frac{\partial \text{Gini}(L_c, F_c)}{\partial \ln L_c} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{yc}} \frac{\partial \sigma_{yc}}{\partial \ln L_c} = \sigma_t \epsilon \sqrt{2} \phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right) > 0, \quad (4.57)$$

where the second expression follows from (4.56). From an urban economics perspective, agglomeration economies disproportionately benefit the most talented individuals: the urban premium increases with talent. From a labor economics perspective, and assuming that observed skills are a good approximation for unobserved talents, this result means that the skill premium increases with city size.

Putting the pieces together, we assume finally that city size and individual talent are log-supermodular as in (4.55) and that the talent distribution is city specific as in Section 4.5.1:

$$\gamma(t, L_c, F_c) = \mathbb{A}_c L_c^{a+\epsilon t}, \quad \text{where } t \sim \mathcal{N}(\mu_{tc}, \sigma_{tc}). \quad (4.58)$$

Then the relationship between urban inequality and city size is the sum of the size and composition effects:

$$\frac{d \text{Gini}(L_c, F_c)}{d L_c} = \frac{\partial \text{Gini}(L_c, F_c)}{\partial L_c} + \frac{\partial \text{Gini}(L_c, F_c)}{\partial \sigma_{tc}} \frac{d \sigma_{tc}}{d L_c} = \sqrt{2} \epsilon \frac{L_c}{\sigma_{tc}} \left(1 + \ln L_c \frac{d \ln \sigma_{tc}}{d \ln L_c} \right) \phi\left(\frac{\sigma_{yc}}{\sqrt{2}}\right),$$

where the second equality follows from (4.54), (4.57), and (4.58). Both terms are positive if $d \sigma_{tc} / d L_c > 0$. The solid line in the left panel in Figure 4.5 reports the empirical counterpart to this expression.⁶³

4.5.3 Selection and urban inequality

So far, we have allowed urban inequality to depend on the talent composition of cities, city size, or both. There was no selection. In order to study the relationship between selection and urban inequality, we introduce selection in a simple way by imposing the following set of assumptions. Assume first that selection takes a simple form, where the earnings of agents endowed with a talent above some threshold t_c take the functional form in (4.51) and are zero otherwise:

$$\gamma(t, t_c, L_c) = \begin{cases} 0 & \text{if } t \leq t_c \\ \mathbb{A}_c t^a L_c^\epsilon & \text{if } t > t_c. \end{cases} \quad (4.59)$$

We refer to the fraction of the population earning zero, $\Phi_c(t_c)$, as the “failure rate” in city c . Second, we rule out sorting and assume that the composition of talent is invariant across cities—that is, $F_c = F$, for all c —and that talents are log-normally distributed as in

⁶³ The empirical relationship between urban density and inequality is less clear. Using worker micro data and different measures of earnings inequality from 1970 to 1990—including one that corrects for observable individual characteristics—Wheeler (2004) documents a robust and significantly *negative* association between MSA density and inequality, even when controlling for a number of other factors. This suggests that workers in the bottom income quintile benefit more from *density* than workers in the top income quintile, which maps into smaller earnings inequality in denser cities.

(4.52). Third, we assume that the conditional distribution of talent above the survival selection cutoff t_c is reasonably well approximated by a Pareto distribution with shape parameter $k > 1$:

$$F(t|t \geq t_c) = 1 - \left(\frac{t_c}{t}\right)^k. \quad (4.60)$$

We use this approximation for two related reasons. First, a Pareto distribution is a good approximation of the upper tail of the log-normal distribution in (4.52)—and this is precisely the tail of interest here. Second, the Gini coefficient associated with (4.59) and (4.60) obeys a simple functional form,

$$\text{Gini}(t_c, L_c) = \Phi(t_c) + \frac{1}{2ak-1} [1 - \Phi(t_c)] = \frac{1 + 2(ak-1)\Phi(t_c)}{2ak-1}, \quad (4.61)$$

whereas the Gini coefficient associated with the conditional log-normal $\Phi(t|t \geq t_c)$ does not. The first term in (4.61) is the decomposition of the Gini coefficient into the contributions of the zero-earners and of the earners with a talent above the cutoff t_c , respectively. The term $1/(2ak-1)$ is the Gini coefficient computed among the subpopulation of agents with a talent above t_c . Note that this formula for the Gini coefficient is valid only if $ak > 1$ because any Gini coefficient belongs to the unit interval by definition. It follows by inspection of the second term of (4.61) that the Gini coefficient increases with the extent of selection as captured by $\Phi(t_c)$.

We propose a model of urban systems that fits the qualitative properties of this reduced-form model in Behrens and Robert-Nicoud (2014b). Preferences are quasi-linear and quadratic and t is Pareto distributed as in Melitz and Ottaviano (2008). *Ex ante* homogeneous workers locate in cities with possibly heterogeneous $\mathbb{A}_c$. Cities endowed with a large $\mathbb{A}_c$ attract more workers in equilibrium. In turn, large urban markets are more competitive and a smaller proportion of workers self-select into entrepreneurship as a result—that is, the failure rate $\Phi(t_c)$ increases with city size. This is related to our fact 4 (selection) for the United States and is consistent with the empirical findings of Di Addario and Vuri (2010) and Sato et al. (2012) for Italy and Japan, respectively. Recalling that workers are homogeneous prior to making their location decision in Behrens and Robert-Nicoud (2014b), we find that returns to successful entrepreneurs increase with city size. This latter effect is absent in (4.59) but is accounted for in the model we develop in Section 4.5.2.

We can finally compute the relationship between urban inequality and city size in the absence of sorting and agglomeration effects as follows:

$$\frac{d\text{Gini}(t_c, L_c)}{dL_c} = \frac{\partial \text{Gini}(t_c, L_c)}{\partial t_c} \frac{dt_c}{dL_c} = 2\phi(t_c) \frac{ak-1}{2ak-1} \frac{dt_c}{dL_c},$$

which is positive if and only if $dt_c/dL_c > 0$, and where we have made use of the partial derivative of (4.61) with respect to t_c . The interaction between selection and size may thus be conducive to the pattern illustrated in Figure 4.5. Behrens et al. (2014c) show that the equilibrium relationship between urban selection and city size depends on the modeler's choice of the functional forms for preferences. It can even be nonmonotonic in theory, thus suggesting that the impacts of size on inequality could also be nonmonotonic.

4.6. CONCLUSIONS

We have extended the canonical urban model along several lines to include heterogeneous workers, firms, and sites. This framework can accommodate all key stylized facts in Section 4.2 and it is useful to investigate what heterogeneity adds to the big picture. Two direct consequences of worker and firm heterogeneity are sorting and selection. These two mechanisms—and their interactions with agglomeration economies and locational fundamentals—shape cities' productivity, income, and skill distributions. We have also argued that more work is needed on the general equilibrium aspects of urban systems with heterogeneous agents. Though difficult, making progress here is key to obtaining a full story about how agents sort across cities, select into occupations, and reap the benefits from and pay the costs of urban size. The first article doing so (albeit in a two-city environment) was that of Davis and Dingel (2013). We use this opportunity to point out a number of avenues along which urban models featuring selection and sorting with heterogeneous agents need to be extended. First, we need models where sorting and nontrivial selection effects interact with citywide income effects and income distributions. This is important if we want to understand better how sorting and selection affect inequalities in cities, and how changes in the urban system influence the macro economy at large. Unfortunately, modeling sorting and selection in the presence of income distributions and nontrivial income effects is a notoriously difficult task. This is probably one explanation for the strong reliance on representative agent models, which, despite their convenience, do not teach us much when it comes to sorting, selection, and inequality. A deeper understanding of the interactions between selection and sorting should also allow us to think better about empirical strategies aimed at disentangling them.

Second, in the presence of heterogeneous agents, the *within-city* allocation of those agents becomes an interesting topic to explore. How do agents organize themselves in cities, and how does heterogeneity across and within cities interact to shape the outcomes in the urban system? There is a large literature on the internal structure of cities, but that literature typically deals with representative agents and is only interested in the implications of city structure for agglomeration economies, land rents, and land use (Beckman,

1976; Fujita and Ogawa, 1982; Lucas and Rossi-Hansberg, 2002; Mossay and Picard, 2011). Extending that literature to include heterogeneous agents seems important to us. For example, if agents sort themselves in specific ways across cities—so that richer agents compete more fiercely for good locations and pay higher land rents—real income inequality in cities may be very different from nominal income inequality. The same holds true for different cities in the urban system, and understanding how heterogeneous agents allocated themselves across and within cities is key to understanding the income and inequality patterns we observe. Davis and Dingel (2014) provide a first step in that direction.

Third, heterogeneous firms and workers do not really interact in urban models. Yet, there is a long tradition in labor economics that deals with that interaction (see, e.g., Abowd et al., 1999). There is also a growing literature in international trade that investigates the consequences of the matching between heterogeneous firms and workers (Helpman et al., 2010). Applying firm-worker matching models to an urban context seems like a natural extension, and may serve to understand better a number of patterns we see in the data. For example, Mion and Naticchioni (2009) use matched employer–employee data for Italy and interpret their findings as evidence for assortative matching between firms and workers.⁶⁴ Yet, this assortative matching is stronger in smaller and less dense markets, thus suggesting that matching quality is less important in bigger and denser markets. Theory has, to the best of our knowledge, not much to say about those patterns, and models with heterogeneous workers and firms are obviously required to make progress in that direction.

Lastly, the attentive reader will have noticed that our models depart from the canonical framework of Henderson (1974) by not including transportation or trade costs, so the *relative location* of cities is irrelevant. Multicity trade models with heterogeneous mobile agents are difficult to analyze, yet progress needs to be made in that direction to understand better spatial patterns, intercity trade flows, and the evolution of the urban system in a globalizing world. In a nutshell, we need to get away from models where trade is either prohibitively costly or free. We need to bring back space into urban economic theory, just as international trade brought back space in the 1990s. The time is ripe for *new urban economics* featuring heterogeneity and transportation costs in urban systems.

ACKNOWLEDGMENTS

We thank Bob Helsley for his input during the early stages of the project. Bob should have been part of this venture but was unfortunately kept busy by other obligations. We further thank our discussant, Don Davis, and the editors Gilles Duranton, Vernon Henderson, and Will Strange for extremely valuable comments and suggestions. Théophile Bougna provided excellent research assistance. K. B. and R. -N. gratefully acknowledge financial support from the CRC Program of the Social Sciences and Humanities Research Council of Canada for the funding of the Canada Research Chair in Regional Impacts of Globalization.

⁶⁴ The PAM between firms and workers, or its absence, is a difficult and still open issue in labor economics.

REFERENCES

- Abdel-Rahman, H.M., 1996. When do cities specialize in production? *Reg. Sci. Urban Econ.* 26, 1–22.
- Abdel-Rahman, H.M., Anas, A., 2004. Theories of systems of cities. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, North-Holland, pp. 2293–2339.
- Abdel-Rahman, H.M., Fujita, M., 1993. Specialization and diversification in a system of cities. *J. Urban Econ.* 3, 189–222.
- Abowd, J.M., Kramarz, F., Margolis, D.N., 1999. High-wage workers and high-wage firms. *Econometrica* 67, 251–333.
- Aitchison, J., Brown, J.A.C., 1963. *The Lognormal Distribution*. Cambridge Univ. Press, Cambridge, UK.
- Albouy, D., Seegert, N., 2012. *The Optimal Population Distribution Across Cities and the Private-Social Wedge*. Univ. of Michigan, processed.
- Albouy, D., Behrens, K., Robert-Nicoud, F.L., Seegert, N., 2015. Are cities too big? Optimal city size and the Henry George theorem revisited, in progress.
- Arthur, W.B., 1994. *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press, Ann Arbor, MI.
- Bacolod, M., Blum, B.S., Strange, W.C., 2009a. Skills in the city. *J. Urban Econ.* 65, 136–153.
- Bacolod, M., Blum, B.S., Strange, W.C., 2009b. Urban interactions: soft skills vs. specialization. *J. Econ. Geogr.* 9, 227–262.
- Bacolod, M., Blum, B.S., Strange, W.C., 2010. Elements of skill: traits, intelligences, and agglomeration. *J. Reg. Sci.* 50, 245–280.
- Baldwin, R.E., Okubo, T., 2006. Heterogeneous firms, agglomeration and economic geography: spatial selection and sorting. *J. Econ. Geogr.* 6, 323–346.
- Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage gap. *Rev. Econ. Stud.* 79, 88–127.
- Baum-Snow, N., Pavan, R., 2014. Inequality and city size. *Rev. Econ. Stat.* 95, 1535–1548.
- Becker, G.S., Murphy, K.M., 1992. The division of labor, coordination costs, and knowledge. *Q. J. Econ.* 107, 1137–1160.
- Becker, R., Henderson, J.V., 2000a. Intra industry specialization and urban development. In: Huriot, J.M., Thisse, J.F. (Eds.), *The Economics of Cities*. Cambridge University Press, Cambridge.
- Becker, R., Henderson, J.V., 2000b. Political economy of city sizes and formation. *J. Urban Econ.* 48, 453–484.
- Beckman, M.J., 1976. Spatial equilibrium in the dispersed city. In: Papageorgiou, Y.Y. (Ed.), *Mathematical Land Use Theory*. Lexington Books, Lexington, MA.
- Behrens, K., 2007. On the location and lock-in of cities: geography vs transportation technology. *Reg. Sci. Urban Econ.* 37, 22–45.
- Behrens, K., Murata, Y., 2007. General equilibrium models of monopolistic competition: a new approach. *J. Econ. Theory* 136, 776–787.
- Behrens, K., Robert-Nicoud, F.L., 2014a. Equilibrium and optimal urban systems with heterogeneous land, in progress.
- Behrens, K., Robert-Nicoud, F.L., 2014b. Survival of the fittest in cities: urbanisation and inequality. *Econ. J.* 124 (581), 1371–1400.
- Behrens, K., Lamorgese, A.R., Ottaviano, G.I.P., Tabuchi, T., 2009. Beyond the home market effect: market size and specialization in a multi-country world. *J. Int. Econ.* 79, 259–265.
- Behrens, K., Mion, G., Murata, Y., Südekum, J., 2013. *Spatial frictions*. Univ. of Québec at Montréal; Univ. of Surrey; Nihon University; and Univ. of Duisburg-Essen, processed.
- Behrens, K., Duranton, G., Robert-Nicoud, F.L., 2014a. Productive cities: sorting, selection and agglomeration. *J. Pol. Econ.* 122, 507–553.
- Behrens, K., Mion, G., Murata, Y., Südekum, J., 2014b. Trade, wages, and productivity. *Int. Econ. Rev.* (forthcoming).
- Behrens, K., Pokrovsky, D., Zhelobodko, E., 2014c. Market size, entrepreneurship, and income inequality. *Technical Report*, Centre for Economic Policy Research, London, UK Discussion Paper 9831.
- Bleakley, H., Lin, J., 2012. Portage and path dependence. *Q. J. Econ.* 127, 587–644.
- Campbell, J.R., Hopenhayn, H.A., 2005. Market size matters. *J. Industr. Econ.* LIII, 1–25.

- Combes, P.P., Gobillon, L., 2015. The empirics of agglomeration economies. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, North-Holland, pp. 247–348.
- Combes, P.P., Duranton, G., Gobillon, L., 2008. Spatial wage disparities: sorting matters! *J. Urban Econ.* 63, 723–742.
- Combes, P.P., Duranton, G., Gobillon, L., Puga, D., Roux, S., 2012. The productivity advantages of large cities: distinguishing agglomeration from firm selection. *Econometrica* 80, 2543–2594.
- Combes, P.P., Duranton, G., Gobillon, L., 2014. The Costs of Agglomeration: Land Prices in French Cities. University of Pennsylvania, Wharton School, in progress.
- Costinot, A., 2009. An elementary theory of comparative advantage. *Econometrica* 77, 1165–1192.
- Couture, V., 2014. Valuing the Consumption Benefits of Urban Density. University of California Berkeley, processed.
- Davis, D.R., Dingel, J.I., 2013. A Spatial Knowledge Economy. Columbia University, processed.
- Davis, D.R., Dingel, J.I., 2014. The comparative advantage of cities. NBER Working paper 20602. National Bureau of Economic Research.
- Davis, J.C., Henderson, J.V., 2008. The agglomeration of headquarters. *Reg. Sci. Urban Econ.* 38, 445–460.
- Davis, D.R., Weinstein, D.E., 2002. Bones, bombs, and break points: the geography of economic activity. *Am. Econ. Rev.* 92, 1269–1289.
- Dekle, R., Eaton, J., 1999. Agglomeration and land rents: Evidence from the prefectures. *J. Urban Econ.* 46, 200–214.
- Desmet, K., Henderson, J.V., 2015. The geography of development within countries. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, North-Holland, pp. 1457–1517.
- Desmet, K., Rappaport, J., 2013. The settlement of the United States, 1800 to 2000: the long transition towards Gibrat's law. Discussion Paper 9353, Centre for Economic Policy Research, London, UK.
- Desmet, K., Rossi-Hansberg, E., 2013. Urban accounting and welfare. *Am. Econ. Rev.* 103, 2296–2327.
- Di Addario, S., Vuri, D., 2010. Entrepreneurship and market size: the case of young college graduates in Italy. *Labour Econ.* 17 (5), 848–858.
- Diamond, R., 2013. The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000. Stanford University, processed.
- Duranton, G., 2006. Some foundations for zipf's law: product proliferation and local spillovers. *Reg. Sci. Urban Econ.* 36, 542–563.
- Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. *Am. Econ. Rev.* 97, 197–221.
- Duranton, G., Puga, D., 2000. Diversity and specialisation in cities: why, where and when does it matter? *Urban Stud.* 37, 533–555.
- Duranton, G., Puga, D., 2001. Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am. Econ. Rev.* 91, 1454–1477.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, North-Holland, pp. 2063–2117.
- Duranton, G., Puga, D., 2005. From sectoral to functional urban specialisation. *J. Urban Econ.* 57, 343–370.
- Eeckhout, J., 2004. Gibrat's law for (all) cities. *Am. Econ. Rev.* 94, 1429–1451.
- Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. *J. Pol. Econ.* 122, 554–620.
- Ellison, G., Glaeser, E.L., 1999. The geographic concentration of industry: does natural advantage explain agglomeration? *Am. Econ. Rev. Pap. Proc.* 89, 311–316.
- Ellison, G.D., Glaeser, E.L., Kerr, W.R., 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. *Am. Econ. Rev.* 100, 1195–1213.
- Ethier, W., 1982. National and international returns to scale in the modern theory of international trade. *Am. Econ. Rev.* 72, 389–405.
- Forslid, R., Okubo, T., 2014. Spatial relocation with heterogeneous firms and heterogeneous sectors. *Reg. Sci. Urban Econ.* 46, 42–56.
- Fujita, M., 1989. *Urban Economic Theory*. MIT Press, Cambridge, MA.

- Fujita, M., cois Thisse, J.F., 2013. *Economics of Agglomeration: Cities, Industrial Location, and Globalization*, second ed. Cambridge University Press, Cambridge, MA.
- Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Reg. Sci. Urban Econ.* 12, 161–196.
- Gabaix, X., 1999. Zipf's law for cities: an explanation. *Q. J. Econ.* 114, 739–767.
- Gabaix, X., Ibragimov, R., 2011. Rank-1/2: a simple way to improve the OLS estimation of tail exponents. *J. Bus. Econ. Stat.* 29, 24–39.
- Gabaix, X., Ioannides, Y.M., 2004. The evolution of city size distributions. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, North-Holland, pp. 2341–2378.
- Gaubert, C., 2014. *Firm Sorting and Agglomeration*. Princeton University, processed.
- Glaeser, E.L., 2008. *Cities, Agglomeration, and Spatial Equilibrium*. Oxford University Press, Oxford, UK.
- Glaeser, E.L., Gottlieb, J.D., 2009. The wealth of cities: agglomeration economies and spatial equilibrium in the United States. *J. Econ. Liter.* 47, 983–1028.
- Glaeser, E.L., Kerr, W.R., 2009. Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? *J. Econ. Manag. Strateg.* 18, 623–663.
- Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *J. Urban Econ.* 63, 1–24.
- Glaeser, E.L., Resseger, M., Tobia, K., 2009. Inequality in cities. *J. Reg. Sci.* 49 (4), 617–646.
- Glaeser, E.L., Kolko, J., Saiz, A., 2001. Consumer city. *J. Econ. Geogr.* 1, 27–50.
- Grossman, G.M., 2013. Heterogeneous workers and international trade. *Rev. World Econ.* 149, 211–245.
- Helpman, E., 1998. The size of regions. In: Pines, D., Sadka, E., Zilcha, I. (Eds.), *Topics in Public Economics*. Cambridge University Press, Cambridge, UK, pp. 33–54.
- Helpman, E., Itzhak, O., Redding, S.J., 2010. Inequality and unemployment in a global economy. *Econometrica* 78, 1239–1283.
- Helsley, R.W., Strange, W.C., 2011. Entrepreneurs and cities: complexity, thickness, and balance. *Reg. Sci. Urban Econ.* 44, 550–559.
- Helsley, R.W., Strange, W.C., 2014. Coagglomeration, clusters, and the scale and composition of cities. *J. Pol. Econ.* 122 (5), 1064–1093.
- Henderson, J.V., 1974. The sizes and types of cities. *Am. Econ. Rev.* 64, 640–656.
- Henderson, J.V., 1988. *Urban Development: Theory, Fact and Illusion*. Oxford University Press, New York, NY.
- Henderson, J.V., 1997. Medium size cities. *Reg. Sci. Urban Econ.* 27, 583–612.
- Henderson, J.V., Ono, Y., 2008. Where do manufacturing firms locate their headquarters? *J. Urban Econ.* 63, 431–450.
- Henderson, J.V., Venables, A.J., 2009. The dynamics of city formation. *Rev. Econ. Dyn.* 12, 233–254.
- Hendricks, L., 2011. The skill composition of US cities. *Int. Econ. Rev.* 52, 1–32.
- Holmes, T.J., Sieg, H., 2014. Structural estimation in urban economics. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, North-Holland.
- Holmes, T.J., Stevens, J.J., 2014. An alternative theory of the plant size distribution, with geography and intra- and international trade. *J. Pol. Econ.* 122 (2), 369–421.
- Hopenhayn, H.A., 1992. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica* 60, 1127–1150.
- Hsu, W.T., 2012. Central place theory and city size distribution. *Econ. J.* 122, 903–922.
- Jacobs, J., 1969. *The Economy of Cities*. Vintage, New York, NY.
- Kim, S., 1989. Labor specialization and the extent of the market. *J. Pol. Econ.* 97, 692–705.
- Kline, P., Moretti, E., 2014. People, places, and public policy: some simple welfare economics of local economic development programs. *Ann. Rev. Econ.* 6 (1), 629–662.
- Krugman, P.R., 1980. Scale economies, product differentiation, and the pattern of trade. *Am. Econ. Rev.* 70, 950–959.
- Krugman, P.R., 1991. Increasing returns and economic geography. *J. Pol. Econ.* 99, 483–499.
- Lee, S., 2010. Ability sorting and consumer city. *J. Urban Econ.* 68, 20–33.
- Lee, S., Li, Q., 2013. Uneven landscapes and city size distributions. *J. Urban Econ.* 78, 19–29.
- Lucas Jr., R.E., 1978. On the size distribution of business firms. *Bell J. Econ.* 9, 508–523.

- Lucas Jr., R.E., Rossi-Hansberg, E., 2002. On the internal structure of cities. *Econometrica* 70, 1445–1476.
- Marshall, A., 1890. *Principles of Economics*, eighth ed. Macmillan and Co., Ltd, London, UK, (1920) edition.
- Matano, A., Naticchioni, P., 2012. Wage distribution and the spatial sorting of workers. *J. Econ. Geogr.* 12, 379–408.
- Melitz, M.J., 2003. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica* 71, 1695–1725.
- Melitz, M.J., Ottaviano, G.I.P., 2008. Market size, trade and productivity. *Rev. Econ. Stud.* 75, 295–316.
- Melitz, M.J., Redding, S.J., 2014. Heterogeneous firms and trade. In: Helpman, E., Gopinath, G., Rogoff, K. (Eds.), *Handbook of International Economics*, vol. 4. Elsevier, North-Holland, pp. 1–54.
- Melo, P.C., Graham, D.J., Noland, R.B., 2009. A meta-analysis of estimates of urban agglomeration economies. *Reg. Sci. Urban Econ.* 39, 332–342.
- Michaels, G., Rauch, F., Redding, S.J., 2012. Urbanization and structural transformation. *Q. J. Econ.* 127, 535–586.
- Mion, G., Naticchioni, P., 2009. The spatial sorting and matching of skills and firms. *Can. J. Econ.* 42, 28–55.
- Moretti, E., 2004. Human capital externalities in cities. In: Henderson, J.V., cois Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, North-Holland, pp. 2243–2291.
- Mori, T., Turrini, A., 2005. Skills, agglomeration and segmentation. *Eur. Econ. Rev.* 49, 201–225.
- Mori, T., Nishikimi, K., Smith, T.E., 2008. The number-average size rule: a new empirical relationship between industrial location and city size. *J. Reg. Sci.* 48, 165–211.
- Mossay, P., Picard, P.M., 2011. On spatial equilibria in a social interaction model. *J. Econ. Theory* 146, 2455–2477.
- Mrázová, M., Neary, J.P., 2013. *Selection Effects with Heterogeneous Firms*. University of Surrey and Oxford University, processed.
- Murata, Y., 2003. Product diversity, taste heterogeneity, and geographic distribution of economic activities: market vs. non-market interactions. *J. Urban Econ.* 53, 126–144.
- Nocke, V., 2006. A gap for me: entrepreneurs and entry. *J. Eur. Econ. Assoc.* 4, 929–956.
- Okubo, T., Picard, P.M., cois Thisse, J.F., 2010. The spatial selection of heterogeneous firms. *J. Int. Econ.* 82, 230–237.
- Ossa, R., 2013. A gold rush theory of economic development. *J. Econ. Geogr.* 13, 107–117.
- Ota, M., Fujita, M., 1993. Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Reg. Sci. Urban Econ.* 23, 695–729.
- Ottaviano, G.I.P., 2012. Agglomeration, trade, and selection. *Reg. Sci. Urban Econ.* 42, 987–997.
- Piketty, T., 2014. *Capital in the 21st Century*. Harvard University Press, Cambridge, MA.
- Puga, D., 2010. Themagnitude and causes of agglomeration economies. *J. Reg. Sci.* 50, 203–219.
- Redding, S.J., 2012. *Goods trade, factormobility and welfare*. Technical Report, National Bureau for Economic Research, Cambridge, MA, NBER Discussion Paper.
- Rosen, S., 1981. The economics of superstars. *Am. Econ. Rev.* 71, 845–858.
- Rosenthal, S.S., Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., cois Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 1. Elsevier, North-Holland, pp. 2119–2171.
- Rosenthal, S.S., Strange, W.C., 2008a. Agglomeration and hours worked. *Rev. Econ. Stat.* 90, 105–118.
- Rosenthal, S.S., Strange, W.C., 2008b. The attenuation of human capital spillovers. *J. Urban Econ.* 64, 373–389.
- Rossi-Hansberg, E., Wright, M.L.J., 2007. Urban structure and growth. *Rev. Econ. Stud.* 74, 597–624.
- Rossi-Hansberg, E., Sarte, P.D., Owens III, R., 2009. Firm fragmentation and urban patterns. *Int. Econ. Rev.* 50, 143–186.
- Rozenfeld, H.D., Rybski, D., Gabaix, X., Makse, H.A., 2011. The area and population of cities: new insights from a different perspective on cities. *Am. Econ. Rev.* 101, 2205–2225.
- Saiz, A., 2010. The geographic determinants of housing supply. *Q. J. Econ.* 125, 1253–1296.
- Sato, Y., Tabuchi, T., Yamamoto, K., 2012. Market size and entrepreneurship. *J. Econ. Geogr.* 12, 1139–1166.

- Sattinger, M., 1993. Assignments models of the distribution of earnings. *J. Econ. Liter.* 31, 831–880.
- Syversen, C., 2004. Market structure and productivity: a concrete example. *J. Pol. Econ.* 112, 1181–1222.
- Syversen, C., 2007. Prices, spatial competition and heterogeneous producers: an empirical test. *J. Ind. Econ.* LV, 197–222.
- Tabuchi, T., cois Thisse, J.F., 2002. Taste heterogeneity, labor mobility and economic geography. *J. Dev. Econ.* 69, 155–177.
- Venables, A.J., 2011. Productivity in cities: self-selection and sorting. *J. Econ. Geogr.* 11, 241–251.
- Vermeulen, W., 2011. Agglomeration Externalities and Urban Growth Controls. SERB Discussion Paper 0093, Spatial Economics Research Centre, London School of Economics.
- Vives, X., 2001. *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press, Cambridge, MA.
- Wheeler, C.H., 2001. Search, sorting, and urban agglomeration. *J. Lab. Econ.* 19, 879–899.
- Wheeler, C.H., 2004. Wage inequality and urban density. *J. Econ. Geogr.* 4, 421–437.
- Wrede, M., 2013. Heterogeneous skills and homogeneous land: segmentation and agglomeration. *J. Econ. Geogr.* 13, 767–798.
- Zhelobodko, E., Kokovin, S., Parenti, M., cois Thisse, J.F., 2012. Monopolistic competition: beyond the constant elasticity of substitution. *Econometrica* 80, 2765–2784.

This page intentionally left blank

CHAPTER 5

The Empirics of Agglomeration Economies

Pierre-Philippe Combes^{*,†,‡}, Laurent Gobillon^{‡,§,¶,||}

^{*}Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Marseille, France

[†]Economics Department, Sciences Po, Paris, France

[‡]Centre for Economic Policy Research (CEPR), London, UK

[§]Institut National d'Etudes Démographiques, Paris, France

[¶]Paris School of Economics, Paris, France

^{||}The Institute for the Study of Labor (IZA), Bonn, Germany

Contents

5.1. Introduction	248
5.2. Mechanisms and Corresponding Specifications	252
5.2.1 Static agglomeration effects and individual skills	252
5.2.1.1 <i>Separate identification of skills and local effects</i>	252
5.2.1.2 <i>Heterogeneous impact of local effects</i>	260
5.2.2 Dynamic impact of agglomeration economies	262
5.2.3 Extending the model to local worker–firm matching effects	266
5.2.4 Endogenous intertemporal location choices	268
5.3. Local Determinants of Agglomeration Effects	270
5.3.1 Density, size, and spatial extent of agglomeration effects	271
5.3.2 Industrial specialization and diversity	274
5.3.3 Human capital externalities	278
5.4. Estimation Strategy	282
5.4.1 Wages versus TFP	282
5.4.2 Endogeneity issues	284
5.4.3 Dealing with endogenous local determinants	286
5.4.3.1 <i>Local fixed effects</i>	286
5.4.3.2 <i>Instrumentation with historical and geological variables</i>	287
5.4.3.3 <i>Generalized method of moments</i>	289
5.4.3.4 <i>Natural experiments</i>	290
5.4.4 Tackling the role of firm characteristics	292
5.4.5 Other empirical issues	294
5.4.5.1 <i>Spatial scale</i>	294
5.4.5.2 <i>Measures of observed skills</i>	295
5.4.5.3 <i>Functional form and decreasing returns to agglomeration</i>	296
5.4.5.4 <i>Spatial lag models</i>	297
5.5. Magnitudes for the Effects of Local Determinants of Productivity	298
5.5.1 Economies of density	298
5.5.2 Heterogeneous effects	303
5.5.3 Spatial extent of density effects	306

5.5.4	Market access effect evaluated using natural experiments	307
5.5.5	Specialization and diversity	309
5.5.6	Human capital externalities	310
5.5.7	Developing economies	311
5.6.	Effects of Agglomeration Economies on Outcomes Other Than Productivity	314
5.6.1	Industrial employment	315
5.6.1.1	<i>From productivity externalities to employment growth</i>	315
5.6.1.2	<i>Total employment, specialization, diversity, and human capital</i>	319
5.6.1.3	<i>Dynamic specifications</i>	321
5.6.2	Firms' location choices	322
5.6.2.1	<i>Strategies and methodological concerns</i>	322
5.6.2.2	<i>Discrete location choice models</i>	324
5.6.2.3	<i>Firm creation and entrepreneurship</i>	327
5.7.	Identification of Agglomeration Mechanisms	328
5.7.1	Labor mobility, specialization, matching, and training	329
5.7.2	Industrial spatial concentration and coagglomeration	331
5.7.3	Case studies	336
5.8.	Conclusion	338
	Acknowledgments	340
	References	341

Abstract

We propose an integrated framework to discuss the empirical literature on the local determinants of agglomeration effects. We start by presenting the theoretical mechanisms that ground individual and aggregate empirical specifications. We gradually introduce static effects, dynamic effects, and workers' endogenous location choices. We emphasize the impact of local density on productivity, but we also consider many other local determinants supported by theory. Empirical issues are then addressed. The most important concerns are about endogeneity at the local and individual levels, the choice of a productivity measure between wages and total-factor productivity, and the roles of spatial scale, firms' characteristics, and functional forms. Estimated impacts of local determinants of productivity, employment, and firms' location choices are surveyed for both developed and developing economies. We finally provide a discussion of attempts to identify and quantify specific agglomeration mechanisms.

Keywords

Agglomeration gains, Density, Sorting, Learning, Location choices

JEL Classification Codes

R12, R23, J31

5.1. INTRODUCTION

Ongoing urbanization is sometimes interpreted as evidence of gains from agglomeration that dominate its costs, otherwise firms and workers would remain sparsely distributed. One can imagine, however, that the magnitude of agglomeration economies depends on

the type of workers and industries, as well as on the period and country. This is a first motivation to quantify agglomeration economies precisely, which is the general purpose of the literature reviewed in this chapter. Moreover, firms' and workers' objectives, profit and utility, are usually not in line with collective welfare or the objective that some policy makers may have in particular for productivity or employment. Even if objectives were identical, individual decisions may not lead to the collective optimum as firms and workers may not correctly estimate social gains from spatial concentration when they choose their location. Generally speaking, an accurate estimation of the magnitude of agglomeration economies is required when one tries to evaluate the need for larger or smaller cities. If one were to conclude that the current city size distribution is not optimal, such an evaluation would be necessary for the design of policies (such as taxes or regulation) that should be implemented to influence agents' location choices toward the social optimum. Lastly, many *a priori* aspatial questions can also be indirectly affected by the extent to which firms and workers relocate across cities, as for instance, inequalities among individuals and the possible need for policies to correct them. Inequality issues might be less severe when workers are mobile and they rapidly react to spatial differences in the returns to labor. Addressing such questions requires beforehand a correct assessment of the magnitude of agglomeration economies.

Agglomeration economies is a large concept that includes any effect that increases firms' and workers' income when the size of the local economy grows. The literature proposes various classifications for the different mechanisms behind agglomeration economies, from [Marshall \(1890\)](#), who divides agglomeration effects into technological spillovers, labor pooling, and intermediate input linkages, to the currently most used typology proposed by [Duranton and Puga \(2004\)](#), who rather consider sharing, matching, and learning effects. Sharing effects include the gains from a greater variety of inputs and industrial specialization, the common use of local indivisible goods and facilities, and the pooling of risk; matching effects correspond to improvement of either the quality or the quantity of matches between firms and workers; learning effects involve the generation, diffusion, and accumulation of knowledge. Ultimately one would like an empirical assessment of the respective importance of each of these components. Unfortunately, the literature has not reached this goal yet, and we will see that there are only rare attempts to distinguish the various channels behind agglomeration economies. They are mostly descriptive and we present them at the end of this chapter. We choose rather to detail the large literature that tries to evaluate the overall impact on local outcomes of spatial concentration, and of a number of other characteristics of the local economy, such as its industrial structure, its labor force composition, or its proximity to large locations. In other words, what is evaluated is the impact on some local outcomes of local characteristics that shape agglomeration economies through a number of channels, not the channels themselves. Local productivity and wages have been the main focus of attention, but we also present the literature that studies how employment and firm location decisions are influenced by local characteristics.

When estimating the overall impact of a local characteristic, such as the impact of local employment density on local productivity, one cannot know whether the estimated effect arises mostly from sharing, matching, or learning mechanisms, or from all of them simultaneously. Most positive agglomeration effects can also turn negative above some city size threshold, or can induce some companion negative effects, and one cannot say whether some positive effects are partly offset by negative ones, as only the total net impact is evaluated. Moreover, while some mechanisms imply immediate static gains from agglomeration, other effects are dynamic and influence local growth. We take into account all these theoretical issues in our framework of analysis, as this is required to correctly choose relevant empirical specifications, correctly interpret the results, and discuss estimation issues. Crucially, even if the effects of mechanisms related to agglomeration economies are not identified separately, knowing, for instance, by how much productivity increases when one increases the number of employees per square meter in a city is crucial for the understanding of firms' and workers' location choices or for the design of economic policies.

We will see that the role of local characteristics is already not that trivial to evaluate. Beyond some interpretation issues that we will detail, the main difficulty arises from the fact that one does not seek to identify correlations between local characteristics and a local outcome but seeks to identify causal impacts. Basic approaches can lead to biased estimates because of endogeneity concerns at both the local level and the individual level. Endogeneity issues at the local level arise from either aggregate missing variables that influence both local outcomes and local characteristics, or reverse causality as better average local outcomes can attract more firms and workers in some locations, which in turn affects local characteristics. Endogeneity issues at the individual level occur when workers self-select across locations according to individual factors that cannot be controlled for in the specification, typically some unobserved abilities, or when they choose their location according to their exact individual outcome that depends on individual shocks possibly related to local characteristics. Dealing with these various sources of endogeneity is probably the area where the literature has made the greatest progress over the last decade. It is not possible anymore to evaluate the determinants of local outcomes without addressing possible endogeneity issues. Therefore, we largely discuss the sources of endogeneity and the solutions proposed in the literature.

Since various agglomeration mechanisms are at work and the impact of many local characteristics on different local outcomes has been studied, it is necessary to first clarify the theories that are behind the specifications estimated in the literature. [Section 5.2](#) starts from a simple model and the corresponding specification that emphasizes the determinants of local productivity. This model is then progressively extended to encompass additional mechanisms, moving from static specifications to dynamic frameworks, while stressing the role of individual characteristics and individual location choices. This approach helps to clarify some of the endogeneity issues. [Section 5.3](#) presents all the local

characteristics whose impact on productivity is studied in the literature, and relates them to theory. With such a theoretical background in mind, we systematically discuss a series of empirical issues in [Section 5.4](#), mostly endogeneity concerns at the local and individual levels, as well as the solutions proposed to tackle them. We also discuss the choice of a productivity measure between wages and total-factor productivity (TFP), and the roles of spatial scale, firms' characteristics, and functional forms. The magnitudes of estimated agglomeration effects on productivity are presented in [Section 5.5](#), which covers in particular the effect of density, its spatial extent, and some possible heterogeneity of the impact across industries, skills, and city sizes. [Section 5.5](#) also presents the results of some recent studies that use a structural approach or exploit natural experiments, as well as results on the role of the industrial structure of the local economy (namely, industrial specialization and diversity) and human capital externalities. Recent results for developing economies are detailed separately as the magnitudes are often not the same as for developed countries and their study is currently being expanded. In [Section 5.6](#), estimated agglomeration effects on employment and firms' location choices instead of productivity are discussed, after starting with considerations related to theory and the choice of a relevant empirical specification. Finally, [Section 5.7](#) presents attempts to identify the channels through which agglomeration economies operate. The identification of such channels is one of the current concerns in the literature.

The organization of our chapter does not follow the development of the field over time. The literature started with the ambitious goal of estimating the impact of a large number of local determinants on employment growth at the city-industry level ([Glaeser et al., 1992](#); [Henderson et al., 1995](#)). However, acknowledging some possibly serious interpretation and endogeneity concerns, the literature then became more parsimonious, focusing on static agglomeration effects on local productivity only (see [Ciccone and Hall, 1996](#); [Glaeser and Maré, 2001](#); [Combes et al., 2008a](#)). This was also made possible thanks to the availability of new datasets with a panel dimension at the individual level. More recent contributions incorporate additional effects such as the dynamic ones already suggested in the previous literature (see [de la Roca and Puga, 2012](#)), or consider richer frameworks through structural models involving endogenous location choices and different sources of heterogeneity across firms and workers (see [Gould, 2007](#); [Baum-Snow and Pavan, 2012](#)). We choose to start with a simple but rigorous framework to analyze the effects of local determinants of productivity, which we then extend. Most of the contributions in the literature are ultimately encompassed, and this includes earlier ones focusing on employment growth. When referring to magnitudes of the effects, we focus more particularly on contributions later than those surveyed in [Rosenthal and Strange \(2004\)](#), but we refer to earlier contributions when they are useful for our discussion.

Still, there are a number of related topics that we do not cover, mostly because they involve too much material and the handbook editors made the choice of devoting

separate chapters to them. In particular, a specific case where the effect of an agglomeration mechanism can be identified is technological spillovers and the links between agglomeration and innovation. This topic is covered by [Carlino and Kerr \(2015\)](#), who also discuss the literature on agglomeration and entrepreneurship, as it is often grounded on technological spillovers. Similarly, we do not cover the literature on the interactions between agglomeration economies and place-based policies, since it is considered in [Neumark and Simpson \(2015\)](#). Finally, we do not present the various attempts made to measure spatial concentration. Nevertheless, we refer to spatial concentration indices in the last part of the survey as some articles use them in regressions to attempt to identify mechanisms of agglomeration economies.

5.2. MECHANISMS AND CORRESPONDING SPECIFICATIONS

It is not possible to discuss the estimation of agglomeration economies without first clarifying the theories and underlying mechanisms that are assessed empirically by the literature. This section presents these theories so that we can then correctly interpret estimates and discuss possible estimation issues.

5.2.1 Static agglomeration effects and individual skills

5.2.1.1 *Separate identification of skills and local effects*

The earlier literature studies agglomeration economies at an aggregate spatial level, the region or the city. An outcome in a local market is typically regressed on a vector of local variables. In this section, we focus mostly on the impact of the logarithm of density on the logarithm of workers' productivity, measured by nominal wage. This corresponds to the relationship considered by [Ciccone and Hall \(1996\)](#), who had a large impact on the recent evolution of the literature. The role of other local determinants such as market access, industrial diversity, or specialization has also been considered, and will be detailed in [Section 5.3](#). Other local outcomes such as industry employment growth or firms' location choices will be discussed in [Section 5.6](#).

Let us first consider a setting without individual heterogeneity among firms and workers. Let $Y_{c,t}$ be the output of a representative firm located in market c at date t . The firm uses two inputs, labor $L_{c,t}$, and other factors of production $K_{c,t}$, such as land, capital, or intermediate inputs. The profit of the firm is given by

$$\pi_{c,t} = p_{c,t}Y_{c,t} - \omega_{c,t}L_{c,t} - r_{c,t}K_{c,t}, \quad (5.1)$$

where $p_{c,t}$ is the price of the good produced, $\omega_{c,t}$ is the wage rate in the local labor market, and $r_{c,t}$ is the unit cost of nonlabor inputs. Suppose that the production function is of the Cobb–Douglas type and can be written as

$$Y_{c,t} = \frac{A_{c,t}}{\alpha^\alpha (1-\alpha)^{1-\alpha}} (s_{c,t} L_{c,t})^\alpha K_{c,t}^{1-\alpha}, \quad (5.2)$$

where $0 < \alpha < 1$ is a parameter, $A_{c,t}$ is the local TFP, and $s_{c,t}$ corresponds to local labor skills. As long as all local firms and workers are assumed to be identical, these quantities depend on c and t only. In turn, this is also the case for $p_{c,t}$, $w_{c,t}$, and $r_{c,t}$. In a competitive equilibrium, an assumption we discuss below, the first-order conditions for the optimal use of inputs reduce to

$$w_{c,t} = \left(p_{c,t} \frac{A_{c,t}}{(r_{c,t})^{1-\alpha}} \right)^{1/\alpha} s_{c,t} \equiv B_{c,t} s_{c,t}. \quad (5.3)$$

The local average nominal wage depends on labor skills, $s_{c,t}$, as well as on a composite local productivity effect, $B_{c,t}$. This equation is enough to encompass almost all agglomeration effects that the literature has considered. If one goes back as far as [Buchanan \(1965\)](#), cities are places where firms and consumers share indivisible goods such as airports, universities, and hospitals, which generate a first type of agglomeration economies. In that case, the composite labor productivity effect, $B_{c,t}$, and therefore the local average wage, are higher in larger cities because $A_{c,t}$ is larger owing to the presence of local (public) goods. This corresponds to a first type of pure local externality in the sense that it is not mediated by the market. A second type of pure local externality, very different in nature, emerges when spatial concentration induces local knowledge spillovers that make firms more productive, as put forward in early endogenous growth models such as that of [Lucas \(1988\)](#). Again, this type of mechanism makes $A_{c,t}$ larger in larger cities. For the moment, we implicitly assume that all these effects are instantaneous and affect only current values of $A_{c,t}$. This is an important restriction that we discuss further below.

Economists have also emphasized a number of agglomeration mechanisms operating through local markets, sometimes referred to as “pecuniary externalities.” Because access to markets is better in larger cities, the price of goods there, $p_{c,t}$, can be higher, and the costs of inputs, $r_{c,t}$, lower. Both effects again make $B_{c,t}$ larger.¹ Ultimately, one would like to assess separately whether pure externalities or local market effects have the most significant role effect on local productivity, or whether, among market effects, local

¹ When a firm sells to many markets, $p_{c,t}$ corresponds to the firm’s average income per unit sold, which encompasses trade costs, and the present analysis can easily be extended, as shown by [Combes \(2011\)](#). Let $Y_{c,r,t}$ denote the firm’s exports to any other market r . The output value is the sum of the value of sales in all markets, $p_{c,t} Y_{c,t} = \sum_r (p_{c,r,t} - \tau_{c,r,t}) Y_{c,r,t} = (\sum_r (p_{c,r,t} - \tau_{c,r,t}) \phi_{c,r,t}) Y_{c,t}$, where $p_{c,r,t}$ is the firm’s price in market r , $\tau_{c,r,t}$ represents trade costs paid by the firm to sell in market r , and $\phi_{c,r,t} = \frac{Y_{c,r,t}}{Y_{c,t}}$ is its share of output that is sold there. As a result, $p_{c,t} = \sum_r (p_{c,r,t} - \tau_{c,r,t}) \phi_{c,r,t}$ is the average of the firm’s prices over all its markets net of trade costs and weighted by its share of sales in each market. The closer to large markets the firm is, the lower the trade costs and the higher this average price. Similarly, when firms buy inputs from many markets, the closer these markets are, the lower the firms’ average unit cost of inputs, $r_{c,t}$.

productivity gains arise from price effects mostly related to goods or inputs. However, such assessments are difficult, and a large part of the empirical literature on agglomeration economies simply quantifies the overall impact on productivity of characteristics of the local economy. The previous discussion shows, in particular, that the positive correlation between wages and density can result from pure externalities as well as effects related to good or input prices.

Furthermore, city size generates not only agglomeration economies but also dispersion forces. Typically, the cost of inputs that are not perfectly mobile, $r_{c,t}$, land at one extreme, is higher in larger cities. If competition is tough enough relative to the benefits from market access in large cities, the price of goods there, $p_{c,t}$, can be lower than in smaller cities. Congestion on local public goods can also emerge, which reduces $A_{c,t}$. Note also that if local labor markets are not competitive, the right-hand side in Equation (5.3) should be multiplied by a coefficient that depends on the local bargaining power of workers. If workers have more bargaining power in larger cities, their nominal wages are higher, and this constitutes an agglomeration effect. Alternatively, a lower bargaining power in larger cities is a dispersion force. The correlation between wage and density reflects only the overall impact of both agglomeration economies and dispersion forces. While the net effect of spatial concentration can be identified, this is not the case for the channels through which it operates. Conversely, if one wants to quantify independently the impact of market effects operating through $r_{c,t}$ and $p_{c,t}$, a strategy is required involving controls for pure externalities arising, for instance, from the presence of local public goods or local spillovers.

One can also consider the inclusion of controls for dispersion forces if data on local traffic congestion or housing/land prices, for instance, are available. This is a start to disentangling agglomeration economies and dispersion forces. Importantly, the motivation for introducing housing/land prices is their influence on the costs of inputs and not compensation for low or high wages in equilibrium such that workers are indifferent between places as in [Roback \(1982\)](#). Indeed, we are focusing here on the determinants of productivity and not on equilibrium relationships. Typically, land price is expected to have a negative impact on nominal wages in accordance with Equation (5.3), while the equilibrium effect implies a positive correlation between the two variables. As wages and land prices are simultaneously determined in equilibrium, controlling for land or housing prices can lead to serious endogeneity biases that are difficult to deal with (see the discussion in [Section 5.4](#)). This suggests that if land represents a small share of input costs, which is usually the case, it is probably better not to control for its price in regressions.

Testing the relevance of a wage compensation model and quantifying real wage inequalities between cities are interesting questions but they require considering simultaneously the roles of nominal wages, costs of living, and amenities. These questions are addressed in a burgeoning literature ([Albouy, 2009](#); [Moretti, 2013](#)), which we briefly discuss in the conclusion. As far as the effect of agglomeration economies on productivity

only is concerned, the nominal wage constitutes the relevant dependent variable and there is no need to control for land prices as illustrated by our model.

Let us turn to the role of local labor skills, captured in Equation (5.3) by $s_{c,t}$. If workers have skills that are not affected by their location, typically inherited from their parents or acquired through education, one definitively does not want to include the effect of skills among agglomeration economies, since it corresponds to a pure composition effect of the local labor force and not an increase in productivity due to local interactions between workers. It is possible that, for reasons not related to agglomeration economies, higher skills are over-represented in cities. This can arise, for instance, if skilled workers value city amenities (related, for instance, to culture or nightlife) more than unskilled ones do or if, historically, skilled people have located more in larger cities and transmit part of their skills to their children who stay there. If the estimation strategy does not control for the selection of higher skills in cities, other local variables such as density capture their role, and the impact of agglomeration economies can be overstated. Alternatively, it is also possible that people are made more skilled by cities, through stronger learning effects in larger cities, or that skilled people generate more local externalities, as suggested by Lucas (1988). In that case, not controlling for the skill level in the city is the correct way to capture the total agglomeration effect due to a larger city size. *A priori*, both the composition effect and the agglomeration effect can occur, and a local measure of skills or education captures both. The aggregate approach at the city level discussed here does not consider individual heterogeneity and does not allow the separate identification of the two effects. This is its first important limit, and an individual data approach is more useful for that purpose, as detailed below.

Finally, a crucial issue is the time span of agglomeration effects. One can accept that productivity and then wages adjust quickly to variations in market-mediated agglomeration effects (operating through changes in $r_{c,t}$ and $p_{c,t}$), but they definitely do not for variations of most pure local externalities that can affect $A_{c,t}$ and $s_{c,t}$. Therefore, the literature tends to distinguish between static and dynamic agglomeration effects. When agglomeration effects are static, $B_{c,t}$ is immediately affected by current values of local characteristics but not by earlier values. This means that a larger city size in a given year affects local productivity only in that year, and that any future change in city size will instantaneously translates into a change in local productivity. By contrast, recent contributions simultaneously consider some possible long-lasting effects of local characteristics that are called dynamic effects. We focus here on static affects and introduce dynamic effects from Section 5.2.2 onward.

Let us turn now to a first empirical specification encompassing static agglomeration effects where the logarithm of the composite productivity effect, $B_{c,t}$, is specified in reduced form as a function of the logarithm of local characteristics and some local unobserved effects. Average local skills, $s_{c,t}$, are specified as a log-linear function of local education and again some local unobserved terms. The sum of all unobserved components is supposed

to be a random residual denoted $\eta_{c,t}$. Denoting $y_{c,t}$ as the measure of the local outcome, here the logarithm of local wage, we obtain from Equation (5.3) the specification

$$y_{c,t} = Z_{c,t}\gamma + \eta_{c,t}, \quad (5.4)$$

where $Z_{c,t}$ includes local variables for both the local composite productivity component and skills. If explanatory variables reduce to the logarithm of density and local skills variables capturing only skill composition effects, and that there is no correlation between the random component and explanatory variables, then the ordinary least squares (OLS) estimate of the elasticity of productivity with respect to density is a consistent measure of total net agglomeration economies. This elasticity is crucial from the policy perspective even if the channels of agglomeration economies and dispersion forces are not identified. For instance, a value for the elasticity of the local outcome with respect to density of 0.03 means that a city twice as large (knowing that a factor of 10 is often obtained for the inter-quartile of local density in many countries) has $2^{0.03} - 1 \approx 2.1\%$ greater productivity, because of either pure local externalities or market agglomeration effects that dominate dispersion effects of any kind.

As mentioned in Section 5.1, the usual goal of the empirical works is to identify causal impacts—that is, what would be the effect on local outcomes of changing some of the local characteristics. Beyond other endogeneity concerns discussed below, a first issue with specification (5.4) is that density can be correlated with some of the local unobserved skill components entering the residual. For instance, proxies for local skills such as diplomas may not be enough to capture all the skills that affect productivity. If unobserved skills are randomly distributed across locations, the OLS estimate of the density parameter is a consistent estimator of the magnitude of agglomeration economies. Alternatively, if unobserved skills are correlated with density, there is an endogeneity issue and the OLS estimate is biased.

Unobserved skills can be taken into account with individual panel data. This requires us to extend our setting to the case where workers are heterogeneous. We assume now that local efficient labor is given by the sum of all efficient units of labor provided by heterogeneous workers—that is, $s_{c,t}L_{c,t} = \sum_{i \in \{c,t\}} s_{i,t} \ell_{i,t}$, where $\ell_{i,t}$ is the number of working hours provided by individual i and $s_{i,t}$ is individual efficiency at date t . The wage bill is now $\sum_{i \in \{c,t\}} w_{i,t} \ell_{i,t}$, where $w_{i,t}$ is the individual wage. Profit maximization leads to

$$w_{i,t} = B_{c,t} s_{i,t}. \quad (5.5)$$

Let $X_{i,t}$ be time-varying observed individual characteristics and u_i be an individual fixed effect to be estimated. We make the additional assumption that individual efficiency can be written as the product of an individual-specific component, $\exp(X_{i,t}\theta + u_i)$, and a residual, $\exp(\epsilon_{i,t})$, reflecting individual- and time-specific random effects. Here, u_i captures the effects of individual unobserved skills which are supposed to be constant over time. Taking the logarithm of (5.5) and using the same specification of agglomeration effects as for (5.4) gives

$$y_{i,t} = u_i + X_{i,t}\theta + Z_{c(i,t),t}\gamma + \eta_{c(i,t),t} + \epsilon_{i,t}, \quad (5.6)$$

where $y_{i,t}$ is the individual local outcome, here the logarithm of individual wage at date t , and $c(i,t)$ is the labor market where individual i is located at date t . Note that we implicitly assume a homogeneous impact of local characteristics γ across all workers, areas, and industries. Heterogeneous impacts are considered in [Section 5.2.1.2](#). For now, we consider that individual fixed effects are here only to capture unobservable skills, although we will discuss in [Section 5.2.2](#) the fact that they can also capture learning effects that may depend on city size.

The use of individual data and the introduction of an individual fixed effect in specification (5.6) were first proposed by [Glaeser and Maré \(2001\)](#), and this should largely reduce biases due to the use of imperfect measures of skills. Most importantly, the individual fixed effect makes it possible to control for all the characteristics of the individual shaping skills that do not change over time and the effect of which can be considered to be constant over time. They include education, which is often observable, but also many other characteristics that are more difficult to observe, such as the education of parents and grandparents, the number of children in the family, mobility during childhood, and personality traits. Since the individual fixed effects are allowed to be correlated with local variables such as density, one can more safely conclude that the effects of local characteristics do not capture some composition effects owing to sorting on the individual characteristics.

The second advantage of individual data is that the local average of any observed individual characteristic can be introduced in the set of local variables simultaneously with the individual characteristic itself or with the individual fixed effect. In particular, while the individual fixed effect controls for the individual level of education, one can consider in $Z_{c,t}$ the local share of any education level to assess whether highly skilled workers exert a human capital local externality on other workers.² The estimated effects of local variables such as density then correspond to agglomeration economies other than education externalities. As discussed above, such a distinction cannot be made when using aggregate data.

The sources of identification of local effects can be emphasized by considering specification (5.6) in first difference, which makes the unobserved individual effect disappear. For simplicity's sake, consider only two terms in the individual outcome specification such that $y_{i,t} = Z_{c(i,t),t}\gamma + u_i$, where $Z_{c,t}$ includes only density. For individuals staying in the same local market c at two consecutive dates, the first difference of outcome is given by $y_{i,t} - y_{i,t-1} = (Z_{c,t} - Z_{c,t-1})\gamma$, and time variation of density within the local market participates in the identification of the density effect, γ . For individuals moving from market c to market c' , we have $y_{i,t} - y_{i,t-1} = (Z_{c',t} - Z_{c,t-1})\gamma$, and both spatial and time variations of density contribute to identifying the density effect. If there is no mover,

² The interpretation based on externalities requires further caution. It is discussed in [Section 5.3.3](#).

agglomeration economies are still identified, but from time variations for stayers only. This is because there is a single parameter to estimate, and averaging the first-differenced outcome equation of stayers at the local-time level, one gets $Z \times (T - 1)$ independent relationships, where Z is the number of local markets.

Note that we assume for the moment that the specification is the same for stayers and movers—that is, that the individual parameters θ , the effects of local characteristics γ , and the distributions of random components are identical. Should this assumption be questioned, one could choose to estimate (5.6) separately on the subsamples of stayers and movers since identification is assured for each subsample, and one could in turn use the separate estimates to test the assumption of homogeneity across the two groups.

Specification (5.6) can be estimated directly by OLS once it has been written in first difference (or projected in the within-individual dimension) to remove the individual fixed effects, but the computation of standard errors is an issue. Indeed, the covariance matrix has a complex structure owing to unobserved local effects and the mobility of workers across labor markets. For mobile individuals, the first difference of the specification includes two different unobserved local shocks, $\eta_{c',t}$ and $\eta_{c,t-1}$, and the locations of those shocks (c and c') vary across mobile individuals, even for those initially in the same local market because they may not have the same destination after they move. There is thus no way to sort individuals properly to get a simple covariance matrix structure and to cluster standard errors at each date by location. It is tempting to ignore unobserved local effects, but this can lead to important biases of the estimated standard errors for effects of local variables, as shown by [Moulton \(1990\)](#).

Alternatively, it is possible to use a two-step procedure that both solves this issue and has the advantage of corresponding to a more general framework. Consider the following system of two equations:

$$y_{i,t} = u_i + X_{i,t}\theta + \beta_{c(i,t),t} + \epsilon_{i,t}, \quad (5.7)$$

$$\beta_{c,t} = Z_{c,t}\gamma + \eta_{c,t}, \quad (5.8)$$

where $\beta_{c,t}$ is a local-time fixed effect that captures the role of any location-time variable whether it is observed or not. The introduction of such fixed effects capturing local unobserved components makes the assumption of independently distributed individual shocks more plausible. The specification is also more general since it takes into account possible correlations between local-time unobserved characteristics and individual characteristics. There are thus fewer possible sources of biases, and this in turn should lead to a more consistent evaluation of the role of local characteristics.

Estimating this model is more demanding in terms of identification, and having movers between locations is now required. Assume for simplicity's sake that the first equation of the model is given by $y_{i,t} = \beta_{c(i,t),t} + u_i$. When one rewrites this specification in first difference for nonmovers and movers, one gets $y_{i,t} - y_{i,t-1} = \beta_{c,t} - \beta_{c,t-1}$ and $y_{i,t} - y_{i,t-1} =$

$\beta_{c',t} - \beta_{c,t-1}$, respectively. There is one parameter $\beta_{c,t}$ to be identified for each location at each date. If there is no mover, one wishes to average the specification at the local-time level for stayers as before but ends up with $(Z - 1) \times T$ independent relationships, whereas there are $Z \times T$ parameters to estimate. In other words, one can identify the time variations of local effects for any location but not their differences between locations.

By contrast, when there are both stayers and movers, identification is assured as can be shown rewriting the specification in difference in differences. The difference of the wage time variation between a mover to c' , denoted i' , and a nonmover i initially in the same location c is given by $(y_{i',t} - y_{i',t-1} - (y_{i,t} - y_{i,t-1})) = \beta_{c',t} - \beta_{c,t}$. For any pair of locations, the difference in wage growth between movers and nonmovers identifies the difference of local effects between the two locations. Moreover, the wage growth of stayers identifies the variation of local effects over time as before. All parameters $\beta_{c,t}$ are finally identified when local markets are well interconnected through stayers and flows of movers, up to one that needs to be normalized to zero as differences do not allow the identification of levels. Interconnection means that any pair of location-time couples, (c, t) and (c', t') , can be connected through a chain of pairs of location-time couples $(j, \tau - 1)$ and (j', τ) such that there are migrants from j to j' between dates $\tau - 1$ and τ if $j \neq j'$, or stayers in j between the two dates if $j = j'$.³ In other words, assuming that there are some migrants between every pair of locations in the dataset, we have $Z^2 \times (T - 1)$ independent relationships and only $Z \times T - 1$ parameters to estimate. Crucially, the assumption that the specification is identical for both movers and stayers is now required, otherwise identification is not possible. Alternatively, more structural approaches can help to some extent to solve the identification issue, and we present them in [Section 5.2.4](#).

Note finally that in practice specification (5.7) is estimated in a first step. Panel data estimation techniques such as within estimation are used because considering a dummy variable for each individual to take into account the fixed effect u_i would be too demanding for a computer. The estimates of $\beta_{c,t}$ are then plugged into Equation (5.8). The resulting specification is estimated in a second stage using linear methods, including one observation for the location-time fixed effect normalized to zero. The sampling error on the dependent variable, which is estimated in the first stage, must be taken into account in the computation of standard errors, and it is possible to use feasible general least squares (see [Combes et al., 2008a](#), for the implementation details). A more extensive discussion on the estimation strategy addressing endogeneity issues is presented in

³ If local markets are not all interconnected, groups of fully interconnected location-time couples must be defined *ex ante* such that location-time fixed effects are all identified within each group up to one being normalized to zero. For more details, the reader may refer to the literature on the simultaneous identification of worker and firm fixed effects in wage equations initiated by [Abowd et al. \(1999\)](#).

Section 5.4, but we first augment the model to consider the role of more sophisticated agglomeration mechanisms.

5.2.1.2 Heterogeneous impact of local effects

The profit maximization we conducted above to ground our specification emphasizes that agglomeration effects may relate to pure externalities, or to good or input price effects. Obviously, the magnitude of these channels may differ across industries. For instance, the impact of density may be greater in high-tech industries owing to greater technological externalities, and good or input price effects depend on the level of trade costs within each industry. The consideration of agglomeration mechanisms that are heterogeneous across industries simply requires extending the specification such that

$$y_{i,t} = u_i + X_{i,t}\theta + Z_{c(i,t),t}\gamma_{s(i,t)} + \eta_{c(i,t),s(i,t),t} + \epsilon_{i,t}, \quad (5.9)$$

where $s(i,t)$ is the industry where individual i works at time t , γ_s is the effect of local characteristics in industry s , and $\eta_{c,s,t}$ is a location-industry-time shock. This specification can be estimated in several ways. The most straightforward one consists in splitting the sample by industry and implementing the approach proposed in Section 5.2.1.1 for each industry separately. Nevertheless, this means that the coefficients of individual explanatory variables as well as individual fixed effects are not constrained to be the same across industries, which may or may not be relevant from a theoretical point of view. This also entails a loss of precision for the estimators. An alternative approach consists in considering among explanatory variables some interactions between density, or any other local characteristic, and industry dummies, and estimating the specification in the within-individual dimension as before to recover their coefficients which are the parameters γ_s .

Again, estimated standard errors may be biased owing to heteroskedasticity arising from location-industry-time random effects, $\eta_{c,s,t}$. To deal with this issue, it is possible to consider a two-step approach which makes use of location-industry-time fixed effects, $\beta_{c,s,t}$, in the following system of equations:

$$y_{i,t} = u_i + X_{i,t}\theta + \beta_{c(i,t),s(i,t),t} + \epsilon_{i,t}, \quad (5.10)$$

$$\beta_{c,s,t} = Z_{c,t}\gamma_s + \eta_{c,s,t}. \quad (5.11)$$

Location-industry-time fixed effects are estimated with OLS once Equation (5.10) has been projected in the within-individual dimension, as done previously when estimating location-time fixed effects. They are identified up to one effect normalized to zero provided that all locations and industries are well interconnected by workers mobile across locations and industries.⁴ Their estimators are plugged into Equation (5.11), which is estimated in a second stage.

⁴ As before, groups of fixed effects should be defined *ex ante* if not all locations and industries are properly interconnected. Of course, the larger the number of industries, the more likely it is that location-industry-time fixed effects are not all identified.

Importantly, introducing the industry dimension increases the number of local characteristics that can have an agglomeration effect. It has become common practice to distinguish between urbanization economies and localization economies. Whereas urbanization economies correspond to externalities arising from characteristics of the location such as density, localization economies correspond to externalities arising from characteristics of the industry within the location. The determinants of agglomeration economies considered in the literature thus depend only on location for urbanization economies and on both location and industry for localization economies. The local determinant of localization economies most often considered is specialization, which is defined as the share of the industry in local employment. While the use of density makes it possible to assess whether productivity increases with the overall size of the local economy, the use of specialization allows the assessment of whether it increases with the local size of the industry in which the firm or worker operates. The pure externalities and market externalities distinguished above can operate at the whole location scale or at the industry-location level. In line with these arguments, one may rather want to estimate in the second step the following specification:

$$\beta_{c,s,t} = Z_{c,t}\gamma_s + W_{c,s,t}\delta_s + \eta_{c,s,t}, \quad (5.12)$$

where $W_{c,s,t}$ are determinants of localization economies including specialization and $Z_{c,t}$ are the determinants of urbanization economies. All the local characteristics considered in the literature are detailed in [Section 5.3](#).

One estimation issue is that the number of fixed effects to estimate in the first stage increases rapidly with the number of locations, and we are not aware of any attempt to estimate the proposed specification. As an alternative, one can mix strategies as proposed by [Combes et al. \(2008a\)](#) and estimate

$$y_{i,t} = u_i + X_{i,t}\theta + \beta_{c(i,t),t} + W_{c(i,t),s(i,t),t}\delta_{s(i,t)} + \epsilon_{i,t}, \quad (5.13)$$

$$\beta_{c,t} = Z_{c,t}\gamma + \eta_{c,t}. \quad (5.14)$$

This model is less general than (5.10) and (5.12) since unobserved location-industry-time effects are not controlled for in the first step, and determinants of urbanization economies are assumed to have a homogeneous impact across industries in the second step (as γ does not depend on the industry). Still, heterogeneous effects of determinants of localization economies are identified in the first stage on top of controlling for unobserved location-time effects.

It is also easy to argue from theory that agglomeration effects are heterogeneous across different types of workers. Some evidence suggests, for instance, that more productive workers are also the ones more able to reap the benefits from agglomeration (see [Glaeser and Maré, 2001](#); [Combes et al., 2012c](#); [de la Roca and Puga, 2012](#)). A specification similar to (5.9) can be used to study, for instance, the heterogeneous effect of density across diplomas. One would simply consider diploma-specific coefficients for density instead of industry-specific ones. However, diplomas usually do not change over time. When a two-step procedure is used, this implies that one diploma-location-time fixed effect must be

normalized to zero for each diploma. The alternative strategy of estimating the two-step procedure on each diploma separately is not much less precise than it was for industries since all the observations for any given individual are in the same diploma subsample, and there is thus a unique individual fixed effect for each worker to be estimated.

However, diplomas may not be enough to fully capture individual skill heterogeneity. One may wish to consider that the effect of density is specific to each individual as in the following specification:

$$y_{i,t} = u_i + X_{i,t}\theta + Z_{c(i,t),t}\gamma_i + \eta_{c(i,t),t} + \epsilon_{i,t}, \quad (5.15)$$

where γ_i is an individual fixed effect. Parameters can be estimated using an iterative procedure.⁵ For a given value of θ , one can regress $y_{i,t} - X_{i,t}\theta$ on $Z_{c(i,t),t}$ for each individual. This gives some estimates for γ_i and u_i . Then, θ is estimated by regressing $y_{i,t} - Z_{c(i,t),t}\gamma_i - u_i$ on $X_{i,t}$. The procedure is repeated using the parameter values from the previous iteration until there is convergence.

One can further extend the model and consider that location in general, and not density alone, has a heterogeneous effect on the local outcome. One considers in this case an interaction term between a local fixed effect and an individual fixed effect. This amounts to saying that it is not the effect of density but rather the combined effect of all local characteristics, whether they are observed or not, which is heterogeneous across individuals. The first step of the two-stage procedure in this case becomes

$$y_{i,t} = u_i + X_{i,t}\theta + \beta_{c(i,t),t} + \delta_{c(i,t),t}\nu_i + \epsilon_{i,t}, \quad (5.16)$$

with the identification restriction that $\sum_i \nu_i = 0$ and one of the local terms $\delta_{c,t}$ is normalized to zero. As before, the specification can be estimated with an iterative procedure. The estimators of parameters $\delta_{c,t}$ are regressed in the second step on local variables to assess the extent to which agglomeration economies influence the local return of unobserved individual characteristics. An additional extension to make the specification even more complete would consist in having the coefficients of individual characteristics depend on the individual. Note that as there are many individual-specific effects entering the model in a nonadditive way, the time span should be large for the estimations to make sense, and there is no guarantee that a large number of periods is enough for the parameters to be properly estimated. In any case, most of the specifications in this last paragraphs are material for future research.

5.2.2 Dynamic impact of agglomeration economies

So far, we have considered that agglomeration economies have an instantaneous effect on productivity and then no further impact in the following periods. In fact, agglomeration economies can be dynamic and can have a permanent impact such as when technological

⁵ This procedure is inspired from Bai (2009), who proposes such a procedure to estimate factor models.

spillovers increase local productivity growth or when individuals learn more or faster in larger cities as suggested by Lucas (1988). One can even argue that an individual moves from a large city to a smaller can transfer part of the individual's productivity gains from agglomeration to the new location and be more productive than other individuals who have not worked in a large city. In that case, dynamic effects operate through the impact of local characteristics on the growth of $A_{c,t}$ and $s_{i,t}$, which are involved in Equation (5.5). One can also consider dynamic effects operating through $p_{c,t}$ and $r_{c,t}$. For instance, agglomeration can facilitate the diffusion of information about the quality of goods and inputs, and this in turn can have an impact on price variations across periods (e.g., when prices are chosen by producers under imperfect competition). Therefore, even if dynamic effects relate more plausibly to technological spillovers and learning effects, market agglomeration economies can also present dynamic features. As a result, the identification issues are like those for static agglomeration economies, and one usually estimates only the overall impact of dynamic externalities and not the exact channel through which they operate. Note that the literature that first tried to identify agglomeration effects on local industrial employment, which dates back to Glaeser et al. (1992) and Henderson et al. (1995), adopts this dynamic perspective from the very beginning. We present this literature in Section 5.6.1.

We explain in this section how the previous productivity specifications can be extended to encompass dynamic effects. The distinction between static and dynamic effects was pioneered by Glaeser and Maré (2001), and we elaborate the discussion below from their ideas and those developed by de la Roca and Puga (2012), which is currently one of the most complete studies on the topic. For a model with static local effects only (disregarding the role of time-varying individual and industry characteristics), written as $y_{i,t} = u_i + \beta_{c(i,t),t} + \epsilon_{i,t}$, the individual productivity growth rate is simply related to the time difference of static effects:

$$y_{i,t} - y_{i,t-1} = \beta_{c(i,t),t} - \beta_{c(i,t-1),t-1} + \epsilon_{i,t}, \quad (5.17)$$

where $\epsilon_{i,t}$ is an error term.⁶ Dynamic local effects in their simplest form are introduced by assuming for $t \geq 1$ that

$$y_{i,t} - y_{i,t-1} = \beta_{c(i,t),t} - \beta_{c(i,t-1),t-1} + \mu_{c(i,t-1),t-1} + \epsilon_{i,t}, \quad (5.18)$$

where $\mu_{c,t-1}$ is a fixed effect for city c at date $t - 1$, which corresponds to the impact of city c on productivity growth between $t - 1$ and t , and thus captures dynamic local effects. Interestingly, this implies

⁶ In this chapter, we consider that $\epsilon_{i,t}$ is a generic notation for the residual and use it extensively in different contexts.

$$y_{i,t} = y_{i,1} + \beta_{c(i,t),t} + \sum_{k=1}^{t-1} \mu_{c(i,t-k),t-k} + \zeta_{i,t}, \quad (5.19)$$

where $\zeta_{i,t}$ is an error term. This equation includes the past values of local effects and shows that dynamic effects, even when they affect only the annual growth rate of a local outcome, do have a permanent impact on its level. Nevertheless, we have made some major assumptions to reach this specification. We now detail them and discuss how to relax them.

A first implicit assumption is that dynamic effects are perfectly transferable over time. For instance, knowledge does not depreciate even after a few years. To consider depreciation, one could introduce in (5.18) some negative effects of past city terms $\mu_{c(i,t-1),t-k}$, $k > 1$ with coefficients lower than 1 in absolute value, and this would lead to an autoregressive specification such that terms $\mu_{c(i,t-1),t-k}$ have an effect attenuated with a time lag when the model is rewritten in level.

Importantly, specification (5.19) makes more sense for individuals who stay in the same location than for movers. Dynamic local effects might also depend on where individuals locate at period t , and therefore on the destination location for movers. Individuals in a large city probably do not benefit from the same productivity gains from learning effects whether they move to an even larger city or to a smaller city (or if they stay where they are). In other words, dynamic gains are not necessarily fully transferable between locations, and the degree of transferability can depend on the characteristics of locations. Therefore, it might be more relevant to assume that dynamic effects depend on both the origin and destination locations and to rewrite the specification of local outcome as

$$y_{i,t} = y_{i,1} + \beta_{c(i,t),t} + \sum_{k=1}^{t-1} \mu_{c(i,t-k),c(i,t),t-k} + \zeta_{i,t}, \quad (5.20)$$

where $\mu_{j,c,\tau}$ is a time-varying fixed effect for being in city j at date $\tau < t$ and in city c at date t . The problem is that the number of parameters to be estimated for dynamic effects becomes very large (the square of the number of locations times the number of years in the panel). Moreover, restrictions on parameters must be imposed for the model to be identified. This can be seen, for instance, when writing the model in first difference for workers staying in the same location between dates $t-1$ and t , for which $c(i,t-1) = c(i,t)$:

$$y_{i,t} - y_{i,t-1} = \beta_{c(i,t),t} - \beta_{c(i,t-1),t-1} + \mu_{c(i,t-1),c(i,t),t-1} + \varepsilon_{i,t}. \quad (5.21)$$

The evolution of the static agglomeration effect cannot be distinguished from the dynamic effect (and this is also true when considering movers instead of stayers). When one observes the productivity variation of stayers, one does not know whether it occurs because static local effects have changed or because some dynamic local effects take place.

de la Roca and Puga (2012) make some assumptions that allow the identification of the model and significantly reduce the number of parameters to be estimated. They assume that static and dynamic effects do not change over time—that is, $\beta_{c,t} = \beta_c$ and

$\mu_{j,c,t-k} = \mu_{j,c}$. Under these assumptions, $\mu_{c,c}$ captures both the dynamic effect and the evolution of static effects. This can be seen from Equation (5.21), where the evolution of static effects would be now fixed to zero. This should be kept in mind when assessing the respective importance of static and dynamic effects, as this cannot be done from the relative explanatory power of β_c and $\mu_{j,c}$. Under these assumptions, it is also possible to rewrite the specification in a more compact form introducing the number of years the individuals have spent in each location:

$$y_{i,t} = u_i + X_{i,t}\theta + \beta_{c(i,t)} + \sum_j \mu_{j,c(i,t)} e_{i,j,t} + \epsilon_{i,t}, \quad (5.22)$$

where $e_{i,j,t}$ is the experience acquired by individual i until period t in city j (the number of years that individual spent there until date t), and $\mu_{j,c}$ captures the value of 1 year of this experience when the worker is located in city c . One can test whether the $\mu_{j,c}$ are statistically different from each other when c varies for given j —that is, whether location-specific experience can be transferred or not transferred to the same extent to any location, as was assumed in (5.19). One can also quantify the respective importance of the effects β_c and $\mu_{c,c}$ keeping in mind that it does not correspond to the respective importance of static and dynamic effects. Earlier attempts to evaluate dynamic effects on wages by Glaeser and Maré (2001), Wheeler (2006), and Yankow (2006) correspond to constrained and simplified versions of this specification, typically distinguishing only the impact on wage growth of moving or not moving to larger cities.

It is then possible in a second stage to evaluate the extent to which dynamic effects depend on the characteristics of the local economy, and to assess whether transferability relates to density of the destination location. One can consider the specification

$$\mu_{j,c} = Z_{j,\bullet}(\psi + Z_{c,\bullet}v) + \zeta_{j,c}, \quad (5.23)$$

where $Z_{j,\bullet}$ is the average over all periods of a vector of location- j characteristics including density. In this specification, the effect of density in the location where learning took place is a linear function of variables entering $Z_{c,\bullet}$ such as density. Clearly, all these dynamic specifications can be extended to encompass some heterogeneity across industries in the parameters of local variables, and possibly some localization effects.

An alternative approach that takes into account time variations in static and dynamic effects may consist in estimating density effects in one stage only, first specifying

$$\beta_{c,t} = Z_{c,t}\gamma + \eta_{c,t}, \quad (5.24)$$

$$\mu_{j,c,t} = Z_{j,t}(\psi + Z_{c,t}v) + \zeta_{j,c,t}, \quad (5.25)$$

and then plugging these expressions into Equation (5.20). This gives a specification where the coefficients associated with the different density terms can be estimated directly with linear panel methods. A limitation of this approach is again that it is difficult to compute standard errors taking into account unobserved local shocks because workers'

moves make the structure of the covariance matrix of error terms intricate when the model is rewritten in first difference or in the within dimension. On the other hand, the separate explanatory power of static and dynamic agglomeration effects is better assessed.

Finally, it is possible to generalize the framework to the case where both static and dynamic effects are heterogeneous across individuals. Specification (5.20) becomes

$$y_{i,t} = u_i + X_{i,t} \theta + \beta_{c(i,t),t} + \delta_{c(i,t),t} \nu_i + \sum_{k=1}^{t-1} \left(\mu_{c(i,t-k),c(i,t),t-k} + \lambda_{c(i,t-k),c(i,t),t-k} r_i \right) + \epsilon_{i,t}, \quad (5.26)$$

where ν_i and r_i are individual fixed effects verifying the identification assumption $\sum_i \nu_i = \sum_i r_i = 0$. Parameters can be estimated by imposing additional identification restrictions such as the fact that static and dynamic effects do not depend on time, and using an iterative procedure as in previous subsections. Note that such a specification has not been estimated yet. One of the best attempts is that of [de la Roca and Puga \(2012\)](#), who restrict the spatial dimension to three classes of city sizes only (which prevents the second-stage estimation and only allows them to compare the experience effect over the three classes). Importantly, they also make the further assumption that the impact of individual heterogeneity is identical for both static and dynamic effects—that is, $\nu_i = r_i$. [D’Costa and Overman \(2014\)](#) attempt to elaborate on the attempt of [de la Roca and Puga \(2012\)](#). They estimate the specification in first differences while allowing for $\nu_i \neq r_i$, but they exclude movers to avoid having to deal with between-city dynamic effects.

5.2.3 Extending the model to local worker–firm matching effects

[Marshall \(1890\)](#) was among the first to emphasize that agglomeration can increase productivity by improving both the quantity and the quality of matches between workers and firms in local labor markets (see [Duranton and Puga, 2004](#), for a survey of this type of mechanism). The better average quality of matches in larger cities can be considered as a static effect captured by the local fixed effects $\beta_{c,t}$ estimated in previous subsections. The matching process in cities can also yield more frequent job changes, which can boost productivity growth. This dynamic matching externality can be incorporated into our framework by considering that at each period t , a worker located in c receives a job offer with probability ϕ_c to which is associated a wage $\tilde{y}_{i,t}$. One assumes that workers change jobs within the local market at no cost and they accept a job offer if the associated wage is higher than the one they would get if they stayed with the same employer. To ease exposition, we suppose that migrants do not receive any job offer at their origin location, but receive one at the destination location once they have migrated. The probability of receiving such an offer is supposed to be the same as that for stayers in this market. We also assume for the moment that there is no dynamic effect other than through

job change. For workers receiving an offer, the wage at time t is $y_{i,t} + \Delta_{i,t}$, where $y_{i,t}$ is given by Equation (5.7) and $\Delta_{i,t} = \max(0, \tilde{y}_{i,t} - y_{i,t})$. The individual outcome is then given by

$$y_{i,t} = u_i + X_{i,t}\theta + \beta_{c,t} + \sum_{\tau=1}^{t-1} 1_{\{O(i,\tau)=1\}} \Delta_{i,\tau} + \epsilon_{i,t}, \quad (5.27)$$

where $O(i, \tau)$ is a dummy variable taking the value 1 if individual i has received a job offer between dates $\tau - 1$ and τ , and 0 otherwise.

For workers keeping the same job in location c between the two dates, there is no dynamic matching gain, and wage growth is given by

$$y_{i,t} - y_{i,t-1} = (X_{i,t} - X_{i,t-1})\theta + \beta_{c,t} - \beta_{c,t-1} + \epsilon_{i,t}, \quad (5.28)$$

where $\epsilon_{i,t} = \epsilon_{i,t} - \epsilon_{i,t-1}$.

For workers changing jobs within location c , improved matching induces a wage premium $\Delta_{i,t}$, and wage growth can be written as

$$y_{i,t} - y_{i,t-1} = (X_{i,t} - X_{i,t-1})\theta + \tilde{\beta}_{c,t} - \beta_{c,t-1} + \nu_{i,t}, \quad (5.29)$$

where $\tilde{\beta}_{c,t} = \beta_{c,t} + E(\Delta_{i,t} | i \in (c, t-1), i \in (c, t))$ is the sum of the local fixed effect for stayers keeping their jobs and the expected productivity gain when changing job, and the new residual is $\nu_{i,t} = \epsilon_{i,t} + \Delta_{i,t} - E(\Delta_{i,t} | i \in (c, t-1), i \in (c, t))$.

For workers changing job between two locations c and c' , wage growth can be expressed as

$$y_{i,t} - y_{i,t-1} = (X_{i,t} - X_{i,t-1})\theta + \beta_{c',t}^c - \beta_{c,t-1} + \nu_{i,t}, \quad (5.30)$$

where $\beta_{c',t}^c = \beta_{c,t} + E(\Delta_{i,t} | i \in (c, t-1), i \in (c', t))$ is the sum of the local fixed effect for stayers keeping their jobs in the destination location and the expected productivity gain when changing jobs from city c to city c' .⁷ This gain may depend on both cities as it could be related, for instance, to the distance between them or their industrial structure.

The difference in local effects from separate wage growth regressions for stayers changing jobs and stayers keeping the same job provides an estimate of the matching effect since $(\beta_{c,t} - \beta_{c,t-1}) - (\beta_{c,t} - \beta_{c,t-1}) = E(\Delta_{i,t} | i \in (c, t-1), i \in (c', t))$. If changing jobs increases productivity through improved matching, this difference should be positive for any location c . If agglomeration magnifies such dynamic matching effects, the probability of changing jobs should increase with density, and the difference $\tilde{\beta}_{c,t} - \beta_{c,t}$ should be larger in

⁷ In fact, workers may move and take a wage cut if they expect future wage gains. This kind of intertemporal behavior cannot be taken into account in a static model as here but it can be taken into account in the dynamic framework developed in the next subsection.

denser areas. More generally, to assess which local characteristics are determinants of dynamic matching effects, one can run the second-step regression:

$$\tilde{\beta}_{c,t} - \beta_{c,t} = Z_{c,t}\Phi + \eta_{c,t}, \quad (5.31)$$

where $Z_{c,t}$ is a vector of local characteristics. Such a model has not been estimated yet, but [Wheeler \(2006\)](#) makes one of the best attempts to do so. Owing to the small size of the dataset, [Wheeler \(2006\)](#) cannot identify the role of local-time fixed effects, but his strategy on the panel of workers changing job is equivalent to directly plugging (5.31), with local market size as the single local characteristic, into the difference between (5.28) and (5.29) to assess by how much the matching effect increases with local market size.

Exploiting wage growth for workers changing both job and city is more intricate, and an important assumption which needs to be made (and was implicitly made in previous sections) is that the location choice is exogenous. In order to get consistent estimates of local effects when movers are used as a source of identification, the location choice should not depend on individual-location shocks on wages conditional on all the explanatory variables and parameters in the model.⁸ This assumption is disputable since workers often migrate because they receive a good job offer in another local labor market, or because they had a bad original match with their firm. By the same token, we can argue that job changes are endogenous for both movers and nonmovers, and this affects the estimates of local effects obtained for specifications in this subsection. As this concern is certainly important, it may be wise to use another kind of approach that explicitly takes into account the endogeneity of location and job choices. This can be done with a dynamic model of intertemporal location choices at the cost of imposing more structure on the specification that is estimated. We now turn to this kind of structural approach, building on the same underlying background.

5.2.4 Endogenous intertemporal location choices

So far, we have considered static and dynamic agglomeration effects within a static framework where workers' location choices are strictly exogenous: Workers do not take into account wage shocks due to localized job opportunities in their migration or job change decisions. When workers do consider alternative job opportunities when making their decisions, it is also likely that they are forward-looking and take into account all future possible outcomes in alternative locations. As shown by [Baum-Snow and Pavan \(2012\)](#), it is possible to introduce static and dynamic agglomeration effects in a dynamic model of location choices that takes these features into account.⁹ Nevertheless, identification is achieved thanks to the structure of the model, and it is sometimes difficult to assess which conclusions

⁸ This assumption is discussed at greater length from an econometric point of view in [Section 5.4.2](#).

⁹ [Gould \(2007\)](#) also proposes a dynamic model where school attendance too is endogenous. See also [Beaudry et al. \(2014\)](#) for a dynamic model with search frictions and wage bargaining with static agglomeration effects but no dynamic agglomeration effects.

would remain under alternative assumptions. For simplicity's sake, we present the main mechanisms of the model for employed workers and consider that there is no unemployment and no consumption amenities, these assumptions being relaxed in [Baum-Snow and Pavan \(2012\)](#). Unemployment can easily be added by considering that there is an additional state for workers and there are exogenous mechanisms (such as job destructions and job offers) leading to transitions between states. Consumption amenities can be considered by including location-specific utility components that do not affect local wages.

Individual unobserved heterogeneity is modeled as draws in a discrete distribution (instead of individual fixed effects). There are H types of workers indexed by $h = 1, \dots, K$. Worker i getting a job in location c draws a job match $\varsigma_{i,c}$ in a distribution which is specific to the location. For a given job, the match is drawn once and for all and does not vary over time. The wage of worker i of type $h(i)$ located in market c and occupying a job with match $\varsigma_{i,c}$ is a variant of Equation (5.22) given by

$$y_{i,c,t}(\varsigma_{i,c}) = X_{i,t}\theta + \beta_{h(i),c,t} + \sum_j \mu_{h(i),j,c} e_{i,j,t} + \varsigma_{i,c} + \epsilon_{i,c,t}, \quad (5.32)$$

where $\beta_{h,c,t}$ is a static location effect depending on the worker type, $\mu_{h,j,c}$ is a location-specific experience effect depending on the worker type, and $\epsilon_{i,c,t}$ is a white noise. Note that whereas the wage depends on the draw of the white noise, we do not index the wage by it to keep the notation simple. A crucial difference from the specifications in previous sections is that we now have a specification for the potential outcome in any location c at each date. Therefore, the wage is now indexed by c , and we write $y_{i,c,t}$ for any potential wage instead of $y_{i,t}$ as previously for the realized one.

The intertemporal utility and location choice are determined in the following way. Consider worker i of type $h(i)$ located in city c at period t . The worker earns a wage $y_{i,c,t}$ and, at the end of the period, has the possibility to move to another job within the same location or to a different location. Migration to another location can be achieved only if the worker gets a job offer in that location (as we have ruled out unemployment for simplicity). The probability of receiving a job within location c for a worker of type h is denoted $\phi_{h,c}$, and the probability of receiving a job in location $j \neq c$ is denoted $\phi_{h,c,j}$. There is a cost C when changing jobs within the local market. If the worker moves between city c and city j , the workers has to pay a moving cost $M_{c,j}$. Let us denote $V_{i,c,t}(\varsigma_{i,c})$ the intertemporal utility of an individual located in city c at time t , and occupying a job with match $\varsigma_{i,c}$. This intertemporal utility can be expressed with the recursive formula

$$\begin{aligned} V_{i,c,t}(\varsigma_{i,c}) = & y_{i,c,t}(\varsigma_{i,c}) + \phi_{h(i),c} E_{\varsigma_c} \max [V_{i,c,t+1}(\varsigma_{i,c}), V_{i,c,t+1}(\varsigma_c) - C] \\ & + \sum_{j \neq c} \phi_{h(i),c,j} E_{\varsigma_j} \max [V_{i,c,t+1}(\varsigma_{i,c}), V_{i,j,t+1}(\varsigma_j) - M_{c,j}], \end{aligned} \quad (5.33)$$

where expectations are computed over the distributions of all future random terms including the matches ς_c when one changes jobs within location and ς_j when one changes jobs by moving to j (but not the realized match $\varsigma_{i,c}$ for the current job). The first term

corresponds to the wage earned at the current location. The second term is the expected outcome associated with a possible offer of a job within the current location. It depends on the probability of receiving a job offer and on the expected future intertemporal utility, which is the one related to the new job if it is worth accepting the offer, or is the one related to the current job otherwise. The third term is the expected outcome associated with a possible job offer in other locations. It depends on the probability of receiving a job offer in every location and on the expected future intertemporal utility related to the location if it is worth moving there, or to the current location otherwise.

The model can be estimated by maximum likelihood after writing the contributions to likelihood of individuals that correspond to their history of events (whether they change jobs, whether they change location, and their wages at each period). The model is parameterized by making some assumptions on the distributions of random and matching components, supposing they follow normal distributions with mean zero and variance to be estimated. Unobserved heterogeneity is modeled through mass points with individuals having some probabilities of being of every type which enter the set of parameters to be estimated. The computation of contributions to likelihood involves the integration over the distribution of unobserved components in line with [Heckman and Singer \(1984\)](#).

Once estimates of the parameters $\beta_{h,c,b}$, $\mu_{h,j,c}$, $\phi_{h,c}$, and $\phi_{h,c,j}$ have been recovered, a variance analysis can be performed to assess the respective importance of static and dynamic local effects, as well as matching effects. Estimated parameters can also be regressed on density (or any other local variable), to evaluate how they vary with changes in the characteristics of locations. In practice, however, the numbers of locations and related parameters are usually too large for the model to be empirically tractable. An alternative is to aggregate locations by quartile of density and consider that each group is a single location in the model. Once the parameters have been estimated, it is possible to assess whether they take larger values for groups of denser locations.

Overall, structural approaches modeling jointly location choices and wages are an interesting tool for taking into account the endogeneity of workers' mobility when assessing the impact of local determinants of agglomeration economies, whereas this has never been properly done with linear panel models. Nevertheless, it comes at the cost of making strong assumptions about the structure of the model, including parametric assumptions about random terms. More details on structural approaches in urban economics are provided by [Holmes and Sieg \(2015\)](#).

5.3. LOCAL DETERMINANTS OF AGGLOMERATION EFFECTS

We have already argued that the literature usually estimates the total net impact of local characteristics related to agglomeration economies rather than the magnitude of agglomeration channels (although there are some tentative exceptions that are presented in

[Section 5.7](#)). The previous section alludes to some of these local characteristics, in particular employment density. This section details the definitions of all the characteristics that have been considered in the literature and explains to what extent they play a role in agglomeration economies. The outcome on which the impacts of local determinants of agglomeration economies are estimated often refers to a particular industry, either because data aggregated by location and industry are used or because one considers individual outcomes of firms or workers in a given industry. Considering this, two types of local characteristics may be included in the specification: those that are not specific to the industry and shape urbanization economies, and those that are specific to the industry and shape localization economies. We show successively how the size of the local market, the industrial structure of the local economy, and the composition of the local labor force can affect agglomeration economies and in turn local outcomes. We will see that in each case there can be both urbanization and localization economies.

5.3.1 Density, size, and spatial extent of agglomeration effects

Equation (5.3) shows which pure and market agglomeration mechanisms involve the size of the local economy. Depending on the mechanism, employment, population, or production can be the most relevant variable to measure local economy size. However, the correlation between these three variables is often too great to allow the identification of their respective effects separately, and one has to restrict the analysis to one of them. The results are, in general, very similar whichever variable is used. Employment is usually preferred to population, first because it better reflects the magnitude of local economic activity, and second because certain other local variables (described below) can be constructed from employment only. Production presents the disadvantage of being more subject to endogeneity issues than employment (see [Section 5.4](#)).

One usually considers models where both productivity and size are measured in a logarithmic specification because this eases interpretations, the estimated parameter being a constant elasticity. This also reduces the possibility of extreme values for the random component of the model and makes its distribution closer to the one of a normal law, which is usually used in significance tests.

[Ciccone and Hall \(1996\)](#) argue that the size of the local economy should be measured by the number of individuals per unit of land—that is, density. Indeed, there is usually a large heterogeneity in the spatial extent of the geographic units that are used, as these units are often based on administrative boundaries. This can also create arbitrary border effects, an issue related to what the literature calls the modifiable areal unit problem—that is, the fact that some conclusions reached by empirical works could depend on the spatial classification used in their analyses, in particular the size and shape of the spatial units. Using density should reduce issues about mismeasurement of the size of the local

economy, which is in line with [Briant et al. \(2010\)](#), who show that using more consistent empirical strategies largely reduces modifiable areal unit problem concerns.

Importantly, from the theory point of view, depending on the microfoundations of pure and market local externalities entering (5.3), either local density or the level of local employment can affect the magnitude of the effects at stake. Therefore, there is no reason to restrict the specification to one variable or the other. Typically, if agglomeration gains outweigh agglomeration costs, one expects, in general, both the density and the size of the local economy to have a positive impact on local productivity. When variables are considered in a logarithmic specification, it is possible and convenient to capture the two effects using density and land area simultaneously (while leaving employment aside). The impact of density, holding land area constant, reflects the gains from increasing either the number of people in the city or the density, while the impact of land area, holding density constant, reflects the gains from increasing the spatial extent of the city (i.e., from increasing both land area and employment proportionally). In a logarithmic specification, any combination of employment and land area identifies the same fundamental parameters but one has to be careful with the interpretation of coefficients, since we have

$$\beta \ln \text{den}_{c,t} + \mu \ln \text{area}_{c,t} = \beta \ln \text{emp}_{c,t} + \varrho \ln \text{area}_{c,t}, \quad \text{with } \varrho = \mu - \beta, \quad (5.34)$$

where $\text{emp}_{c,t}$ is total employment in location c at date t , $\text{area}_{c,t}$ is land area, and $\text{den}_{c,t} = \frac{\text{emp}_{c,t}}{\text{area}_{c,t}}$ is density. This equation shows that whereas the effect of total employment for a given land area and the effect of density for a given land area correspond to the same parameter β , the effect of land area for a given total employment ϱ is equal to the difference between the effect of land area for a given density μ and the effect of density β . In fact, ϱ can be negative even when agglomeration gains result from both density and spatial extent. It would be wrong to conclude that there are agglomeration costs from a negative estimated value, or no agglomeration gains from spatial extent from a nonsignificant estimated coefficient. When density and land area are used, agglomeration gains exist when any of the estimated coefficients is significantly positive.

Firms trade with distant markets, and communication exchanges occur between agents located sometimes quite far apart. A number of studies have attempted to evaluate the spatial extent of local spillovers beyond the strict limits of the local unit. These spillovers can occur for any of the urbanization and localization effects considered in this section, but most contributions in the literature consider them for local size only. Spatial econometric approaches usually consider spillovers for all the local determinants but at the cost of assuming for all of them an identical influence of distance on spillovers, and making it more difficult to deal with endogeneity issues (see [Section 5.4.5.4](#)). A flexible specification where density is considered at various distances from the worker's or firm's location may be envisaged. Typically, one can introduce in the specification many additional variables for density measured at 20, 50, 100, 150, 200 km, etc., from the location. However, there is sometimes not enough variation in the data to identify so

many effects of density. Therefore, some authors follow [Harris \(1954\)](#) and put more constraints on the impact of trade and communication costs by assuming that their impact is proportional to the inverse of distance, which typically leads to Harris's following market potential variable:

$$\text{MP}_{c,t} = \sum_{\ell \neq c} \frac{\text{den}_{\ell,t}}{d_{c,\ell}}, \quad (5.35)$$

where $d_{c,\ell}$ is the distance between location c and location ℓ .

A number of variants for computing market potential exist since one can consider population, employment or production, in level form or in density form, as a measure of market size. Several market potential variables can be considered simultaneously (e.g., one for density and one for land area). One can also refine the way trade and communication costs are assessed by using, instead of as-the-crow-flies distances, real distances by road or real measures of trade and communication costs. Nevertheless, all the corresponding market potential variables are usually highly correlated, as illustrated by [Combes and Lafourcade \(2005\)](#), and the effect of only one of them can actually be identified. If density is used as the measure of the local economy size, computing market potential using densities is more consistent. Importantly, the own location is excluded from formula (5.35) for the Harris market potential to obtain an “external” market potential whose impact can usually be identified separately from the effect of the own location size. In any case, and as for the own density, one cannot say whether the impact of market potential is a market-based effect or a pure externality, and more generally which mechanism is at play.

[Fujita et al. \(1999\)](#) emphasize that in economic geography models based on Dixit–Stiglitz monopolistic competition, local nominal wages are an increasing function of a specific variable, called the “structural market access,” which is closely related to the Harris market potential. Intuitively Dixit–Stiglitz models suggest that Harris's specification needs to be augmented with local price effects to take into account the role of imperfect competition that makes the price of the manufacturing good differ across locations owing to its differentiation affecting both its supply and its demand. In other words, there is now an impact of locations further away through $p_{c,t}$ in (5.3), which is captured by the structural market access variable. Note that the structural market access variable aggregates the effects of sizes of both the own and distant locations, and its computation thus requires a consistent measure of trade costs not only between locations, but also within locations. This is a concern by itself as internal trade costs are usually not available in datasets, and no fully satisfactory solution has been proposed yet to evaluate them. The most frequent strategy for coping with the issue, which is *ad hoc*, consists in assuming that, within a location, trade costs are proportional to the square root of land area.

Interestingly, [Redding and Venables \(2004\)](#) show that in a model where varieties are used as intermediate inputs, another variable very similar to the market access, called the “structural supply access,” determines the price of inputs, $r_{c,t}$, in (5.3). The greater

the supply access, the lower input prices and the higher nominal wages. Owing to the strong link to the theory of structural market access and supply access, which makes them dependent on the elasticity of substitution between varieties, for instance, no empirical counterpart can be directly constructed. [Hanson \(2005\)](#) was the first to suggest using also theory to relate market access to observables, and in particular local housing stocks. [Redding and Venables \(2004\)](#) take another route, where both market and supply accesses are estimated through a first-step trade gravity equation, and their predictors are then used in a second-step wage equation. [Combes and Lafourcade \(2011\)](#) show that a structural specification encompassing the role of market and supply access in agglomeration economies can also be obtained in a Cournot competition setting.

Unfortunately, structural market and supply access are highly correlated in general, precisely because circular causalities related to agglomeration effects lead households, firms, and intermediate input suppliers to choose the same locations.¹⁰ It is therefore difficult to identify their respective effects separately. One also has to keep in mind that the simultaneous presence of knowledge spillovers would suggest adding a standard Harris market potential in the specification in order to simultaneously take into account pure agglomeration effects coming from the local technological level and labor skills, $A_{c,t}$ and $s_{c,t}$. Nevertheless, it is itself highly correlated with the structural market and supply access, and only one of the three variables usually has a significant effect. When structural market access only is considered, one cannot exclude the possibility that it captures agglomeration effects other than those at play in economic geography models à la Dixit and Stiglitz for instance, even if the approach is structural.

5.3.2 Industrial specialization and diversity

The theory used to ground the role of location size on local productivity makes it obvious that most effects should be specific to the industry. They depend on structural parameters such as trade and communication costs, the degree of product differentiation, or the magnitude of increasing returns to scale, which are *a priori* all specific to the industry. This suggests that, when a reduced form approach is used, heterogeneous effects of density, land area, and the Harris market potential across industries could be considered, as suggested in [Section 5.2.1.2](#). In other words, the first way of considering the role of local industrial structure is to investigate industry-specific impacts of determinants of urbanization economies. At the other extreme, theory can be used to construct structural market and supply access variables that are specific to the industry, and which therefore correspond to what is referred to as localization economies. These are agglomeration

¹⁰ Agglomeration economies increase productivity and thus attract firms. This leads to an increase in the demands for local labor and intermediate inputs as well as wages and input prices, which attract workers and input suppliers. In turn, the inflow of workers and suppliers magnifies productivity gains from agglomeration economies, attracting even more firms, and so on.

effects within the industry, the determinants of which are local characteristics that depend not only on location and date but also on industry, the triplet $\{c, s, t\}$ with the previous notation.

Usually, authors do not construct structural market and supply access variables that are specific to the industry because necessary data are not available. Alternatively, one can consider in the specification other variables that characterize the industry within the local economy. One needs to be careful when introducing such variables related to localization economies in addition to the local economy size variables related to urbanization economies. Let us first consider the role of the size of the industry within the location. Typically, if all locations had the same share of all industries, the effect of such a variable would not be identified. A location with larger total employment would have more employment in all industries, and higher productivity in an industry could not be attributed more to higher employment in the industry than to higher total employment. Nevertheless, since localization effects seem to play no role in that case given that all locations have the same industrial composition, one may wish to attribute higher industry productivity in larger cities to higher overall employment in the local economy—that is, to urbanization effects. When the industrial share differs across locations for some industries, total and industrial employment are not proportional across locations, and one is faced with the same identification issue. Industrial employment can generate productivity gains both when it is higher because total employment at the location is higher, and when the share of the industry is higher for given total employment at the location. These two effects are captured by employment in industry s in location c at date t , $\text{emp}_{c,s,t}$, but they can be distinguished by decomposing this employment into the product of its share within the local economy, a variable often labeled specialization (or concentration in [Henderson et al., 1995](#)), and the local size of the economy: $\text{emp}_{c,s,t} = \text{spe}_{c,s,t} \text{emp}_{c,t}$, with

$$\text{spe}_{c,s,t} = \frac{\text{emp}_{c,s,t}}{\text{emp}_{c,t}}.$$

To ease interpretation, [Combes \(2000\)](#) argues that in a specification in logarithmic form, one has to consider total employment (or employment density) next to specialization. Both these variables are expected to have a positive impact, when there are urbanization and localization economies respectively.

Because all variables are in logarithmic form, the same parameters would also be identified if total employment (or density) and industrial employment (not specialization) were considered. However, one needs again to be careful with interpretations. We have

$$\beta \ln \text{emp}_{c,t} + \vartheta \ln \text{spe}_{c,s,t} = \varrho \ln \text{emp}_{c,t} + \vartheta \ln \text{emp}_{c,s,t}, \quad \text{with } \varrho = \beta - \vartheta. \quad (5.36)$$

This equation shows that whereas the effect of specialization for a given total employment and the effect of industrial employment for a given total employment take the same value ϑ , the effect of total employment for given industrial employment ϱ is equal to the difference

between the effect of total employment for a given specialization β and the effect of industrial employment ϑ . A nonsignificant estimate for q , as obtained, for instance, by [Martin et al. \(2011\)](#) for France, does not imply that there is no urbanization effect, but rather means that the effect of specialization and the effect of total employment, which are usually both positive, compensate.¹¹ Finally, note that one could consider the density of industrial employment (rather than its level), as we considered the density of total employment and not its level. We do not advise using this specification as it can lead to the same possible misinterpretations as for the industrial employment level.

[Jacobs \(1969\)](#) made popular the intuition that industrial diversity could be favorable as there could be cross-fertilization of ideas and transmission of innovations between industries. This has been formalized, for instance, by [Duranton and Puga \(2001\)](#), and many summary measures of diversity have been proposed. The most used is probably the inverse of a Herfindahl index constructed from the shares of industries within local employment:

$$\text{div}_{c,t} = \left[\sum_s \left(\frac{\text{emp}_{c,s,t}}{\text{emp}_{c,t}} \right)^2 \right]^{-1}.$$

Since specialization is also introduced in the specification, interpretation is easier if one removes the own industry from the computation of $\text{div}_{c,t}$. In that case, whereas specialization relates to the role of the industry local share, diversity relates to the role of the distribution of employment over all other industries, and the two indices clearly capture two different types of mechanisms. In particular, whereas specialization is a determinant of localization economies, the Herfindahl index is a determinant of urbanization economies. Note that when the number of industries is large, it makes little difference to drop the own industry from computations, and the correlation between the Herfindahl indices obtained with and without the own industry is large.

The Herfindahl index has the bad property of taking values largely influenced by the number of units, industries here, from which it is computed. The range of variations of $\text{div}_{c,t}$ is $[1, S_{c,t}]$, where $S_{c,t}$ is the total number of industries active in location c at date t . When detailed industrial classifications are used, $S_{c,t}$ can vary a lot across locations and the Herfindahl index reflects this number more than the actual distribution of employment between industries. For this reason, [Combes et al. \(2004\)](#) propose assessing the role

¹¹ Earlier contributions by [Glaeser et al. \(1992\)](#) and [Henderson et al. \(1995\)](#) also consider the share and not the level of industrial employment to capture localization economies. However, because these authors study the determinants of industrial employment growth, and not the productivity level, they argue that the level of industrial employment must be introduced simultaneously, and its effect is identified because not all variables are expressed in logarithmic form. In that case, identification is assured thanks only to nonlinearities, and the results can be misleading, as emphasized by [Combes \(2000\)](#). We return to this point in [Section 5.6.1](#).

of industrial diversity by introducing the Herfindahl index in regressions simultaneously with the number of locally active industries meant to capture the unevenness of the distribution of industries over space.

Another solution consists in moving to other types of industrial diversity indices, keeping in mind that all have weaknesses. For example, some authors propose using the so-called Krugman index introduced by [Krugman \(1991a\)](#). The index is sometimes called the Krugman specialization index, which is misleading since it actually measures an absence of diversity, and specialization refers to another concept as we have just seen. The Krugman index is a measure of the distance between the distributions of industry shares in the location and at the global level:

$$\text{K-index}_{c,t} = \sum_s \left| \frac{\text{emp}_{c,s,t}}{\text{emp}_{c,t}} - \frac{\text{emp}_{s,t}}{\text{emp}_t} \right|,$$

where $\text{emp}_{s,t}$ is employment in industry s at the global level and emp_t is total employment.

As the Krugman index can take the value zero, it is not possible to express it in a logarithmic form. A diversity index can be constructed as the logarithm of 1 minus the Krugman index. Note that here diversity is maximal when the local distribution of employment across industries is identical to the global one, while an equal share of employment across all sectors at the local level corresponds to a less diverse situation.

Instead of using own-industry specialization and diversity variables in a specification, one could introduce a full set of variables corresponding to specialization in each industry. The coefficients of these variables could depend both on the that own industry and the industry for which specialization is computed, so that one ends up with a matrix of coefficients. This way one could identify local externalities within each industry and externalities between any two industries (which would not be constrained to be symmetrical). This would possibly correspond more to what [Jacobs \(1969\)](#) had in mind when she said that a number of other industries have a positive effect on the own productivity but certainly not all of them as the diversity indices implicitly assume. The effect of specialization at distant locations could also be assessed by introducing some Harris market potential variables constructed using industrial employment. However, there may be a lack of variation in the data to identify all the effects in these alternative specifications. Endogeneity issues are also magnified, as explained in more detail in [Section 5.4.2](#). All variables should be instrumented at the same time, and this can prove to be very difficult in practice.

Finally, for given local total and industrial employment, another industrial characteristic that may influence the magnitude of localization economies is whether local industrial employment is concentrated in a small number of firms or is evenly split among many firms. Typically large firms could be more able to internalize some of the local effects, while small firms would have more difficulty avoiding outgoing knowledge spillovers but could also simultaneously benefit more from spillovers. The local distribution of firm sizes also influences the degree of competition in local input markets and in local

non-tradable good markets. With this type of intuition in mind, Glaeser et al. (1992) suggest considering the average firm size within the local industry (in fact they consider its inverse) as an additional determinant of localization economies:

$$\text{size}_{c,s,t} = \frac{\text{emp}_{c,s,t}}{n_{c,s,t}},$$

where $n_{c,s,t}$ is the number of firms in industry s in location c at time t . This variable can also be considered simultaneously with a Herfindahl index computed using the shares of firms within local industrial employment as proposed by Combes et al. (2004). This index captures local productive concentration and can be written as

$$\text{pcon}_{c,s,t} = \sum_{j \in \{c,s,t\}} \left(\frac{\text{emp}_{j,t}}{\text{emp}_{c,s,t}} \right)^2,$$

where $\text{emp}_{j,t}$ is the employment of plant j . Note that the range of variations of this variable depends on the number of plants active in the local industry $n_{c,s,t}$, and this number thus needs to be introduced simultaneously in the specification. Alternatively and more intuitively, one may prefer to introduce instead the average firm size, $\text{size}_{c,s,t}$ (as, when expressed in logarithmic form, $\text{spe}_{c,s,t}$, $\text{size}_{c,s,t}$, and $n_{c,s,t}$ are collinear).

Importantly, as $\text{size}_{c,s,t}$ and $\text{pcon}_{c,s,t}$ depend on the location choices of firms and their scale of production, which are directly influenced by the dependent variable (local productivity), their use leads to endogeneity concerns that are more serious than for the other local characteristics. These concerns are discussed in more detail in Section 5.4. Absent a solid instrumentation strategy, one should avoid introducing these determinants of localization economies in the specification.

5.3.3 Human capital externalities

Another strand of the literature has tried to identify human capital externalities. Local productivity is regressed on an indicator of local human capital, typically the share of skilled workers in local employment or the local ratio between the numbers of skilled workers and unskilled workers. Somewhat surprisingly, other local characteristics capturing agglomeration effects are most often not introduced simultaneously in the regressions except in a few cases, such as in Combes et al. (2008a). There is no underlying theoretical reason as we saw that the various agglomeration economy channels may depend on all local characteristics. Furthermore, the human capital variable may be correlated with local characteristics which are not controlled for, such as density, with which it is usually positively correlated, and therefore it does not capture the effect of human capital only.

Another difficulty arises from the fact that, beyond some human capital externalities, the estimated coefficient for the local share of skilled workers captures the imperfect

substitutability between skilled and unskilled workers. When this share increases, both types of workers can benefit from the externalities, but unskilled workers benefit from an extra positive effect because they become relatively less numerous, which increases their marginal productivity. Conversely, skilled workers are negatively affected by this substitution effect. We illustrate this identification issue by considering the following local production function that extends our previous framework:

$$y_{c,t} = [(A_{c,t}^H H_{c,t})^\rho + (A_{c,t}^L L_{c,t})^\rho]^{\frac{\alpha}{\rho}} K_{c,t}^{1-\alpha}, \quad (5.37)$$

where $A_{c,t}^j$ is the productivity of workers with skills j with $j = H$ for high-skilled workers and $j = L$ for low-skilled workers, $H_{c,t}$ is the number of high-skilled workers, $L_{c,t}$ is the number of low-skilled workers, and ρ is a parameter such that $\rho < 1$. The production function is of Cobb–Douglas type in labor and other inputs, $K_{c,t}$, and the labor component is a constant elasticity of substitution (CES) function in high-skilled and low-skilled workers with an elasticity of substitution equal to $-1/(1-\rho)$. As previously, workers are counted in terms of efficient units such that

$$H_{c,t} = \sum_{i \text{ high-skilled} \in \{c,t\}} s_{i,t} \ell_{i,t}, \quad (5.38)$$

$$L_{c,t} = \sum_{i \text{ low-skilled} \in \{c,t\}} s_{i,t} \ell_{i,t}, \quad (5.39)$$

with $\ell_{i,t}$ the number of hours worked and $s_{i,t}$ the number of efficient labor units per hour of individual i at date t . As regards the human capital externality, the ratio between the numbers of high-skilled and low-skilled workers $S_{c,t} = H_{c,t}/L_{c,t}$ is supposed to influence the productivity of workers differently depending on their skills such that

$$A_{c,t}^H = (S_{c,t})^{\gamma^H} \quad \text{and} \quad A_{c,t}^L = (S_{c,t})^{\gamma^L}, \quad (5.40)$$

where γ^j captures the magnitude of human capital externalities for workers with skills j . For simplicity's sake, we assume here that $S_{c,t}$ does not affect any other agglomeration channel—namely, the prices of output and other inputs—and that no other local characteristic plays a role. It is possible to solve for wages at the individual level in the same way we did in [Section 5.2](#) using first-order conditions to determine the optimal use of labor and capital. The wages of high-skilled and low-skilled workers, $w_{i,t}^j$ for $j = H, L$, is obtained as

$$w_{i,t}^H = \frac{\alpha}{(1-\alpha)^{1-1/\rho}} r_{c,t}^{1-1/\rho} p_{c,t}^{1/\alpha} s_{i,t} (A_{c,t}^H)^\rho [(A_{c,t}^H)^\rho + (A_{c,t}^L)^\rho S_{c,t}^{-\rho}]^{\frac{1-\rho}{\rho}}, \quad (5.41)$$

$$w_{i,t}^L = \frac{\alpha}{(1-\alpha)^{1-1/\rho}} r_{c,t}^{1-1/\rho} p_{c,t}^{1/\alpha} s_{i,t} (A_{c,t}^L)^\rho [(A_{c,t}^H)^\rho + (A_{c,t}^L)^\rho S_{c,t}^{-\rho}]^{\frac{1-\rho}{\rho}} S_{c,t}^{1-\rho}. \quad (5.42)$$

The wage elasticities with respect to $S_{c,t}$ for high-skilled and low-skilled workers, respectively, can be derived as

$$\delta_{c,t}^H = \gamma^H - \phi_{c,t}(1 - \rho)(1 + \gamma^H - \gamma^L), \quad (5.43)$$

$$\delta_{c,t}^L = \gamma^L + (1 - \phi_{c,t})(1 - \rho)(1 + \gamma^H - \gamma^L), \quad (5.44)$$

where $\phi_{c,t}$ is the ratio between the wage bill for high-skilled workers and the total wage bill.

Several comments can be made about these elasticities. Most importantly, they capture not only the effect of human capital externalities only but also the degree of substitution between high-skilled and low-skilled workers. Suppose that human capital externalities are present for both types of workers but their impact is greater on high-skilled workers than on low-skilled workers, $\gamma^H > \gamma^L$. In that case, the wage elasticity for low-skilled workers with respect to $S_{c,t}$, $\delta_{c,t}^L$, is always positive as both the externality and the substitution effects increase their productivity. By contrast, the wage elasticity for high-skilled workers, $\delta_{c,t}^H$, may be either positive or negative, as the substitution effect goes in the opposite direction from the externality effect. As acknowledged by [Moretti \(2004a\)](#) and [Ciccone and Peri \(2006\)](#), the magnitude of human capital externalities cannot be recovered from simple regressions of the logarithm of wage on $S_{c,t}$, even when conducted separately for high-skilled and low-skilled workers. However, the specification can be easily augmented to identify both externality and substitution effects.

Wage elasticities $\delta_{c,t}^H$ and $\delta_{c,t}^L$ in (5.43) and (5.44) vary across locations since there is no reason why the wage bill ratio $\phi_{c,t}$ should be constant over space. This suggests regressing the logarithm of wage not only on the human capital variable $S_{c,t}$ but also on its interaction with $\phi_{c,t}$ (while also including in the specification individual fixed effects, individual variables, and local variables affecting other types of agglomeration economies). Regressions should be run separately for high-skilled and low-skilled workers as the coefficients for the two variables are not identical for the two types of workers. According to (5.43) and (5.44), one recovers four coefficients that can be used to estimate the three parameters γ^H , γ^L , and ρ . The model is overidentified, which makes it possible to conduct a specification test.

An alternative approach has been proposed by [Ciccone and Peri \(2006\)](#), but only the average effect of human capital externalities can be recovered and not those specific to each type of worker. We present this approach in a simplified way. [Ciccone and Peri \(2006\)](#) first compute a local average wage weighted by the share of each worker type in local employment, $w_{c,t} = s_{c,t}w_{c,t}^H + (1 - s_{c,t})w_{c,t}^L$, with $s_{c,t}$ the share of high-skilled workers in local employment. The elasticity of this average wage with respect to $S_{c,t}$, holding $s_{c,t}$ constant, is given by

$$\frac{\partial \log w_{c,t}}{\partial \log S_{c,t}} = \phi_{c,t} \gamma^H + (1 - \phi_{c,t}) \gamma^L. \quad (5.45)$$

This relationship is strictly valid for variations over time in the short run in line with the definition of the elasticity. [Ciccone and Peri \(2006\)](#) make the approximation that it can be used to study long-run variations of the logarithm of the wage between two dates t and t' (1970 and 1990 in their application) when the logarithm of $S_{c,t}$ varies while holding constant the local share of workers. More precisely, they first construct a city wage index at date t' considering the local composition of workers at date t :

$$\tilde{w}_{c,t'} = s_{c,t} w_{c,t}^H + (1 - s_{c,t}) w_{c,t}^L. \quad (5.46)$$

The log-wage difference $\log \tilde{w}_{c,t'} - \log w_{c,t}$ is then regressed on $\log S_{c,t'} - \log S_{c,t}$ to recover an effect supposed to be the weighted average of the effects of human capital externalities given by (5.45).

What remains unclear is the source of variations over time of $S_{c,t}$. Holding the share of high-skilled workers in total employment $s_{c,t}$ constant implies that the ratio between the numbers of high-skilled and low-skilled workers, $S_{c,t}$, is constant too. Another issue arises because the right-hand side of (5.45) is considered to be a constant coefficient, whereas it clearly varies across cities since $\phi_{c,t}$ is specific to the city. Finally, even if the wage $\tilde{w}_{c,t'}$ is supposed to be computed with the local composition of workers fixed to its value at date t , its computation involves the wages of both skill groups at date t' , $w_{c,t'}^j$. These are not the wages that workers would have had when holding constant the composition of employment. Indeed the actual variation of wages between the two dates may have been influenced by the changes in the local composition of workers.

The use of a CES production function emphasizes the role of the elasticity of substitution between high-skilled and low-skilled workers, which can be recovered from the estimations. It is possible to conduct a similar analysis with a Cobb–Douglas production function although the elasticity of substitution is then fixed and equal to -1 (in particular, we get a Cobb–Douglas specification in our setting when ρ tends to zero). In that case, local labor cost shares are constant and they are given by the Cobb–Douglas coefficients of the two groups. Nevertheless, the procedure we propose can still be applied if the coefficients of the Cobb–Douglas production function are allowed to differ across locations.

Finally, alternative variables can be considered to measure local human capital externalities, such as the share of high-skilled workers in total employment. The choice of a variable ultimately relies on the choice of an *ad hoc* functional form. For instance, [Moretti \(2004a\)](#) and [Combes et al. \(2008a\)](#) regress the logarithm of individual wages on the local share of high-skilled workers in total employment, instead of the ratio between the numbers of high-skilled and low-skilled workers. Controlling for an individual fixed effect, as

well as individual and local characteristics. Even when the specification is estimated separately for high-skilled and low-skilled workers, the issue remains that only a composite of the externality effect and the substitution effect is identified. To go further and identify separately the two effects, it might be worth augmenting the specifications with the interaction of the human capital variable and the local share of high-skilled workers in the wage bill, as proposed above.

5.4. ESTIMATION STRATEGY

Now that the links between theory and empirical specifications, as well as the interpretation of estimated coefficients, have been clarified, we move to a number of empirical issues. First, we discuss the use of TFP rather than nominal wage as a measure of productivity. We then turn to endogeneity issues which emerge when estimating wage or TFP specifications. We present the solutions proposed in the literature to deal with these issues as well as their limits. We finally discuss a series of other empirical issues regarding spatial scale, functional forms, observed skills measures, and spatial lag models.

5.4.1 Wages versus TFP

So far, we have mostly considered nominal wage at the worker level as our measure of productivity. Alternatively, one may wish to use a measure at the firm level such as output value or value added. It is possible to derive a specification for such a measure that is consistent with the production function used in [Section 5.2](#). Let us rewrite the production function at the firm level as

$$Y_{j,t} = \frac{A_{c,t}}{\alpha^\alpha (1-\alpha)^{1-\alpha}} (s_{j,t} L_{j,t})^\alpha K_{j,t}^{1-\alpha}, \quad (5.47)$$

where j denotes the firm, $Y_{j,t}$ is the firm output, $s_{j,t}$ corresponds to average labor skills, which are allowed to vary across firms, $L_{j,t}$ and $K_{j,t}$ are labor and other inputs, respectively, and $A_{c,t}$ is the technological level supposed to be local (we could alternatively consider that it varies across firms within the same local labor market but this does not change the reasoning and we prefer to stick to a simple specification). The output value is given by $p_{j,t} Y_{j,t}$, where $p_{j,t}$ is the average income of the firm per unit produced (see [footnote 1](#) for more details). The logarithm of TFP can be recovered as

$$\ln p_{j,t} Y_{j,t} - \alpha \ln L_{j,t} - (1-\alpha) \ln K_{j,t} = \ln \frac{p_{j,t} A_{c,t} s_{j,t}^\alpha}{\alpha^\alpha (1-\alpha)^{1-\alpha}}. \quad (5.48)$$

Equation (5.48) for TFP is equivalent to (5.3) in logarithmic form for wage. It can be used to relate the logarithm of TFP (rather than wage) to some local characteristics, density among others, which are determinants of agglomeration economies operating through firm price $p_{j,t}$, average labor skills $s_{j,t}$, and local technological level $A_{c,t}$.

If value added is reported in the dataset instead of output value, intermediate consumption can be taken into account in the production function. For instance, consider that production is Leontieff in intermediate consumption denoted $I_{j,t}$ with share in output a and the Cobb–Douglas function (5.47):

$$Y_{j,t} = \min \left(\frac{I_{j,t}}{a}, \frac{A_{c,t}}{\alpha^\alpha (1-\alpha)^{1-\alpha}} (s_{j,t} L_{j,t})^\alpha K_{j,t}^{1-\alpha} \right). \quad (5.49)$$

Profit maximization yields that intermediate consumption is proportional to production, and this leads to

$$\ln(p_{j,t} Y_{j,t} - \nu_{j,t} I_{j,t}) - \alpha \ln L_{j,t} - (1-\alpha) \ln K_{j,t} = \ln \frac{(p_{j,t} - a \nu_{j,t}) A_{c,t} s_{j,t}^\alpha}{\alpha^\alpha (1-\alpha)^{1-\alpha}}, \quad (5.50)$$

where the left-hand side is TFP measured now in terms of value added, with $\nu_{j,t}$ the unit price of intermediate input. This makes it possible to conduct the analysis in a similar way as when TFP is measured in output value. The interpretation of estimated parameters is slightly different since the output price is now net of the unit cost of intermediate consumption.

There are two important differences with a wage analysis, which arise because the term that depends on local characteristics is $p_{j,t} A_{c,t} s_{j,t}^\alpha$ when one considers TFP in output value, whereas it was $(p_{c,t} A_{c,t} / (r_{c,t})^{1-\alpha})^{1/\alpha} s_{c,t}$ in the case of the nominal wage (see Equation (5.3)). The local cost of inputs other than labor does not enter the expression for output value and the determinants of agglomeration economies only capture effects related to technological level, output price, and average skills. This means that land and housing prices no longer play a role. This is clearly an advantage since we saw that the interpretation of the effect of housing price is difficult for wage regressions, and the use of this price as an explanatory variable raises serious endogeneity concerns. Moreover, the elasticity of agglomeration economies obtained from TFP regressions must be multiplied by 1 over the share of labor in the production function $1/\alpha$ to be directly comparable with the one obtained from wage regressions. For these two reasons, the economic interpretation of the impact of local characteristics is not the same when studying TFP or wages.

It is also important to note that wages are usually only proportional to and not equal to labor productivity by a factor that depends on the local monopsony power of the firm. This proportionality factor may be correlated with some local determinants of agglomeration economies, but one may wish to avoid considering its spatial variations as part of agglomeration effects. This may be the case when differences in local monopsony power result from differences in institutional features, which occur, for instance, between countries, and not from differences in the degree of competition in local labor markets. The use of TFP avoids making any assumption about the relationship between the local monopsony power and agglomeration economies. Finally, note that in the framework proposed here, agglomeration effects may operate at the firm level and not only at the local level as in previous sections, since the output price $p_{j,t}$ and average

labor skills $s_{j,t}$ are now specific to the firm. This may also be considered for wages, but we postpone the related discussion until [Section 5.4.4](#).

Additionally, an empirical concern is that firm TFP, the left-hand side in (5.48), is not directly observable in datasets, and computing its value requires estimating parameter α .¹² However, output, labor, and other inputs are simultaneously determined by the firm, which causes an endogeneity issue that can potentially bias the estimated coefficient obtained from OLS. Several methods have been proposed to estimate α consistently, such as a generalized method of moments (GMM) approach applied to the specification of output value in first difference (to deal with firm unobservables) using lagged values of labor and other inputs as instruments in the spirit of [Arellano and Bond \(1991\)](#) and followers, or sophisticated semiparametric approaches to control for unobservables which make use of additional information on investment ([Olley and Pakes, 1996](#)) or intermediate consumption ([Levinsohn and Petrin, 2003](#)). There is no consensus on a method that would be completely convincing, and robustness checks have to be conducted using several alternative approaches.

Moreover, agglomeration variables may be endogenous too for the reasons we develop in the next subsection, and this issue needs to be addressed. One way to proceed consists in applying a two-stage approach where the production function is estimated in the first stage with one of the alternative methods we have just cited and no local variable is introduced. Local-time averages of residuals are then computed and regressed in a second stage on some local characteristics. We detail below approaches to deal with the endogeneity of local characteristics in the second stage. Alternatively, local-time fixed effects can be introduced in a first stage and their estimators regressed in a second stage, in the spirit of what was proposed for individual wages (see [Combes et al., 2010](#), for more details). This second approach has the advantage of properly controlling at the individual level for unobserved local shocks that may be correlated with firm variables. A last approach consists in estimating a specification of output value $p_{j,t}Y_{j,t}$ including both inputs and local characteristics as explanatory variables, instrumenting variables all at once. This was proposed, for instance, by [Henderson \(2003\)](#), who estimates an output value specification with the GMM.

5.4.2 Endogeneity issues

We now detail the various endogeneity problems that can occur and approaches that have been proposed to solve them. When the effect of local characteristics on individual

¹² One can relax the assumption of constant returns to scale and also estimate parameters for inputs other than labor without requiring that their total share in input costs is equal to $1-\alpha$.

outcome is estimated, endogeneity can occur both at the individual level and at the local economy level. To see this, we rewrite Equation (5.6) as

$$y_{i,t} = u_i + X_{i,t}\theta + \sum_c [Z_{c,t}\gamma + \eta_{c,t}] 1_{\{c(i,t)=c\}} + \epsilon_{i,t}, \quad (5.51)$$

where $1_{\{c(i,t)=c\}}$ is a dummy variable equal to 1 when individual i locates in c at date t . This expression involves local effects related to observables, $Z_{c,t}$, and unobservables, $\eta_{c,t}$, on every local market, and makes explicit the location choice $1_{\{c(i,t)=c\}}$ which is made at the individual level.

There is an endogeneity issue at the local level when a variable in $Z_{c,t}$, density for instance, is correlated with the local random component $\eta_{c,t}$. This can happen because of reverse causality or the existence of some missing local variables that affect directly both density and wages. Reverse causality is an issue when higher local average wages attract workers, as this increases the quantity of local labor and thus density. In that case, one expects a positive bias in the estimated coefficient of density (provided that density has a positive effect on wages owing to agglomeration economies).

There is a missing variable problem when, for instance, some local amenities not included in $Z_{c,t}$ are captured by the local random term and they determine both local density and wages. Productive amenities such as airports, transport infrastructures, and universities increase productivity and attract workers, which makes the density increase. In that case, a positive bias in the estimated coefficient of density is also expected. In line with [Roback \(1982\)](#), consumption amenities such as cultural heritage or social life increase the attractiveness of some locations for workers and thus make density higher. Such amenities do not have any direct effect on productivity, but the increase in housing demand they induce makes land more expensive. As a result, local firms use less land relatively to labor, and this decreases labor productivity when land and labor are imperfect substitutes. This causes a negative bias in the estimated coefficient of density since density is positively correlated with missing variables that decrease productivity.

Finally, the unobserved local term captures among other things the average of individual wage shocks at the local level. This average may depend on density as workers in denser local markets may benefit from better wage offers owing, for instance, to better matching. One may consider that matching effects are part of agglomeration economies and then there is no endogeneity issue. Alternatively, one may be interested solely in the effects of knowledge spillovers and market access for goods captured by density, in which case there is an expected positive bias in the estimated effect of density owing to the contamination by matching mechanisms.

Endogeneity concerns can also arise at the individual level when location dummies $1_{\{c(i,t)=c\}}$ are correlated with the individual error term $\epsilon_{i,t}$. This occurs when workers sort across locations according to individual characteristics not controlled for in the specification such as some of their unobserved abilities. We emphasize in [Section 5.2.1](#) the

importance of considering individual fixed effects u_i to capture the role of any individual characteristic constant over time. However, workers might still sort across space according to some time-varying unobserved characteristics entering $\epsilon_{i,t}$.

Endogeneity at the individual level also emerges when workers' location choices depend on the exact wage that they get in some local markets, typically when they receive job offers associated with known wages. Notice that this type of bias is closely related to matching mechanisms although there is here an individual arbitrage between locations, whereas the matching effects mentioned earlier rather refer to a better average situation of workers within some local markets. Importantly, as long as individual location decisions depend only on the explanatory terms introduced in the specification, which can go as far as the individual fixed effect, some time-varying individual characteristics such as age, and a location-time fixed effect, there is no endogeneity bias. [Combes et al. \(2011\)](#) detail these endogeneity concerns.

5.4.3 Dealing with endogenous local determinants

The literature has mostly addressed endogeneity issues at the local level using several alternative strategies. A simple approach consists in including time-invariant local fixed effects in specifications estimated on panel data to deal with missing local variables that are constant over time. Some authors instrument the local determinants of agglomeration economies using additional variables such as local historical or geological variables. Estimations with GMM, where lagged values of local determinants themselves are used for instrumentation, have been considered too but their validity relies on stronger assumptions. Finally, other articles exploit natural experiments involving a shock on local characteristics related to agglomeration economies. This section examines these various strategies. The reader may also refer to the chapter by [Baum-Snow and Ferreira \(2015\)](#) for additional considerations on causality.

By contrast, we are not aware of nonstructural contributions dealing with endogeneity at the individual level, to the extent that some concerns would remain in the most complete specifications including both individual and location-time fixed effects. Structural approaches considering dynamic frameworks like those presented in [Section 5.2.4](#) are clearly a natural way to consider endogenous individual location choices.

5.4.3.1 Local fixed effects

One reason why local determinants of agglomeration economies can be endogenous is that some missing variables determine them simultaneously with the local outcome. In particular, this is the case when there are missing amenities that affect both local productivity and the local population. A strategy for coping with this issue when panel data are at hand is to include time-invariant local fixed effects in the estimated specification. There are several reasons why this strategy may not work well. First, it does not deal with missing variables that evolve over time: for instance, new airports or stations are built or

improved over the years depending precisely on their local demand and the performance of local firms and workers. Second, time-invariant local fixed effects do not help in solving the endogeneity issue due to reverse causality, such that higher expected wages or productivity in a location attract more firms and workers. Third, identification relies on time variations of the local outcome and local determinants of agglomeration economies only. If the variations of local determinants are mismeasured, which is likely to happen as local determinants are often computed from samples of limited size and variations are often considered only in the short run because the time span of panels is, in general, quite short, estimated effects can be highly biased because of measurement errors. This kind of problem can be particularly important for local characteristics which vary little across time—for instance, because the economy is close to a spatial equilibrium.¹³ Their effect is difficult to identify separately from the role of permanent characteristics that affect productivity without being related to agglomeration economies. Nevertheless, one can try to identify their effect by using an instrumentation strategy applied to a specification in level.

5.4.3.2 Instrumentation with historical and geological variables

An alternative strategy for coping with endogeneity at the local level consists in finding instruments that deal with both reverse causality and missing amenities. Instruments should verify two conditions: relevance and exogeneity. Instruments are relevant when they are correlated with the instrumented variables $Z_{c,t}$, and they are exogenous when they are not correlated with the aggregate random term $\eta_{c,t}$. Two necessary conditions for exogeneity are that instruments are not correlated with missing local variables and not determined by the outcome.

Several sets of instruments have been proposed. The first one consists of historical instruments and more particularly long lagged values of variables measuring agglomeration economies (see [Ciccone and Hall, 1996](#); [Combes et al., 2008a](#)). Historical values of population or density are usually considered to be relevant because local housing stock, office buildings, and factories last over time and create inertia in the local population and economic activity. If the lags are long enough (say, 150 years), instruments are believed to be exogenous because of changes in the type of economic activities (agriculture to manufacturing then services) and sometimes wars that reshaped the area under study. Local outcomes today are therefore unlikely to be related to components of local outcomes a long time ago that probably affected the historical population. However, there are local permanent characteristics that may have affected past location choices and still affect local productivity today, such as the centrality of the location in the country, a suitable climate, or geographical features such as access to the coast or the presence of a large

¹³ This does not necessarily mean that they do not shape the magnitude of agglomeration economies.

river. If these features are not properly controlled for in regressions, the local historical population may not be exogenous.

The second set of instruments consists of geological variables related to the subsoil of the location (see [Rosenthal and Strange, 2008](#); [Combes et al., 2010](#)). These variables typically describe soil composition, depth to rock, water capacity, soil erodibility, and seismic and landslide hazard. They are believed to be relevant because the characteristics of soils were important for agriculture centuries ago, even millennia ago, and manufacturing and services have since developed where human settlements were already located. They are believed to be exogenous because people may have had only a negligible effect on soil and geology, and these do not influence the productivity of most modern activities.

Some authors argue that consumption amenities can be used as instruments since according to the [Roback \(1982\)](#) model, they are relevant because they attract workers and therefore determine the local population, and they are exogenous as they would not directly affect local productivity. This is not certain, however, because the inflow of workers puts pressure on local land markets, which in turn gives firms incentives to substitute labor for land in the production process, as we have argued above. As a result, productivity can be affected and consumption amenities are not exogenous. Therefore, we advocate using consumption amenities as control variables rather than as instruments when they are available in datasets.

In practice, historical variables are usually found to be extremely relevant instruments, in particular past population, indicating major inertia in the distribution of population over space. Geological variables are also found to be relevant but to a lesser extent, and their power to explain instrumented variables is not very high. Exogeneity can only be properly tested by confronting different sets of instruments with each other, under the assumption that at least one set of instruments is valid. Indeed, the Sargan exogeneity test implicitly compares the estimators obtained with all the alternative combinations of instruments. The test is passed when these estimators are not significantly different from each other. One has to make the assumption that at least one set of instruments is valid such that the instrumental variable estimator obtained with that set of instruments is consistent. Otherwise, the test could be passed with all instruments being invalid and the instrumental variable estimators obtained with the different combinations of instruments all converging to the same wrong value. As an implication, making an exogeneity test using only very similar instruments (e.g., population 150, 160, and 180 years ago) is not appropriate since the estimated coefficient could be biased the same way in all cases and the overidentification test would then not reject exogeneity. An overidentification test using different types of instruments which are not of the same nature is more meaningful. For instance, it is likely that historical and geological variables satisfy this property: even if geology initially influenced people's location choices a very long time ago, many other factors have also determined the distribution of the population across space since

then and make the local historical population a century ago less related to local geology. Some authors, such as [Stock and Yogo \(2005\)](#), have started to develop weak instrument tests that assess whether different instruments have enough explanatory power of their own and can be used together to conduct meaningful overidentification tests. Such tests should be reported systematically.

Lastly, since [Imbens and Angrist \(1994\)](#), it has been emphasized that instrumentation identifies a local average treatment effect only—that is, an effect specific to the instruments chosen, and not necessarily the average treatment effect. Some differences between the two occur when instruments differently weight observations, locations here, in regressions. For instance, the current total population may be instrumented with the historical urban population rather than the historical total population because of data availability issues (see [Combes et al., 2008a](#)). In that case, the instrument is more relevant for locations with a current population which is large. Indeed, the instrument takes the value zero for all locations with no urban population a long time ago, and varies for locations of large size with positive urban population a while ago. Overall, this also argues for considering different sets of instruments, testing whether they lead to similar estimates as mentioned earlier, and keeping in mind the arguments developed here for the interpretation of different estimates.

5.4.3.3 Generalized method of moments

A third strategy that has been used to cope with endogeneity issues when having panel data is to use a GMM approach to estimate the specification in first difference while using lagged values of variables as instruments, both in level and in first difference. Two main types of specification involving determinants of agglomeration economies have been estimated that way: dynamic specifications of employment at the city-industry level ([Henderson, 1997](#); [Combes et al., 2004](#)) and static or dynamic specifications of TFP or wages ([Henderson, 2003](#); [Mion, 2004](#); [Graham et al., 2010](#); [Martin et al., 2011](#)). As detailed in [Section 5.4.1](#), articles on productivity typically specify in logarithmic form the firm production or value added as a function of labor, other inputs (usually physical capital), local variables determining agglomeration economies, possibly earlier in time, and a firm fixed effect capturing time-invariant firm and local effects. The specification is rewritten in first difference between t and $t - 1$ to eliminate the firm fixed effect. A similar strategy is implemented at the local level when no firm-level data are available. When the effects of all variables are estimated in a single step, first differences of labor, capital, and local variables are simultaneously instrumented by their past values in $t - k$, with $k \geq 2$, and/or by their past levels. When a two-step strategy is implemented such that a TFP specification is first estimated and then either local-time averages of residuals or local-time fixed effects are regressed on local characteristics in a second step, the same kind of instrumentation can be implemented at each step. Lastly, an alternative approach has been proposed by [Graham et al. \(2010\)](#), who specify a vector autoregressive model

where the first equation relates current labor productivity to its past values and those of local characteristics, and additional equations relate current values of local characteristics to their past values and those of productivity. All equations are simultaneously estimated with dynamic GMM, and Granger tests are used to assess the presence of reverse causality between productivity and local characteristics.

As detailed in [Section 5.6.1](#), studies of employment dynamics specify city-industry employment at time t as a function of its lags at times $t - 1, \dots, t - k$, with $k \geq 1$, other time-varying local characteristics, and a city-industry fixed effect. Lags of the dependent variable capture both mean-reversion and agglomeration size effects as argued by [Combes et al. \(2004\)](#), while local characteristics capture other types of agglomeration economies.¹⁴ Again the specification is rewritten in first difference between t and $t - 1$, and first-differenced lags of city-industry population are instrumented with past levels before $t - k$, with $k \geq 3$, and other local variables with their value in $t - 2$.

The approach is valid when the two conditions of relevance and exogeneity of instruments are verified. The relevance of instruments is usually not an issue as there is some inertia in local variables and the time span is usually short (a couple of decades at most). Exogeneity can be the most problematic issue. Take the example of city-industry employment $y_{z,s,t}$ written in first difference $\Delta y_{z,s,t} = y_{z,s,t} - y_{z,s,t-1}$ and regressed on its lagged value $\Delta y_{z,s,t-1}$. The practice consists in instrumenting $\Delta y_{z,s,t-1}$ with the past level $\Delta y_{z,s,t-2}$. The exogeneity condition is not verified if the shock in the outcome specification—say, $\nu_{z,s,t}$ —is serially correlated. This causes the shock in first difference $\Delta \nu_{z,s,t}$ to be correlated with the past employment level $y_{z,s,t-2}$. For instance, industry-city shocks probably last several years, and the exogeneity condition is thus unlikely to hold. One may wish to use as instruments more remote past levels $y_{z,s,t-k}$, with k much larger than 2 to attenuate the bias, but this strategy will also probably fail when the data span 15 or 20 years only. A common practice for testing the validity of the exogeneity condition is to use several lags of the outcome before $t - 1$ as instruments and conduct a Sargan overidentification exogeneity test. This practice is dubious since the test relies on instruments all from the same source, the dependent variable itself. As suggested earlier, variables of a different kind should be used as instruments together with past values of the outcome for the overidentification test to be meaningful. Overall, we advise against relying on approaches based on GMM with lagged values as instruments to identify the role of local determinants on local outcomes.

5.4.3.4 Natural experiments

Another strategy for dealing with an endogenous local determinant consists in exploiting the context of a natural experiment that has induced a sizeable localized shock on that determinant which is not directly related to the outcome variable. The general idea of the approach is to evaluate the effect of the variable from the comparison of the average

¹⁴ Note that there are also specific interpretation issues that are discussed in [Section 5.6.1](#).

variation in outcome in places which have experienced the shock with the average variation in outcome in comparable places which have not experienced the shock. Sometimes, the quantitative value of the shock is not known, and only its effect (i.e., the change in the agglomeration determinant times the coefficient of the variable) is identified. To see this, consider the aggregate model:

$$\beta_{c,t} = Z_{c,t}\gamma + \theta_c + \eta_{c,t}, \quad (5.52)$$

where $\beta_{c,t}$ is a local outcome such as a location–time fixed effect estimated in the first step on individual data, $Z_{c,t}$ includes the local characteristics that determine agglomeration effects, and θ_c is a location fixed effect capturing among others the role of local time-invariant characteristics. A common practice is to make the city fixed effect disappear by rewriting the model in first difference:

$$\Delta\beta_{c,t} = \Delta Z_{c,t}\gamma + \Delta\eta_{c,t}. \quad (5.53)$$

Beyond the fact that controlling for time-invariant local effects can raise measurement issues as discussed above, another problem is that the variation in local variable $\Delta Z_{c,t}$ may be correlated with the variation in residual $\Delta\eta_{c,t}$ because of unobserved time-varying amenities or reverse causality. This problem can be circumvented in the case of a natural experiment. Consider that there is a subset denoted tr (for “treated”) of N_{tr} locations experiencing a shock, or “treatment,” that affects the local variable from date τ onward such that $Z_{c,t} = \bar{Z}_{c,t} + \phi 1_{\{t \geq \tau\}}$, where $\bar{Z}_{c,t}$ is the value of the local variable in the absence of the shock, and $1_{\{t \geq \tau\}}$ is a dummy for being affected by the shock. Consider also that there is a subset denoted ntr (for “nontreated”) of N_{ntr} locations that do not experience any shock from date τ onward. The difference-in-differences estimator of the effect of the shock between dates $\tau - 1$ and τ is the difference between the average outcomes of the treated and nontreated locations, given by

$$\widehat{\phi\gamma} = \frac{1}{N_{\text{tr}}} \sum_{c \in \text{tr}} \Delta\beta_{c,\tau} - \frac{1}{N_{\text{ntr}}} \sum_{c \in \text{ntr}} \Delta\beta_{c,\tau}. \quad (5.54)$$

This estimator converges to the true effect of the shock $\phi\gamma$ provided that the numbers of locations in the treated and nontreated groups tend to infinity and that there is similarity between treated and nontreated locations in terms of the growth of local variables and shocks in the absence of treatment:

$$E[\Delta\bar{Z}_{c,t} | c \in \text{tr}] = E[\Delta\bar{Z}_{c,t} | c \in \text{ntr}] \quad \text{and} \quad E[\Delta\eta_{c,t} | c \in \text{tr}] = E[\Delta\eta_{c,t} | c \in \text{ntr}]. \quad (5.55)$$

Note that when the value of the shock ϕ is observed, it is then possible to recover the marginal impact of the local variable, γ .

The challenge when using a natural experiment is to find a control group which is similar to the treated group such that locations in the two groups would have experienced similar variations in local characteristics absent the shock and such that their unobserved

characteristics would have evolved similarly (condition 5.55). If this is not the case, strategies based on matching can lead to further comparability between the two groups, or regression discontinuity approaches can be used to identify the effect of treatment locally.

A limitation when exploiting a natural experiment, in particular when using these two complementary strategies, is that external validity is not certain. The shock may be specific to a particular context, and locations in the treated and nontreated groups may not be representative of the overall set of cities. Therefore, the estimator obtained from the natural experiment may not correspond to the average effect of the shock for the whole set of cities.

Some articles such as those by [Hanson \(1997\)](#), [Redding and Sturm \(2008\)](#), and [Greenstone et al. \(2010\)](#) have achieved some success in using natural experiments when studying the effect of local determinants of agglomeration economies on outcomes of firms. We detail their strategies and conclusions in [Section 5.5.4](#) concerning the results obtained in the literature.

5.4.4 Tackling the role of firm characteristics

We have so far considered a production function where the TFP of firms is influenced by location but not by any intrinsic characteristic of firms. It is possible to argue though that firms differ in their management teams, with some being more efficient than others, and this creates some heterogeneity in productivity. Moreover, there can be some sorting of firms across space depending on management efficiency—for instance, with firms with the better management teams being created in larger locations. International trade models with heterogeneous firms also imply that only the most able firms can survive in larger markets (see, e.g., [Melitz and Ottaviano, 2008](#)) owing to competition effects that are not related to agglomeration gains. If such firm selection effects exist and firm heterogeneity is not properly taken into account, estimated effects of local characteristics such as city size are biased.

Heterogeneity in firm productivity can be taken into account in the specifications of firm output value derived in [Section 5.4.1](#) by making the TFP specific to the firm rather than to the area in the same way we did for output and input prices. A possible way of taking into account firm heterogeneity in wage regressions is to include firm fixed effects in wage specifications such as (5.6), which becomes

$$y_{i,t} = u_i + v_{j(i)} + X_{i,t}\theta + Z_{c(i,t),t}\gamma + \eta_{c(i,t),t} + \epsilon_{i,t}, \quad (5.56)$$

where $j(i)$ is the firm of individual i and v_j is a firm fixed effect. Two estimation issues need to be discussed. First, it is never possible to control properly for all productive amenities by including explanatory variables at the local level in the regression. Firm fixed effects are thus bound to capture the effect of any omitted local variable not varying over time, and they thus cannot simply be interpreted as firm effects. From a theoretical point of

view, this is crucial when trying to interpret the correlation between worker and firm fixed effects. This correlation does not necessarily capture the effect of a worker–firm match, but could also capture the effect of a worker–area match with some sorting of firms depending on unobserved local characteristics.

Second, it is difficult, if not impossible, to take into account time-varying local unobservables in the computation of standard errors. Indeed, the two-step approach proposed in [Section 5.2.1.1](#) cannot be applied since local-time fixed effects cannot be identified separately from firm fixed effects. This occurs because firms do not move across space and the local average of their effects is then confounded with local effects. The larger the unobserved local effects, the larger the possible bias in standard errors derived from least squares estimation. Some determinants of agglomeration economies could appear to have a significant effect, whereas they would not have a significant effect if unobserved local effects were properly considered.

An alternative approach consists in introducing proxies in the specification for firm characteristics related, for instance, to management or organization, instead of firm fixed effects. One can then apply the two-stage approach to properly take into account local unobservables in the computation of standard errors. Such proxies are hard to find, however, and when estimations are conducted in a single step, firm variables may also capture the effects of local unobservables, which can be due to agglomeration economies. In particular, some authors use firm size as a regressor and do not control for local-time fixed effects (see, e.g., [Mion and Naticchioni, 2009](#)). Firm size may capture not only firm productivity but also agglomeration gains from increasing returns to scale due to a better market access. One may try to distinguish firm productivity by rather using firm size centered with respect to its local average. Another clear limitation to controlling for firm size is that it depends on time-dependent shocks that also affect wages. This causes a simultaneity bias in the estimations. Note that all these issues are common to most firm observed characteristics.

Firm heterogeneity can itself be used to distinguish agglomeration effects from competition effects as proposed by [Combes et al. \(2012b\)](#). That article considers a value-added specification where only labor, capital, and skills are introduced. Firm TFP is measured with the residual computed at the firm level. An economic geography model with heterogeneous firms shows that a test for the presence of agglomeration and competition effects can then be conducted by comparing firms' TFP distributions in small and large cities. If the distribution in large cities is a right-shifted version of the distribution in small cities, all firms in large cities benefit from agglomeration effects. If the distribution in large cities is rather a left-truncated version of the distribution in small cities, competition is fiercer in large cities, which leads to a larger share of the least productive firms being unable to survive there. Estimations from French data taking into account both the right-shift and left-truncation transformations support the presence of agglomeration effects but not the presence of competition effects.

5.4.5 Other empirical issues

5.4.5.1 *Spatial scale*

Articles differ in the spatial scale at which the impact of local determinants is measured. There are two main reasons for that: there is no real consensus on the spatial scope at which each agglomeration mechanism takes place, and any local determinant captures, in general, several mechanisms, the relative intensity of which can differ across spatial scales. Theory makes it clear that the spatial scope of agglomeration effects depends on their type. For instance, whereas technological spillovers often require face-to-face contacts, other agglomeration effects such as input–output linkages could take place at a larger scale such as the region. The issue is in fact more complicated as changing the size of the spatial units usually involves changing their shape, and both changes create modifiable areal unit problems, which were mentioned above. However, [Briant et al. \(2010\)](#) show in the particular case of the effect of local density on individual wages that changing shapes is of secondary importance for the estimates compared with taking into account individual unobserved heterogeneity with individual fixed effects. Changing the size of units has a slightly larger effect but an order of magnitude lower than biases related to misspecifications. Hence, choosing the right specification when measuring the impact of local characteristics appears to be more important than choosing the right spatial units. In practice, differences in estimates when the spatial scale varies can give a clue to the various agglomeration mechanisms at play at the various scales. Knowledge spillovers, human capital externalities, and matching effects should be the most prevalent agglomeration forces at short distances—say, within cities or even neighborhoods. By contrast, the effects of market access for both final and intermediate goods emphasized by economic geography models should be the main agglomeration forces driving differences in local outcomes at a larger scale, such as the region.

Keeping these remarks in mind, some articles have tried to evaluate the spatial extent of the impacts of local characteristics, and the scale at which they are the strongest. A common approach is to consider an individual or location defined at a fine scale and to draw rings with increasing radius around it. The value of any local characteristic can be computed using only locations within each ring separately. The spatial extent of agglomeration effects related to the local characteristic is then tested by including within the same specification its values for all rings. Among the first studies using this strategy on US data, [Rosenthal and Strange \(2003\)](#) were aiming at explaining local firm creation and [Desmet and Fafchamps \(2005\)](#) were aiming at explaining local employment. In [Rosenthal and Strange \(2003\)](#), local activity is considered to be located within 1 mile of the zip code centroid, and three rings around it are considered. The first ring contains activities located between 1 and 5 miles, the second between 5 and 10 miles, and the third between 10 and 15 miles. In [Desmet and Fafchamps \(2005\)](#), the first ring contains activities located between 0 and 5 km from the county, the second between 5 and

10 km, the third between 10 and 20 km, and so on every 10 km up to 100 km. Agglomeration effects are considered to attenuate with distance when a decreasing impact is obtained the further away the rings are from the location. The spatial scope of agglomeration effects is given by the distance after which the local characteristic does not have a significant effect anymore. It can happen that agglomeration effects first increase with distance before decreasing. The turning point gives the spatial scale at which they are the strongest.

5.4.5.2 Measures of observed skills

Individual skills are not evenly distributed across locations. [Combes et al. \(2008a\)](#) show, for instance, that individual fixed effects and location fixed effects obtained from the estimation of a wage equation from French data are largely positively correlated. The uneven distribution of traits, intelligence, and education is documented for the United States by [Bacolod et al. \(2010\)](#). [Bacolod et al. \(2009a\)](#) show that city size is positively correlated with cognitive and people skills, but is negatively correlated with motor skills and physical strength. [Bacolod et al. \(2009b\)](#) also provide evidence that workers in the right tail of the people skill distribution in large cities have higher skills than those in small cities, and that the least skilled are less skilled in large cities than in small cities. This is in line with [Combes et al. \(2012c\)](#), who measure skills with individual fixed effects, and [Beckhout et al. \(2014\)](#), who measure skills with diplomas. Both articles conclude that there is a distribution of skills with larger variance and shifted to the right in larger cities. As discussed above, skills have two specific roles to play when estimating the effects of agglomeration economies on an economic outcome. First, skills can themselves be one of the determinants of agglomeration economies. Second, there can be some sorting of skills across locations, and it is important to control for this to avoid biases when measuring the impact of local characteristics related to agglomeration economies.

As mentioned above, it is possible to keep the form of skills unspecified in wage equations by introducing individual fixed effects when using panel data. This has the two drawbacks that one has to rely on mobile individuals for identification, and individual characteristics that matter for productivity cannot be identified. This strategy cannot be implemented when panel data are not available, but various measures of observed skills can be used at the cost of not controlling for unobservable individual characteristics. There is a long tradition in labor economics of using obvious measures such as diplomas or years of schooling, and we mention [Duranton and Monastiriotis \(2002\)](#) for the United Kingdom and [Wheaton and Lewis \(2002\)](#) for the United States as two early attempts that followed that route. It is also tempting to use the socioprofessional category, “occupation,” which is often recorded in labor force surveys. It captures the exact job done by workers and part of the effects of the past career, and may thus be considered as a measure that should be more correlated with current skills than education.

On the other hand, there is an endogeneity concern since occupation is attached to the job and is jointly determined with the wage. There is no obvious solution for this endogeneity issue, except to use a more structural approach that would jointly model wages and occupational choice.

An interesting alternative is to introduce measures of traits and intelligence. [Bacolod et al. \(2009a, 2010\)](#) build on psychological approaches and use detailed occupations from the *Dictionary of Occupational Titles* to construct such measures using information on job requirements and principal component analysis. They end up with four indices related to cognitive skills, people skills, motor skills, and physical strength. It is possible to assess how individuals score on these four dimensions from the job they have just after completion of their education. [Bacolod et al. \(2009a\)](#), in line with studies in labor economics, also use the Armed Forces Qualification Test, the Rotter index, and the SAT scores for college admission in the United States to control further for worker ability and better capture the quality of education. Some attempts have also been made to use other indirect proxies to control for skills. [Fu and Ross \(2013\)](#) use dummies for locations of residence, with the idea that the choice of a residential location is based on tastes, which are themselves likely to be partially correlated with individual productivity. At the same time, the location of residence can be endogenous as it is chosen while taking into account the location of the workplace and the wage.

5.4.5.3 Functional form and decreasing returns to agglomeration

Most articles estimate a log-linear relationship between local outcome and local characteristics. When the elasticity is between 0 and 1, this corresponds to a function in levels which is concave but nondecreasing. This is an approximation and there is no theoretical reason why the relationship between the logarithm of local outcome and the logarithm of local determinants should be linear. Theory rather predicts that the marginal returns to agglomeration should decrease with city size, for instance, because local congestion increases as the city grows. Gains from human capital externalities from the first skilled workers in a location may be rather large, but the more numerous skilled workers are, the lower the marginal gain from one additional skilled worker. A similar line of argument may hold for most technological spillovers. Economic geography models with variable markups and strategic interactions, such as the one proposed by [Combes and Lafourcade \(2011\)](#), do present the feature that in the short run gains from agglomeration dominate costs as long as the asymmetry between locations is not too large, but further agglomeration in the largest locations can lead to a reverse result. As illustrated in [Section 5.2.1](#), local productivity is negatively affected through some channels, such as the increase of land prices with the population, whatever the city size. This kind of effect can become dominant when cities are very large. More generally, one expects gains from agglomeration to increase and be concave with a steep slope at the beginning, and costs to increase and be convex with an initial slope close to zero. In that case, the difference between the

two is concave and bell shaped. The relationship between the determinants of agglomeration economies, in particular population size, and local outcomes is then expected to decrease beyond some threshold.

The simplest way to test for the presence of non-log-linear relationships consists in augmenting the specification with the square of the logarithm of local determinants, but more complex functions of local determinants such as higher-order polynomials can also be used. For instance, [Au and Henderson \(2006b\)](#) regress the value added of a city on a nonlinear specification of its size using a sample of Chinese cities. [Graham \(2007\)](#) develops an original strategy based on a translog production function and two measures of effective urban density. Effective density is computed as a market potential function using either straight-line distances or generalized transport costs that consider road traffic congestion. Corresponding measures are used to estimate the magnitude of diminishing returns from agglomeration—that is, the concave impact of density, and its link with transport congestion. Note finally that the presence of concave effects can be studied for other local characteristics and outcomes. For instance, [Martin et al. \(2011\)](#) quantify the nonlinear effect of specialization on firm value added. Overall, the literature is rather suggestive of diminishing returns to agglomeration (see [Section 5.5](#)). In practice, when estimating a nonlinear effect, one should always check that the support of observations covers the whole interval where the nonlinear effect is interpreted. Otherwise, interpretation is based on extrapolation rather than an empirical feature of the data.

5.4.5.4 Spatial lag models

There is a strand in spatial econometrics considering that spatial lag models can be informative on the effect of local determinants of agglomeration economies. In these models, a local outcome is regressed on a weighted average of neighbors' outcomes or on a weighted average of neighbors' exogenous characteristics, or both, where weights decrease with distance, and the spatial correlation of residuals is sometimes taken into account (see [Lesage and Pace, 2009](#), for details). The weighted averages of neighbors' outcomes or characteristics are considered to capture agglomeration effects. It is now standard to estimate this kind of model with maximum likelihood. An important limitation to this approach is that the model is identified as a result of parametric assumptions, in particular as regards the impact of space on agglomeration effects and the distribution of residuals.

As emphasized by [Gibbons and Overman \(2012\)](#), spatial specifications face a reflection problem à la Manski, which is known to be very difficult to deal with properly. For instance, consider the case where individual wage is regressed on neighbors' composition in terms of diplomas because one expects human capital externalities to spill over the boundaries of spatial units. This composition may be endogenous as highly educated workers may be attracted to the vicinity of workers earning high wages, in particular because they can finance local public goods.

The reflection problem is usually addressed in spatial econometrics by using spatial lags of higher order as instruments, in the spirit of panel estimation strategies which consist in instrumenting variables by long time lags of their first difference. However, this kind of approach relies on assumptions on the extent of spatial effects. Indeed, one needs to assume that these effects involve only close neighbors, whereas more distant neighbors do not have any direct effect on the outcome, which is the reason why they can be used to construct instruments verifying the exclusion restriction. Nevertheless, it is possible that neighbors located further away also directly affect the outcome, and the instruments are thus invalid. An additional issue is that the validity of instruments cannot be properly assessed using an overidentification test as all instruments are built from the same underlying variables, computed at various distances but fundamentally affected by common shocks.

Overall, the main identification concern remains: one needs to find a strategy to identify the effect of local determinants of agglomeration economies using a natural experiment or valid instruments, and unfortunately spatial lag models are of no help for that. [Corrado and Fingleton \(2012\)](#), [Gibbons and Overman \(2012\)](#), [McMillen \(2012\)](#), and [Gibbons et al. \(2015\)](#) propose a more thorough discussion of the concerns regarding spatial econometrics.

5.5. MAGNITUDES FOR THE EFFECTS OF LOCAL DETERMINANTS OF PRODUCTIVITY

Previous sections presented relevant strategies that could be used to estimate the impact of local determinants of agglomeration economies, and clarified the underlying econometric assumptions and interpretations. Contributions in the literature rarely adopt exactly these empirical strategies and often use variants. This makes it rather difficult to compare their results and it can sometimes explain discrepancies in their conclusions. We survey these contributions as well as their results, and try to emphasize the main assumptions that are made in the estimation strategies in light of previous sections. We first present the large body of articles on the average impact of density on productivity. We then turn to the scarce articles estimating heterogeneous effects across city sizes, workers' skills, or industries. We also review contributions on the spatial extent of agglomeration effects, which include some using natural experiments to address endogeneity issues. Results on specialization, diversity, and human capital externalities are then described, and a final section is devoted to the results obtained for developing countries.

5.5.1 Economies of density

It is now established that the local density of economic activities increases the productivity of firms and workers. This conclusion emerges from a large number of studies mentioned below. Some of them use aggregate data and regress the logarithm of regional

wage or TFP on the current logarithm of employment or population density. Typical values for the elasticity when controlling for some local variables but disregarding both reverse causality and individual unobserved heterogeneity to deal with spatial sorting are between 0.04 and 0.07. The estimates are rather diverse because different countries, industries, or periods of time are considered, as emphasized by [Melo et al. \(2009\)](#). Some studies estimate even larger magnitudes but usually use fewer control variables. The elasticity range 0.04–0.07 implies that when the density is twice as great, productivity is between 3 and 5% higher. Density in the last decile in developed countries is usually at least two to three times greater than in the first decile, and may even be 15 times greater (when considering European regions, or regions within some countries). The productivity gap associated with the interdecile difference may be as large as 20%.

Correcting for aggregate endogeneity is generally found to have a small effect on elasticities. Instrumentation decreases them by 10–20%, and sometimes leaves the estimates unaffected or may even make them increase slightly. By contrast, using individual data and introducing individual fixed effects to control for spatial selection can change the estimated elasticity of productivity with respect to density much more. This elasticity can be divided by a factor larger than 2 and can reach a value typically around 0.02. As detailed below, depending on the country and on the precise method used to control for skills (individual fixed effect or observed skills variables), the magnitude of the sorting bias can differ significantly.

Turning to specific estimates, the two benchmark studies using aggregate data for the United States—those of [Ciccone and Hall \(1996\)](#) and [Rosenthal and Strange \(2008\)](#) for the years 1988 and 2000, respectively—report similar values for the elasticity of productivity with respect to density, at around 0.04–0.05. The first study uses historical variables (e.g., lagged population, lagged population density, or lagged railroad network) as instruments for density and the second study uses geological variables (seismic and landslide hazard, percentage of area underlain sedimentary rock). In both cases, instrumentation barely affects estimates, and if anything, slightly increases the elasticity of productivity with respect to density.

Some studies attempt to estimate this elasticity for European regions. [Ciccone \(2002\)](#) replicates [Ciccone and Hall \(1996\)](#) on NUTS 3 regions in France, Germany, Italy, Spain, and the United Kingdom. His main instrument is land area, which is not very convincing since we argue in [Section 5.3.1](#) that land area can have a direct effect on productivity. He gets an elasticity of around 0.05 for 1992. Interestingly, he also finds no evidence that agglomeration effects significantly differ across countries. Two more recent studies extend the set of countries considered in the analysis, although at the cost of using larger spatial units. [Brühlhart and Mathys \(2008\)](#) consider 245 NUTS 2 regions in 20 western and eastern European countries, with data on the 1980–2003 period for western European countries but only on the 1990–2003 period for eastern European countries, and eight broad industries covering both manufacturing and financial services. They consider first

differences and resort to GMM to deal with endogeneity issues in the estimations. Unfortunately, the results seem to differ widely depending on the empirical strategy they adopt. Still, they estimate quite large agglomeration gains with a long-run elasticity of productivity with respect to density reaching 0.13. Interestingly, the strength of agglomeration effects seems to have increased over time. This result is consistent with economic geography models that predict a bell-shaped curve for trade costs versus agglomeration gains. The European economy, which has experienced a decline in trade costs over the last decades, appears to lie on the right-hand side of the curve, where agglomeration effects are reinforced when trade costs become smaller. [Foster and Stehrer \(2009\)](#) obtain estimates closer to those of [Ciccone \(2002\)](#) when using a panel of over 255 NUTS 2 regions in 26 European countries for the 1998–2005 period that covers six industries, including “agriculture, forestry and fishing,” which is not considered by [Brülhart and Mathys \(2008\)](#). They also obtain the further result of a larger magnitude of agglomeration economies for new member states than for old ones. Nevertheless, they use land area as the only exogenous instrument, as in [Ciccone \(2002\)](#), and consider that the regional skill composition is exogenous, which is not very convincing. [Marrocu et al. \(2013\)](#) further extend the number of countries, regions, and time span while leaving aside the endogeneity issues, and conclude that specialization gains would be more prevalent in new member states and diversity would be more prevalent in older ones.

A number of early studies estimate agglomeration economies for separate countries on either wages or TFP aggregated by region. We do not summarize the results of all these studies as they have already been covered by [Rosenthal and Strange \(2004\)](#). We rather focus on recent articles that use richer datasets at the individual level that include workers’ or firms’ precise location.

[Glaeser and Maré \(2001\)](#) were the first to evaluate agglomeration effects on wages net of individual fixed effects, the analysis being conducted on US data. Unfortunately, the size of their dataset does not allow them to evaluate the elasticity of wages with respect to density but allows them to evaluate only the impact of a couple of dummies for city size. For the same reason, it is also difficult to compare the magnitude of the effects estimated by [Wheeler \(2006\)](#) and [Yankow \(2006\)](#), still from US data, with the magnitudes in the rest of the literature. [Combes et al. \(2008a\)](#) are able to estimate the effect of density on wages across all French cities at the individual level while considering individual fixed effects and taking into account aggregate endogeneity with the two-step estimation procedure involving instrumentation that is described in [Section 5.2.1.1](#). They find an elasticity of wages with respect to density of around 0.030, which is half that obtained when individual unobserved heterogeneity is not taken into account. Using a more elaborate instrumentation strategy, [Combes et al. \(2010\)](#) obtain a value of 0.027. This figure is very close to the one obtained for Spain by [de la Roca and Puga \(2012\)](#) when they do not control for dynamic agglomeration effects, which is 0.025. [Mion and Naticchioni \(2009\)](#) replicate the strategy of [Combes et al. \(2008a\)](#) with Italian data and get an even

smaller estimate of 0.01, which is still significantly different from zero. From UK data, [D'Costa and Overman \(2014\)](#) get an elasticity of 0.016, and from Dutch data, [Groot et al. \(2014\)](#) get 0.021, controlling for many individual variables and city-industry-time fixed effects but not individual fixed effects.¹⁵

[Combes et al. \(2008a\)](#) also show that individual abilities do not distribute randomly across locations. Workers who have higher skills are more often located in productive cities, which are denser. The correlation between individual and area fixed effects is 0.29, and the correlation between individual fixed effects and density is as high as 0.44. This is the fundamental reason why controlling for individual characteristics has so much influence on the estimate of the elasticity of productivity with respect to density. [Mion and Naticchioni \(2009\)](#) find that sorting is slightly weaker in Italy, as they obtain a correlation between individual fixed effects and density of 0.21. There is also some evidence of spatial sorting in Spain as shown by [de la Roca and Puga \(2012\)](#) when dynamic agglomeration effects are not taken into account, and in the United Kingdom as shown by [D'Costa and Overman \(2014\)](#) when both static and dynamic effects are considered.

The role of skills has been debated further by [de la Roca and Puga \(2012\)](#), who show from Spanish data that the explanatory power of individual fixed effects largely falls once dynamic agglomeration effects are taken into account in the specification. As detailed in [Section 5.2.2](#), dynamic effects are captured with variables measuring the time spent in different classes of city size. When these variables are not included in the specification, having spent more time in larger cities is captured by the individual fixed effect. The inclusion of city experience variables allows [de la Roca and Puga \(2012\)](#) to disentangle the effects of individual skills captured by individual fixed effects from dynamic agglomeration gains. In order to assess the magnitude of dynamic gains, [de la Roca and Puga \(2012\)](#) consider a quantity defined at the city level as the sum of the time-invariant city fixed effect and the effect of experience accumulated in the city for a worker who stayed there for 7 years (which is the average length of time for workers in their sample). The elasticity of this quantity with respect to density that captures both static and dynamic agglomeration effects is 0.049, which is almost twice as large as the elasticity of city fixed effects evaluated as 0.025. This indicates major dynamic gains which would be even larger for more able workers as shown by the estimation of a specification allowing for an interaction between the individual fixed effect and city experience. Perhaps surprisingly, dynamic gains are found to be independent of the size of the city to which workers move subsequently. There would thus be a transferability of learning effects, which is homogeneous across locations.

¹⁵ In contrast with these references, when considering individual data on siblings from the United States, [Krashinsky \(2011\)](#) finds that the average urban wage premium becomes nonsignificant when introducing family fixed effects because there is a sorting of families across urban areas.

Following an empirical strategy close to that of [de la Roca and Puga \(2012\)](#), [D’Costa and Overman \(2014\)](#) show for the United Kingdom that dynamic effects are also present but weaker than in Spain. In particular, dynamic gains appear to be one shot only, the first year of stay in a city, and do not cumulate over time (except for the youngest workers, below 21 years old). These results are consistent with those of [Faberman and Freedman \(2013\)](#), who study the impact of the age of firms on earnings returns to density with US data and find that almost all of the gains occur at the birth of firms. The structural exercise conducted by [Baum-Snow and Pavan \(2012\)](#) allows them to consider endogenous individual location choices, static and dynamic heterogeneous agglomeration gains, and matching effects. Their conclusions for the United States are similar to those for Spain. Both static and dynamic gains from agglomeration are present, static gains being more important to explain differences between small and medium cities, and dynamic gains playing a more significant role to explain differences between medium-sized and large cities. Conversely, individual sorting and matching effects play a secondary role in the city wage premium.

Owing to computation limits, many studies consider only classes of city size and not all the cities separately. Moreover, in [de la Roca and Puga \(2012\)](#), the heterogeneous individual impact of dynamic agglomeration economies is supposed to be identical to the direct effect of individual skills, and static agglomeration effects are not allowed to be specific to skills, whereas in [D’Costa and Overman \(2014\)](#), both static and dynamic agglomeration effects are homogeneous across workers. Lastly, considering time-invariant city fixed effects makes the city experience component also capture the time evolution of static agglomeration gains. Other recent attempts that consider both static and dynamic effects in specifications closer to those of [Glaeser and Maré \(2001\)](#) include the work of [Lehmer and Möller \(2010\)](#), who find for Germany that only dynamic effects occur once firm size and individual fixed effects are taken into account, [Carlsen et al. \(2013\)](#), who find for Norway that static gains are homogeneous across education levels, while dynamic ones increase with education, and [Wang \(2013\)](#), who finds for the United States that both static and dynamic gains are present and that they are stronger for younger and more educated workers. To conclude, [de la Roca and Puga \(2012\)](#) and [Baum-Snow and Pavan \(2012\)](#) pioneered the simultaneous study of static and dynamic agglomeration effects on wages, while taking into account the observed and unobserved heterogeneity of workers. Further investigation along the lines suggested in [Section 5.2](#) constitutes an appealing avenue of research.

As discussed in [Section 5.4.1](#), it is worth studying TFP rather than wages since it is a direct measure of productivity that can sometimes be computed at the firm or establishment level, keeping in mind that interpretations change. On the other hand, no convincing method has been proposed to control for individual skills when estimating agglomeration effects on TFP even with individual data at hand, and we have seen that sorting according to skills can induce considerable biases. [Henderson \(2003\)](#) for the

United States and [Cingano and Schivardi \(2004\)](#) for Italy were among the first to study firm-level TFP. However, their assessment of possible endogeneity biases is only partial. [Henderson \(2003\)](#) uses GMM techniques to instrument both input use and local variables, with the caveats we mentioned in [Section 5.4.3.3](#). [Cingano and Schivardi \(2004\)](#) take into account the endogeneity of input use only, through the implementation of the Olley–Pakes estimation procedure. [Graham \(2009\)](#) provides estimates for the United Kingdom based on firm-level TFP data but he instruments neither input use nor local effects. [Di Giacinto et al. \(2014\)](#) assess the respective impact of locating in an urban area and in an industrial district on firm-level TFP in Italy, while instrumenting input use but not the size of the local economy, which is also included as a control. As regards France, [Combes et al. \(2010\)](#) estimate firm TFP with the Olley–Pakes estimation procedure among others and use the estimates to construct a local measure of TFP, which is then regressed on density while using historical and geological variables as instruments. [Martin et al. \(2011\)](#) rather rely on GMM using lagged values of explanatory variables as instruments. To the best of our knowledge, a large number of European countries, including Germany and Spain, have not yet benefited from specific estimates of agglomeration effects on TFP.

Studies on TFP usually conclude that there are significant agglomeration gains in firm productivity, even if some authors who simultaneously control for the level of industrial employment (not its share) wrongly reach the conclusion of their absence (see the discussion in [Section 5.3.2](#)). [Melo et al. \(2009\)](#) show that elasticities of TFP with respect to density are on average estimated to be larger than those obtained for wages, typically around 50% larger, and so are they in [Combes et al. \(2010\)](#), where both types of estimates are computed on the same dataset and endogeneity is taken into account using the same instruments. Indeed, [Combes et al. \(2010\)](#) get an elasticity of TFP with respect to density of 0.035–0.040, whereas they obtain 0.027 for the elasticity of wages. According to our basic model, it is difficult to interpret the difference between the two types of estimates. In wage equations, all the effects are rescaled by the share of labor in the production function. Moreover, agglomeration economies percolating through the cost of inputs other than labor, such as land and intermediate inputs, affect wages but not TFP (see [Section 5.4.1](#)). A further possible reason for the difference in estimates obtained from wage and TFP regressions is that no one has managed to successfully control for individual skills when working on TFP. Taking properly into account workers' unobserved heterogeneity in TFP estimations is an avenue for future research.

5.5.2 Heterogeneous effects

As explained in [Section 5.4.5.3](#), the impact of local characteristics on productivity should be bell shaped as agglomeration gains are increasing and concave, while agglomeration costs are increasing but convex. Variations in the marginal effects of local characteristics

are a first type of heterogeneity. For instance, the gain from increasing city size could be positive and large for small cities, and turn negative for very large ones, predictions that need to be investigated, for instance, to assess whether or not the size of cities is optimal.

Most studies do not report an estimated degree of concavity for agglomeration effects. Exceptions include the study of [Au and Henderson \(2006b\)](#), who estimate for China a bell-shaped relationship between the productivity and size of cities and conclude that most cities lie on the left-hand-side of the peak—that is, they are too small to achieve the highest level of productivity. For the United Kingdom, [Graham \(2007\)](#) develops an original strategy based on road traffic congestion to estimate the diminishing returns of agglomeration effects and their link with transport congestion. Five of nine industries present concave effects of density. Furthermore, it is shown that when congestion is taken into account, the elasticity with respect to density increases in seven of the nine industries. This is in line with expectations since in the absence of controls, the elasticity with respect to density reflects the overall net impact of density, taking into account both positive and negative effects. In the United Kingdom, congestion is shown to represent up to 30% of the agglomeration effect.

Agglomeration effects can also be heterogeneous across industries as the strength of agglomeration economies depends on industry characteristics. Nevertheless, estimations by industry remain scarce. One reason may be that the design of the empirical model, and in particular the search for valid instruments, has to be done industry by industry. Another reason is the lack of availability of local data per industry. The works of [Brühlhart and Mathys \(2008\)](#) and [Foster and Stehrer \(2009\)](#) are notable exceptions, but these works are at the European regional level and do not control for individual effects. They find significant agglomeration effects in all but one of the industries they consider. The exception is agriculture, in which regional density has a negative impact, a result that is fairly intuitive. Given the share of land in agricultural production and the fact that land prices increase with density, less dense places clearly represent the best alternative for productivity in this industry. [Morikawa \(2011\)](#) estimates from firm-level data the elasticity of firm TFP with respect to density for detailed services industries in the United States without using instruments. He finds large elasticities ranging from 0.07 to 0.15. In their meta-analysis, [Melo et al. \(2009\)](#) conclude that on average agglomeration effects tend to be stronger in manufacturing industries than in service industries.

Some studies have tried to evaluate the extent to which agglomeration economies are stronger for some types of workers or firms. For instance, [Bacolod et al. \(2009b\)](#) and [Abel et al. \(2012\)](#) for the United States, [Di Addario and Patacchini \(2008\)](#) for Italy, and [Groot and de Groot \(2014\)](#) for the Netherlands confirm the intuition that returns to education are higher in cities. This is also found for the United States by [Lindley and Machin \(2014\)](#), who then assess to what extent the change in wage inequality across states over the 1980–2010 period arises from a shift in skill composition and a variation in education-specific returns to agglomeration economies. Firms in industries that are more

skill intensive should be concentrated where returns to education are higher, the larger cities, and this is observed by [Elvery \(2010\)](#) for US metropolitan areas. The study by [Lee \(2010\)](#) is one of the rare studies to exhibit an industry in which the urban wage premium is found to decrease with skills, the health-care sector in the United States. He explains his result by labor supply effects for high-skilled health-care employees as surgeons, dentists, or podiatrists, who would be more attracted by urban life than nurses or massage therapists, and this would put a downward pressure on their wages in larger cities.

Using a structural approach controlling for endogenous location choices, [Gould \(2007\)](#) shows that both static and dynamic agglomeration gains are present for white-collar workers but not for blue-collar workers. [Matano and Naticchioni \(2012\)](#) reach a similar conclusion after performing quantile regressions on Italian data and controlling for sorting on unobservable worker characteristics. They find that agglomeration effects appear to strengthen along the wage distribution. This is in line with the conclusions of [Combes et al. \(2012b\)](#), who use the full distribution of firm-level TFP in France to show that the most efficient firms gain more from density than the least efficient ones. For instance, firms in the last quartile of productivity gain three times more from density than those in the first quartile. It is also found that the largest establishments gain more from density. The benefits are 50% greater for establishments with more than 100 workers than those with 6–10 workers. Going in the opposite direction, [Henderson \(2003\)](#) and [Martin et al. \(2011\)](#) conclude that specialization effects are larger for smaller firms, but these two articles measure specialization with the level and not share of industrial employment. Therefore, they partially confound density and the specialization effects as explained in [Section 5.3.2](#).

Other authors have investigated the sources of heterogeneous productivity gains from agglomeration, but rarely take into account simultaneously the endogeneity issues related to reverse causality and missing local variables. For instance, [Rosenthal and Strange \(2003\)](#) using US data find that the number of hours worked decreases with density for nonprofessionals but increases for professionals, and the effect is stronger for young workers. Moreover, the number of hours worked by young professionals is particularly sensitive to the proximity of other young professionals. [Bacolod et al. \(2009a\)](#) investigate which skills have returns positively related to city size. They conclude that only cognitive and social skills are better rewarded in large cities, while motor skills and physical strength are rewarded less well. In line with these results, [Andersson et al. \(2015\)](#) find that it is only for nonroutine jobs that there are gains from agglomeration in Sweden once the spatial sorting of skills is taken into account.

There is also scarce evidence of heterogeneous agglomeration gains across demographic groups. [Phimister \(2005\)](#) estimates gender differences in city size premium from UK data, controlling for individual fixed effects but without taking into account endogeneity issues. He finds a larger urban premium for women, especially for those who are married or cohabiting. [Ananat et al. \(2013\)](#) investigate differences across races in the United States while

controlling for unobserved worker heterogeneity through residential location choices as in [Fu and Ross \(2013\)](#) but without dealing with endogeneity issues at the local level. They find that agglomeration effects are heterogeneous across races, the black–white wage gap increasing by 2.5% when there are 1 million more inhabitants in the city.

5.5.3 Spatial extent of density effects

The rapid spatial decay of agglomeration effects is another robust finding in the literature. Agglomeration economies do not spill much over space. For the advertising agency industry, [Arzaghi and Henderson \(2008\)](#) provide evidence of an extremely fast spatial decay of agglomeration effects that are shown to occur primarily within 500 m. This decay is certainly too extreme to be representative of more standard industries but, still, effects are rarely found to be significant beyond 100 km, and the threshold is often lower.

The first way to assess the spatial extent of agglomeration effects consists in considering a single market potential variable that encompasses both the own location size and the sizes of other locations. As detailed in [Section 5.3.1](#), one can consider the Harris market potential, which is simply the sum over all spatial units, including the own location, of their size (or density) divided by the distance between the location and the unit considered. More structural forms of market potential from economic geography models can also be used. Importantly, in all cases, one implicitly assumes a quite strong spatial decay of agglomeration effects. For instance, when trade costs are inversely related to distance, the impact on a location of the economic activity located 20 km away is four times lower than that of activity located 5 km away, it is 10 times lower at 100 km than at 10 km, and so on. The positive effect of the economic size of distant locations and the spatial decay of this effect are rarely rejected empirically. For instance, [Head and Mayer \(2006\)](#) in a study on European NUTS 2 regions obtain, when neither local skills nor endogeneity are taken into account, that both the Harris market potential and a structural market potential significantly increase regional wages, the two variables having a similar explanatory power. [Holl \(2012\)](#) assesses the effect of a Harris market potential based on distance through the real road network for which the historical population, geology, and historical transport networks are used as instruments. He finds a positive effect of this market potential on regional wages in Spain. Structural articles following [Hanson \(2005\)](#), such as the two early replications by [Mion \(2004\)](#) for Italy and [Brakman et al. \(2004\)](#) for Germany, confirm the positive impact of structural market potential on regional wages, even if sorting on skills is not always taken into account and endogeneity concerns are not always fully addressed. [Brakman et al. \(2006\)](#), [Breinlich \(2006\)](#), [Brakman et al. \(2009\)](#), and [Bosker et al. \(2010\)](#) find evidence of a positive effect of structural market potential on GDP per capita for NUTS 2 European regions. [Fallah et al. \(2011\)](#) show for US metropolitan areas

that the impact of the structural market potential is stronger at the top of the wage distribution. Some other contributions for developing countries are discussed in [Section 5.5.7](#).

Assessing separately the role of the own density and market potential definitely makes more sense if different local externalities operate at different distances. External market potential (which excludes the own size or density) is most often found to have a significant positive effect on local productivity when it is introduced in addition to density in the specification. For instance, [Combes et al. \(2008a, 2010\)](#) find that both variables have a significant positive effect in France, even when they are both instrumented and individual unobserved heterogeneity is taken into account. For NUTS 2 European regions, [Foster and Stehrer \(2009\)](#) introduce next to density a measure of market potential with a spatial decay of agglomeration economies arising from other regions of exponential form—that is, with a decline that is even sharper than the inverse of distance. When trying exponential functions with various coefficients, they find that only those with the strongest spatial decay exhibit significant effects. Note that, in general, introducing the external market potential in regressions only slightly reduces the impact of the own density.

The second strategy for assessing the spatial decay of agglomeration economies consists in introducing in the specification variables for the economic size of distant locations. [Ciccone \(2002\)](#) finds for NUTS 3 European regions that production in neighboring regions has a positive impact on local productivity. He does not report the magnitude of the coefficient however, and he does not test for an impact of regions located further away. [Rice et al. \(2006\)](#) find for UK regions that agglomeration economies attenuate sharply with distance. Distant markets do affect local wages and productivity, but markets located 40–80 min away have one-quarter the effect of those located less than 40 min away, and markets located 80–120 min away have no significant impact. [Rosenthal and Strange \(2008\)](#) obtain even larger spatial gradients when estimating the effect of employment concentration in rings around location on wages in US cities. The effect of the 0–5-mile ring is four to five times larger than the effect of the 5–25-mile ring. Turning to the outer rings (25–50 miles and 50–100 miles), they find that the effects are even smaller and very often not significantly different from zero. The spatial pattern obtained for Italy by [Di Addario and Patacchini \(2008\)](#) is consistent with this one since the impact of local population size is strongest between 0 and 4 km and is not significant anymore beyond 12 km.

5.5.4 Market access effect evaluated using natural experiments

As our chapter shows, strategies used to tackle endogeneity issues are not always convincing, and in some cases, authors do not even attempt to tackle them. A few recent publications propose using natural experiments as a source of variation in the local economy

size to circumvent endogeneity problems. [Greenstone et al. \(2010\)](#) test the presence of agglomeration effects on firm TFP by exploiting the arrival of large plants in some given US counties. Such plants affect the intensity of agglomeration economies, although it is not possible to quantitatively assess the exact magnitude of the shocks. The key idea for finding a relevant control group for counties receiving a large plant is to rely on a real estate journal, *Million Dollar Plants*, that gives for any large plant created the county that the plant ultimately chose (the winner) and the counties that survived a long selection process but were ultimately not selected (the runners-up). [Greenstone et al. \(2010\)](#) show that on average runner-up counties have characteristics similar to those of winners. The effect of plant arrivals on incumbent plants is studied in a panel including both winner and runner-up counties but not others. Firm TFP is regressed on an interaction term between a dummy for being in the winner group and a dummy for the dates after the arrival of the large plant. The estimated coefficient of this interaction corresponds to the difference-in-differences estimator. It is found to be significantly positive and sizeable, especially for incumbent plants sharing similar labor and technology pools with the new plant. Whereas the empirical strategy is quite convincing for identifying the effect of arriving plants, the link between the arrival of plants and changes in the intensity of agglomeration spillovers remains unknown (see the argument in [Section 5.4.3.4](#)). Moreover, external validity is far from certain since only a small subsample of counties is studied.

Articles exploiting natural experiments to evaluate the effect of market potential typically use the opening and closing of frontiers that prevent firms or cities from interacting with neighbors. An early example is given by [Hanson \(1997\)](#), who studies the effect of the trade reform in Mexico in the 1980s that turned the country from a closed economy to an economy open to trade with foreign countries, and in particular with the United States. The opening of the frontiers has increased the market potential, especially for firms close to the Mexican–US border. It is shown that the opening of frontiers attracted firms close to this border, whereas the concentration of firms in the capital city Mexico, which is located at a distance from this border, decreased. A more recent interesting use of a natural experiment is provided by [Redding and Sturm \(2008\)](#), who study the effect of the division of Germany in 1949 on the growth of cities on the western side of the West German–East German border.¹⁶ The border cut their access to cities on the eastern side and thus decreased their market potential. The effect on cities located further away from the border should have been smaller as they had better access to other cities in western Europe. Consequently, [Redding and Sturm \(2008\)](#) compare the population growth of western cities close to the border with that of western cities far from the border, the two groups of cities having the same population trends before

¹⁶ Note that the outcome here is city growth and not productivity as in other contributions surveyed in this section. This is because we chose to review all significant articles using natural experiments at the same place. Other results on city growth are reviewed in [Section 5.6](#).

the division of the country. This is done in the same spirit as [Greenstone et al. \(2010\)](#), by restricting the sample to western cities and regressing city growth on an interaction term between a dummy for being close to the West German–East German border and a dummy for dates after 1949. It is found that division of Germany led to a substantial relative decline of population growth for cities close to the border.¹⁷ The effect is larger for smaller cities, which is expected since they have a smaller own market and rely more on other city markets. An interesting additional exercise would be to assess to what extent the division of Germany decreased the value of a market potential index and to deduce from this measure of the shock and the difference-in-differences estimator a value for the elasticity of population growth with respect to market potential. This coefficient could be compared with the one obtained using a more standard least squares instrumentation approach.

5.5.5 Specialization and diversity

We now review articles evaluating the effect of localization economies on local productivity. The main variable used for that purpose is specialization, which is computed as the share of the industry in the local economy. Its effect on local productivity is assessed while controlling for the size or density of total activity. In many studies, when density and specialization are simultaneously introduced, both are found to have a significant positive effect on productivity. For instance, [Cingano and Schivardi \(2004\)](#) show that this is the case in Italy when industries are pooled together. They also find that the spatial decay is very strong, since specialization in neighboring regions has no impact on local productivity. For France, [Combes et al. \(2008a\)](#) find that the effect of specialization, estimated on wages separately for each industry, is significantly positive for 94 industries out of 99. Its magnitude is larger in business services and in two high-tech industries, medical instruments and artificial fibers. This is intuitive since such industries could face stronger technological spillover effects. These results confirm those of [Henderson \(2003\)](#) for the United States, where a larger effect of specialization is found in high-tech industries. [Martin et al. \(2011\)](#) obtain a significant positive effect of specialization on firm productivity in France that becomes negative above a certain level of specialization, which is consistent with the presence of concave localization effects. From European data, [Brühlhart and Mathys \(2008\)](#) find a negative impact of own-industry density on output per worker in the industries they study, with the notable exception of financial services. Using a spatial variance analysis, [Combes et al. \(2008a\)](#) show that whereas total

¹⁷ A follow-up study ([Ahlfeldt et al., 2012](#)) shows that the division and reunification of Berlin had a significant effect on the gradient of land prices and employment in West Berlin close to the former main concentration of economic activity in East Berlin but a negligible effect along other more economically remote sections of the Berlin Wall.

employment density explains a large share of spatial disparities in productivity, the explanatory power of specialization remains small.

Following both the intuition of [Jacobs \(1969\)](#) and the central role of preference for diversity in many economic geography models, another appealing variable to explain productivity is the overall industrial diversity of the location. However, its estimated effect has been shown to be not robust. It is sometimes significantly positive, sometimes significantly negative, and often not significant at all, as, for example, for France in both [Combes et al. \(2008a, 2010\)](#), for Italy in [Cingano and Schivardi \(2004\)](#), and for the United States in [Henderson \(2003\)](#). Even if there are interesting intuitions behind diversity variables, no effect seems to be at play. This may be due to the way diversity is measured, since it is often through a Herfindahl or Krugman specialization index computed from the industry shares in the local economy using a rather aggregate industry classification. Moreover, some industries may benefit from a group of other industries but usually not from all industries as assumed in the Herfindahl index. To tackle this issue, [Moretti \(2004b\)](#) uses a measure of proximity between industries and finds for the United States that spillovers between economically close industries are larger than spillovers between economically distant industries, and this better matches what Jacobs had in mind.

5.5.6 Human capital externalities

We have already emphasized that the local share of professionals or highly educated workers has many effects on productivity that can be difficult to disentangle. First, when using data aggregated at the city level or the region level, one cannot identify separately the direct composition effect of skilled workers on average productivity and their human capital externality effect. When using individual data, one can assess the role of the local share of skilled workers on individual productivity, while simultaneously taking into account the direct composition effect by introducing individual variables or individual fixed effects. Nevertheless, [Section 5.3.3](#) shows that the local share of skilled workers captures not only the externality effect but also a substitution effect, which is positive for unskilled workers and negative for skilled workers.

There has been a debate since the beginning of this millennium on the existence and magnitude of local human capital externalities. While [Moretti \(2004a,b\)](#) find significant positive effects of human capital measures, [Ciccone and Peri \(2006\)](#) rather obtain an estimate that is not significant. It is difficult to make a conclusive case for either side. [Moretti \(2004a\)](#) implements the now standard approach of regressing the individual wage on the share of college-educated workers, but this share captures both the externality and substitution effects. This is also the case in [Moretti \(2004b\)](#) when studying TFP rather than wages. On the other hand, [Ciccone and Peri \(2006\)](#) use a shift-share approach supposed to control for substitution effects, but the sources of identification remain unclear as

explained in [Section 5.3.3](#). Importantly, no article simultaneously controls for the presence of possible gains from density, whereas density is usually positively correlated with local human capital.

Other articles mostly use the same approach as [Moretti \(2004a\)](#) and obtain similar results. [Rosenthal and Strange \(2008\)](#) find the same positive effect of the local share of college-educated workers in the United States. Considering this share at various distances from each worker location, they also find that the effects of human capital externalities attenuate sharply with distance. The effect of the share of college-educated workers in the 0–5-mile ring around the location is 3.5 times larger than the effect of this share in the 5–25-mile ring. These results are consistent with those of [Fu \(2007\)](#), who finds for the Boston Metropolitan Area using data on census blocks that human capital externalities decrease quickly beyond 3 miles.

For Europe, [Rice et al. \(2006\)](#) assess the role of the local share of workers with degree-level qualifications in the United Kingdom and find that it has a positive effect on wages and productivity. However, since the specification is estimated not at the individual level but rather at the local level, it is not possible to quantify separately the composition and externality effects. This is possible for France, and [Combes et al. \(2008a\)](#) find a positive effect of the local share of professionals within the industry on individual wages, even after controlling for individual fixed effects and age, as well as location–time fixed effects that capture in particular the effect of density. Similarly, [Rodríguez-Pose and Tselios \(2012\)](#) find a positive impact of the regional levels of education on individual earnings for European regions while using individual data and controlling for individual characteristics and region–time fixed effects.

Interestingly, when both productivity and wage data are available, one can evaluate how much of the productivity gains due to agglomeration are transformed into wage gains for workers. While this has not been done for Europe, [Moretti \(2004b\)](#) finds for the United States that estimated productivity differences between cities with high human capital and low human capital are similar to observed differences in wages of manufacturing workers, indicating an almost complete transfer of human capital effects to workers. Since unobserved worker heterogeneity is not controlled for in that study, the similarity between the productivity and wage differences can also result from a composition effect affecting both wages and TFP.

5.5.7 Developing economies

We now present empirical results on the presence of agglomeration economies in some developing countries. The related literature is recent, and research needs to be pursued to gain knowledge on additional countries. The effect of market size on wages has been studied for China, India, and Colombia. Panel data are usually not available, and it is thus, generally not possible to take into account unobserved individual heterogeneity. Differences between individuals are rather taken into account through individual explanatory

variables such as qualification, gender, age, and sometimes occupation or the type of firm where the individual is employed. Overall, market size is found to have a larger effect than in developed countries. [Combes et al. \(2013\)](#), for instance, study the effect of density on individual wages in 87 Chinese prefecture cities, using as instruments for density the peripherality, the historical status of the city, and the distance to historical cities. The elasticity of wages with respect to density is found to be 0.10–0.12, around three times larger than in developed countries. [Chauvin et al. \(2014\)](#) evaluate the effect of density on individual annual earnings in India at the district level and also find a large elasticity of around 0.09–0.12. [Duranton \(2014\)](#) investigates the impact of population on individual wages in Colombia while controlling for area at the local labor market level (which amounts to investigating the effect of density). Instrumentation is conducted using historical populations or soil characteristics (erodibility and fertility). The estimated elasticity is 0.05, and thus lower than in China and India, but still large compared with estimates for developed countries.

Other measures of productivity have been used in studies at the aggregate level. [Henderson et al. \(2001\)](#) evaluate the effect of city population on value added per worker in Korea for 5 industry groups and 50 cities using panel data over the 1983–1993 period. They do not find evidence of a size effect for any industry, but their results are based on time evolutions without instrumentation for the endogeneity of the city population. Similarly, [Lee et al. \(2010\)](#) find that population density does not have any significant effect on establishment-level output per worker in Korea when estimating a specification where local fixed effects and control variables are considered. [Au and Henderson \(2006a\)](#) and [Au and Henderson \(2006b\)](#) study at the city level the effect of total employment and its square on output per worker in China in the 1990s, using as instruments urban plans not related to output and urban amenity variables. They control for the local shares of manufacturing and services, and the shape of the total employment effect is allowed to vary with these shares. They find a concave effect of total employment on output per worker. The vast majority of Chinese cities appear to have a size of less than 50% of the peak, where agglomeration economies are the most important. This can be explained by the *hukou* system that restricts workers' social rights mostly to their birthplace and thus limits their mobility, especially in the 1990s, when it was strictly enforced.

There are also a couple of publications on firm productivity. [Lall et al. \(2004\)](#) study the effect of urban density on firm productivity in India for 11 industries considered separately, estimating jointly a production function and a cost function. The effect is found to be significantly positive in one industry only. [Saito and Gopinath \(2009\)](#) quantify the impact of regional population on firm TFP in the food industry in Chile, estimating a production function using the Levinsohn–Petrin approach. The elasticity is found to be significantly positive, at around 0.07. In both articles, the authors do not deal with the endogeneity of local determinants of agglomeration economies.

The role of market potential is considered along with the size of the local economy by some of the previous articles. [Lall et al. \(2004\)](#) study the impact of the Harris market

potential in India, an originality of their work being the use of accurate transport times rather than distances in the construction of their market potential variable. This variable includes the own location, and its effect is found to be negative but nonsignificant for several industries. Other articles conduct similar exercises but remove the own area from the computation of the market potential measure to disentangle the size effects from the local economy and external markets. Interestingly, [Duranton \(2014\)](#) obtains a significantly negative sign for the effect of external market potential on wages in Colombia. An explanation may be that when workers are perfectly mobile as in [Krugman \(1991b\)](#), the spatial equilibrium without full agglomeration implies lower nominal wages in larger regions to compensate for the better market access that decreases the prices of consumption goods. [Combes et al. \(2013\)](#) find no significant effect of market potential on wages in China once it is instrumented simultaneously with other local determinants, whereas [Au and Henderson \(2006a\)](#) find a positive effect on output per worker but the variable is not instrumented.

Some articles have adopted quasi-structural approaches inspired by [Redding and Venables \(2004\)](#) and [Hanson \(2005\)](#) to focus on the effects on wages of structural market access and supplier access that are derived from economic geography models. This has the limitation that the own area is involved in the construction of the access variables and the effect of the own local economy size cannot be identified separately from the effects of external market and supplier access. [Amiti and Cameron \(2007\)](#) study the effect of both access variables on wages at the firm level in Indonesia, but without being fully structural in their construction and without using instruments to take into account endogeneity issues. Both market and supplier access are found to have a positive effect. Only 10% of the market access effect goes above 108 km, and only 10% of the supplier access effect goes above 262 km.

[Fally et al. \(2010\)](#) evaluate the impact of market and supplier access on individual wages in Brazil using a two-stage approach. First, a wage equation including state-industry fixed effects and individual characteristics is estimated in the spirit of [Combes et al. \(2008a\)](#) but at the industry level and without individual fixed effects since only cross-section data are available. In a second step, estimated state-industry fixed effects are regressed on structural measures of market and supplier access. These measures are obtained following strictly the strategy proposed by [Redding and Venables \(2004\)](#) where market and supplier access are recovered from the estimates of the trade flow specification derived from a economic geography model. An originality is that trade flows are measured at the industry level, which allows the construct of the access variables for each industry separately, whereas other articles only use aggregate flows and therefore construct only aggregate access variables.¹⁸ Both market and supplier access variables are found to have a significant positive effect on wages when estimations are conducted using OLS.

¹⁸ The second-step estimation could have been for each industry separately, as proposed in [Section 5.2.1](#), but pooling all industries together was preferred, possibly because the number of locations (27 states) is small.

The supplier access variable is then removed from the specification and only the market access variable is instrumented (both variables rarely have simultaneously a significant effect owing to their high correlation). Market access is found to keep its significant positive impact on wages.

Finally, [Hering and Poncet \(2010\)](#) evaluate the effect of market access on individual wages in 56 Chinese cities. They also follow the strategy proposed by [Redding and Venables \(2004\)](#) to build the market access variable but they do not consider the role of supplier access at all. Labor skills are captured by individual observed characteristics and a single-step estimation strategy is used. [Hering and Poncet \(2010\)](#) instrument market access by centrality indices and find a significant positive effect which is larger for skilled workers.

Note that in all these contributions, structural access variables are the only local determinants of agglomeration economies considered in the specifications. Therefore, their impacts cannot be identified separately from the effects of other local determinants not derived from economic geography models if these other determinants are correlated with access variables, which can occur in particular when distance plays a similar role in the attenuation of their effects.

Finally, some articles have studied local determinants of agglomeration economies other than market size. [Henderson et al. \(2001\)](#) assess the effect of industrial specialization (measured with industry local employment) on productivity growth in Korea. They find some evidence of localization economies for all the industry groups they consider, the magnitude of the effects being similar to those for the United States. [Lopez and Suedekum \(2009\)](#) are interested in localization economies and agglomeration spillovers on TFP for establishments in Chile. They consider both downstream and upstream spillovers between firms related by input–output relationships. They find a positive effect of the number of intraindustry establishments consistent with the presence of localization effects and a positive effect of the number of establishments in upstream industries consistent with unidirectional agglomeration spillovers. [Saito and Gopinath \(2009\)](#) evaluate the impact of diversity, measured by a Herfindahl index, on firm TFP in the food industry in Chile, but find no significant effect. Endogeneity of local determinants and spatial sorting of workers are considered in none of these articles.

5.6. EFFECTS OF AGGLOMERATION ECONOMIES ON OUTCOMES OTHER THAN PRODUCTIVITY

Although the most straightforward interpretations are made for the effects of local variables on local productivity, a rather large literature has attempted to identify the role of agglomeration economies on local outputs other than productivity. These outputs include employment or employment growth, and firm location decisions. We now turn to this literature and relate it to the same theoretical framework as the one we developed

for productivity. This allows us to emphasize difficulties that are encountered when interpreting the results. Nevertheless, we survey the results that have been obtained over the last decade.

5.6.1 Industrial employment

We first focus on the local determinants of local industrial employment. We provide a theoretical background to specifications estimated in the literature, comment on the interpretations that can be made for the estimated coefficients, and finally present the results obtained in related articles.

5.6.1.1 From productivity externalities to employment growth

The two early studies that initiated the empirical evaluation of agglomeration economies in the 1990s, those of Glaeser et al. (1992) and Henderson et al. (1995), do not directly focus on the determinants of local productivity but focus rather on those of local employment growth at the industry level. A possible reason is that data on wages or TFP at fine geographical levels such as cities or local labor markets were less available than today, and this is even more the case for individual data. At the same time, employment is, by itself, a local outcome of interest, especially for policymakers, when, for instance, regional unemployment disparities are large as in Europe.

We develop a theoretical framework similar to the one used for productivity in order to ground employment equations and to allow for relevant interpretations of the effects found in this literature. As will become clear below, it is necessary to rely on a production function at the industry level with nonconstant returns to scale and we consider

$$Y_{c,s,t} = \frac{A_{c,s,t}}{\alpha_1^{1-\alpha_2} \alpha_2^{\alpha_2}} (s_{c,t} L_{c,s,t})^{\alpha_1} K_{c,s,t}^{\alpha_2}, \quad (5.57)$$

where $\alpha_1 + \alpha_2 < 1$. The first-order conditions equalizing the return of inputs to their marginal productivity are

$$w_{c,s,t} = \frac{\alpha_1 p_{c,s,t} A_{c,s,t}}{\alpha_1^{1-\alpha_2} \alpha_2^{\alpha_2}} s_{c,s,t}^{\alpha_1} L_{c,s,t}^{\alpha_1-1} K_{c,s,t}^{\alpha_2}, \quad (5.58)$$

$$r_{c,t} = \frac{\alpha_2 p_{c,s,t} A_{c,s,t}}{\alpha_1^{1-\alpha_2} \alpha_2^{\alpha_2}} s_{c,s,t}^{\alpha_1} L_{c,s,t}^{\alpha_1} K_{c,s,t}^{\alpha_2-1}. \quad (5.59)$$

Substituting into (5.59) the expression of capital given by (5.58) leads to

$$L_{c,s,t} = \left(\frac{p_{c,s,t} A_{c,s,t} s_{c,s,t}^{\alpha_1}}{w_{c,s,t}^{1-\alpha_2} \alpha_2^{\alpha_2} r_{c,s,t}} \right)^{1/1-\alpha_1-\alpha_2}. \quad (5.60)$$

We first leave aside the role of wages, which will be discussed below. Making the same assumptions as in Section 5.2 on how local characteristics determine $p_{c,s,t}$, $A_{c,s,t}$, and $r_{c,s,t}$,

we can use Equation (5.60) to motivate an empirical specification where the logarithm of local industry employment (instead of wage) is expressed as a function of local variables such as local density, land area, and specialization:

$$\ln L_{c,s,t} = \beta \ln \text{den}_{c,t} + \mu \ln \text{area}_{c,t} + \vartheta \ln \text{spe}_{c,s,t} + \nu_{c,s,t}. \quad (5.61)$$

First notice that, as in the case of productivity, the exact channel of agglomeration economies cannot be identified since local characteristics determining agglomeration effects may have an impact on employment not only through technological progress, but also through input prices and goods prices. Importantly, the role of specialization cannot be identified since the dependent variable, industrial employment, is a log-linear combination of specialization and density, and terms have to be rearranged to avoid redundancy. This identification issue is the reason why the production function was specified at the industry level. By contrast, the role of other local variables can still be studied since (5.61) implies

$$\ln L_{c,s,t} = \frac{\beta - \vartheta}{1 - \vartheta} \ln \text{den}_{c,t} + \frac{\mu - \vartheta}{1 - \vartheta} \ln \text{area}_{c,t} + \nu_{c,s,t}. \quad (5.62)$$

The impact of the remaining local determinants is now net of the impact of specialization, and cannot be identified separately from it.¹⁹ It was initially suggested in the literature that the static agglomeration effect related to specialization could be identified using nonlinearities by also including in (5.61) the level of specialization in addition to its logarithm as an extra local variable. However, this makes interpretations difficult, especially when the two effects are estimated with different signs as, for instance, in Henderson et al. (1995). Parametric identification relying only on specific functional forms should be avoided.

Glaeser et al. (1992) propose rewriting (5.60) in first difference and then considering that the growth rate of local variables instead of their level is a function of the levels of local determinants. They interpret local variables as determinants of technological progress, but these variables also capture the role of agglomeration economies operating through goods and input prices as shown by (5.60). Specialization can now be included among local characteristics, and its effect is identified separately. The corresponding specification is given by

$$\ln L_{c,s,t} - \ln L_{c,s,t-1} = \tilde{\beta} \ln \text{den}_{c,t-1} + \tilde{\mu} \ln \text{area}_{c,t-1} + \tilde{\vartheta} \ln \text{spe}_{c,s,t-1} + \epsilon_{c,s,t}. \quad (5.63)$$

The coefficients of local variables capture dynamic agglomeration effects such as improved learning but not the impact of static ones as in (5.62).

¹⁹ Firm-level data would make it possible to identify the effect of industry employment by regressing firm employment on industry employment, in a way analogous to how individual wages allowed us to identify the role of individual skills separately from human capital externalities. This has not been done before to the best of our knowledge.

When there is time autocorrelation of residuals, it is possible to derive from (5.62) a dynamic specification of local-industry employment similar to (5.63) even if there are no static and dynamic agglomeration effects. Suppose for instance that $\nu_{c,s,t}$ follows an AR(1) process such that

$$\nu_{c,s,t} = (1 - \rho) \nu_{c,s,t-1} + \varepsilon_{c,s,t}, \quad (5.64)$$

where $0 < \rho < 1$ and the residuals $\varepsilon_{c,s,t}$ are identically and independently distributed. When there is no agglomeration effect such that Equation (5.62) reduces to $\nu_{c,s,t} = \ln L_{c,s,t}$ and if we take into account the fact that $L_{c,s,t} = \text{den}_{c,t} \cdot \text{area}_{c,t} \cdot \text{spe}_{c,s,t}$, equation (5.64) implies

$$\begin{aligned} \ln L_{c,s,t} - \ln L_{c,s,t-1} &= -\rho \ln L_{c,s,t-1} + \varepsilon_{c,s,t} \\ &= -\rho \ln \text{den}_{c,t-1} - \rho \ln \text{area}_{c,t-1} - \rho \ln \text{spe}_{c,s,t-1} + \varepsilon_{c,s,t}, \end{aligned} \quad (5.65)$$

which involves the same explanatory variables as (5.63) but with coefficients constrained to be the same and negative. This suggests that when a specification such as (5.63) is estimated, it is possible to obtain negative coefficients for local variables even in the presence of dynamic agglomeration economies, and negative signs have indeed been obtained in the literature.

Taking all the intuitions in (5.61), (5.63), and (5.65) together, one may consider a specification with static and dynamic agglomeration effects (as we did for productivity in Section 5.2.2), as well as time autocorrelation of residuals, which leads to

$$\begin{aligned} \ln L_{c,s,t} - \ln L_{c,s,t-1} &= -\rho \ln L_{c,s,t-1} + \beta (\ln \text{den}_{c,t} - \ln \text{den}_{c,t-1}) \\ &\quad + \mu (\ln \text{area}_{c,t} - \ln \text{area}_{c,t-1}) + \vartheta (\ln \text{spe}_{c,s,t} - \ln \text{spe}_{c,s,t-1}) \\ &\quad + \tilde{\beta} \ln \text{den}_{c,t-1} + \tilde{\mu} \ln \text{area}_{c,t-1} + \tilde{\vartheta} \ln \text{spe}_{c,s,t-1} + \varepsilon_{c,s,t}. \end{aligned} \quad (5.66)$$

This specification involves time variations of static effects, dynamic effects, and inertia in industrial employment due to the time autocorrelation of residuals.²⁰ Rearranging terms to eliminate current and past specialization (as their coefficients are not identified), we finally get

$$\begin{aligned} \ln L_{c,s,t} - \ln L_{c,s,t-1} &= \frac{\tilde{\vartheta} - \rho}{1 - \vartheta} \ln L_{c,s,t-1} + \frac{\beta - \vartheta}{1 - \vartheta} \ln \text{den}_{c,t} + \frac{\mu - \vartheta}{1 - \vartheta} \ln \text{area}_{c,t} \\ &\quad + \frac{\tilde{\beta} - \beta + \vartheta - \tilde{\vartheta}}{1 - \vartheta} \ln \text{den}_{c,t-1} + \frac{\tilde{\mu} - \mu + \vartheta - \tilde{\vartheta}}{1 - \vartheta} \ln \text{area}_{c,t-1} + \varepsilon_{c,s,t}, \end{aligned} \quad (5.67)$$

²⁰ This specification is not completely consistent with all the specifications above. It is possible to derive a specification which is consistent but it is much more intricate.

which is a specification close to the one estimated by Henderson (1997) and Combes et al. (2004). Alternatively, one can replace past industrial employment $L_{c,s,t-1}$ by $\text{den}_{c,t-1} \cdot \text{area}_{c,t-1} \cdot \text{spe}_{c,s,t-1}$ to rather consider a specification with past specialization although the same parameters are identified.

Unfortunately, the five coefficients in Equation (5.67) are combinations of the seven parameters of interest. It is thus difficult to interpret the estimated coefficients even if one is able to deal with the endogeneity of right-hand-side variables. For instance, a negative impact of past industrial employment is compatible not only with the presence of inertia in the series together with a positive static effect of specialization, but also with a negative static effect of specialization. Similarly, a positive impact of past local determinants is not incompatible with a negative impact of some static or dynamic agglomeration effects. As there are more parameters of interest than estimated coefficients, the different effects cannot be disentangled. The model could be augmented with other local characteristics such as market potential or diversity, and more lags of industrial employment, using statistical tests to determine how many lags should finally be kept. However, the same identification issues would remain as the impact of these variables would mix again static and dynamic effects.

Another point that we have not discussed so far about Equation (5.60) is that the local wage (or local wage growth if the dependent variable is employment growth) should be used as a control variable in the empirical specification if one wishes to restrict the interpretation of the effects of local characteristics to their role in $p_{c,s,t}$, $A_{c,s,t}$ and $r_{c,s,t}$ only (consistent with the analysis on productivity) and avoid considering their role in $w_{c,s,t}$. Since one estimates a labor demand equation, the local wage is expected to have a negative effect on local employment. For given wages, agglomeration effects increase labor demand, and therefore we expect a positive effect of density, area, and market potential among other factors on local employment as in the case of productivity.

However, controlling for wages means that only a partial equilibrium effect of agglomeration economies is captured. It corresponds to the direct impact of agglomeration economies on labor demand but it does not capture the feedback effects on this demand resulting from the wage change induced by agglomeration. Moreover, from the econometric point of view, controlling for wages raises serious additional endogeneity issues, on top of those described above when the dependent variable measures productivity.

One can choose not to control for the local wage but then the impact of local characteristics on local employment operates not only through $p_{c,s,t}$, $A_{c,s,t}$ and $r_{c,s,t}$ but also through $w_{c,s,t}$ and the effect through the wage is negative. Typically, agglomeration economies raise nominal wages, which in turn yield a decrease in labor demand. The overall impact of agglomeration economies on employment is now ambiguous, and in particular it can be negative. On the one hand, agglomeration economies that increase $p_{c,s,t}$ and $A_{c,s,t}$ and decrease $r_{c,s,t}$ tend to positively affect employment; on the other hand,

they also increase $w_{c,s,t}$, which tends to negatively affect employment. When the effect of density on local employment is found to be negative, one does not know if density has a negative effect on productivity, and therefore a negative effect on employment because productivity is positively related to employment, or if density has a positive effect on productivity, which in turn has a positive effect on wages, themselves affecting employment negatively. For instance, [Cingano and Schivardi \(2004\)](#) get opposite signs for some of the common determinants of productivity and employment, on the basis of the same Italian dataset. This suggests that a positive effect of agglomeration economies on local productivity can actually turn into a negative effect on local employment, an issue that was initially raised by [Combes \(2000\)](#).

Finally, [Combes et al. \(2004\)](#) also propose breaking down local employment into two terms, employment per firm and the local number of firms:

$$\ln L_{c,s,t} = \ln \left(\frac{L_{c,s,t}}{n_{c,s,t}} n_{c,s,t} \right) = \ln \frac{L_{c,s,t}}{n_{c,s,t}} + \ln n_{c,s,t}, \quad (5.68)$$

where $n_{c,s,t}$ is the local number of firms within the industry. One can evaluate separately the impact of local characteristics on average employment in existing firms and on the number of firms. Indeed, urbanization and localization variables can have different effects on the intensive and extensive margins of employment. In first differences, the analysis indicates whether agglomeration economies have the same or opposite effects on internal firm growth and on external growth, or whether the effects are stronger for one or the other employment growth components. Finally, note that some authors evaluate the effect of local human capital on employment growth in the spirit of what has been done for productivity, as, for instance, by [Simon \(2004\)](#) for the United States, and by [Suedekum \(2008, 2010\)](#) for Germany. The interpretation is again blurred by the existence of substitution effects between high-skilled and low-skilled workers as discussed in [Section 5.3.3](#).

5.6.1.2 Total employment, specialization, diversity, and human capital

The explanatory variables introduced into employment growth regressions are usually very similar to those considered in productivity regressions, except that local density is replaced by local total employment. Estimated specifications generally involve dynamic agglomeration effects following (5.63) but not static effects. Results for the effect of total employment on industrial employment growth clearly illustrate the diversity of results obtained in the literature on local employment growth. Beyond the fact that samples for different countries and periods are used, the previous section illustrates how the use of different specifications changes the interpretation of estimated effects. For instance, [Combes \(2000\)](#) finds for France that the local market size has a positive effect on industrial employment growth for manufacturing industries but a negative effect for service industries. [Viladecans-Marsal \(2004\)](#) finds for Spain that the effect on industrial employment is

not significant for three of six industries, while it has a bell-shaped effect in the three other industries. [Blien et al. \(2006\)](#), who extend the analysis of [Blien and Suedekum \(2005\)](#), obtain for Germany that local market size plays a positive role on industrial employment growth for both manufacturing and service activities. There are two recent studies on Italy, one that pools together manufacturing and service industries ([Mameli et al., 2008](#)) and one that focuses on business services ([Micucci and Di Giacinto, 2009](#)). Both conclude that total employment has a positive impact on industrial employment growth.

As we mentioned above, the question of the spatial decay of agglomeration effects is crucial. For the United States, [Desmet and Fafchamps \(2005\)](#) consider the impact on local employment growth of total employment and industrial employment share at various distances from the location. They show that for nonservice industries, such as manufacturing and construction, the effects are negative for distances below 20 km, but are slightly positive for distances between 20 and 70 km. This is consistent with employment moving away from city centers with high aggregate employment to nearby locations. Service industries exhibit a different pattern for the effect of total employment: the coefficients are positive at distances below 5 km, and are slightly negative at distances between 5 and 20 km. This is consistent with employment growing faster in city centers and more slowly in nearby areas. Unfortunately, this question has rarely been addressed for European economies. [Viladecans-Marsal \(2004\)](#) studies the effect on industrial employment of the local characteristics of neighboring cities in Spain. She finds the effects of total local employment and employment in neighboring locations to be significant in two of the six industries she considers. In the same vein, and still with Spanish data, [Solé-Ollé and Viladecans-Marsal \(2004\)](#) show that growth of the central municipality within metropolitan areas has a positive effect on growth in the suburbs. [Micucci and Di Giacinto \(2009\)](#) also find for Italy a significant impact of distant locations on local employment growth.

The impact of diversity on productivity has been found to be not robust, and this is also true for its effect on industrial employment growth. Whereas [Glaeser et al. \(1992\)](#) find a positive impact of diversity (measured by the share of the five largest industries within the city) on industrial employment growth, [Henderson et al. \(1995\)](#), who use a Herfindahl index over all local industries, obtain a significant positive effect in a couple of high-tech industries only. For France, [Combes \(2000\)](#) finds that the same diversity index has a positive impact on employment growth in service industries but a negative one in most manufacturing industries, although it is positive for a few of them. For Spain, [Viladecans-Marsal \(2004\)](#) finds a positive static effect on employment for three industries but a negative effect for some others and a nonsignificant effect for two of them. For Germany, [Blien et al. \(2006\)](#) find that diversity has a positive effect on employment growth in both manufacturing and service industries, the effect being strong in manufacturing industry. Diversity is also found to have a significant positive impact in Italy according to [Mameli et al. \(2008\)](#).

The impact of specialization is difficult to assess because its effect on agglomeration economies cannot be disentangled from the mean reversion process of industrial employment as shown earlier. The impact of specialization is found to be negative in both manufacturing and service industries in France by [Combes \(2000\)](#), in Germany by [Blien et al. \(2006\)](#), and in Italy by [Mameli et al. \(2008\)](#). This result may arise from strong mean reversion that more than compensates for positive agglomeration effects. [Van Soest et al. \(2006\)](#) obtain a positive effect of specialization in the Netherlands, but the impact is very local and dies out quickly with distance.

[Glaeser et al. \(1992\)](#) popularized the use of the local average size of firms in industry as a determinant of localization economies as discussed in [Section 5.3.2](#). Both [Combes \(2000\)](#) for France and [Blien et al. \(2006\)](#) for Germany find that the presence of larger firms reduces employment growth in both manufacturing and service industries. To refine the role of local firm size, [Combes \(2000\)](#) introduces a local Herfindahl index of firm size heterogeneity. He finds that the local concentration of employment within large firms is also detrimental to local growth. Therefore, in France, the local market structure that fosters employment growth the most appears to be small firms of even size. A further example of the difficulty of interpreting the findings of this literature is given by [Mameli et al. \(2008\)](#), who show from Italian data that the effect of most local determinants on local employment is not very robust, in the sense that their sign changes depending on the industrial classification which is used.

Finally, local human capital is found to positively affect total employment growth, both in the United States by [Simon \(2004\)](#) and in Germany by [Suedekum \(2008\)](#). However, the latter study emphasizes that mostly unskilled employment growth is favored, which is consistent with the presence of strong substitution effects between the two groups of workers and weak agglomeration effects.

5.6.1.3 Dynamic specifications

A crucial question is the time needed for a determinant of agglomeration economies to have a sizeable effect. The availability of panel datasets has generated a series of articles that estimate jointly the dynamics of both the dependent local variable and local determinants of agglomeration economies in specifications with multiple lags involving both static and dynamic agglomeration effects. In other words, instead of estimating the specifications described in [Section 5.6.1](#), researchers estimate full autoregressive models, as initially proposed by [Henderson \(1997\)](#) for US cities. Once this kind of model has been estimated, short-run effects of local determinants can be distinguished from their long-run effects.

For instance, [Blien et al. \(2006\)](#) show that in Germany the impact of diversity dies out quickly over time, in both the manufacturing sector and the service sector. This means that diversity has no long-run effects. Similarly, the effect of local firm size is significant in

the short run but not in the long run in the two sectors. As mentioned above, [Combes et al. \(2004\)](#) propose decomposing industrial employment into average employment per firm and the number of firms in the local industry. They then estimate from French data a vector autoregressive model involving these two dependent variables (this approach has been replicated with German data by [Fuchs, 2011](#)). It is found that the local determinants of the growth of existing firms are not necessarily the same as those that promote the creation of new firms. Overall, there is a greater inertia in the adjustment process in the United States than in France and Germany. Lagged values stop being significant after 1 year of lag for France and Germany. This is starkly at odds with the 6- or 7-year significant lags found in [Henderson \(1997\)](#) for the United States.

Unfortunately, as emphasized in [Section 5.6.1.1](#), interpretations of estimated coefficients in terms of static and dynamic agglomeration effects remain very difficult because both types of effect can enter each estimated coefficient. Moreover, even if the structure of vector autoregressive models makes them rather suited to deal with endogeneity concerns by using dynamic panel estimation techniques, the application of such techniques is debatable in the context of agglomeration effects as argued in [Section 5.4.3.3](#). Ultimately, the literature using dynamic specifications remains descriptive and is not really able to provide causal interpretations of the effects in terms of agglomeration economies.

5.6.2 Firms' location choices

Rather than assessing the impact of local determinants of agglomeration economies on productivity or industrial employment, some authors have tried to evaluate the impact of these determinants on the location choices of firms. Firms should locate where their expected profit is the highest. As profit increases with productivity, the local determinants of productivity should also affect firm location choices. This is the intuition motivating the approaches presented in this subsection. They lead to applications usually relating to location choices of foreign direct investments (FDIs) or determinants of firm creation.

5.6.2.1 *Strategies and methodological concerns*

To assess the role of local determinants of firm location choices, [Carlton \(1983\)](#) proposes using the discrete choice modeling strategy developed by [McFadden \(1974\)](#). The idea is that, for any given firm, the value of each location depends on a deterministic local profit and an idiosyncratic component. The local profit is supposed to be the same for all firms, but the idiosyncratic component varies across firms (and components are identically and independently distributed across locations for a given firm). This prevents firms from all choosing the same location, which would not correspond to reality. Assuming that idiosyncratic components follow extreme value laws, the firm location choice follows a logistic model, or logit model, which is quite easy to estimate.

Economic geography models predict how firms distribute themselves across space according to local profits, which are nonzero in the short run under imperfect

competition. The location choice thus depends on the same quantities as those that enter the productivity equation (5.50) (the prices of goods and intermediate inputs, the technological level of the firm, and workers' efficiency) as well as the nominal wage. As a result, any of the urbanization and localization variables which enter the empirical specification of productivity can be included in a specification explaining firm location choices. However, interpretations are even more difficult than in the case of industrial employment, as there are direct and indirect effects which sometimes go in opposite directions. Indeed, profits depend not only on productivity but also on input use and output quantity, which are themselves influenced by agglomeration effects but are not introduced in the regression. One can also choose whether or not to control for the local level of wages, but interpretations then differ as in the case of industrial employment. Therefore, proposing correct and precise interpretations is difficult because many effects are at play, and they interfere in nonlinear ways to shape local profits.

Furthermore, almost all the local variables explaining location choices can be considered to be endogenous, precisely owing to the location choices of both firms and workers. This induces reverse causality affecting most local determinants of agglomeration economies. Unfortunately, this kind of issue is tackled even less often in empirical studies on firm location choices than in the literature on the local determinants of productivity and employment. At best, authors lag explanatory variables by one period of time, which is certainly not enough to correct for any endogeneity bias that may occur. To cope with the problem of omitted local variables, some authors include regional dummies at a geographical scale larger than the one considered for location choices, while others exploit time series and introduce local fixed effects. The same important caveats appear as for productivity studies, and they are detailed in [Section 5.4.3](#).

For all these reasons, the literature on firm location choices has to be considered as mostly descriptive. A safer route to assess the role of agglomeration effects on firm location choices would probably be to consider much more structural approaches, which however present the drawback of considering a more limited number of agglomeration channels.

Besides these limits, it is possible to enrich the approach when studying the location choices of firms among places in several countries using a nested logit model involving several stages. For instance, firms first choose the country to which they will locate and then, conditional on this choice, choose the region or city within the country. Two additive random components are now considered, one specific to the region and one specific to the country, and they are assumed to be independent. This structure produces a total random component correlated between regions within a given country, and the correlation can be estimated simultaneously with the other parameters in the model. In fact, the effects of local determinants of location choices at the different spatial scales are evaluated separately, once the geographical decomposition of the whole territory has been chosen (e.g., countries or continents, divided themselves into regions or cities). The nested logit approach has the advantage of limiting the number of possible locations

considered for a firm's choice at a given stage. This can be a desirable feature considering current computer capacities, especially if some fixed effects (for industries or other geographical scales) are introduced in the model. These estimation strategies have been considered in empirical studies that take either a reduced form approach, such as [Carlton \(1983\)](#), or a more structural approach where firm location choices are part of an economic geography model, such as [Head and Mayer \(2004\)](#).

Research based on discrete location choice models has primarily been applied to FDI because the determinants underlying their location decisions are more discernible than those of domestic firms, which are less footloose. In particular, location choices are made by multinational firms in a relatively short period of time, without bearing the weight of historical contingencies like national firms. This makes them more appropriate candidates to test for the presence of agglomeration effects. An alternative approach adopted in a number of articles consists in considering the number of firm entries in a region as the dependent variable, and studying its determinants with a simple Tobit approach, or a count model such as the Poisson model or the negative binomial model, or even with a linear model. The Tobit model takes into account the left censorship of the dependent variable but considers that this variable is continuous. The main advantage of count models is that there is no computational limit on the number of alternatives such as in the logit model. However, there are strong distributional assumptions on residuals. The standard linear model does not impose any assumption on the distribution of residuals and is very flexible for the number of covariates that can be considered, but it ignores the discrete nature of the data and left censoring.

5.6.2.2 Discrete location choice models

Among early studies on the effect of local economy characteristics on location choices of FDI, [Head et al. \(1999\)](#) focus on the determinants of firm location choices between the 50 states of the continental United States, while [Guimaraes et al. \(2000\)](#) conduct a similar exercise for the 275 regions in Portugal, which are much smaller. Because of the urban and regional perspective of our survey, we do not discuss studies on location choices between countries. It may be noted, however, that their findings do not significantly differ from those for location choices within a country even if the nature of the underlying agglomeration economies is likely to differ.

As predicted by theory, the first factor that is almost systematically found to have a positive effect on location choices of FDI is the size of the local economy. For instance, market size is measured with local total income in [Head et al. \(1999\)](#), and with two variables, manufacturing and services employment, in [Guimaraes et al. \(2000\)](#). Among other determinants of firm location choices is market access. [Guimaraes et al. \(2000\)](#) consider the distance to the main cities in Portugal as a proxy. At the European level, [Head and Mayer \(2004\)](#) compare the performance of Harris and structural market

potential variables in explaining the location choices of Japanese affiliates across European regions at the NUTS 2 level. They find that both have a significant positive impact on these choices, even when controlling for a substantial number of other variables. [Basile et al. \(2008\)](#) analyze the location choices of multinational firms of various nationalities in 50 regions in eight EU countries. External market potential is found to have a significant positive effect as well as the own region total value added, which is considered simultaneously. However, both effects appear to be mainly driven by location choices of European multinationals, and they are not significant for non-European ones.

The positive impact of market potential seems to be fairly universal, and it is confirmed when data are disaggregated along various dimensions. For instance, [Crozet et al. \(2004\)](#) find a positive effect on FDI in France whatever the country of origin of firms. When studying FDI in Germany, [Spies \(2010\)](#) always finds a positive effect of market potential when conducting estimations for each industry separately. [Pusterla and Resmini \(2007\)](#), who focus on FDI in the NUTS 2 regions in four eastern European countries, find that both local manufacturing employment and market potential variables positively affect FDI, although most of the impact is on low-tech industries and not on high-tech ones.

As in the literature on productivity determinants, the functional form chosen for the role of distance in the market potential—the inverse of distance in most cases—assumes a fast spatial decay of agglomeration effects. The role of proximity has been further investigated. [Basile \(2004\)](#), for instance, finds a negative effect on FDI of agglomeration in adjacent provinces in Italy, while at the same time agglomeration in the own province has a positive effect. Interestingly, foreign acquisitions can be distinguished from greenfield investments. The effect of the local number of establishments is found to be significantly positive only for foreign acquisitions. However, local demand measured by electricity consumption, which is also introduced into the specification, has a positive influence on the two types of firms. Greenfield investments are more appealing for evaluating the role of agglomeration effects because firms have more freedom in their location choices.

This literature almost systematically considers the role of a variable absent from local productivity or growth estimations: past foreign presence in the region. This variable can have effects going in opposite directions. On the one hand, it may attract future FDI because it reflects unobservable characteristics of the region that are also beneficial to new FDI, or because it reflects an existing business network that may be useful to new FDI. On the other hand, past foreign presence may have a negative impact on new FDI because of competition effects. From a theoretical point of view, it is also difficult to assess how such a variable interferes with other local determinants of agglomeration economies, in particular the size of the local economy. As always, absent relevant instruments and natural experiments, identifying causal effects is very difficult.

Current FDI is shown to be positively correlated with previous FDI. For instance, past FDI is found to attract Japanese affiliates in European regions (Head and Mayer, 2004), and to induce both acquisitions and greenfield investments in Italy (Basile, 2004). Past investment also has an influence in both low-tech and high-tech industries in Germany (Spies, 2010), eastern European countries (Pusterla and Resmini, 2007), and Ireland (Barrios et al., 2006). Basile et al. (2008) find for European regions a positive effect of foreign presence on both European and non-European FDI. Crozet et al. (2004) study FDI in France by the country of origin and find a positive effect of past presence for specific countries only, the largest effects being observed for Japan, the United Kingdom, Belgium, and the United States. Finally, Devereux et al. (2007) find a positive effect of past foreign investment in the United Kingdom on both new investment by domestic firms and FDI, the effect being larger for FDI. The role of social and business networks has also been indirectly investigated through variables such as the distance to the home country or headquarters, which is found to have a negative impact on FDI in France by Crozet et al. (2004) and on European FDI in European regions by Basile et al. (2008). Generally, sharing a common language also has the expected positive effect on FDI, and this can be interpreted as indirect evidence of the presence of communication externalities.

As for productivity, authors also study the effect of local industry characteristics on location choices. FDI is fairly systematically found to be positively correlated with specialization, usually measured by the local count of domestic firms in the industry at the European level (Head and Mayer, 2004), or within countries such as in Portugal (Guimaraes et al., 2000), France (Crozet et al., 2004), or the United Kingdom (Devereux et al., 2007). Devereux et al. (2007) also find a positive impact of local industrial diversity. For Ireland, Barrios et al. (2006) find that diversity has had a significantly positive impact on FDI since the 1980s, but not before, and only for high-tech firms for which specialization has no impact. Conversely, whereas diversity does not matter for low-tech firms, specialization has a positive impact on low-tech FDI. Hilber and Voicu (2010) find for Romania that both domestic and foreign industry-specific agglomeration measures positively affect FDI, but only the effect of domestic agglomeration is robust to the introduction of regional fixed effects. The same is found for the effect of domestic industry-specific agglomeration in neighboring regions. The positive effect of diversity that is estimated without regional fixed effects is found to be not robust to their introduction.

Guimaraes et al. (2000) distinguish between the impact of manufacturing and service concentration, and find a larger impact from service concentration. This result was confirmed in later studies, in particular for eastern European regions. According to Cieřlik (2005), service concentration has a significant positive large effect on FDI in Poland at the NUTS 3 level (49 regions), and the same is found for Romania at the NUTS 3 level (21 regions) by Hilber and Voicu (2010), even when region fixed effects are included in the specification. As an example, an increase of 10.0% in the density of service employment in a Romanian region makes the average Romanian region 11.9% more likely to attract a foreign investor.

As we can see, there are a variety of results that emphasize effects going more or less in the same direction but that remain difficult to compare (because authors usually estimate different specifications) and interpret (because of both the large number of possible effects and the possible presence of reverse causality).

These issues are even more important when studying the role of local labor markets in FDI as has been done in the literature. In particular, the impact of local labor costs has been investigated, but a significant concern is that authors are rarely able to control simultaneously for the local quality of labor. The labor cost per efficient unit of labor would be predicted by theory to influence location choices, but only the nominal cost is, in general, available. When labor efficiency is not taken into account, a positive impact of wages on the choice of a location may reflect the presence of high-skilled workers. Moreover, wages are simultaneously determined with firm location choices, and this endogeneity issue is usually not addressed. The endogeneity issue may be even more important when the local unemployment rates are introduced into the specification and microfoundations of the specification are even more unclear. A high local unemployment rate may reflect a large labor supply, and thus low wages or, on the contrary, wages that are too high and cause unemployment. Ultimately, owing to the lack of theoretical background for empirical specifications, we think that little can be learned from the impact of these variables. This is why we do not detail here their estimated effects, and we believe that a better use of theory will be required to really investigate the role of local labor markets.

5.6.2.3 Firm creation and entrepreneurship

Some recent literature argues that the location choices of new entrepreneurs and their determinants are worth studying because they should be more informative on the role and magnitude of agglomeration effects than the location choices of new plants by existing firms, as these choices are influenced by the locations of existing establishments of these firms. Unfortunately, as pointed out by [Glaeser et al. \(2010b\)](#), the literature on this topic is relatively small. Some contributions relate to the literature on innovations, and are surveyed in [Carlino and Kerr \(2015\)](#). We describe here some contributions that describe the determinants of firm creations in a more general way.

Among articles on the United States, [Rosenthal and Strange \(2003\)](#) show that firm creation is more important when the own-industry employment located within the first mile is larger, but the effect then vanishes rapidly with distance. Indeed, the impact within the first mile is 10–1000 times larger than the impact 2–5 miles away. They do not find any robust impact of urbanization on firm creation. [Glaeser and Kerr \(2009\)](#) propose disentangling among plant creations those that do not result from existing firms, as this is a better measure of entrepreneurial activity. The local level of activity appears to favor entrepreneurship, as it goes along with the presence of many small local suppliers. [Glaeser et al. \(2010a\)](#) find not that there are higher returns where entrepreneurs settle but that entrepreneurs rather choose places where there are larger local entrepreneurial

pools. Using the same dataset, and in the spirit of articles on determinants of local industrial employment, [Delgado et al. \(2010\)](#) augment the specification with dynamic effects and argue that mean reversion effects coexist with agglomeration gains.

Among contributions on other countries, [Figueiredo et al. \(2002\)](#) investigate the location choices of entrepreneurs in Portugal. Interestingly, they are able to distinguish between native and non-native entrepreneurs, and agglomeration effects are found only for non-natives. At a fine geographical scale, [Arauzo-Carod and Viladecans-Marsal \(2009\)](#) show for Spain that firm creation increases with own-industry previous entries. The effect is larger, the higher the technological level of the industry. Finally, [Harada \(2005\)](#) and [Sato et al. \(2012\)](#) find for Japan that a larger market size increases the willingness to become an entrepreneur, and that the effect is U shaped for the share of individuals that become entrepreneurs eventually. Put differently, people are more often entrepreneurs in both large and small locations. By contrast, [Addario and Vuri \(2010\)](#) find that population density reduces the probability of being an entrepreneur in Italy even if entrepreneurs' earnings are larger in denser areas.²¹

Overall, there is a great variety of results, which may be related to the estimation of different specifications and the way endogeneity issues are handled, especially as these issues are not always addressed. Still, once the burgeoning literature on location choices of entrepreneurs is better related to theory, and takes better into account spatial sorting and reverse causality, it should deliver interesting conclusions on the local determinants of entrepreneurship.

5.7. IDENTIFICATION OF AGGLOMERATION MECHANISMS

The literature assessing the effects of local determinants of agglomeration economies on local outcomes estimates the overall net impacts of local variables, but it does not enter the black box of the underlying mechanisms at stake. Some attempts to identify some of these mechanisms have been made recently in three directions. A series of articles focuses on job search and matching effects, and evaluates whether agglomeration effects on productivity are related to the way local labor markets operate. Other authors have taken an indirect route by testing whether industrial spatial concentration or firms co-location relates to industry characteristics associated with the Marshallian three broad families of agglomeration mechanisms: labor pooling, knowledge spillovers, and input–output linkages. Lastly, a couple of case studies have been proposed to quantify specific agglomeration effects.

²¹ There is also recent literature on developing countries (see [Ghani et al., 2013, 2014](#)).

5.7.1 Labor mobility, specialization, matching, and training

Some of the gains from agglomeration arise from an increase in job mobility and better matching between workers and firms. Some studies assess whether agglomeration increases the frequency of workers' moves between firms, industries, or occupations, as well as the chances for the unemployed of finding a job. [Freedman \(2008\)](#) studies the effect of specialization on workers' job mobility and earnings dynamics for the software publishing industry in one anonymous state using a US longitudinal matched employer–employee dataset. Higher specialization in a 25 km radius increases the chances of moving between two software jobs. A wage regression also shows that specialization within a 25 km radius lowers the initial wage but is also associated with a steeper wage profile leading to a wage premium.

Using the National Longitudinal Survey of Youth, [Wheeler \(2008\)](#) evaluates the effect of local population, density, and diversity on mobility between industries depending on the number of previous job moves. When looking at a sample of first job changes, he finds that industry changes occur more often in large and diverse local markets than in small and nondiversified ones. Once several jobs have been held, the positive relationship becomes negative. As workers in large markets also tend to experience fewer job changes overall, the evidence is consistent with agglomeration facilitating labor market matching. In a similar spirit, [Bleakley and Lin \(2012\)](#) study the effect of the metropolitan area employment density on occupation and industry changes using US data. They instrument current local density with historical local density and current density at the state level. The rate of transitions of occupation and industry is found to be lower in denser markets, but the result is reversed for younger workers, which is consistent with the interpretation of [Wheeler \(2008\)](#). The local employment share in the own industry or the own occupation also has a negative effect on industry and occupation changes.

The effects of agglomeration variables on the job search process is investigated by [Di Addario \(2011\)](#) for Italy. She estimates the effects of local population and specialization on the probabilities for nonemployed individuals of searching for a job and becoming employed. Agglomeration variables are instrumented with historical population, seismic hazard, and soil characteristics. Overall, the results show that a larger local population and location in an industrial district or superdistrict increase the probability of being employed. Conversely, the impact of any variable on search behavior is found to be zero.

Some authors have investigated whether matches between workers and firms are more productive in larger/denser areas. Some approaches used to evaluate the effect of matching on productivity in a static framework are discussed in [Section 5.2.3](#). In an application, [Wheeler \(2006\)](#) finds that wage growth is more important in large cities than in small ones and that this difference is mostly related to differences in wage growth when changing jobs. This is consistent with better matching in larger cities. However, this study does not take into account the endogeneity of job and location mobility.

This can be done using a more structural approach as explained in [Section 5.2.4](#). [Baum-Snow and Pavan \(2012\)](#) estimate a structural model and find that match quality contributes little to the observed city size premium, in comparison with other static and dynamic agglomeration effects. Differences in the conclusions may be due to differences in the structure of the static and dynamic models, and more specifically how the endogeneity of individual choices is handled.

Alternative static approaches have been proposed to assess the role of match quality. [Andersson et al. \(2007\)](#) use matched worker–firm panel data on California and Florida to estimate a wage equation involving worker and firm fixed effects. They then compute for each county the correlation across firms between the firm fixed effect and the average worker fixed effect within the firm. The correlation is regressed at the county level on the average firm fixed effect, average worker fixed effect, and density. The estimated coefficient of density is found to be positive and significant, indicating improved matching in denser areas. [Figueiredo et al. \(2014\)](#) evaluate the effect of density on matches between workers and firms using Portuguese employer–employee panel data. Their empirical strategy has two stages. First, they estimate a wage equation involving worker, firm, and match effects. Second, estimated match effects are regressed on explanatory variables including, in particular, density and specialization, as well as worker and firm fixed effects. The estimated effect of density in the second stage is not significant. The effect of specialization is significantly positive at the 10% level only. What remains unclear is to what extent the sole match effect captures all complementarity effects between workers and firms. Wage is expressed in logarithmic form in the first-stage specification, which means that the exponentiated product of worker and firm fixed effects also captures complementarities.

Finally, [Andini et al. \(2013\)](#) assess for Italy whether there is an effect of density (and classification into an industrial district) on worker and firm individual measures of labor pooling. Density is measured at the local labor market level, and is instrumented using historical values. The individual outcomes are the change of employer or type of work, or both, workplace learning, past experience, training by the firm, skill transferability, difficulty of replacing the worker or finding another job, measures of specialization, and the appropriateness of experience and education. The firm outcomes are the share of terminations that are voluntary, the share of vacancies filled from workers previously employed in the same industry, and the number of days needed to train key workers, a measure of appropriateness of a new worker in terms of education and experience. Overall, the results support theories of labor pooling, but the evidence is weak, possibly owing to the small size of the datasets. In particular, there is some evidence of a positive effect of agglomeration on turnover, on-the-job training, and improvement of job matches.

Another possible mechanism that might lead to higher productivity in cities is task specialization. The underlying idea is that there are benefits to the division of labor, and this division is limited by the extent of the market. The division of labor is then expected to be greater in larger markets. There are a few bits of research on the

relationship between the division of labor and city size. [Duranton and Jayet \(2011\)](#) study this relationship using information on more than 5 million workers in 454 occupations and 114 sectors extracted from the 1990 French census. It is shown that even after the uneven distribution of industries across cities has been taken into account, larger cities exhibit a larger share of workers in scarcer occupations. For example, the difference between Paris and the smallest French cities is around 70%. For Germany, [Kok \(2014\)](#) shows that the specialization of jobs and the required level of cognitive skills increase with city size. To our knowledge, the links between city size, the division of labor, and productivity have not yet been investigated.

Lastly, some authors have investigated whether knowledge spillovers arise from the mobility of workers between firms within the same local labor market. [Serafinelli \(2014\)](#) shows that in the region of Veneto, Italy, hiring a worker with experience at highly productive firms significantly increases the productivity of other firms. According to his results, worker flows explain around 15% of the productivity gains experienced by other firms when a new highly productive firm is added to a local labor market. [Combes and Duranton \(2006\)](#) propose a model in which firms choosing their location anticipate that they can improve their productivity by poaching workers from other firms. However, their workers can be poached too unless they are paid higher wages, which makes firms' production costs higher. Some authors have proposed testing this story indirectly by studying how training within firms varies with city size, the alternative to training being to poach workers who have already been trained from other firms. [Brunello and Gambarotto \(2007\)](#) for Italy, [Brunello and Paola \(2008\)](#) for the United Kingdom, and [Muehlemann and Wolter \(2011\)](#) for Switzerland show that indeed there is less on-the-job training in larger markets, and this is particularly true in the United Kingdom.

Overall, the literature on mobility, job search, and training comprises interesting attempts to determine the agglomeration mechanisms that relate to the labor market. It remains mostly descriptive though and would gain from considering approaches more grounded in theory.

5.7.2 Industrial spatial concentration and coagglomeration

Another strand of the literature has tried to identify the separate role of the three main types of mechanisms underlying agglomeration economies according to [Marshall \(1890\)](#): knowledge spillovers, labor pooling, and input–output linkages. For that purpose, a couple of articles augment the specifications of employment or firm creation presented in Section 5.6 with variables that should capture these three types of mechanisms. A larger number of articles, which we present first, compute spatial indices of concentration or coagglomeration for every industry, and then regress them on industry characteristics related to the three families of mechanisms. As analyses usually do not rely on a precise theoretical framework, this literature is for the moment mostly descriptive.

Kim (1995) was among the first to compute a spatial concentration index for some industries, in his case the Gini spatial concentration index (see Combes et al., 2008b), and regress it on industry characteristics and more particularly on average firm size. His purpose was to test the intuition that industries with stronger increasing returns to scale, which should be characterized by larger firms in equilibrium, are spatially more concentrated. The spatial concentration index is computed for a division of the United States into 9 large regions, for 20 industries, and for 5 points in time over the 1880–1987 period. The share of raw materials in production is introduced in the specification supposedly to control for the impact of comparative advantages on spatial concentration, and industry fixed effects are used to capture the role of industry effects that are constant over time.

There are major limitations to this kind of empirical strategy. Even simple economic geography models show that increasing returns to scale interact with trade costs and the degree of product differentiation to fix the degree of spatial concentration in equilibrium (see Combes et al., 2008b). However, only one industry characteristic among these three is introduced in the specification. It is thus necessary to make the strong assumption that either the two other characteristics are not correlated with the first one or they are sufficiently invariant over time to be captured by industry fixed effects. If trade costs and product differentiation indices were available, considering them in the specification would certainly not be straightforward since theoretical models usually predict highly nonlinear relationships between outcomes and underlying parameters. Introducing these characteristics as additional separate linear explanatory variables could be too extreme a simplification. Similarly, comparative advantage theory stresses the role of the interaction between factor intensity in the production function and regional factor endowments. Controlling for factor intensity but not for the distribution of endowments over space leads to ignoring the mechanism that generates regional specialization. Lastly, some mechanisms affecting spatial concentration, such as knowledge spillovers and labor pooling, are not taken into account either.

Further studies have tried to assess the role of additional agglomeration mechanisms by augmenting the estimated specification.²² The attempt by Rosenthal and Strange (2001) is an interesting one in this direction. The spatial concentration measure is the Ellison and Glaeser (1997) index computed for four-digit manufacturing industries in the United States. Variables for the three types of mechanisms are considered. Input sharing is measured by the shares of manufacturing and nonmanufacturing inputs in shipments. Knowledge spillovers are captured by innovations per dollar of shipment. Alternatively, some other authors also use R&D expenses. The measures of labor pooling are the value of shipments less the value of purchased inputs divided by the number of workers, the share of management workers, and the share of workers with at least a bachelor degree. These measures remain far from the intuition that industries with specific

²² They also use more detailed data, albeit on a shorter period of time.

needs for some labor skills gain more than others from concentrating. A number of other control variables are introduced, many of which relate to primary input use with the purpose of capturing again comparative advantage effects. As only cross-section data are available, industry fixed effects can be introduced only at the three-digit level and not at the four-digit level. The Ellison and Glaeser index takes into account in its construction an index of productive concentration that closely relates to the industry average plant size. Therefore, it is not clear whether or not one should control for firm size, and [Rosenthal and Strange \(2001\)](#) choose to leave it out of the specification.

The results obtained by [Rosenthal and Strange \(2001\)](#) are typical of this kind of study. Whereas labor pooling has a positive effect, knowledge spillovers have a positive impact on spatial concentration only when they are measured at a small scale (the zip code). Reliance on manufactured inputs affects agglomeration at the state level but not at a smaller scale. By contrast, reliance on service inputs has a negative effect on agglomeration at the state level. [Overman and Puga \(2010\)](#) propose an alternative indirect measure of labor market pooling. It is based on the assumption that a labor pool of workers with adequate skills allows firms to absorb productivity shocks more efficiently. Using UK establishment-level panel data, they construct an establishment-level measure of idiosyncratic employment shocks and average it across time and establishments within the industry. They find that industries that experience more volatility are more spatially concentrated.

Long ago, [Chinitz \(1961\)](#) suggested that examining the degree of coagglomeration of industries depending on their characteristics is another way to test for the presence of agglomeration economies. This approach is implemented in a systematic way by [Ellison et al. \(2010\)](#), who study the extent to which US manufacturing industries locate close to one another. The idea is to compute an index of coagglomeration between two industries and to regress it on measures of proximity between the two industries in terms of labor pooling, knowledge spillovers, and input–output linkages. Labor pooling is measured with the correlation of occupation shares between the two industries. Alternatively, some authors use a measure of distance between the distributions of these shares in the two industries. The share of input from the other industry and the share of output to the other industry are used as proxies for input and output linkages. Technological proximity is measured by two types of variables. The first type uses the shares of R&D flowing to and from the other industry. The second type uses patent citations of one industry made by the other industry. Such variables are, in general, not symmetrical. For instance, the first industry can cite the second industry more than the second industry cites the first industry. Therefore, it is the maximum value of the variable for the two industries that is used in the regressions.

Importantly, in order to control for comparative advantage effects, [Ellison et al. \(2010\)](#) introduce among the explanatory variables a coagglomeration index of spatial concentration due to natural advantages, which is an extension of the natural advantages spatial concentration index proposed by [Ellison and Glaeser \(1999\)](#). Results are also

provided for alternative coagglomeration indices. Indeed, a standard index such as the one of Ellison and Glaeser considers a classification of spatial units across which the economic activity is broken down and measures the concentration in these units. A limitation is that the relative location of units and the distances that separate them are not taken into account. As a result, the index is invariant up to any permutation of the units. For instance, it takes the same values if one relocates all units with large amounts of activity close to the center of the economy or if one locates them at the periphery. Alternative measures of spatial concentration and coagglomeration have been developed by [Duranton and Overman \(2005\)](#) to deal with this issue. They are based on the distribution of distances between establishments and can be computed for any spatial scope. One can assess whether there is concentration for a distance between establishments of 5 miles, 10 miles, and so on. [Ellison et al. \(2010\)](#) also estimate their specifications using the Duranton and Overman index computed for a distance of 250 miles. Finally, since explanatory variables are computed from the same quantities as the dependent variable, there might be endogeneity issues, and [Ellison et al. \(2010\)](#) propose instrumenting explanatory variables with similar variables constructed from UK data instead of US data.

The results give some support to the three types of agglomeration mechanisms. The largest effect is obtained for input–output linkages, followed by labor pooling. [Kolko \(2010\)](#) conducts a similar exercise for both manufacturing and service industries, using as additional measures of the links between industries variables related to the volume of interindustry trade. He studies both agglomeration and coagglomeration at various spatial scales: zip code, county, metropolitan area, and state. The limitations are that he does not use distance-based concentration indices such as the Duranton and Overman index, he does not control for spatial concentration due to natural advantages, and he does not deal with endogeneity issues using instrumentation. Ultimately, trade between industries appears to be the main driver of industry coagglomeration for both manufacturing and services. More precisely, service industries that trade with each other are more likely to collocate in the same zip-code area, although not in the same county or state; by contrast, manufacturing industries that trade with each other are more likely to collocate in the same county or state but not in the same zip-code area. Input sharing also positively affects coagglomeration for both manufacturing and services at any spatial level, and this is true for occupational similarity to some extent as a positive effect is found but only for services and at the zip-code level. As regards spatial concentration, labor pooling is the only variable having a significant impact. Its effect is positive but occurs in the manufacturing sector only.

[Kerr and Kominers \(2015\)](#) further study the determinants of spatial concentration in the spirit of [Ellison et al. \(2010\)](#). They compute the Duranton and Overman spatial concentration index for different industries and different distances. Values are pooled together and then regressed on dummies for distances interacting with an industry measure of knowledge spillovers, and then alternatively an industry measure of labor pooling.

The proxies used for these determinants are slightly different from those in other studies. As regards knowledge spillovers, [Kerr and Kominers \(2015\)](#) consider the citation premium for 0–10 miles relative to 30–150 miles. Labor pooling is captured by a Herfindahl index of occupational concentration computed over 700 categories. Most estimated coefficients obtained for interactions with dummies for distances decrease with distance, and they are significantly different from zero for short distances only. This suggests that establishments in industries with shorter knowledge spillovers or more labor pooling are more concentrated. Similar results are obtained whether one uses US data or UK data to compute measures of knowledge spillovers and labor pooling. Nevertheless, estimations for these two channels of agglomeration economies are conducted separately without confronting them in a single regression. Finally, estimated coefficients for interactions between dummies for distances and dependency on natural advantages tend to increase with distance and are significant for large enough distances only. This is consistent with the intuition that industries more dependent on natural advantages are more dispersed.

A difficulty faced by this literature is that the dependent variable is a complex function of certain quantities, such as local industrial employment, which relate to the quantities describing firms and establishments within the industry that are used in the construction of explanatory variables. Therefore, it is not easy to argue about expected effects of explanatory variables in equilibrium, and this makes interpretations difficult. In light of this difficulty, [Dumais et al. \(1997\)](#) in a section not included in [Dumais et al. \(2002\)](#) propose re-examining the literature on industrial employment in order to assess the role of some specific agglomeration channels. They consider a specification where local industrial employment is used as the dependent variable instead of an index of spatial concentration in the industry. Proxies for Marshallian externalities are constructed at the local level using the following strategy. Measures of proximity between industries as regards knowledge spillovers, labor pooling, and input and output linkages are computed at the national level. For a given type of agglomeration channel, the local variable for an industry is then computed as the sum over all other industries in their proximity weighted by the share of these industries in the location. These local variables are also sometimes interacted with some of the local determinants of industrial employment presented in [Section 5.6.1](#). All these terms serve as explanatory variables in the specification of local industrial employment.

Recently, a similar strategy has been implemented by [Jofre-Montseny et al. \(2011\)](#) to determine the effects of the different types of agglomeration economies on the location of new firms in Spain at the municipality level and city level.²³ In the same vein, [Jofre-Montseny et al. \(2014\)](#) estimate from Spanish data, for each industry separately, a firm location model with two main local explanatory variables, local employment within the industry and in other industries. The industry-specific estimates for these

²³ Articles using the same strategy but for the study of agglomeration economies on TFP include those of [Rigby and Essletzbichler \(2002\)](#), [Baldwin et al. \(2010\)](#), [Drucker and Feser \(2012\)](#), and [Ehrl \(2013\)](#).

two variables are then regressed on industry characteristics with proxies for knowledge spillovers, labor pooling, input sharing, and energy and primary input use. We emphasized above the difficulty in interpreting estimates of employment growth specifications, while [Jofre-Montseny et al. \(2014\)](#) propose further extending these specifications by introducing interactions between local determinants and factors influencing the different agglomeration forces at the industry level. Such extended empirical frameworks are necessarily even more ambiguous and difficult to interpret than the basic employment growth specifications that we discussed in [Section 5.6.1](#).

Overall, this strand of literature is an interesting effort to identify the mechanisms underlying agglomeration economies. Ultimately though, it is very difficult to give a clear interpretation of the results, and the conclusions are mostly descriptive. This is due to the weak links between estimated specifications and theoretical models. Another concern is whether the right measure of concentration or coagglomeration has been chosen. The exact properties of concentration indices, even measures à la [Duranton and Overman \(2005\)](#), still need to be established. Moreover, one needs to assume that industry characteristics used as explanatory variables really capture the mechanisms they are meant to, and have additive linear effects, whereas this is not certain. For instance, according to theory, two industries sharing inputs have more incentive to colocate when trade costs for these inputs are large. In that perspective, variables capturing input–output linkages should be caused to interact with a measure of trade costs, but this is not done in the literature. Finally, there are probably some endogeneity issues since the dependent variable and the explanatory variables are usually computed from the same quantities. However, the presence and channels of endogeneity are difficult to assess, and it is hard to conclude that some instruments are valid, as estimated specifications have usually not been derived from any precise theoretical framework. On the other hand, since the overall impact of agglomeration on productivity can be evaluated with reasonable confidence nowadays as we emphasized in previous sections, we think that investigating the relative magnitude of agglomeration channels is an important and promising avenue for future research. The descriptive evidence presented in this subsection could be used to build theoretical models from which specifications could be derived, allowing the identification of agglomeration channels and strategies to tackle endogeneity concerns. Structural approaches applied to case studies, which are presented in the next subsection, constitute some first steps in that direction.

5.7.3 Case studies

Some specific mechanisms of agglomeration economies can be assessed through case studies of firms or industries for which the nature of possible density effects are known and can be specified.

An interesting structural attempt to evaluate the importance of agglomeration economies in distribution costs is proposed by [Holmes \(2011\)](#). The study focuses on the

diffusion of Wal-Mart across the US territory and considers the location and timing of the opening of new stores. These new stores may sell general merchandise and, if they are supercenters, they may also sell food. When operating a store, Wal-Mart gets merchandise sales revenues but incurs costs that include not only wages, rent, and equipment costs, but also fixed costs. These fixed costs depend on the local population density as well as the distance to the nearest distribution center for general merchandise and, possibly, the distance to the nearest food distribution center. Higher store density usually goes along with shorter distance from distribution centers. When opening a new store, Wal-Mart faces a trade-off between savings from a shorter distance to distribution centers and cannibalization of existing stores. The estimation strategy to assess the effects of population density and proximity to distribution centers is the following. The choice of consumers across shops is modeled and demand parameters are estimated by fitting the predicted merchandise and food revenues with those observed in the data. An intertemporal specification of the Wal-Mart profit function taking into account the location of shops is then considered. In particular, this function depends on revenues net of costs, which include wages, rent, and equipment costs as well as fixed costs. For a given location of shops, net revenues can be derived from the specification of demand, where parameters have been replaced by their first-stage estimators. To estimate parameters related to fixed costs, [Holmes \(2011\)](#) then considers the actual Wal-Mart choices for store openings as well as deviations in which the opening dates of pairs of stores are reordered. Profit derived for an actual choice of store openings must be at least equal to that of deviations. This gives a set of inequalities that can be brought to the data in order to estimate bounds for the effects of population density and distance to distribution centers. It is estimated that when a Wal-Mart store is closer by 1 mile to a distribution center, the company enjoys a yearly benefit that lies in a tight interval around \$3500. This constitutes a measure of the benefits of store density.

The benefits from economies of density in agriculture related to the use of neighboring land parcels are evaluated by [Holmes and Lee \(2012\)](#). When using a particular piece of equipment, a farmer can save on setup costs by using it across many fields located close to each other. Moreover, if a farmer has knowledge of a specific crop, it is worth planting that crop in adjacent fields, although this may be at the expense of reducing the crop diversity that can be useful against risks. The analysis is conducted on planting decisions in the Red River Valley region of North Dakota, for which there are a variety of crops and years of data on crop choice collected by satellites. More precisely, the focus is on quarter sections which are 160-acre square parcels. These sections can be divided into quarters of 40 acres, each designed as a field. The empirical strategy relies on a structural model where farmers maximize their intertemporal profit on the four quarters of their parcels, choosing for each quarter the extent to which they cultivate a given crop (rather than alternative ones). Production depends on soil quality and the quantity of investment in a particular kind of equipment useful to cultivate the specific crop but which has a cost. It is possible to show that because of economies of density arising from the use of the

specific piece of equipment on all quarters, the optimal cultivation level for a crop on a quarter depends not only on the soil quality of this quarter but also on that of the other quarters. The specification can be estimated and parameters can be used to assess the importance of economies of density. Results show that there is a strong link between quarters of the same parcel. If economies of density were removed, the long-run planting level of a particular crop would fall by around 40%. Two-thirds of the actual level of crop specialization can be attributed to natural advantages and one-third can be attributed to economies of density.

5.8. CONCLUSION

Most of the literature identifies the overall impact of local determinants of agglomeration economies, but not the role of specific mechanisms that generate agglomeration effects. This is already a crucial element when assessing the role of cities. Major progress has been made in dealing with spatial sorting of workers and firms as well as endogeneity issues due to missing variables and reverse causality, especially when assessing the effect of density on productivity.

We developed a consistent framework that encompasses both the early attempts to estimate agglomeration effects using aggregate regional data and more sophisticated strategies using individual data, recently including some structural approaches. This allowed us to discuss most empirical issues and the solutions that have been proposed in the literature. We also presented the attempts to study the determinants of other local outcomes—namely, employment and firm location choices—but more investigations are still needed. For instance, further theoretical and empirical clarifications would be useful when studying the determinants of local employment in order to better disentangle the short-term dynamics from long-term effects, and the respective role of labor demand and supply. The determinants of firm location choices have benefited so far from a very limited treatment of selection and endogeneity issues. Surprisingly, the impact of agglomeration economies on unemployment has received little attention and deserves more work at least from a European perspective as regional disparities in unemployment rates there remain large. Finally, identifying the channels of agglomeration economies is also clearly important, but the related literature remains limited except for some contributions on innovation that are surveyed in [Carlino and Kerr \(2015\)](#). Meaningful strategies relying on sound theoretical ground to provide an empirical assessment of channels of agglomeration economies are still needed, and current evidence while being interesting is rather descriptive.

Some researchers have started to investigate routes complementary to those mentioned in this chapter. First, the existence of a spatial equilibrium implies that agglomeration costs are a necessary counterpart of agglomeration gains. This prediction is

supported by [Gibbons et al. \(2011\)](#), who show that in Great Britain there is an almost one-for-one relationship between local housing costs and nominal earnings, which are higher in larger cities, once the effects of housing quality and workers skills are taken into account. Second, some authors have gone a step further by looking at the implications in terms of welfare of the simultaneous presence of agglomeration costs and gains. However, some effects have not yet been considered in the analyses, whereas they have some importance from a policy perspective. For instance, considering how city size affects environmental concerns or road congestion costs is important for designing urban policies that improve welfare.

There have been only a few early independent attempts to evaluate agglomeration costs, and they are for developing countries only ([Thomas, 1980](#); [Richardson, 1987](#); [Henderson, 2002](#)). Recently, housing and land prices have started to be investigated more systematically, although articles usually rely for their analyses on datasets that are not comprehensive. There are a few rare exceptions, such as [Davis and Heathcote \(2007\)](#) and [Davis and Palumbo \(2008\)](#) on the whole United States, or [Combes et al. \(2012a\)](#) on the determinants of land prices in French urban areas. This last article estimates the elasticity of land prices with respect to city population, from which the elasticity of urban costs is recovered. Its magnitude is found to be similar to that of the elasticity of agglomeration gains on productivity. [Albouy and Ehrlich \(2013\)](#) replicate the approach to investigate the determinants of land prices in US metropolitan areas. Finally, some authors have tried to exploit natural or controlled experiments, such as [Rossi-Hansberg et al. \(2010\)](#), who use residential urban revitalization programs implemented in Richmond, Virginia, to evaluate the effect of housing externalities on land value.

Housing is not the only good whose price varies across locations, but little is known for other types of goods. Using barcode data on purchase transactions, [Handbury and Weinstein \(2015\)](#) and [Handbury \(2013\)](#) assess how prices of grocery products vary with city size. [Handbury and Weinstein \(2015\)](#) find that raw price indices slightly increase with city size, and this would constitute an additional source of agglomeration costs for households. However, this result is obtained before correcting prices for quality differences across varieties and before taking into account effects related to preferences for diversity that are present when considering CES utility functions. Once these are taken into account, price indices decrease with city size. This is the typical agglomeration gain that can be found in economic geography models with mobile workers à la [Krugman \(1991b\)](#). The price index decrease is due mostly to a much larger number of available varieties in larger cities, but is also due to a higher quality of varieties sold there. [Handbury \(2013\)](#) allows preferences to differ between rich and poor households, and obtains the further result that the price index decreases with city size only for rich households but increases for poor ones. Clearly, investigating further these types of agglomeration effects is high on the agenda.

Lastly, since there is evidence that gains and costs from agglomeration as well as location choices differ across types of workers, there is a need to consistently reintroduce space in welfare analyses when one wishes to assess individual or household inequalities. [Moretti \(2013\)](#) shows that real wage disparities between skilled and unskilled workers have increased less over the last 30 years than what nominal wage disparities would suggest, once the increase in the propensity of skilled workers compared with unskilled workers to live in larger cities has been taken into account. Indeed, the increase in the difference in housing costs between skilled and unskilled workers represents up to 30% of the increase in the difference in nominal wages. [Albouy et al. \(2013\)](#) show that Canadian cities with the highest real wage differ for English speakers and French speakers.

However, this type of real wage computation does not consider differences in amenity endowments across cities and possible differences in the valuation of amenities across worker groups. As workers are mobile, differences in real wages across locations should reflect to some extent differences in amenity value (see [Roback, 1982](#)). [Albouy et al. \(2013\)](#) show that indeed the real wage they compute for Canadian cities is slightly correlated with arts and climate city ratings. For the United States, [Albouy \(2008\)](#) and [Albouy \(2009\)](#) find that the most valuable cities have coastal proximity, sunshine, and mild seasons. These findings are in line with those of [Desmet and Rossi-Hansberg \(2013\)](#), who use a slightly more general model calibrated on US data to assess the welfare impact of eliminating differences in amenities or frictions (within-city commuting time, local taxes, government expenditure) between cities. [Diamond \(2013\)](#) takes into account workers' heterogeneity and shows that the increased skill sorting in the United States is partly due to the endogenous increase in amenities within high-skill cities.

Some recent theoretical contributions such as those of [Behrens et al. \(2014\)](#), [Eeckhout et al. \(2014\)](#), and [Behrens and Robert-Nicoud \(2014\)](#) suggest that sorting and disparities are worth studying simultaneously within and between cities. [Glaeser et al. \(2009\)](#) and [Combes et al. \(2012c\)](#) show that indeed larger cities present larger dispersions of wages and skills, respectively, in the United States and France. [Baum-Snow and Pavan \(2013\)](#) further document the emergence of both within-city and between-city inequalities in wages and skills in the United States. A full empirical welfare assessment of both within-city and between-city disparities considering agglomeration costs and benefits, heterogeneous workers that are imperfectly mobile, and amenity data in addition to productivity measures as well as land and housing prices is a challenge for future research.

ACKNOWLEDGMENTS

We are grateful to Gilles Duranton, Vernon Henderson, Jeffrey Lin, Steve Ross, and William Strange, as well as participants at the handbook conference at the Wharton School of the University of Philadelphia for useful comments and discussion. Financial support from the Agence Nationale de la Recherche in France, Grants ANR-11-BSH1-0014 and ANR-12-GLOB-0005, is gratefully acknowledged.

REFERENCES

- Abel, J.R., Dey, I., Gabe, T.M., 2012. Productivity and the density of human capital. *J. Reg. Sci.* 52, 562–586.
- Abowd, J.M., Kramarz, F., Margolis, D.N., 1999. High wage workers and high wage firms. *Econometrica* 67, 251–333.
- Addario, S.D., Vuri, D., 2010. Entrepreneurship and market size. The case of young college graduates in Italy. *Labour Econ.* 17, 848–858.
- Ahlfeldt, G., Redding, S., Sturm, D., Wolf, N., 2012. The economics of density: evidence from the Berlin-Wall. CEP Discussion Papers 1154.
- Albouy, D., 2008. Are big cities really bad places to live? Improving quality-of-life estimates across cities. Working paper 14472, National Bureau of Economic Research.
- Albouy, D., 2009. What are cities worth? Land rents, local productivity, and the capitalization of amenity values. Working paper 14981. Revised 2014, National Bureau of Economic Research.
- Albouy, D., Ehrlich, G., 2013. The distribution of urban land values: evidence from market transactions. Mimeograph, University of Illinois.
- Albouy, D., Leibovici, F., Warman, C., 2013. Quality of life, firm productivity, and the value of amenities across Canadian cities. *Can. J. Econ.* 46, 379–411.
- Amiti, M., Cameron, L., 2007. Economic geography and wages. *Rev. Econ. Stat.* 89, 15–29.
- Ananat, E., Fu, S., Ross, S.L., 2013. Race-specific agglomeration economies: social distance and the black-white wage gap. Working paper 18933, National Bureau of Economic Research.
- Andersson, F., Burgess, S., Lane, J.I., 2007. Cities, matching and the productivity gains of agglomeration. *J. Urban Econ.* 61, 112–128.
- Andersson, M., Klaesson, J., Larsson, J.P., 2015. The sources of the urban wage premium by worker skills: spatial sorting or agglomeration economies? *Pap. Reg. Sci.*, forthcoming.
- Andini, M., de Blasio, G., Duranton, G., Strange, W., 2013. Marshallian labour market pooling: evidence from Italy. *Reg. Sci. Urban Econ.* 43, 1008–1022.
- Arauzo-Carod, J.M., Viladecans-Marsal, E., 2009. Industrial location at the intrametropolitan level: the role of agglomeration economies. *Reg. Stud.* 43, 545–558.
- Arellano, M., Bond, S., 1991. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev. Econ. Stud.* 58, 277–297.
- Arzaghi, M., Henderson, J.V., 2008. Networking off Madison Avenue. *Rev. Econ. Stud.* 75, 1011–1038.
- Au, C., Henderson, J., 2006a. How migration restrictions limit agglomeration and productivity in China. *J. Dev. Econ.* 80, 350–388.
- Au, C.C., Henderson, V., 2006b. Are Chinese cities too small? *Rev. Econ. Stud.* 73, 549–576.
- Bacolod, M., Blum, B.S., Strange, W.C., 2009a. Skills in the city. *J. Urban Econ.* 65, 136–153.
- Bacolod, M., Blum, B.S., Strange, W.C., 2009b. Urban interactions: soft skills versus specialization. *J. Econ. Geogr.* 9, 227–262.
- Bacolod, M., Blum, B.S., Strange, W.C., 2010. Elements of skills: traits intelligences, education, and agglomeration. *J. Reg. Sci.* 50, 245–280.
- Bai, J., 2009. Panel data models with interactive fixed effects. *Econometrica* 77, 1229–1279.
- Baldwin, J.R., Brown, W.M., Rigby, D.L., 2010. Agglomeration economies: microdata panel estimates from Canadian manufacturing. *J. Reg. Sci.* 50, 915–934.
- Barrios, S., Görg, H., Strobl, E., 2006. Multinationals' location choice, agglomeration economies, and public incentives. *Int. Reg. Sci. Rev.* 29, 81–107.
- Basile, R., 2004. Acquisition versus greenfield investment: the location of foreign manufacturers in Italy. *Reg. Sci. Urban Econ.* 34, 3–25.
- Basile, R., Castellani, D., Zanfei, A., 2008. Location choices of multinational firms in Europe: the role of EU cohesion policy. *J. Int. Econ.* 74, 328–340.
- Baum-Snow, N., Ferreira, F., 2015. Causal inference in urban economics. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, vol. 5A. North-Holland, Amsterdam.

- Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage gap. *Rev. Econ. Stud.* 79, 88–127.
- Baum-Snow, N., Pavan, R., 2013. Inequality and city size. *Rev. Econ. Stat.* 93, 1535–1548.
- Beaudry, P., Green, D.A., Sand, B., 2014. Spatial equilibrium with unemployment and wage bargaining: theory and estimation. *J. Urban Econ.* 79, 2–19.
- Behrens, K., Robert-Nicoud, F., 2014. Survival of the fittest in cities: urbanisation and inequality. *Econ. J.* 12 (581), 1371–1400.
- Behrens, K., Duranton, G., Robert-Nicoud, F., 2014. Productive cities: sorting, selection, and agglomeration. *J. Polit. Econ.* 122, 507–553.
- Bleakley, H., Lin, J., 2012. Thick-market effects and churning in the labor market: evidence from US cities. *J. Urban Econ.* 72, 87–103.
- Blien, U., Suedekum, J., 2005. Local economic structure and industry development in Germany, 1993–2001. *Econ. Bull.* 17, 1–8.
- Blien, U., Suedekum, J., Wolf, K., 2006. Productivity and the density of economic activity. *Labour Econ.* 13, 445–458.
- Bosker, M., Brakman, S., Garretsen, H., Schramm, M., 2010. Adding geography to the new economic geography: bridging the gap between theory and empirics. *J. Econ. Geogr.* 10, 793–823.
- Brakman, S., Garretsen, H., Schramm, M., 2004. The spatial distribution of wages: estimating the Helpman-Hanson model for Germany. *J. Reg. Sci.* 44, 437–466.
- Brakman, S., Garretsen, H., Schramm, M., 2006. Putting new economic geography to the test: free-ness of trade and agglomeration in the EU regions. *Reg. Sci. Urban Econ.* 36, 613–635.
- Brakman, S., Garretsen, H., Van Marrewijk, C., 2009. Economic geography within and between European nations: the role of market potential and density across space and time. *J. Reg. Sci.* 49, 777–800.
- Breinlich, H., 2006. The spatial income structure in the European Union—what role for economic geography? *J. Econ. Geogr.* 6, 593–617.
- Briant, A., Combes, P.P., Lafourcade, M., 2010. Does the size and shape of geographical units jeopardize economic geography estimations? *J. Urban Econ.* 67, 287–302.
- Brühlhart, M., Mathys, N.A., 2008. Sectoral agglomeration economies in a panel of European regions. *Reg. Sci. Urban Econ.* 38, 348–362.
- Brunello, G., Gambarotto, F., 2007. Do spatial agglomeration and local labor market competition affect employer-provided training? Evidence from the UK. *Reg. Sci. Urban Econ.* 37, 1–21.
- Brunello, G., Paola, M.D., 2008. Training and economic density: some evidence from Italian provinces. *Labour Econ.* 15, 118–140.
- Buchanan, J.M., 1965. An economic theory of clubs. *Economica* 32, 1–14.
- Carlino, G., Kerr, W., 2015. Agglomeration and innovation. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, vol. 5A. North-Holland, Amsterdam.
- Carlsen, F., Rattsø, J., Stokke, H., 2013. Education, experience and dynamic urban wage premium. Department of Economics Working paper 142013, Norwegian University of Science and Technology.
- Carlton, D., 1983. The location and employment choices of new firms: an econometric model with discrete and continuous endogenous variables. *Rev. Econ. Stat.* 65, 440–449.
- Chauvin, J.P., Glaeser, E., Tobio, K., 2014. *Urban Economics in the US and India*. Harvard University.
- Chinitz, B., 1961. Contrasts in agglomeration: New-York and Pittsburgh. *Am. Econ. Rev.* 51, 279–289.
- Ciccone, A., 2002. Agglomeration effects in Europe. *Eur. Econ. Rev.* 46, 213–227.
- Ciccone, A., Hall, R.E., 1996. Productivity and the density of economic activity. *Am. Econ. Rev.* 86, 54–70.
- Ciccone, A., Peri, G., 2006. Identifying human capital externalities: theory with an application to US cities. *Rev. Econ. Stud.* 73, 381–412.
- Ciéslík, A., 2005. Regional characteristics and the location of foreign firms within Poland. *Appl. Econ.* 37, 863–874.
- Cingano, F., Schivardi, F., 2004. Identifying the sources of local productivity growth. *J. Eur. Econ. Assoc.* 2, 720–742.
- Combes, P.P., 2000. Economic structure and local growth: France, 1984–1993. *J. Urban Econ.* 47, 329–355.
- Combes, P.P., 2011. The empirics of economic geography: how to draw policy implications? *Rev. World Econ.* 147, 567–592.

- Combes, P.P., Duranton, G., 2006. Labour pooling, labour poaching, and spatial clustering. *Reg. Sci. Urban Econ.* 36, 1–28.
- Combes, P.P., Lafourcade, M., 2005. Transport costs: measures, determinants, and regional policy implications for France. *J. Econ. Geogr.* 5, 319–349.
- Combes, P.P., Lafourcade, M., 2011. Competition, market access and economic geography: structural estimation and predictions for France. *Reg. Sci. Urban Econ.* 41, 508–524.
- Combes, P.P., Magnac, T., Robin, J.M., 2004. The dynamics of local employment in France. *J. Urban Econ.* 56, 217–243.
- Combes, P.P., Duranton, G., Gobillon, L., 2008a. Spatial wage disparities: sorting matters! *J. Urban Econ.* 63, 723–742.
- Combes, P.P., Mayer, T., Thisse, J.F., 2008b. *Economic Geography: The Integration of Regions and Nations*. Princeton University Press, New Jersey.
- Combes, P.P., Duranton, G., Gobillon, L., Roux, S., 2010. Estimating agglomeration effects with history, geology, and worker fixed-effects. In: Glaeser, E.L. (Ed.), *Agglomeration Economics*. Chicago University Press, Chicago, IL, pp. 15–65.
- Combes, P.P., Duranton, G., Gobillon, L., 2011. The identification of agglomeration economies. *J. Econ. Geogr.* 11, 253–266.
- Combes, P.P., Duranton, G., Gobillon, L., 2012a. The costs of agglomeration: land prices in French cities. Discussion Paper 9240, Centre for Economic Policy Research.
- Combes, P.P., Duranton, G., Gobillon, L., Puga, D., Roux, S., 2012b. The productivity advantages of large markets: distinguishing agglomeration from firm selection. *Econometrica* 80, 2543–2594.
- Combes, P.P., Duranton, G., Gobillon, L., Roux, S., 2012c. Sorting and local wage and skill distributions in France. *Reg. Sci. Urban Econ.* 42, 913–930.
- Combes, P.P., Démurger, S., Li, S., 2013. Urbanisation and migration externalities in China. Discussion Paper 9352, Centre for Economic Policy Research.
- Corrado, L., Fingleton, B., 2012. Where is the economics in spatial econometrics? *J. Reg. Sci.* 52, 210–239.
- Crozet, M., Mayer, T., Mucchielli, J.L., 2004. How do firms agglomerate? A study of FDI in France. *Reg. Sci. Urban Econ.* 34, 27–54.
- Davis, M.A., Heathcote, J., 2007. The price and quantity of residential land in the United States. *J. Monet. Econ.* 54, 2595–2620.
- Davis, M.A., Palumbo, M.G., 2008. The price of residential land in large US cities. *J. Urban Econ.* 63, 352–384.
- D’Costa, S., Overman, H., 2014. The urban wage growth premium: sorting or learning? *Reg. Sci. Urban Econ.* 48, 168–179.
- de la Roca, J., Puga, D., 2012. Learning by working in big cities. Discussion Paper 9243, Centre for Economic Policy Research.
- Delgado, M., Porter, M.E., Stern, S., 2010. Clusters and entrepreneurship. *J. Econ. Geogr.* 10, 495–518.
- Desmet, K., Fafchamps, M., 2005. Changes in the spatial concentration of employment across US counties: a sectoral analysis 1972–2000. *J. Econ. Geogr.* 5, 261–284.
- Desmet, K., Rossi-Hansberg, E., 2013. Urban accounting and welfare. *Am. Econ. Rev.* 103, 2296–2327.
- Devereux, M.P., Griffith, R., Simpson, H., 2007. Firm location decisions, regional grants and agglomeration externalities. *J. Public Econ.* 91, 413–435.
- Di Addario, S., 2011. Job search in thick markets. *J. Urban Econ.* 69, 303–318.
- Di Addario, S., Patacchini, E., 2008. Wages and the city. Evidence from Italy. *Labour Econ.* 15, 1040–1061.
- Diamond, R., 2013. The determinants and welfare implications of US workers’ diverging location choices by skill: 1980–2000. Stanford University, Mimeograph.
- Di Giacinto, V., Gomellini, M., Micucci, G., Pagnini, M., 2014. Mapping local productivity advantages in Italy: industrial districts, cities or both? *J. Econ. Geogr.* 2, 365–394.
- Drucker, J., Feser, E., 2012. Regional industrial structure and agglomeration economies: an analysis of productivity in three manufacturing industries. *Reg. Sci. Urban Econ.* 42, 1–14.
- Dumais, G., Ellison, G., Glaeser, E.L., 1997. Geographic concentration as a dynamic process. Working paper 6270, National Bureau of Economic Research.

- Dumais, G., Ellison, G., Glaeser, E.L., 2002. Geographic concentration as a dynamic process. *Rev. Econ. Stat.* 84, 193–204.
- Duranton, G., 2014. Agglomeration effects in Colombia. Mimeograph, Wharton University.
- Duranton, G., Jayet, H., 2011. Is the division of labour limited by the extent of the market? Evidence from French cities. *J. Urban Econ.* 69, 56–71.
- Duranton, G., Monastiriotis, V., 2002. Mind the gaps: the evolution of regional earnings inequalities in the UK 1982–1997. *J. Reg. Sci.* 42, 219–256.
- Duranton, G., Overman, H.G., 2005. Testing for localization using microgeographic data. *Rev. Econ. Stud.* 72, 1077–1106.
- Duranton, G., Puga, D., 2001. Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am. Econ. Rev.* 91, 1454–1477.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam, pp. 2063–2117.
- Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. *J. Polit. Econ.* 122, 554–620.
- Ehrl, P., 2013. Agglomeration economies with consistent productivity estimates. *Reg. Sci. Urban Econ.* 43, 751–763.
- Ellison, G., Glaeser, E.L., 1997. Geographic concentration in US manufacturing industries: a dartboard approach. *J. Polit. Econ.* 105, 889–927.
- Ellison, G., Glaeser, E.L., 1999. The geographic concentration of industry: does natural advantage explain agglomeration? *Am. Econ. Rev. Pap. Proc.* 89, 311–316.
- Ellison, G., Glaeser, E.L., Kerr, W.R., 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. *Am. Econ. Rev.* 100, 1195–1213.
- Elvery, J.A., 2010. City size and skill intensity. *Reg. Sci. Urban Econ.* 40, 367–379.
- Faberman, J., Freedman, M., 2013. The urban density premium across establishments. Working paper 2013-01, Federal Reserve Bank of Chicago.
- Fallah, B.N., Partridge, M.D., Olfert, M.R., 2011. New economic geography and US metropolitan wage inequality. *J. Econ. Geogr.* 46, 865–895.
- Fally, T., Paillacar, R., Terra, C., 2010. Economic geography and wages in Brazil: evidence from micro-data. *J. Dev. Econ.* 91, 155–168.
- Figueiredo, O., Guimarães, P., Woodward, D., 2002. Home-field advantage: location decisions of Portuguese entrepreneurs. *J. Urban Econ.* 52, 341–361.
- Figueiredo, O., Guimarães, P., Woodward, D., 2014. Firm-worker matching in industrial clusters. *J. Econ. Geogr.* 14, 1–19.
- Foster, N., Stehrer, R., 2009. Sectoral productivity, density and agglomeration in the Wider Europe. *Spat. Econ. Anal.* 4, 427–446.
- Freedman, M., 2008. Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *J. Urban Econ.* 64, 590–600.
- Fu, S., 2007. Smart café cities: testing human capital externalities in the Boston metropolitan area. *J. Urban Econ.* 61, 86–111.
- Fu, S., Ross, S.L., 2013. Wage premia in employment clusters: how important is worker heterogeneity? *J. Labor Econ.* 31, 271–304.
- Fuchs, M., 2011. The determinants of local employment dynamics in Western Germany. *Empir. Econ.* 40, 177–203.
- Fujita, M., Krugman, P.R., Venables, A.J., 1999. *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press, Cambridge, MA.
- Ghani, E., Kerr, W., O'Connell, S., 2013. Local industrial structures and female entrepreneurship in India. *J. Econ. Geogr.* 13, 929–964.
- Ghani, E., Kerr, W., O'Connell, S., 2014. Determinants of entrepreneurship in India. *Reg. Stud.* 48, 1071–1089.
- Gibbons, S., Overman, H.G., 2012. Mostly pointless spatial econometrics. *J. Reg. Sci.* 52, 172–191.
- Gibbons, S., Overman, H.G., Resende, G., 2011. Real earnings disparities in Britain. Discussion Paper 65, Spatial Economic Research Center.

- Gibbons, S., Overman, H.G., Patacchini, E., 2015. Spatial methods. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, vol. 5A. North-Holland, Amsterdam.
- Glaeser, E.L., Kerr, W.R., 2009. Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? *J. Econ. Manag. Strateg.* 18, 623–663.
- Glaeser, E.L., Maré, D.C., 2001. Cities and skills. *J. Labor Econ.* 19, 316–342.
- Glaeser, E.L., Kallal, H., Scheinkman, J.A., Schleifer, A., 1992. Growth in cities. *J. Polit. Econ.* 100, 1126–1152.
- Glaeser, E.L., Resseger, M., Tobio, K., 2009. Inequality in cities. *J. Reg. Sci.* 49, 617–646.
- Glaeser, E.L., Kerr, W.R., Ponzetto, G.A.M., 2010a. Clusters of entrepreneurship. *J. Urban Econ.* 67, 150–168.
- Glaeser, E.L., Rosenthal, S.S., Strange, W.C., 2010b. Urban economics and entrepreneurship. *J. Urban Econ.* 67, 1–14.
- Gould, E., 2007. Cities, workers, and wages: a structural analysis of the urban wage premium. *Rev. Econ. Stud.* 74, 477–506.
- Graham, D.J., 2007. Variable returns to agglomeration and the effect of road traffic congestion. *J. Urban Econ.* 62, 103–120.
- Graham, D.J., 2009. Identifying urbanisation and localisation externalities in manufacturing and service industries. *Pap. Reg. Sci.* 88, 63–84.
- Graham, D.J., Melo, P.S., Jiwattanakulpaisarn, P., Noland, R.B., 2010. Testing for causality between productivity and agglomeration economies. *J. Reg. Sci.* 50, 935–951.
- Greenstone, M., Hornbeck, R., Moretti, E., 2010. Identifying agglomeration spillovers: evidence from winners and losers of large plants openings. *J. Polit. Econ.* 118, 536–598.
- Groot, S.P.T., de Groot, H.L.F., 2014. Estimating the skill bias in agglomeration externalities and social returns to education: evidence from Dutch matched worker-firm micro-data. Discussion Paper 2014-088, Tinbergen Institute.
- Groot, S.P.T., de Groot, H.L.F., Smit, M.J., 2014. Regional wage differences in the Netherlands: micro-evidence on agglomeration externalities. *J. Reg. Sci.* 54, 503–523.
- Guimaraes, P., Figueiredo, O., Woodward, D., 2000. Agglomeration and the location of foreign direct investment in Portugal. *J. Urban Econ.* 47, 115–135.
- Handbury, J., 2013. Are poor cities cheap for everyone? Non-homotheticity and the cost of living across US cities. Mimeograph, Wharton University.
- Handbury, J., Weinstein, D., 2015. Goods prices and availability in cities. *Rev. Econ. Stud.*, forthcoming.
- Hanson, G.H., 1997. Increasing returns, trade, and the regional structure of wages. *Econ. J.* 107, 113–133.
- Hanson, G.H., 2005. Market potential, increasing returns, and geographic concentration. *J. Int. Econ.* 67, 1–24.
- Harada, N., 2005. Potential entrepreneurship in Japan. *Small Bus. Econ.* 25, 293–304.
- Harris, C., 1954. The market as a factor in the localization of industry in the United States. *Ann. Assoc. Am. Geogr.* 44, 315–348.
- Head, K., Mayer, T., 2004. Market potential and the location of Japanese investment in the European Union. *Rev. Econ. Stat.* 86, 959–972.
- Head, K., Mayer, T., 2006. Regional wage and employment responses to market potential in the EU. *Reg. Sci. Urban Econ.* 36, 573–595.
- Head, K., Ries, J.C., Swenson, D.L., 1999. Attracting foreign manufacturing: investment promotion and agglomeration. *Reg. Sci. Urban Econ.* 29, 197–218.
- Heckman, J., Singer, B., 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 2, 271–320.
- Henderson, J., 1997. Externalities and industrial development. *J. Urban Econ.* 42, 449–470.
- Henderson, V., 2002. Urban primacy, external costs, and the quality of life. *Resour. Energy Econ.* 24, 95–106.
- Henderson, J., 2003. Marshall's economies. *J. Urban Econ.* 53, 1–28.
- Henderson, J., Kuncoro, A., Turner, M., 1995. Industrial development in cities. *J. Polit. Econ.* 103, 1067–1090.

- Henderson, V., Lee, T., Lee, Y.J., 2001. Scale externalities in Korea. *J. Urban Econ.* 49, 479–504.
- Hering, L., Poncet, S., 2010. Market access and individual wages: evidence from China. *Rev. Econ. Stat.* 92, 145–159.
- Hilber, C.A.L., Voicu, I., 2010. Agglomeration economies and the location of foreign direct investment: empirical evidence from Romania. *Reg. Stud.* 44, 355–371.
- Holl, A., 2012. Market potential and firm-level productivity in Spain. *J. Econ. Geogr.* 12, 1191–1215.
- Holmes, T.J., 2011. The diffusion of Wal-Mart and economies of density. *Econometrica* 79, 253–302.
- Holmes, T.J., Lee, S., 2012. Economies of density versus natural advantage: crop choice on the Back Forty. *Rev. Econ. Stat.* 94, 1–19.
- Holmes, T., Sieg, H., 2015. Structural estimation in urban economics. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, vol. 5A. North-Holland, Amsterdam.
- Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. *Econometrica* 62, 467–475.
- Jacobs, J., 1969. *The Economy of Cities*. Random House, New York.
- Jofre-Montseny, J., Marín-López, R., Viladecans-Marsal, E., 2011. The mechanisms of agglomeration: evidence from the effect of inter-industry relations on the location of new firms. *J. Urban Econ.* 70, 61–74.
- Jofre-Montseny, J., Marín-López, R., Viladecans-Marsal, E., 2014. The determinants of localization and urbanization economies: evidence from the location of new firms in Spain. *J. Reg. Sci.* 54, 313–337.
- Kerr, W., Kominers, S.D., 2015. Agglomerative forces and cluster shapes. *Rev. Econ. Stat.*, forthcoming.
- Kim, S., 1995. Expansion of markets and the geographic distribution of economic activities: the trends in US regional manufacturing structure, 1860–1987. *Q. J. Econ.* 110, 881–908.
- Kok, S., 2014. Town and city jobs: your job is different in another location. *Reg. Sci. Urban Econ.* 49, 58–67.
- Kolko, J., 2010. Urbanization, agglomeration, and coagglomeration of service industries. In: Glaeser, E.L. (Ed.), *The Economics of Agglomeration*. National Bureau of Economic Research, Cambridge, MA, pp. 151–180.
- Krashinsky, H., 2011. Urban agglomeration, wages and selection: evidence from samples of siblings. *Labour Econ.* 18, 79–92.
- Krugman, P.R., 1991a. *Geography and Trade*. MIT Press, Cambridge, MA.
- Krugman, P.R., 1991b. Increasing returns and economic geography. *J. Polit. Econ.* 99, 484–499.
- Lall, S.V., Shalizi, Z., Deichmann, U., 2004. Agglomeration economies and productivity in Indian industry. *J. Dev. Econ.* 73, 643–673.
- Lee, S., 2010. Ability sorting and consumer city. *J. Urban Econ.* 68, 20–33.
- Lee, B.S., Jang, S., Hong, S.H., 2010. Marshall's scale economies and Jacobs' externality in Korea: the role of age, size and the legal form of organisation of establishments. *Urban Stud.* 47, 3131–3156.
- Lehmer, F., Möller, J., 2010. Interrelations between the urban wage premium and firm-size wage differentials: a microdata cohort analysis for Germany. *Ann. Reg. Sci.* 45, 31–53.
- Lesage, J., Pace, R.K., 2009. *Introduction to Spatial Econometrics*. CRC Press, New York.
- Levinsohn, J., Petrin, A., 2003. Estimating production functions using inputs to control for unobservables. *Rev. Econ. Stud.* 70, 317–342.
- Lindley, J., Machin, S., 2014. Spatial changes in labour market inequality. *J. Urban Econ.* 79, 121–138.
- Lopez, R., Suedekum, J., 2009. Vertical industry relations, spillovers, and productivity: evidence from Chilean plants. *J. Reg. Sci.* 49, 721–747.
- Lucas Jr., R.E., 1988. On the mechanics of economic development. *J. Monet. Econ.* 22, 3–42.
- Mameli, F., Faggian, A., McCann, P., 2008. Employment growth in Italian local labour systems: issues of model specification and sectoral aggregation. *Spat. Econ. Anal.* 3, 343–360.
- Marrocu, E., Paci, R., Usai, S., 2013. Productivity growth in the Old and New Europe: the role of agglomeration externalities. *J. Reg. Sci.* 53, 418–442.
- Marshall, A., 1890. *Principles of Economics*. Macmillan, London.
- Martin, P., Mayer, T., Mayneris, F., 2011. Spatial concentration and plant-level productivity in France. *J. Urban Econ.* 69, 182–195.
- Matano, A., Naticchioni, P., 2012. Wage distribution and the spatial sorting of workers. *J. Econ. Geogr.* 12, 379–408.

- McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), *Frontier in Econometrics*. Academic Press, New York, pp. 105–142.
- McMillen, D.P., 2012. Perspectives on spatial econometrics: linear smoothing with structured models. *J. Reg. Sci.* 52, 192–209.
- Melitz, M.J., Ottaviano, G.I.P., 2008. Market size, trade, and productivity. *Rev. Econ. Stud.* 75, 295–316.
- Melo, P.C., Graham, D.J., Noland, R.B., 2009. A meta-analysis of estimates of urban agglomeration economies. *Reg. Sci. Urban Econ.* 39, 332–342.
- Micucci, G., Di Giacinto, V., 2009. The producer service sector in Italy: long term growth and its local determinants. *Spat. Econ. Anal.* 4, 391–425.
- Mion, G., 2004. Spatial externalities and empirical analysis: the case of Italy. *J. Urban Econ.* 56, 97–118.
- Mion, G., Naticchioni, P., 2009. The spatial sorting and matching of skills and firms. *Can. J. Econ.* 42, 28–55.
- Moretti, E., 2004a. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *J. Econom.* 121, 175–212.
- Moretti, E., 2004b. Workers' education, spillovers, and productivity: evidence from plant-level production functions. *Am. Econ. Rev.* 94, 656–690.
- Moretti, E., 2013. Real wage inequality. *Am. Econ. J. Appl. Econ.* 5, 65–103.
- Morikawa, M., 2011. Economies of density and productivity in service industries: an analysis of personal service industries based on establishment-level data. *Rev. Econ. Stat.* 93, 179–192.
- Moulton, B.R., 1990. An illustration of the pitfall in estimating the effects of aggregate variables on micro units. *Rev. Econ. Stat.* 72, 334–338.
- Muehlemann, S., Wolter, S.C., 2011. Firm-sponsored training and poaching externalities in regional labor markets. *Reg. Sci. Urban Econ.* 41, 560–570.
- Neumark, D., Simpson, H., 2015. Place-based policies. In: Duranton, G., Henderson, V., Strange, W. (Eds.), *Handbook of Urban and Regional Economics*, vol. 5A. North-Holland, Amsterdam.
- Olley, G., Pakes, A., 1996. The dynamics of productivity in the telecommunication equipment industry. *Econometrica* 64, 1263–1297.
- Overman, H.G., Puga, D., 2010. Labor pooling as a source of agglomeration: an empirical investigation. In: Glaeser, E.L. (Ed.), *The Economics of Agglomeration*. National Bureau of Economic Research, Cambridge, MA, pp. 133–150.
- Phimister, E., 2005. Urban effects on participation and wages: are there gender differences? *J. Urban Econ.* 58, 513–536.
- Pusterla, F., Resmini, L., 2007. Where do foreign firms locate in transition countries? An empirical investigation. *Ann. Reg. Sci.* 41, 835–856.
- Redding, S., Sturm, D., 2008. The costs of remoteness: evidence from German division and reunification. *Am. Econ. Rev.* 98, 1766–1797.
- Redding, S., Venables, A.J., 2004. Economic geography and international inequality. *J. Int. Econ.* 62, 63–82.
- Rice, P., Venables, A.J., Patacchini, E., 2006. Spatial determinants of productivity: analysis for the regions of Great Britain. *Reg. Sci. Urban Econ.* 36, 727–752.
- Richardson, H.W., 1987. The costs of urbanization: a four-country comparison. *Econ. Dev. Cult. Chang.* 35, 561–580.
- Rigby, D.L., Essletzbichler, J., 2002. Agglomeration economies and productivity differences in US cities. *J. Econ. Geogr.* 2, 407–432.
- Roback, J., 1982. Wages, rents and the quality of life. *J. Polit. Econ.* 90, 1257–1278.
- Rodríguez-Pose, A., Tselios, V., 2012. Individual earnings and educational externalities in the European Union. *Reg. Stud.* 46, 39–57.
- Rosenthal, S.S., Strange, W.C., 2001. The determinants of agglomeration. *J. Urban Econ.* 50, 191–229.
- Rosenthal, S.S., Strange, W.C., 2003. Geography, industrial agglomeration, and agglomeration. *Rev. Econ. Stat.* 85, 377–393.
- Rosenthal, S.S., Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam, pp. 2119–2171.

- Rosenthal, S.S., Strange, W.C., 2008. The attenuation of human capital spillovers. *J. Urban Econ.* 64, 373–389.
- Rossi-Hansberg, E., Sarte, P.D., Owens III, R., 2010. Housing externalities. *J. Polit. Econ.* 118, 485–535.
- Saito, H., Gopinath, M., 2009. Plants self-selection, agglomeration economies and regional productivity in Chile. *J. Econ. Geogr.* 9, 539–558.
- Sato, Y., Tabuchi, T., Yamamoto, K., 2012. Market size and entrepreneurship. *J. Econ. Geogr.* 12, 1139–1166.
- Serafinelli, M., 2014. Good firms, worker flows and local productivity. Mimeograph, University of Toronto.
- Simon, C.J., 2004. Industrial reallocation across US cities, 1977–1997. *J. Urban Econ.* 56, 119–143.
- Solé-Ollé, A., Viladecans-Marsal, E., 2004. Central cities as engines of metropolitan area growth. *J. Reg. Sci.* 44, 321–350.
- Spies, J., 2010. Network and border effects: where do foreign multinationals locate in Germany? *Reg. Sci. Urban Econ.* 40, 20–32.
- Stock, J.H., Yogo, M., 2005. Testing for weak instruments in linear IV regression. In: Andrews, D.W., Stock, J.H. (Eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge University Press, Cambridge, MA, pp. 80–108.
- Suedekum, J., 2008. Convergence of the skill composition across German regions. *Reg. Sci. Urban Econ.* 38, 148–159.
- Suedekum, J., 2010. Human capital externalities and growth of high- and low-skilled jobs. *Jahrb. Nat. Stat.* 230, 92–114.
- Thomas, V., 1980. Spatial differences in the cost of living. *J. Urban Econ.* 8, 108–122.
- Van Soest, D.P., Gerking, S., Van Oort, F.G., 2006. Spatial impact of agglomeration externalities. *J. Reg. Sci.* 46, 881–899.
- Viladecans-Marsal, E., 2004. Agglomeration economies and industrial location: city-level evidence. *J. Econ. Geogr.* 5, 565–582.
- Wang, Z., 2013. Smart city: learning effects and labor force entry. Mimeograph, Brown University.
- Wheaton, W.C., Lewis, M.J., 2002. Urban wages and labor market agglomeration. *J. Urban Econ.* 51, 542–562.
- Wheeler, C.H., 2006. Cities and the growth of wages among young workers: evidence from the NLSY. *J. Urban Econ.* 60, 162–184.
- Wheeler, C., 2008. Local market scale and the pattern of job changes among young men. *Reg. Sci. Urban Econ.* 38, 101–118.
- Yankow, J.J., 2006. Why do cities pay more? An empirical examination of some competing theories of the urban wage premium. *J. Urban Econ.* 60, 139–161.

CHAPTER 6

Agglomeration and Innovation

Gerald Carlino*, William R. Kerr†

*Federal Reserve Bank of Philadelphia, Philadelphia, PA, USA

†Harvard University, Bank of Finland, and NBER, Boston, MA, USA

Contents

6.1. Introduction	350
6.2. What is Innovation?	352
6.2.1 Distinguishing invention from innovation	352
6.2.2 Types of innovation	353
6.2.3 Measuring innovation	354
6.2.3.1 <i>Investments in the innovation process</i>	354
6.2.3.2 <i>Patents and citations</i>	355
6.2.3.3 <i>Literature-based indicators of innovation</i>	357
6.3. Patterns of Agglomeration and Innovation	358
6.3.1 Higher degree of spatial concentration at a point in time	358
6.3.2 Innovation is more localized than other economic forces linked to agglomeration	361
6.3.3 Spatial movements of clusters over time	366
6.4. Formal Theories Linking Agglomeration and Innovation	366
6.4.1 Sharing	368
6.4.2 Matching	369
6.4.3 Knowledge spillovers	371
6.4.3.1 <i>Knowledge spillovers and endogenous growth</i>	372
6.4.3.2 <i>Empirical evidence on localized knowledge spillovers: wages</i>	376
6.4.3.3 <i>Empirical evidence on localized knowledge spillovers: patents</i>	380
6.4.3.4 <i>Empirical evidence on localized knowledge spillovers: other approaches</i>	384
6.4.4 Marshallian equivalence	386
6.4.5 City size and industrial composition	386
6.4.6 "Natural advantages" for innovation	388
6.5. Additional Issues on Innovation and Agglomeration	390
6.5.1 Sustaining mechanisms	390
6.5.2 Policy choices	391
6.5.3 Connections across clusters	392
6.5.4 Innovation in entrepreneurial clusters versus large firms	394
6.5.5 Multiunit firm locations and agglomeration clusters	396
6.6. Conclusions	396
Acknowledgments	397
References	398

Abstract

This chapter reviews academic research on the connections between agglomeration and innovation. We first describe the conceptual distinctions between invention and innovation. We then discuss how these factors are frequently measured in the data and note some resulting empirical regularities. Innovative activity tends to be more concentrated than industrial activity, and we discuss important findings from the literature about why this is so. We highlight the traits of cities (e.g., size and industrial diversity) that theoretical and empirical work links to innovation, and we discuss factors that help sustain these features (e.g., the localization of entrepreneurial finance).

Keywords

Agglomeration, Clusters, Innovation, Invention, Entrepreneurship

JEL Classification Codes

J2, J6, L1, L2, L6, O3, R1, R3

6.1. INTRODUCTION

This chapter summarizes the recent literature on agglomeration and innovation and how agglomeration and innovation relate to economic performance and growth. It is well established that population and economic activity are spatially concentrated. In addition, innovation is more spatially concentrated than manufacturing employment (Audretsch and Feldman, 1996), and there is evidence that research and development (R&D) activity is more concentrated than employment (Buzard et al., 2015). Why is innovative activity clustered? What is the best way to measure this concentration? What is the economic impact of this concentration? These and related questions are covered in this chapter.

Section 6.2 begins with a discussion of invention versus innovation and how these forces are measured. Section 6.3 quickly reviews patterns of innovation and agglomeration. Section 6.4 describes formal theories linking agglomeration and innovation. Particular attention is paid to the role of knowledge spillovers in local areas, and extra attention is devoted to describing recent work to quantify these relationships. This section also makes connections to the models developed in endogenous growth theory. Section 6.5 discusses research on additional factors that work to sustain agglomeration clusters, link global clusters together, promote large versus small company innovation, and deal with similar phenomena. Section 6.6 concludes. In the final section and at several points throughout the chapter, we highlight important areas for future research.

Several key themes weave their way through this review:

- *Innovation that comes in many shapes and sizes, except in economic studies*—Section 6.2.1 provides quick definitions of invention versus innovation versus commercialization, describes underlying varieties such as exploration versus exploitation research or

product versus process work, and so on. Most nuances are then forgotten as the substantial majority of theoretical and empirical studies consider single-dimension outcomes such as patents. Hopefully, future research will develop a richer accounting of the variations of innovation and how these variations relate to the traits of clusters. We do not pretend that this will be easy, due in part to data constraints. More important, an *ad hoc* collection of facts will not be very useful, so instead, we need a structure of innovation outcomes and how they relate to local traits.

- *Agglomeration theories that can be applied especially for innovation clusters but often lack empirical confirmation*—This chapter details the various mechanisms and models that have been proposed for the formation of agglomerations of economic activity. We trace how these models apply to innovation specifically, often conjecturing that rationales such as input sharing and knowledge spillovers apply particularly well in this setting. Yet, our review of the empirical literature often finds limited verification of these theories with respect to innovative industries. This is an important area for future research and one that is facilitated by the many data sets recently developed. A subtheme of the chapter is how the existing evidence is spread across many adjacent fields (e.g., entrepreneurial finance). This porous boundary reflects the complex nature of innovation and requires researchers to cover a larger domain in their work.
- *Moving to the frontier in terms of the urban economics toolkit*—Empirical measurement in urban economic studies has made substantial strides forward in the past two decades. Two prominent examples are the use of continuous-distance metrics for measuring economic activity and the development of microrecord data sets for studying economic and social interactions at detailed levels. Research related to innovation clusters is moving toward this empirical frontier, but the gap is not yet fully closed.
- *Globalization of innovation*—The last part of this chapter describes some key research opportunities, one of which we emphasize here. We have assembled evidence about innovative clusters within countries, and separate lines of economic research consider flows of activity across countries. Increasingly, research in this area is merging these fields by considering the connections of clusters across countries (e.g., the particular flows of activity from Bangalore, India, or Seoul, South Korea, to Silicon Valley). With the ongoing globalization of innovation, theoretical and empirical refinements of this work are very important.

Our review focuses equally on highlighting classic studies and identifying recent contributions. Previous reviews such as [Audretsch and Feldman \(2004\)](#) and [Feldman and Kogler \(2010\)](#) contain more extensive documentation of the older literature. Our review also seeks to highlight how research on innovation in economic geography is now connecting with adjacent fields. This has been a very active research area over the past decade and will continue to be so for some time to come.

6.2. WHAT IS INNOVATION?

This section introduces and distinguishes the concepts of innovation and invention. We describe the typical data sets used to study the phenomena and discuss some of the relevant advantages and limitations that researchers should be aware of.

6.2.1 Distinguishing invention from innovation

It has become popular for researchers and policymakers to talk about the role innovation plays in economic growth. But what is innovation and how does it differ from invention? According to the [Organisation for Economic Co-operation and Development's Oslo Manual \(2005\)](#),

An innovation is the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations.

Invention is a term often used interchangeably with innovation. But invention is a separate item in the dictionary. According to the *New Oxford Dictionary of English*, invention can be defined as

Creating something new that has never existed before.

[Schumpeter \(1939, pp. 84–86\)](#) made a clear distinction between invention and innovation:

The making of the invention and the carrying out of the corresponding innovations are, economically and sociologically, two entirely different things. They may, and often have been, performed by the same person; but this is merely a chance coincidence which does not affect the validity of the distinction. Personal aptitude—primarily intellectual in the case of the inventor, primarily volitional in the case of the businessman who turns the invention into an innovation—and the methods by which the one and the other work, belong to different spheres.

For Schumpeter, invention is the creation of a new product, service, or process, whereas innovation is the commercialization or introduction of that product or service into the market. Many inventions are patented, but most patents never reach the point of commercialization. Those that do often require a long gestation period. On the other hand, innovations are closely linked to commercialization and often do not require a corresponding invention or patents. Innovation takes place when something new or different is brought to fruition. Thus, conceptually and partly by definition, innovation has an immediate effect on welfare through the introduction of new products and improvements in existing products, often yielding a direct effect on regional and national growth.

Despite these nuances, most researchers do not distinguish between invention and innovation. Intuitively, invention seems to be an antecedent of innovation, but the conceptual basis for separating them is not so absolute, as both are necessary and complementary factors in the advancement of technology ([Scherer, 1986](#)). This perspective suggests that the concept of innovation has two parts: the generation of ideas and the conversion of these ideas into useful commercial applications. This is the sense in which most current

researchers use the term “innovation.” For example, the National Science Foundation (NSF) defines innovation as “a series of steps that begins with imagination, and results in the creation of something of value for society.”¹

Despite this blurred boundary, regions within countries represent an important instance in which the distinction between invention and innovation has economic and policy relevance. A region’s growth depends upon the successful commercialization of new products and processes more so than on invention. The idea for a new product may have occurred in one city, but the commercialization of that product may take place in an entirely different location. Thus, the benefits from local public subsidies for basic research may not stimulate growth in targeted communities, except for creating a few jobs for scientists and engineers.

While these distinctions between invention and innovation are important, our review generally focuses on the more inclusive definition of innovation. Likewise, this chapter will consider evidence related to knowledge transfer and spillovers more broadly. While these forces clearly extend beyond innovation *per se*, findings from studies of innovation are often used to describe knowledge flows, and vice versa.

6.2.2 Types of innovation

Innovation can take many forms. At one level, the results from innovative efforts can be physical or abstract, such as computer programs or new business methods versus a higher-definition TV. As we show in [Section 6.2.3](#), our ability to measure these advances differs across fields and over time. At the second and arguably more important level, innovation is a very broad term that includes multiple types of efforts. The types of advances sought by start-up companies can be quite different from those pursued by large universities or corporate R&D labs. The theoretical and empirical literature tends to lump everything together, as our review later conveys, but it is important to outline some of the distinctions made.

The oldest distinction is between incremental innovation and radical innovation. Incremental innovation, sometimes called Usherian innovation, refers to variations on a theme, involving a learning process through cumulative experience and experimentation ([Usher, 1929](#)). One example of Usherian innovation is the introduction of the electronic ignitions that replaced the mechanically timed ignition used to ignite a fuel–air mixture in the internal combustion engine. Alternatively, radical innovation, sometimes called Schumpeterian or disruptive innovation, is trend breaking and renders existing products or methods obsolete ([Schumpeter, 1934, 1939](#); [Christensen, 1997](#)). For example, the transistor transformed the way people listened to music because the transistor led to the miniaturization of the radio, thereby making it portable.

The second distinction focuses on product innovation versus process innovation (e.g., [Cohen and Klepper, 1996b](#)). Product innovation refers to a new or improved product,

¹ The ambiguous definitions are also due in part to empirical considerations. Patent data are a useful, if somewhat limited, source of inventive data. High-quality data for innovation are unfortunately scarce, and thus, most researchers utilize patent data even if their conceptual focus is innovation.

whereas process innovation represents an improvement in a firm's production technology. Innovation also encompasses the introduction of existing products to new locations. The third distinction, which relates to product–process separation, focuses on whether innovations are internal versus external to the firm, also known as exploration versus exploitation (e.g., [March, 1991](#); [Akcigit and Kerr, 2010](#)).

Similar to the distinction between innovation and invention, these differences will mostly fade into the background for the rest of this review. This is not desirable but instead reflects the very limited degree to which these distinctions have been made in an economic geography context. We describe the available studies where applicable, but much more work needs to be done in this vein, as agglomeration will interface in quite different ways with various forms of innovation. For example, [Duranton and Puga \(2001\)](#) found evidence that French entrepreneurs tend to develop prototypes for new products in diversified cities and that these businesses relocate to cheaper specialized locations for mass production once the products are sufficiently mature. This process would suggest that the innovation in diversified cities would be systematically different from that in specialized locations.

6.2.3 Measuring innovation

Most studies measure technological innovation in one of three ways: (1) by the inputs used in the innovation process, such as R&D expenditures or venture capital (VC) investment; (2) by intermediate outputs of the innovation effort, such as the number of patents; or (3) by some final measure of innovative work, such as the count of new product announcements. Each measure has strengths and weaknesses that are outlined below. Studies on innovation tend to use a single indicator of innovation and argue that the chosen measure has fewer shortcomings than other measures.

6.2.3.1 Investments in the innovation process

Given the scarce data on innovative output, early studies frequently used R&D employment and expenditures as a measure of innovative activity. These factors certainly reflect innovative effort, and [Feldman \(1994, pp. 30–31\)](#) provided references to these studies. This approach has lost some favor over time, however, since input-based measures do not capture the efficiency of the innovation process. Moreover, newer measures such as patent data are easier to access and more granular, providing greater empirical traction, including the qualitative importance of the innovations produced. R&D data are particularly difficult to collect at the local level except through confidential surveys such as the NSF R&D Survey.²

² One exception to highlight is that many growth frameworks model firms as choosing their innovation intensity in terms of R&D efforts per sales. In this setting, converting innovation metrics such as patents into R&D-equivalent terms can enable the calibration of models (e.g., [Acemoglu et al., 2014](#)).

While R&D expenditures are less frequently used, a surge of work today considers VC investment by area as a metric for how innovative different regions are. VC firms fund innovative start-up companies in return for equity, with [Gompers and Lerner \(2006\)](#) providing a full introduction. Conceptually, VC-based metrics again reflect an input into the innovative process, rather than an output or end goal. One advantage of VC-based measures is that they are now available at the microlevel through sources such as VentureXpert. Similar to the patent data described next, this affords researchers extensive flexibility in metric design. The central challenge for metrics based upon VC activity is that these investments are concentrated in specific technological areas (e.g., computers, software, and biotechnology) and types of firms (e.g., younger start-ups), thus making them quite incomplete for describing innovation broadly.

6.2.3.2 *Patents and citations*

Most regional studies use patent data to measure innovation, and we accordingly devote more attention to the strengths and limitations of the patent data. Essentially, a patent constitutes a legal right to exclude others from making, using, or selling the patented invention or process for some period of time. Patents provide the economic incentives for inventors to pursue new opportunities by granting a temporary monopoly to successful efforts. To qualify for patent protection, an invention must be deemed useful (utility), new (novelty), and a nonobvious extension of the prior art (nonobviousness).

A patent publication contains a description of the invention, the names of the inventors, the institution (e.g., firm and university) to which the patent is assigned, references (citations) to related inventions or techniques, claims that define the rights of the new invention, and much more. The United States Patent and Trademark Office (USPTO) granted more than 250,000 patents per year in the past few years. [Hall et al. \(2001\)](#) provided more details about these data, and [Griliches \(1990\)](#) surveyed the use of patent data for measures of technological change.

Advantages: An advantage of using patents over R&D expenditures is that they are a direct outcome of the inventive process. Moreover, patent data are available in full (for what is patented) and not subject to confidentiality restrictions or suppressed information. Patent data are released at the microlevel, and thus, researchers can use the granularity to describe inventive activity in great detail (e.g., isolating Microsoft's patents from a specific city in a specific technological area). Due to recent efforts to digitize historical records, patent data are now available over long periods. These advantages and easy access make patent data the most used empirical approach.

Disadvantages: One concern with using patents as a measure of innovation is that patents reflect the first stage of innovation—that is, the invention. To what extent is the location where an invention takes place coincident with the successful commercialization of the invention? [Feldman \(1994\)](#) found a correlation of 0.8 between patents and the locations where new products are introduced to the market. While this correlation is

positive and moderately high, [Feldman and Kogler \(2010\)](#) nonetheless stressed that “studies that draw inferences about innovation by focusing on invention should be interpreted with caution.”

Another concern about using patents as an innovation indicator is that the value of patents is very highly skewed. Most patents are not worth very much, while a few are very valuable (e.g., [Harhoff et al., 1999](#)). If a patent has value, we would expect it to be renewed before the patent expires. [Serrano \(2010\)](#) calculated that 78% of US patents granted during 1983–2001 were not renewed, indicating that most patents are of low value. In the words of [Griliches \(1979\)](#) and [Pakes and Griliches \(1980, p. 378\)](#), “patents are a flawed measure (of innovative output) particularly since not all new innovations are patented and since patents differ greatly in their economic impact.” In addition, patent owners often obtain patents on close substitutes for their patent in an attempt to block follow-on patenting by rival inventors. Rather than adding value, these “patent fences” are strategic and thus skew innovation measures.

Fortunately, researchers can adjust for patent quality in their innovation metrics by weighting patents by the number of citations they receive, similar to what economists do with journal articles. Most researchers exclude self-citations from these counts (i.e., a Microsoft patent that cites another Microsoft patent). An important recent practice is to make these measurements relative to each technology category and application year to adjust for differences in propensities to cite across fields and the truncation biases in citations for the most recent cohorts. As we describe in [Section 6.4.3](#), citations are also used to model knowledge spillovers ([Jaffe et al., 1993](#); [Murata et al., 2014](#); [Buzard et al., 2015](#)).

There are naturally worries about what is not patented. Based on a 1994 Carnegie Mellon University survey of almost 1500 industrial R&D labs in the United States, [Cohen et al. \(2000\)](#) reported that firms typically use a number of mechanisms to protect their returns to invention. In most manufacturing industries, patents are used less frequently than other approaches such as trade secrets and lead-time advantages. [Cohen et al. \(2000\)](#) found that the two most important reasons given by firms for not applying for a patent are the amount of information disclosed in a patent application and the ease of legally inventing around a patent. They also reported that the cost of patent litigation deters smaller firms from pursuing patents.³

Another issue is that there are large differences in the propensity to patent across industries. For example, [Cohen et al. \(2000\)](#) reported that firms in the chemical, drug, mineral product, and medical equipment industries applied for patents for more than two-thirds of their innovations. In contrast, firms in the food, textile, glass, and steel and other metal industries applied for patents on fewer than 15% of their product

³ These selection issues are also critically important across countries. It is frequently noted that local firms are hesitant to file patents in nations with weak intellectual property protections (e.g., China) because of disclosure concerns.

innovations. Quantitative work considering raw differences across industries in patent rates needs to proceed with caution, while growth in innovation rates for industries is more secure.

The final consideration relates to the longitudinal adjustment of what forms of technologies are patentable. For example, one of the strongest trends in the USPTO data is the introduction and growth of software patents. Founded in 1975, Microsoft had just five patents by 1990 and more than \$1 billion in revenue; by 2009, the company held 10,000 patents and over \$58 billion in revenue. More broadly, the rate of patenting for a given set of inputs slowed in some fields over the past three decades, while it accelerated in others. As cities and regions differ in their technical and industrial compositions, a worry exists that the measurement of changes in innovation at the city level will be biased or suffer substantial measurement error due to these underlying forces (e.g., Seattle's rate of innovation appears to skyrocket, but this is mainly due to changes in the legal practice of patenting software). [Lerner and Seru \(2014\)](#) further discuss these challenges and the importance of designing area-based studies to control for the underlying composition of inventive activity.

6.2.3.3 Literature-based indicators of innovation

A third approach is broadly referred to as literature-based indicators, since the metrics are based on product announcements in trade, engineering, and technical publications. [Acs and Audretsch \(1988\)](#), [Audretsch and Feldman \(1996\)](#), [Feldman and Audretsch \(1999\)](#), and [Acs et al. \(2002\)](#), among others, used new product announcement data available from the US Small Business Administration (SBA) as an indicator of innovative activity for state and metropolitan areas in the United States. An advantage that new product announcement data have relative to R&D expenditures and patents as measures of innovation is that they document the commercialization of a new product.⁴

Although the new product announcement data are interesting, they are subject to several important qualifications. First, the data are available only for 1982 and do not allow any follow-up analysis. According to [Acs and Audretsch \(1988\)](#), the data largely consist of product innovations to the exclusion of other important types of innovation, such as process, services, and managerial innovations. Another concern is that the new product announcements tend to be issued by the marketing departments of companies; therefore, these announcements are not subject to the same scrutiny as patents are to determining the originality and contribution of the product. Also, the announcement data may suffer from selection bias, as journal editors of trade publications may select innovations that they think are influential.

⁴ [Capello and Lenzi \(2014\)](#) develop a measure of innovation based on Eurostat's Community Innovation Survey of firms introducing new products and processes. Many R&D surveys are now including questions of this variety.

In sum, the various indicators of innovative activity have strengths and weaknesses. Few studies that look at the spatial scope of innovation have used multiple indicators to see if the results are robust to indicator choice. Is there a strong correlation among the various indicators? Are the findings using patents similar to those based on R&D expenditures or to findings based on new product announcement data? One possibility for future research is to develop a composite indicator based on two or three of the individual indicators ([Hagedoorn and Cloudt, 2003](#)). It is also hoped that the ongoing explosion in data collection will afford new measures for future studies.

6.3. PATTERNS OF AGGLOMERATION AND INNOVATION

There is an abundance of evidence that innovative activity is spatially concentrated. We quickly review some basic evidence, not attempting to catalog all of the measurements that have been made. Three points are generally made: (1) innovation is more concentrated at a point in time than general economic activity; (2) it appears more localized than other economic forces linked to agglomeration; and (3) there are some spatial movements of clusters over time, although the new clusters remain spatially concentrated (Detroit circa 1900 has remarkable parallels to Silicon Valley today).

6.3.1 Higher degree of spatial concentration at a point in time

Each of the innovation measures discussed in the prior section displays substantial spatial concentration. For R&D activity, [Buzard and Carlino \(2013\)](#) showed that the spatial concentration of establishments undertaking R&D efforts is more pronounced than it is for firms generally. [Figure 6.1](#) is taken from [Buzard and Carlino \(2013\)](#). [Buzard and Carlino \(2013\)](#) show, in particular, that R&D activity for most industries tends to be concentrated in the Northeast Corridor, around the Great Lakes, in California's Bay Area, and in Southern California. This concentration is not limited to the United States. For example, [Carrincazeaux et al. \(2001\)](#) found that six regions in France account for 75% of all corporate R&D workers, compared with 45% of the production workers.

VC investments and patents are similarly concentrated, as [Figure 6.2](#) from [Chatterji et al. \(2014\)](#) shows. During the 1990s, three-quarters of the US population resided in metropolitan areas. In contrast, 92% of the patents were granted to residents of metropolitan areas, and virtually all VC investments were made in major cities. [Bairoch \(1988\)](#) showed that patent originations tend to be concentrated in large cities. [Fornahl and Brenner \(2009\)](#) found that patents tend to be concentrated in 11 of the 97 German regions they considered.

In terms of final outputs, [Acs et al. \(1994\)](#) found that the introduction of new products is more spatially concentrated than patents. [Feldman and Audretsch \(1999\)](#) found that less than 4% of the product innovations took place outside of metropolitan areas and that half of the new product innovations in 1982 occurred in just four metropolitan areas (Boston,

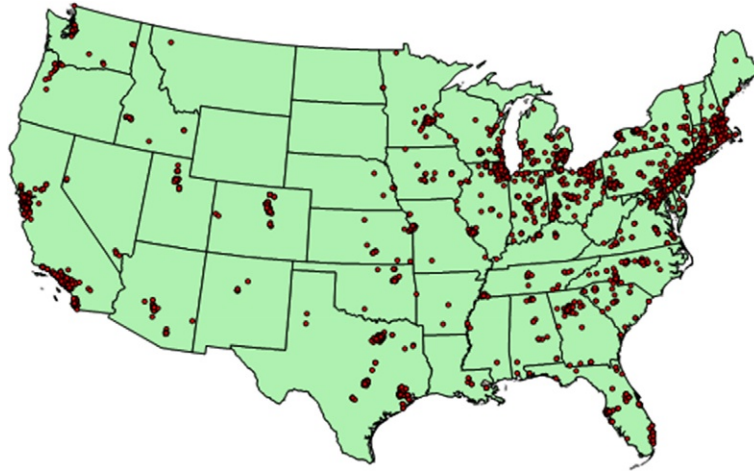


Figure 6.1 The location of R&D labs in 1998. Each dot on the map represents the location of a single R&D lab. In areas with a dense cluster of labs, the dots tend to sit on top of one another, representing a spatial cluster of labs. *Notes: The figure is taken from Buzard and Carlini (2013). Source data are from the Directory of American Research and Technology (1998).*

New York City, San Francisco, and Los Angeles). Feldman (1994) attributed this strong concentration of the commercialization of innovation to the need for specialized business services (e.g., firms specializing in market research and product testing, specialized patent lawyers, and the availability of financing) and similar infrastructure.

Studies have gone beyond these raw statistics to compare the concentration of innovation with the underlying spatial concentration of economic activity. Krugman (1991) and Audretsch and Feldman (1996) used a “locational Gini coefficient” to study the spatial clustering of manufacturing industries. A locational Gini coefficient shows how similar (or dissimilar) the location pattern of employment in a given industry is from the location pattern of aggregate employment. Let s_{ij} represent location i ’s share of employment in industry j , and let x_i represent location i ’s share of aggregate employment. The spatial Gini coefficient for industry j is defined as $G_j = \sum_i (x_i - s_{ij})^2$. $G_j = 0$ indicates that

employment in industry j is no more or less spatially concentrated than aggregate employment, and $G_j > 0$ implies that employment in industry j is overconcentrated. Audretsch and Feldman (1996) applied this approach at the state level using SBA data.

One problem with the locational Gini coefficient is that it may spuriously indicate the localization of an industry resulting from the lumpiness of plant employment. Ellison and Glaeser (1997) improved on the locational Gini coefficient by offering an alternative index (the EG index) that controls for the industrial organization of an industry. Conceptually, the EG index compares the degree of spatial concentration of manufacturing employment in industry j with what would arise if all plants in the industry were

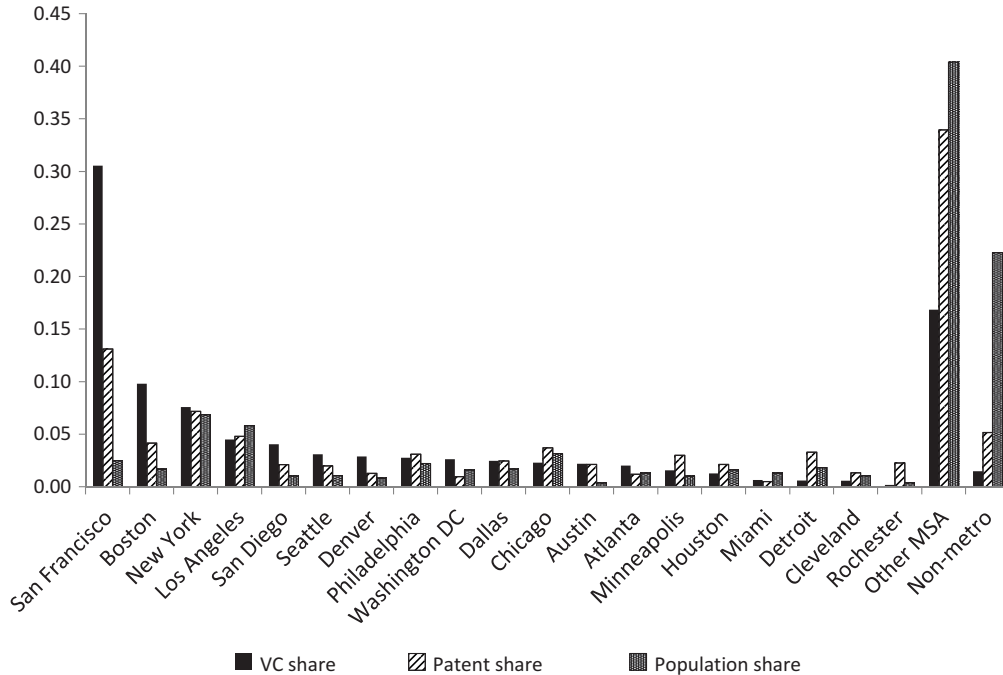


Figure 6.2 Spatial concentration of US patenting and venture capital shares over the 1990–2005 period. VC calculations use the share of deals over the 1990–2005 period. Patent calculations use the share of granted patents applied for from each city during 1990–2005. The population share is from 1999. *Notes: The figure is taken from Chatterji et al. (2014). Source data are from VentureXpert, USPTO patent data, and county-level population statistics.*

randomly distributed across locations. Using this approach, Ellison and Glaeser (1997) and Rosenthal and Strange (2001) found evidence of the geographic concentration of employment in many US manufacturing industries. Although the EG index is widely adopted in the literature, few studies use it to determine the agglomerative forces for innovative activity; one exception is Buzard and Carlino (2013).

The locational Gini coefficient and the EG index suffer from important aggregation issues. (1) The first is known as the modifiable area unit problem (MAUP).⁵ These metrics depend upon the boundaries used to demarcate regions, and conclusions may differ if counties versus states, for example, are used as boundaries. The MAUP grows in severity as the level of aggregation increases. (2) A related issue is referred to as “border effects”—each region is considered an exclusive zone, and the closeness of regions is not factored in.

⁵ See Briant et al. (2010) and Menon (2012) for a discussion of the MAUP and its implications for spatial concentration measures. Duranton and Overman (2005) suggested five properties for a good index of concentration. The index should (1) be comparable across industries, (2) control for overall concentration of the industry, (3) control for industrial concentration, (4) be unbiased with respect to scale and aggregation, and (5) test for the significance of the results.

Thus, while Philadelphia County and Montgomery County border each other and have activity spilling across them, they are treated as being as distant from each other as they are from Los Angeles County. These partitions often lead to underestimations of concentration. (3) The earlier metrics provide index values but do not assess the statistical significance of their results. Without further statistical analyses, it is unclear whether the concentrations reported are significantly different from ones that might result even if the locations of economic activity resulted from random draws. Recently, [Cassey and Smith \(2014\)](#) have developed a procedure to simulate confidence intervals for statistical tests of EG indices.

In response to these issues, [Duranton and Overman \(2005\)](#) developed a continuous-distance metric that does not rely on fixed spatial boundaries. This approach provides researchers with a range of scales in continuous space, rather than at a single predefined scale, on which to measure effects. In addition, these techniques utilize Monte Carlo methods to determine whether the number of plants at a given distance is significantly different from the number found if their locations were randomly chosen. The power of these techniques comes with two significant costs. First, researchers must access data on individual plant locations, which can be challenging. Second, the calculation of these metrics takes substantially more time and computational power.

A number of more recent researchers have used continuous methods to measure the spatial concentration of innovative activity (e.g., [Inoue et al., 2013](#); [Murata et al., 2014](#); [Buzard et al., 2015](#); [Kerr and Kominers, 2015](#)), which we recommend. For example, [Buzard et al. \(2015\)](#) used point patent methods (Ripley's k -function analysis) to analyze locational patterns of R&D labs over a range of spatial scales (e.g., within a half mile, one mile, five miles, etc.). This approach allows them to consider the spatial extent of the agglomeration of R&D labs and to measure how rapidly the clustering of labs attenuates with distance. It is important to note that they look for geographic clusters of labs that represent statistically significant departures from spatial randomness using simulation techniques. In sum, continuous approaches appear to be the preferred method for measuring the geographic concentration of innovative activity going forward, although we later discuss how political boundaries are still important when using these techniques.

6.3.2 Innovation is more localized than other economic forces linked to agglomeration

[Marshall \(1890\)](#) first elaborated on how many different forces give rise to spatial concentration. As depicted in the next section, innovation is related to multiple Marshallian channels, but it also has the closest connection with knowledge spillovers. Comparative studies often find that knowledge spillovers operate at the smallest spatial scales of the types of agglomeration forces. [Rosenthal and Strange \(2001\)](#) regressed the spatial concentration of an industry (based on an EG index) on characteristics designed to measure the individual forces of agglomeration. The analysis is carried out separately at the zip code, county, and state levels. They found that knowledge spillovers positively affect

agglomeration only at the zip code level, with other forms of agglomeration such as labor pooling operating at broader levels. [Ellison et al. \(2010\)](#) exploited patterns of industry colocation to measure the relative importance of the various mechanisms using discrete and continuous-distance measures. They related coagglomeration levels to the extent to which industry pairs share goods, workers, and knowledge. They found evidence for all three mechanisms, and knowledge spillovers again are the most localized. Using a new metric for analysis of location patterns within cities, [Billings and Johnson \(2014\)](#) also emphasize colocation for knowledge sharing. [Capello and Lenzi \(2014\)](#) find that the growth benefits from new knowledge are spatially concentrated when considering 262 regions in the EU.

Complementing these regional calculations, several studies estimate spillovers based on production function methods. This approach again emphasizes the exceptional concentration of knowledge flows and innovative activity. [Figure 6.3](#) illustrates results

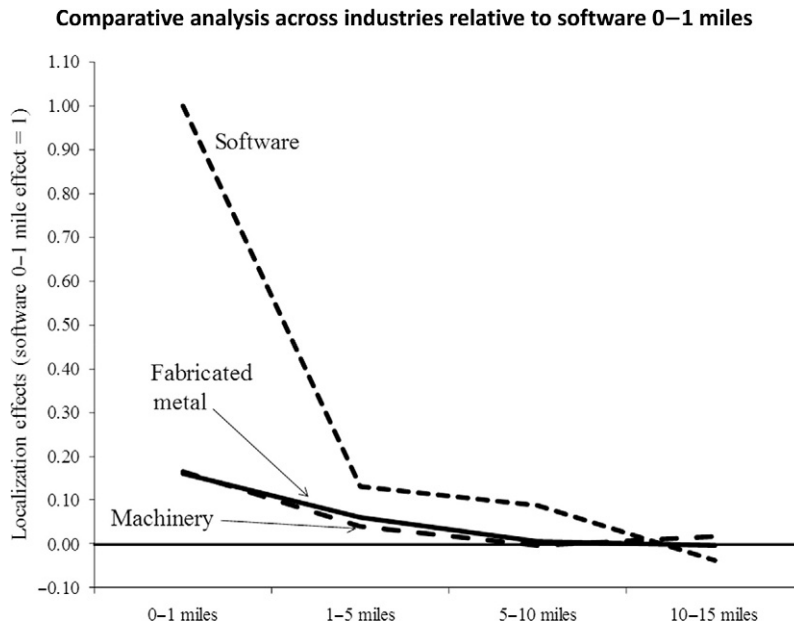


Figure 6.3 Localization effects by distance between establishments-comparative analysis across industries relative to software 0–1 miles. The horizontal axis measures distance between two establishments in the industry. The vertical axis measures localization effects, with the scale set such that the localization effect for the software industry in the 0–1 mile distance is equal to one. Clustering is more important for the software industry than for the fabricated metal or machinery industries, particularly at the 0–1 mile distance. In all industries, the benefit of proximity to firms in one's own industry drops off sharply when located 1–5 miles apart rather than 0–1 miles apart. Once firms are 10–15 miles apart, there are no more differential benefits due to site selection within the city, although generalized city-level benefits may still exist. *Notes: The figure is derived from localization results presented in [Rosenthal and Strange \(2003\)](#).*

presented in a comparative study of software and manufacturing industries by [Rosenthal and Strange \(2003\)](#). To allow easy interpretation of their findings, the vertical axis in the exhibit is set such that the spillover benefits in the software industry within one mile are equal to a value of one; all other spillover effects are measured relative to this value. The horizontal axis shows the distance between any two establishments in the industry.

For all industries, the localization effects of being near similar businesses decay rapidly with distance within cities—the positive localization effect from being within one mile of another company in one’s own industry is at least 10 times greater than the positive effect realized when locating two to five miles away from the said company. After five miles, the benefits continue to decay with distance, albeit at a less dramatic rate. By 10 miles, there are no more within-city localization benefits. This statement does not preclude general spillover effects that all firms within the city enjoy from higher rates of industrial concentration; it only means that a firm’s specific site within the city no longer governs access to the benefits. It is important to note that [Rosenthal and Strange \(2003\)](#) identified dramatic differentials across industries. The decay functions of two representative manufacturing industries, fabricated metal and machinery, have a similar shape to the decay function of the software industry. The most dramatic difference, however, is that the magnitude of spillover effects in the two manufacturing examples tend to be only 20% or so of the software effect at any distance.

While the declines over the first mile are dramatic in [Rosenthal and Strange \(2003\)](#), they can be even sharper in a very highly networked industry. [Arzaghi and Henderson \(2008\)](#) studied location choices of ad agencies in Manhattan. They showed that for an ad agency, knowledge spillovers and the benefits of networking with nearby agencies are extensive, but the benefits dissipate quickly with distance from other ad agencies and are gone after roughly a half mile. [Figure 6.4](#) plots the decay function in localization effects using this study’s results, with the localization effect for ad agencies for 0–250 m being set equal to a value of one. The decline in [Figure 6.4](#) is much more dramatic, with the localization effect losing 80% of its value by the time two firms are 500 m apart. Rather than the 10 miles of the previous example, the cluster benefits for ad agencies run out at 750 m! This study provides one of the sharpest depictions of tightly concentrated knowledge flows developed to date, and we believe that it represents an important precedent for future research related to innovation more directly. We discuss these features and especially studies building upon patents, in more detail as we introduce direct theories in the next section.

While these two approaches—regional-based evidence and production function regressions—end up with similar conclusions about the rapid attenuation of knowledge flows, they embody very different spatial scales and are not immediately reconcilable with each other. For example, can city- or county-level comparisons of software versus machinery industrial concentration yield much insight if the main clustering effect in both industries is only over one mile?

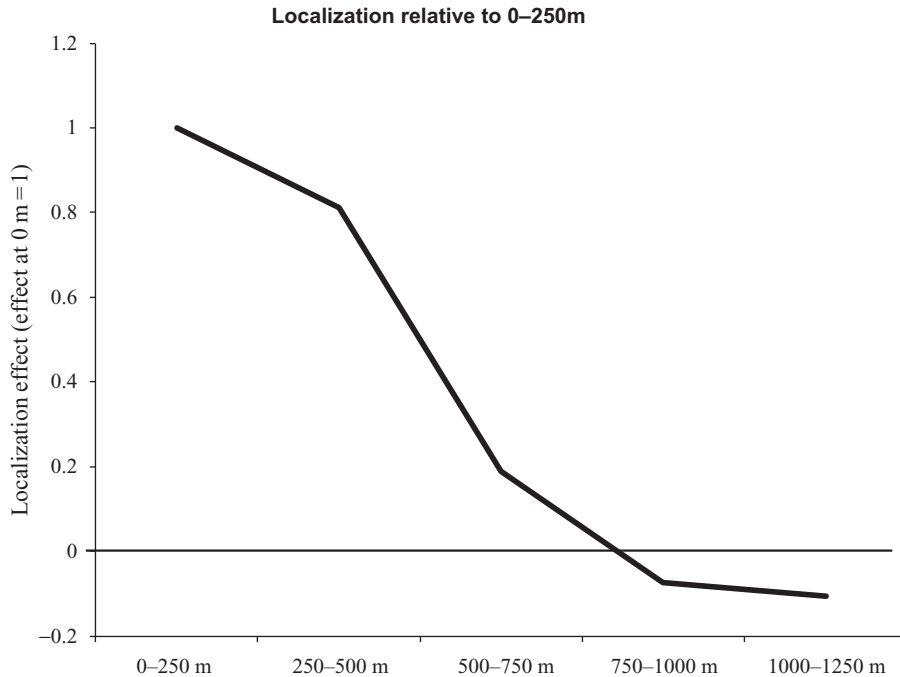


Figure 6.4 Localization effects among ad agencies in Manhattan-localization relative to 0–250 m. The horizontal axis measures the distance between two establishments in the industry. The vertical axis measures localization effects, with the scale set such that the localization effect at 0–250 m is equal to 1. Compared with [Figure 6.3](#), the localization effects in the ad agencies attenuate extremely fast, reaching 0 by 750 m. Ad agencies and the people who work in them place a very high importance on networking, collaboration and information sharing, and knowledge spillovers, all of which are much more effective over very short distances. *Notes: The figure is derived from localization results presented in [Arzaghi and Henderson \(2008\)](#).*

[Kerr and Kominers \(2015\)](#) consider this question using variation across patent clusters. A central element of their model and empirics is illustrated in [Figure 6.5](#), which depicts technology flows in Silicon Valley. Downtown San Francisco and Oakland, CA, are to the north and off of the map. The triangle in the bottom-right corner of the map is the core of Silicon Valley. This core contains three-quarters of the industrial patents filed from the San Francisco Bay Area and 18 of the top 25 zip codes in terms of patenting. Plotted on the map are three of the four largest zip codes for patenting in the San Francisco area that are outside of the core. Each focal zip code is marked with a box, and the other points of the shape are the three zip codes that firms in the focal zip code cite most often in their work. Zone 1, which covers Menlo Park, extends deepest into the core. Zone 2, for Redwood City, CA, shifts up and encompasses Menlo Park and Palo Alto but has less of the core. Zone 3, which covers South San Francisco, further shifts out and brushes the core.

Top patenting zip codes outside of the core and their sourcing zones

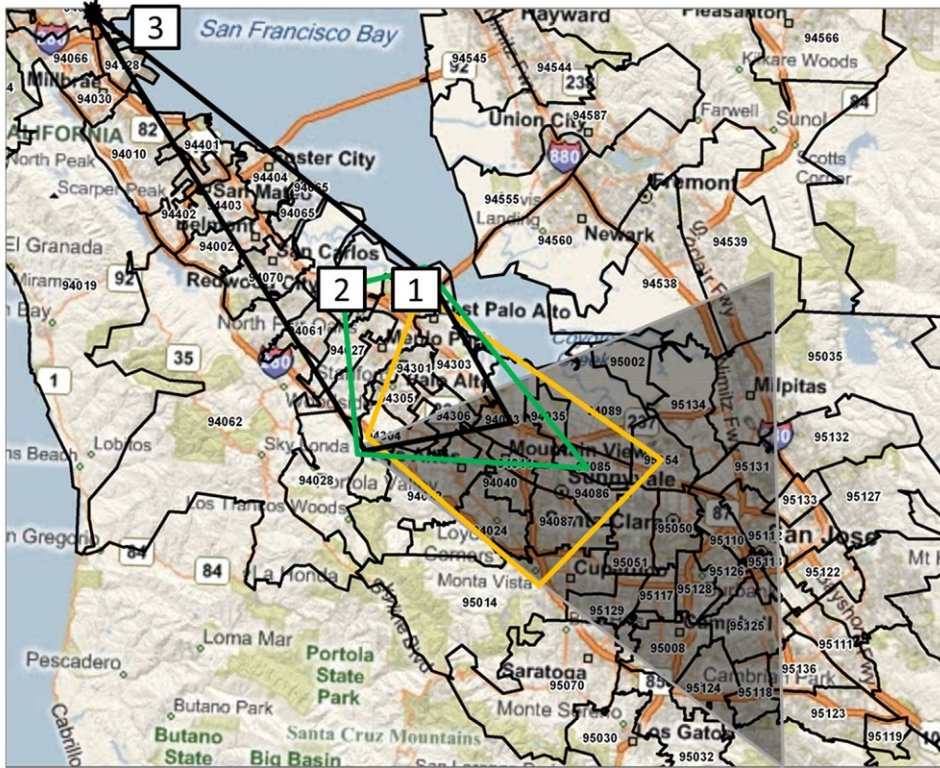


Figure 6.5 Technology sourcing from Silicon Valley. Top patenting zip codes outside of the core and their sourcing zones. The exhibit shows the shapes of technology spillover zones in the area surrounding Silicon Valley. The core of Silicon Valley is depicted by the shaded triangle. The Silicon Valley core contains 76% of the patenting for the San Francisco region. This map describes the technology sourcing for three of the four largest zip codes for patenting *not* included in the core itself. Technology sourcing zones are determined through patent citations. The boxes indicate the focal zip codes, and the shape of each technology sourcing zone is determined by the three zip codes that firms in the focal zip code cite most in their work. The yellow (gray in the print version) zone (1) for Menlo Park extends deepest into the core. The green (dark gray in the print version) zone (2) for Redwood City shifts up and encompasses Palo Alto but less of the core. The black zone (3) for South San Francisco further shifts out and brushes the core. These technology zones are characterized by small, overlapping regions. None of the technology sourcing zones traverse the whole core, and only the technology zone with the closest zip code (Menlo Park) reaches far enough into the core to include the area of the core where the greatest number of patents occur. Transportation routes and geographic features influence the shapes and lengths of these sourcing zones. Notes: The figure is taken from [Kerr and Kominers \(2015\)](#).

These technology zones are characterized by small, overlapping regions. None of the technology sourcing zones traverse the whole core, much less the whole cluster, and only the closest zip code (Menlo Park) even reaches far enough into the core to include the area of Silicon Valley where the greatest number of patents is issued. While technology sourcing for individual firms is localized, the resulting cluster extends over a larger expanse of land. [Kerr and Kominers \(2015\)](#) use this insight to build a model of clusters comprising small, overlapping regions. This model reconciles the area-based approaches with the firm-level analyses by showing that agglomerative forces with longer spillover distances at the firm level yield macrostructures characterized by fewer, larger, and less dense clusters. [Kerr and Kominers \(2015\)](#) find that these predictions empirically hold true when comparing the clusters of different technologies with each other. Extended out of sample, this logic also provides a foundation for why innovation clusters can be especially concentrated in small geographic regions.

6.3.3 Spatial movements of clusters over time

While most thoughts of innovation clusters today naturally begin with Silicon Valley, it is important to recall that innovation clusters do move over time. [Lamoreaux et al. \(2004\)](#) described the role of Cleveland as a prominent and frontier hub for innovation related to the second industrial revolution, including early forms of angel financing and incubators. [Klepper \(2010\)](#) similarly described how Detroit circa 1900 has remarkable parallels to Silicon Valley today. [Saxenian \(1994\)](#) also prominently discussed the migration of the semiconductor industry from Boston's Route 128 to Silicon Valley in the 1970s and 1980s. [Kerr \(2010a\)](#) described the specific movement of technologies to new breakthrough centers following the [Duranton \(2007\)](#) model. With the development of better long-run patent data, we hope that further regularities can be defined that enhance this case study evidence.

6.4. FORMAL THEORIES LINKING AGGLOMERATION AND INNOVATION

Often, innovative activity takes place inside industrial clusters, such as in the semiconductor industry in Silicon Valley, in the financial services and advertising industries in New York City, in the entertainment industry in Los Angeles, and in the production of country music in Nashville, TN, to name just a few such clusters. Industrial clusters not only are a source for innovation leading to productivity growth but also tend to stimulate the formation of new businesses, leading to ongoing growth in entrepreneurial activity ([Porter, 1998](#)).

How do we account for this geographic concentration of innovative activity? Much of the theoretical literature on urban agglomeration economies has focused on externalities in the production of goods and services rather than on invention itself. Nevertheless,

the three formal mechanisms primarily explored in this literature—sharing, matching, and knowledge spillovers—are also relevant for innovative activity.⁶

This section discusses models and evidence on these traditional channels as they relate to innovation. The next section will discuss additional topics and adjacent fields that are more specific to innovation processes. We argue in this section that the traditional Marshallian externalities, even beyond knowledge spillovers, are especially important for the spatial concentration of innovative activity, over and above that evident for production. In addition, the “natural advantages” of an area, such as its unique culture and institutions, explain the location of some innovative activity (Saxenian, 1994).

This chapter focuses mostly on evidence related to innovative clusters and local interactions across firms. These firms are usually presumed to be of small size relative to local aggregate activity, taking the local conditions around them as given. In Section 6.5, we return to this question and discuss research on whether innovation is concentrated in a single large firm, spread across many local firms, or something in between.

For this setting, an innovation production function is a convenient way to demonstrate the effects that these agglomeration forces have on innovative activity: $I_{ic} = g(A_c)F(RD_{ic}, K_{ic})$, where I indexes the innovative output of firm i in city c ; RD and K represent R&D and human capital usage, respectively, by firm i ; and $g(A_c)$ represents agglomeration economies that are external to firm i but internal to the firm’s city.⁷ If firm i wanted to double its production of I , it could do so by doubling RD and K . It is important to note that it is unnecessary to double agglomeration economies in the city. If the agglomeration economies also doubled, the production of I would more than double.

One baseline empirical strategy is to model $g(A_c)$ as a function of urban size or density. This is roughly the approach taken by Carlino et al. (2007), who found that the rate of patenting per capita—or patent intensity—is about 20% higher in a metropolitan area with twice the employment density (jobs per square mile) of another metro area.⁸ While this elasticity is interesting, the approach does not tell which of the mechanisms are responsible for the increased patenting activity found in denser locations or anything about the relative importance of the various mechanisms; this requires theoretical models of how each of the mechanisms operate. We consider each in turn using the Duranton and Puga (2004) taxonomy.

⁶ These themes are developed by Duranton and Puga (2004). Marshallian economies are often expressed as customer–supplier linkages, labor pooling, etc., and Duranton and Puga (2004) describe the theoretical advantages from formulating their common principles, such as sharing, matching, etc. Recent surveys of the empirical literature on agglomeration economies include Eberts and McMillen (1999), Rosenthal and Strange (2004), and Combes and Gobillon (2015) provide a comprehensive update in this handbook.

⁷ See Audretsch and Feldman (2004) for further discussion of knowledge production functions in a regional setting.

⁸ Carlino and Hunt (2009) found that the elasticity of patenting with respect to density remains about 0.2 regardless of whether they use unweighted patents or citation-weighted patents.

6.4.1 Sharing

The sharing of common inputs depends on the existence of economies of scale in the production of these inputs at a given location. Thick factor markets can arise when innovative activity clusters locally. These clusters allow each of their members to benefit as if they had greater scale (Porter, 1998). Thick local markets allow for the efficient sharing of a variety of specialized inputs and access to a pool of specialized and experienced workers and to a variety of specialized business services (e.g., patent attorneys, commercial labs for product testing, and trade organizations). Knowledge can flow rapidly through these linkages (Porter, 1990). The ability to outsource locally allows firms to lower production costs by maintaining leaner inventories, and Porter (1998) argued that this effect is especially important for “advanced and specialized inputs involving embedded technology, information, and service content.” Industrial clusters allow members to experiment at lower costs, and if they decide to innovate, firms can externally source what they require to quickly implement their innovations.

Helsley and Strange (2002) developed an interesting dynamic model of innovation in which a dense network of input suppliers facilitates innovation by lowering the costs needed to bring new ideas to realization. The model is dynamic, since previous innovations in an area produce inputs that make entrepreneurs more productive, leading to continuing growth in entrepreneurial activity. Gerlach et al. (2009) demonstrated that firms invest more in R&D and take greater risk in R&D choices in clusters compared with spatially isolated firms. They also demonstrated that similar firms choose different R&D projects when located in clusters, which creates a diversified portfolio of R&D investments.

The benefits that firms in a given industry gain from colocating are often enhanced using production technology based on modularity. Modularity is the method of making complex products or creating processes from smaller subsystems (modules) that are developed by a network of independent firms. Under this system, different suppliers are responsible for separate modules with the understanding that each supplier will follow “design rules” that ensure that the modules fit and work together (Baldwin and Clark, 1997).⁹ An important aspect of modularity is that innovative activity becomes decentralized across many firms instead of being concentrated in a single vertically integrated firm. The rate of technical progress can be enhanced by decentralization, since independent firms can focus more fully on innovation to their specific components compared with

⁹ Modularity as a method of production has been around for some time. A classic example is automobile manufacturing in Detroit, in which car manufacturers purchased components such as car bodies, transmissions, and brakes from largely local independent suppliers. More recently, computer system manufacturers have almost entirely relied on a network of local independent suppliers of modules in the production of final products. In related theoretical work, Hellmann and Perotti (2011) depicted conditions under which innovative ideas are better developed in thick markets of innovative firms versus within larger organizations.

the divided attention each component receives when firms are vertically integrated. [Sturgeon \(2002\)](#) stressed that modular production allows colocated firms to use the benefits of spatial proximity to establish and manage global-scale production networks. Using evidence from case studies and interviews, [Saxenian \(1991\)](#) concluded that Silicon Valley-based computer system manufacturers became more efficient by spreading the costs and risks associated with the rapid changes in product designs and technologies by adopting modularity techniques.

In general, there is little empirical evidence on the importance of input sharing for firms engaged in innovative activity, even if the relationship is intuitive. [Feldman \(1994\)](#) showed that the presence of a local network of firms in related industries helps to refine new innovations and that the existence of specialized business services helps firms bring their ideas to fruition. [Audretsch and Feldman \(1996\)](#) found a greater tendency for innovative activity to colocate in industries in which skilled labor is an important input. We discuss the sharing of common labor pools in more detail later.

What is better established is the development and sharing of specialized business services. This has been especially true with the case of entrepreneurial finance (e.g., angels and VC). Traditional sources of financing, such as bank loans, may be unavailable to innovative start-ups due to their high risk, large financing requirements, and asymmetrical information, especially in high-tech industries ([Gompers and Lerner, 2001](#)). As a result, VC organizations tend to invest locally in order to monitor their investments and to provide operating assistance to these firms (e.g., [Gompers and Lerner, 2001](#); [Horvath, 2001](#)). [Kolympiris et al. \(2011\)](#) found that VC investment within the biotech sector is particularly concentrated in areas of a 10 mile radius or less. These investors illustrate input sharing in many ways—their business models depend upon a certain scale and diversity of local opportunities, the value-added benefits they offer to portfolio companies frequently stem from this local networking (e.g., introductions to key employees to hire), and so on.¹⁰

6.4.2 Matching

A second theory argues that thicker markets improve the quality of matches in local labor markets. Models that support this theory include [Helsley and Strange \(1990, 2002\)](#), [Wheeler \(2001\)](#), [Berliant et al. \(2006\)](#), [Strange et al. \(2006\)](#), and [Papageorgiou \(2013\)](#). For example, in [Berliant et al. \(2006\)](#), workers in large cities can be more selective in forming matches because the opportunity cost of waiting for a prospective partner is lower. Despite the fact that people are more selective, on average, they form matches more quickly. As a result, the average output from matches is higher, and a higher share

¹⁰ [Samila and Sorenson \(2011\)](#) provided evidence that VC investment has positive growth spillovers at the MSA level in the United States, but there is much more research that can be done on how VC activity connects with local economic development.

of the workforce is engaged in productive matches. [Strange et al. \(2006\)](#) found that firms with an (exogenous) uncertain demand for highly specialized workers benefit from locating in specialized industrial clusters, since these clusters provide a more diverse range of worker skills to meet unforeseen or challenging opportunities.¹¹

Also, specialized workers can readily find new positions without having to change locations in innovation clusters. For example, there is evidence of job hopping in Silicon Valley ([Fallick et al., 2006](#)) and in software publishing ([Freedman, 2008](#)). This research shows that mobility increases with industrial concentration. [Fallick et al. \(2006\)](#) found much greater mobility of college-educated males employed in the computer industry in Silicon Valley compared with the interfirm mobility of similarly educated workers in the computer industry in other areas outside of California. Part of this turnover could be induced as firms and workers seek better matches. Moreover, [Fallick et al. \(2006\)](#) described the important link between this mobility and the modularity described earlier.¹²

While labor market pooling may reduce firm costs, leading to greater profit, competition for workers and worker turnover can have the opposite effect ([Matouschek and Robert-Nicoud, 2005](#); [Combes and Duranton, 2006](#); [Gerlach et al., 2009](#)). Labor market pooling exposes firms to poaching of vital workers by rival firms and increases labor costs to retain employees. [Matouschek and Robert-Nicoud \(2005\)](#) analyzed the role that investments in human capital have on the location decisions of firms. They show that whether human capital investments act as a force for or against geographic concentration depends on who is making the investments. If firms undertake the investment in human capital, they are less likely to agglomerate in an attempt to protect their investments from rival firms. Firms are more likely to agglomerate when workers undertake the investment in human capital.

In the model of [Rotemberg and Saloner \(2000\)](#), workers are more likely to undertake investment in human capital when firms are agglomerated. In a small labor market, workers who invest in acquiring skills have weak bargaining power with local employers unless they are willing to undertake costly relocations. This setting can generate holdup problems by a monopsonistic employer attempting to extract the surplus from the workers' investments, thereby reducing the original incentive to invest in skill development. In thick labor markets, the presence of a large number of potential employers mitigates the holdup problem.

¹¹ Interestingly, [Strange et al. \(2006\)](#) found that when firms are involved with creating new products or processes, they tend to locate in large cities but not in industrial clusters.

¹² Mobility could be unusually high in Silicon Valley because of the unenforceability of noncompete clauses under California law ([Gibson, 1999](#)). It is important to note that [Fallick et al. \(2006\)](#) found that employee turnover in other industries is no higher in California than in other locations, suggesting that noncompete clauses are not the primary reason for the job hopping observed in California. We return to noncompete clauses in [Section 6.5](#).

There is a lack of empirical evidence on these issues, and the endogeneity of location and investment choices makes these models difficult to test. One mentioned route is to exploit exogenous cross-country differences in legislation requiring firms to provide worker training. For example, German firms are required to provide some general training for their workers. Evidence for the [Matouschek and Robert-Nicoud \(2005\)](#) hypothesis would be found if firms in Germany are more dispersed than they would be in countries in which firms are not required to provide such training, all other things being equal. Tests might also be feasible through investments in immigrant workers in settings in which the worker is tied to his or her employer for a period of time (e.g., the United States' H-1B visa program).

6.4.3 Knowledge spillovers

The third strand of the theory argues that the geographic concentration of people and jobs in cities facilitates the spread of tacit knowledge, which is not easily codified or transferred through written documents. While the exact mechanism is not well identified in theory, the underlying idea articulated in [Marshall \(1890\)](#) is that the geographic proximity to knowledge facilitates the transfer of such knowledge and unplanned or serendipitous interactions among individuals fostering the exchange of information among workers and firms.

Steve Jobs understood that people's physical proximity is a key ingredient in the innovative process when he helped to design the layout of Pixar Animation Studios. The original plan called for three buildings, with separate offices for animators, scientists, and executives. Jobs instead opted for a single building, with a vast atrium at its core. To ensure that animators, scientists, and executives frequently interacted and exchanged ideas, Jobs moved the mailboxes, the cafeteria, and the meeting rooms to the center of the building. Pixar has been incredibly innovative, as demonstrated by its many important breakthroughs in the application of computer graphics to filmmaking. Pixar has also been quite productive, generating 14 feature films that have earned 27 Academy Awards.

Information sharing has become a hallmark of Silicon Valley firms. Facebook, Google, and Twitter, for example, have designed their work and recreational spaces to enhance the interaction and the sharing of ideas among workers. [Liu \(2010\)](#) quantified how the physical layout of a biotech firm shaped its internal knowledge flows, and [Olson and Olson \(2003\)](#) described the sharp attenuation of collaboration and distance. Information sharing is not just a Silicon Valley phenomenon, and the value of face-to-face contact for innovation has long been recognized. In the 1940s, Bell Labs' (AT&T's R&D lab) new building was specifically designed so that everyone would interact with one another ([Gertner, 2012](#)). In seventeenth-century England, people gathered in coffeehouses to share ideas, with different coffeehouses attracting specialized clienteles. The London Stock Exchange began life in 1698 in a coffeehouse in which merchants met, while a

second coffeehouse frequented by shippers and traders became recognized as the place to obtain marine insurance and gave rise to Lloyd's of London.

While theories of knowledge spillovers were originally developed to explain the concentration of industries in general, they are particularly important in explaining the clustering of innovative activity. R&D, more than most industries, depends on new knowledge. Often, the latest knowledge about technological developments is valuable to firms but only for a short time, and the reciprocal exchange of information among colocated firms engaged in innovation can reduce uncertainty (Feldman, 1994). Thus, it behooves innovative firms to locate near sources of information and each other.

This extended section reviews knowledge spillovers in three steps. We first outline its connection with growth theory and some of the economic geography models that have been influential in this context. These models identify a Mincerian wage relationship that has been frequently used in empirical work, and the second part highlights this body of work. We then review the large body of work using patents and other techniques.

6.4.3.1 Knowledge spillovers and endogenous growth

Solow (1957) showed that the accumulation of physical capital is insufficient for long-run growth of per capita output. In Solow's model, the long-run rate of growth of per capita output is determined largely by an exogenous rate of technical progress, providing no explanation for productivity improvements. Since the rate of productivity growth is the most important determinant of long-run growth, treating such an important factor as given leaves many unanswered questions.

Beginning in the mid-1980s, economists developed theories of endogenous growth in which cities often play a prominent role. Lucas (1988) emphasized the role that the stock of human capital in cities plays in economic growth. While the channels through which knowledge spillovers are transmitted are not well articulated, the basic idea is that skilled or educated workers are better able to receive new knowledge from others. The acquisition of knowledge is related to cities since the dense concentration of people and firms in cities creates an environment in which new ideas travel quickly. Thus, cities facilitate knowledge spillovers. The idea that the exchange of knowledge among individuals enhances productivity can be expressed by allowing production to depend on aggregate human capital in a city. Let Y_{ic} denote a homogeneous final good produced by worker i in city c according to

$$Y_{ic} = H_c^\delta h_{ic},$$

where h_{ic} represents the human capital of worker i and H_c^δ is aggregate human capital in the city. Notice that there are constant returns to scale in the human capital of worker i , and increasing returns are associated with aggregate human capital at the city level. The parameter δ reflects the importance of an external return in production. The aggregate measure of human capital can be expressed as

$$H_c^\delta = \left\{ \frac{1}{n} \sum_{n=1}^N h_{ic}(n)^\rho \right\}^{\delta/\rho},$$

where $h_{ic}(n)$ is the human capital of the n th worker and ρ is a parameter governing how the amount of the human capital of different workers is aggregated to form the city's stock of human capital. When $\rho = 1$, the average level of human capital in a city is what matters when creating the externality. When $\rho > 0$, the skills of the most talented workers create the externality. If the most talented workers sort into the largest cities, as [Glaeser and Saiz \(2004\)](#) found, externalities will not be equalized across cities. In most empirical specifications of Lucas' model, $\rho = 1$ is assumed.

Assuming the zero-profit condition holds, one obtains a relationship between the nominal wage of workers and the aggregate measure of human capital that can be expressed in a Mincerian wage regression:

$$\ln W_{ic} = \delta \ln H_c + \nu \ln h_{ic} + \varepsilon_{ic},$$

where the parameter of interest is δ .¹³ Later, we will review wage-based evidence in this regard. It is important to highlight that this specification leaves many open issues: First, endogeneity bias is an important concern when estimating δ , since aggregate schooling may be determined simultaneously with wages. Second, the specification does not depict whether the learning effect is within industries of cities or across them. Third, most empirical studies consider quantitative measures of human capital, such as years of schooling, but differences in the quality of human capital are not captured. Fourth, with the previously mentioned formulation, the spillovers are static.¹⁴ Fifth, factors beyond knowledge spillovers can give rise to these Mincerian wage regressions, an issue that we articulate more fully when reviewing these studies, and so, evidence derived from these approaches is consistent with knowledge spillovers but not complete proof of them.

[Eaton and Eckstein \(1997\)](#) embedded the human capital accumulation structure of [Lucas \(1988\)](#) within a system of cities framework. Cities grow at a common rate in equilibrium, with the relative size for a given city being determined by the environment for learning that the city provides. In [Eaton and Eckstein \(1997\)](#), the agglomeration externality is static. [Glaeser \(1999\)](#) proposed a model with a dynamic externality in which the

¹³ [Duranton and Puga \(2014\)](#) provided a microfoundation for this model in which the externality in human capital arises through entrepreneurship. If entrepreneurs are overrepresented in the population of more educated workers, then cities with relatively more educated workers will be more entrepreneurial and grow faster. See also [Jones \(2005\)](#).

¹⁴ [Lucas \(1988\)](#) proposed a model with dynamic externalities. In Lucas' models, individuals decide how much time they will devote to current production and how much time they will spend acquiring skills. Growth is now driven by an externality in the accumulation of human capital in cities. In this setup, the human capital externality is the engine of growth and an agglomeration economy. [Duranton and Puga \(2014\)](#) provided further details.

direct interactions among workers in cities are the basis for accumulation and diffusion of knowledge. In the model, young and unskilled workers learn through interactions with older and skilled workers. Cities provide greater opportunities for younger workers to meet and interact with older workers. [Black and Henderson \(1999\)](#) developed a model of human capital accumulation and population growth in cities. In the model, human capital externalities make workers more productive in larger cities. Workers devote some of their time to the accumulation of human capital, which, in turn, makes cities more attractive and leads to population growth. The growth in population, in turn, reinforces the externality in human capital. [Black and Henderson's \(1999\)](#) model comprises two sectors in which cities are specialized in a particular sector. Although cities differ along a variety of dimensions, all cities grow at the same rate in the long run. [de la Roca and Puga \(2012\)](#) used Spanish data on individual worker characteristics and their employment histories to investigate the urban wage premium and found that workers receive an earnings premium in larger cities. They also find evidence of a dynamic externality in that workers tend to accumulate more valuable knowledge in larger cities, leading to faster earnings growth.

Most recently, [Davis and Dingel \(2013\)](#) developed a spatial equilibrium framework to show why skill premiums are higher in large cities. They developed a system of cities model in which the exchange of ideas is costly and leads to agglomeration economies. The model has tradable and nontradable goods, labor is the only factor of production, and idea exchanges are local. Agents allocate their time according to the expected gains from exchanging knowledge, where the gains are greater in places where prospective partners are more numerous and of higher ability. Agents with the greatest ability to receive new knowledge are most willing to pay for this knowledge. This framework leads to cities of various sizes, with larger cities exhibiting greater skill premiums and productivity and thus offering not only higher wages but also higher housing costs.

In another branch of the literature, [Romer \(1986, 1990\)](#) formalized the relationship between knowledge and economic growth. Romer divided inputs into two categories: rivalrous inputs and nonrivalrous ideas. The notion is that once a new design is introduced by one firm, the design is subsequently available to all other firms. This nonrivalrous nature of knowledge is central for economic growth in Romer's model. Consider the constant returns to scale production function of a representative firm i at time t :

$$Y_{it} = K_{it}^{\alpha} (A_t L_{it})^{1-\alpha},$$

where Y indexes the output of firm i , A represents the aggregate amount of knowledge available to all firms at time t , and K and L are the inputs of capital and labor used by firm i . If firm i wanted to double its production of Y , it could do so by doubling K and L . Due to the nonrivalrous nature of knowledge, it is not necessary to double the stock of knowledge. If the stock of knowledge is also doubled, the production of Y would more than double. Therefore, incorporating the stock of knowledge into the production function leads to increasing returns to scale because of the nonrivalrous nature of knowledge.

One concern is that in a competitive market, once a new idea is discovered, the marginal cost to supplying the idea to other users is zero. But private firms undertake costly R&D to profit from this research. It is unlikely that R&D will be undertaken if a firm cannot profit from this research. [Romer \(1990\)](#) addressed this problem by treating knowledge as a quasi-public good in that, although new knowledge is nonrivalrous, the owner of a new idea can exclude others from its use through patents and trade secret protection, for example. The rents that the owners of patents receive compensate them for the cost of discovering new ideas.

The idea that patent and trade secret protection gives firms receiving such protection monopoly power suggests that models that depart from perfect competition are needed. This is accomplished in the innovation-based strand of the literature that formalizes Schumpeter's idea of growth through creative destruction ([Grossman and Helpman, 1991](#); [Aghion and Howitt, 1992](#); [Aghion et al., 2014](#)). In these models, innovation takes place along a "quality ladder" that is driven by temporary monopoly profits that innovating entrepreneurs require as compensation for the cost of innovating. Every product has its own current point on a quality ladder, and firms undertake R&D to improve the quality of a product. Once a product is introduced, however, the technology diffuses, and rival entrepreneurs can produce the same quality product at a lower cost and hence offer a lower selling price. Entrepreneurs can recapture market advantages through innovation, which leads to climbing up the quality ladder.

[Kelly and Hageman \(1999\)](#) and [Duranton \(2007\)](#) extended the quality ladder approach to a regional setting. In [Duranton \(2007\)](#), cities are composed of industries, and production for an industry is concentrated at the site of the latest breakthrough invention. When a new discovery takes place for an industry in another city, the industry relocates to the new center. This quality ladder model provides a solid foundation for explaining the fast churn of industries across cities, the slower movement of cities up and down the size distribution (as industry-level shocks are partially netted out), and the stable long-term distribution of city sizes. [Kerr \(2010a\)](#) provided evidence of the [Duranton \(2007\)](#) model when looking at the locations of the breakthrough patents by technology field during the 1975–1985 period and the subsequent growth of technologies across locations. The link of these breakthrough locations to subsequent production is less established.

While cities have this long-standing theoretical connection to growth, an open question is the quantitative importance of the externalities associated with human capital in cities for national economic growth. [Duranton and Puga \(2014\)](#) develop a system of city models in the tradition of [Henderson \(1974\)](#) and address this question. Using recent estimates found in the literature for the elasticity of agglomeration economies and the elasticity of congestion cost with respect to city size, they calculate that cities account for almost two-thirds of aggregate output growth. While this calibration exercise is tantalizing, it remains speculative and will hopefully be the groundwork for complete assessments.

One potential pitfall should be noted about the connections between endogenous growth theory and studies of innovation clusters. Researchers need to think hard about the role of mobility in their setting and the resulting equilibrium conditions (e.g., Glaeser, 2008). In many core growth models, population is fixed because the presumed application is to a country, and thus, the impact of innovation is pushed toward the income per capita development. Cities can adjust their population, and the impact of innovation may come through population growth instead. A full articulation requires thinking about mobility costs, housing prices, amenities, and similar factors alongside the standard issues such as regional convergence (e.g., Magrini, 2004). It is likewise important to consider how the diffusion of ideas across cities and nonpermanent relocations can influence the degree to which the cities in question are being governed by the logic and assumptions built into growth models.¹⁵

6.4.3.2 Empirical evidence on localized knowledge spillovers: wages

Since knowledge spillovers are invisible, empirical verification of them relies on indirect evidence. The two main approaches in regional studies are Mincerian wage regressions and analyses of patent citations. We start with the wage analyses that are motivated by the theoretical models of city and growth. From the outset, we caution that these studies relate local human capital stocks to wages, and thus, while they are consistent with knowledge spillovers, they are not conclusive evidence for them. Combes and Gobillon (2015) discuss this issue extensively in their chapter in this handbook.

Many studies use metrics of educational attainment as a proxy for the human capital stock of cities. The conceptual idea is that a greater share of educated workers in a city makes other workers in that city more productive. The share of the adult population with a college education differs dramatically across cities, with a threefold range evident in 2010 between the high of 28% in Raleigh, NC, and the low of 9% in Visalia, CA. Among employed workers specifically, Moretti (2012) showed that there is an even larger five-fold differential. Most studies look for this increased productivity by examining worker wages, quantifying the additional earnings that similar workers (e.g., in terms of their age, education, occupation, industry, and experience) receive as the share of college graduates in their city increases.

Using the 1980 census data, Rauch (1993) estimated that each additional year of average education for a worker in a city increases expected wages 3–5%. As already noted, endogeneity is an important concern, since higher wages might cause individuals to get more schooling instead of the other way around. This reverse causality would lead to an

¹⁵ As an example, some studies apply growth theory frameworks to primary metropolitan statistical areas (PMSAs) in the United States. PMSAs separate San Jose, CA, home of Silicon Valley, from San Francisco. This separation is clearly inappropriate, as the tech cluster covers the whole region. In fact, many tech firms in Silicon Valley offer shuttle buses for employees who wish to live in downtown San Francisco.

upward bias in ordinary least squares (OLS) estimates of the social return to schooling. Alternatively, OLS estimates of the social return to schooling can be biased downward if highly skilled workers are attracted to amenity-rich locations, since these workers accept lower wages in return for greater amenities. In addition, there could be unobserved factors that correlate with measures of city-level human capital that independently affect productivity and wages (e.g., good local governance), leading to biased estimates.

To address endogeneity bias, [Acemoglu and Angrist \(2000\)](#) used compulsory K–12 school attendance laws that existed in an individual's state at the time of his birth and child labor laws as instruments for determining a state's average level of schooling. They also use an individual's quarter of birth as an instrument for individual-level schooling. Instrumental variable estimates suggest that a 1-year increase in average schooling is associated with about a 1–2% increase in average wages compared with a similar OLS regression that yielded a 7% external return. One caveat to these findings is that they empirically identify the social return associated with the supply of workers with secondary schooling, as opposed to the supply of the most educated workers frequently emphasized in the theory on human capital spillovers. A second issue is that states are not the best geographic scale for estimating the social returns to education. These externalities are certainly stronger at the city or subcity level.

[Moretti \(2004a\)](#) estimated the social return associated with the share of an MSA's population with a college degree during the period 1979–1994. [Moretti \(2004a\)](#) dealt with endogeneity by using the presence of a land-grant college in the MSA as an instrument for college share.¹⁶ He finds that a 1% increase in the share of college graduates increases the wages of college graduates by 0.5%. The wages of high school graduates increase by 1.5%, and the wages of high school dropouts increase by almost 2%. One issue with this approach is that the spatial distribution of land-grant universities is highly uneven ([Shapiro, 2006](#)). Also, reverse causality remains a concern. [Shapiro \(2006\)](#) pointed out that it is hard to compare correlations between college attainment and land-grant schools today with correlations from the past because the Census Bureau did not begin asking about educational attainment before the 1940 census, “by which time the land-grant schools were already of significant size.”

[Ciccone and Peri \(2006\)](#) pointed out that the findings reported so far may overstate the social returns to human capital, as the estimates confound human capital externalities with factor demand effects. They showed that an increased share of highly skilled workers in a city alters the city's skill composition such that an increase in a city's average wage can occur in the absence of human capital spillovers. [Ciccone and Peri \(2006\)](#) developed an

¹⁶ To account for unobserved worker characteristics, [Moretti \(2004a\)](#) used longitudinal data from the National Longitudinal Survey of Youths. Moretti reported that omitted worker characteristics are not a major source of bias and uses lagged age structures to account for unobserved MSA-specific demand shocks.

approach that holds the labor force skill mix constant over time. They found no evidence of a return to a 1-year increase in average schooling during the period from 1970 to 1990 once they account for imperfect substitutability between education groups. [Combes and Gobillon \(2015\)](#) raise important concerns regarding this approach including that the evolution of wages in a local area may be influenced by changes in the local skill composition of workers.¹⁷

As pointed out by [Duranton \(2006\)](#) and [Henderson \(2007\)](#), the literature on the external returns to education may still be upwardly biased, since much of the work does not control for urban agglomeration economies. [Glaeser and Saiz \(2004\)](#) found that skilled workers sort into larger cities, indicating that the estimated effects of human capital on individual worker wages may be picking up urban-scale effects. In fact, when [Rauch \(1993\)](#) included metropolitan land area (an instrument for metropolitan population scale) in his regressions, the coefficient on the standard metropolitan statistical area average education variable in the wage equation is still positive but only marginally significant (see table 1, column 6 in Rauch). [Ciccone and Peri's \(2006\)](#) empirical analysis includes aggregate metropolitan employment, which can be interpreted as an agglomeration effect. Interestingly, as [Henderson \(2007\)](#) pointed out, although [Ciccone and Peri \(2006\)](#) found no evidence of a human capital externality, they do find evidence of an urban agglomeration effect, underscoring the need to distinguish these effects.

[Rosenthal and Strange \(2008\)](#) used the 2000 census data to estimate the impact of both agglomeration economies and very localized knowledge spillovers on individual worker wages. They recognize that knowledge spillovers attenuate rapidly within a few miles from the source of the externality and thus pay close attention to the rate at which the “wage–agglomeration” relationship decays with distance. Their approach involves differencing and the use of instrumental variables to address endogeneity and measurement error in their agglomeration and human capital variables. First, they found that the spatial concentration of employment (the measure of agglomeration economies) within a five mile radius is positively related to wages. Second, they found that swapping 50,000 workers with less than a college degree within five miles with workers who have at least a college degree increases the wage of a typical worker by 6–15%. Finally, agglomeration and human capital effects attenuate sharply with distance. These are important findings in that they recognize that the human capital externalities are certainly stronger at the “neighborhood” level and that they will be substantially underestimated at even the

¹⁷ [Combes and Gobillon \(2015\)](#) provide a thorough discussion of the empirical issues associated with the imperfect substitutability between skill groups and approaches for dealing with the issue. [Bacolod et al. \(2010\)](#) pointed out that important aspects of skills can be missed under the usual approach of equating a worker's skills to education. In their framework, individual characteristics such as intelligence, personality, and dexterity interact with a local environment to produce skills. They found that education is an important part of the process in skill determination; however, a worker's innate abilities, along with urban agglomeration, contribute to the development of skills.

metropolitan area level. As we will see, a growing number of studies report that knowledge spillovers attenuate rapidly with increasing distance from the source of that knowledge.

An important issue that is largely ignored in this literature strand is that theory suggests that productivity differentials should be capitalized in land rents and in wages. The degree to which the externality is capitalized in wages or rent depends on the elasticities in local land and labor markets. [Rauch \(1993\)](#) and [Conley et al. \(2003\)](#) are the only studies that allow for the capitalization of the education externality into both wages and rents and find a positive effect on rents. The finding of a small or zero externality when simply looking at wages should not be interpreted as evidence that the externality is small or even zero; it is imperative to know the magnitude of human capital spillovers (the sum of its effect on wages and rents). Knowing the extent to which spillovers are capitalized into either wages or rents also has important welfare and policy implications. If the externality is largely capitalized in land values, then landowners will benefit, and a tax on land to finance education can be efficient. However, if some regions benefit more than others from increasing returns due to knowledge spillovers, a positive correlation could exist between public spending on higher education and the net exodus of skilled people (college graduates) in those regions with decreasing returns to higher education.¹⁸

Another issue is that static Mincerian regressions do not analyze growth in wages, but a primary advantage of cities (particularly bigger ones) is that they facilitate learning. These dynamic advantages lead to greater human capital formation and perhaps more experimentation and innovation. [Glaeser and Maré \(2001\)](#) were among the first to show that workers learn more quickly in dense metropolitan areas. The dynamic externality is such that the initial effect on wages is small upon workers arriving in a new city but

¹⁸ Recently, researchers have demonstrated how ignoring rent differentials across cities (a measure of cost of living differentials) can bias estimates of real income inequality. Several studies document that an increased share of college graduates sorted into high-wage US cities between 1980 and 2000 ([Diamond, 2012](#); [Moretti, 2013](#); [Lindley and Machin, 2014](#)). As college workers are increasingly located in expensive cities, [Moretti \(2013\)](#) calculated that at least 22% of the increase in spatial college wage premium between 1980 and 2000 can be explained by geographic differences in the cost of living. One possibility is that college graduates receive greater utility from city amenities than other workers, and the higher costs reflect access to these amenities. Alternatively, the relative demand for college graduates may have increased in high-cost locations as a result of an increase in the productivity of more skilled workers. If so, the benefits college graduates receive from the spatial college wage premium is somewhat offset by higher living costs. [Moretti \(2013\)](#) found that the geographic sorting by skill groups is mostly driven by changes in their relative demands and concluded that price-adjusted welfare inequality is lower than nominal income inequality. [Diamond \(2012\)](#) also studied changes in welfare inequality during the period from 1980 to 2000. In contrast to [Moretti \(2013\)](#), Diamond found that despite higher rents, the well-being inequality gap between college graduates and high school graduates is actually 20% greater than the observed gap as a result of the increased utility college graduates derive from the consumption of localized amenities. That is, the growth in welfare inequality between skill groups is greater than the growth in income inequality between these two groups.

accumulates over time. More recently, several studies confirm faster wage growth in larger cities (de la Roca and Puga, 2012; Baum-Snow and Pavan, 2013; Wang, 2014). de la Roca and Puga (2012) used longitudinal data for a sample of Spanish workers during the period 2004–2009 to investigate sources of urban wage premium. They found during the 7 years of their sample that one-half of the urban wage premiums are static, received upon arriving in a city, while the other half accumulate over time associated with the dynamic benefits from learning. Wang (2014) finds that college-educated workers who spend their early years in large cities tend to have faster wage growth.

In summary, this literature has progressed substantially in a very short period. From the perspective of this chapter, the most needed area of further inquiry is a sharper connection of wage levels and growth to knowledge spillovers in cities. To provide empirical traction and identification, such analyses will likely need to look at microareas within cities, exploit administrative data on wages, and combine these data with additional information about knowledge flows. These ingredients are fortunately becoming available, even to the point of imagining future studies of wages, learning, and knowledge flows within individual buildings and city blocks.

6.4.3.3 Empirical evidence on localized knowledge spillovers: patents

A second empirical route for studying knowledge flows looks at patents. Citations made in patent documents have received particular attention, as they directly relate to knowledge flows, and this literature has exploded in size since the digitization of patent publications in the late 1990s. Jaffe, Trajtenberg, and Henderson (1993, hereafter JTH) is a particularly famous paper. JTH study whether inventors are more likely to cite other inventors who are geographically close to them, making use of the detailed location information available on patents for inventors. If knowledge spillovers are localized within a given metropolitan area, then citations to patents within a given metropolitan area should come disproportionately from other inventors who are located within that metropolitan area.

JTH examine two samples of originating patents—950 patents from 1975 and 1450 patents from 1980. They also consider institutional variation, separating university patents, those from the top 200 R&D performing firms, and those from other domestic corporate applicants. From the start, JTH are concerned that technologically related activity may be clustered geographically for reasons unrelated to knowledge spillovers. For example, the semiconductor industry could be concentrated in Silicon Valley because of the presence of VC firms, which might lead to a spurious inference about the role of knowledge spillovers. So, for each citation, JTH select a control citation that is technologically similar to the original citation and from the same time period.

This matched-sample approach tests whether citations are more localized relative to what would be expected based on the preexisting spatial distribution of technological activity. JTH calculate the geographic matching frequency between the citing patents

and the originating patents and compare these with the control sample. JTH find a significant “home bias” in patent citations. Excluding self-citations, citations are two to three times (for the corporate samples) to six times (for the university sample) more likely than control patents to come from the same metropolitan area. Since the observed citation-linked patent pairs are found to be in closer spatial proximity than are the control patent pairs, JTH take this as strong evidence that knowledge spillovers are geographically concentrated.¹⁹

This seminal study led to significant further inquiry and refinements. A first line of work seeks to clarify whether patent citations really measure knowledge flows. Jaffe et al. (2000) investigated this issue using a survey of inventors. They not only found evidence of communications among inventors but also noted that citations are a “noisy signal of the presence of spillovers.” Citations can be listed on patents by inventors, or they can be added by patent examiners reviewing and approving the patent application. Since 2000, patent publications separate these types of citations. Thompson (2006) found that citations added by inventors are much more spatially localized around the inventor than those citations added by examiners, which is particularly powerful evidence of the localized knowledge flows.

A second branch of inquiry involves the identification of the control group. Thompson and Fox-Kean (2005, hereafter TFK) revisited the JTH study and found that the results are sensitive to the way the control patents are selected. JTH draw control patents using three-digit technological classifications. This approach works to the extent that the three-digit level represents appropriately aggregated technology fields; if in fact, subfields are prominent and are themselves clustered for various reasons, then the control patents do not adequately account for underlying technology clustering. Continuing the earlier example, the three-digit semiconductors technology class may itself be composed of several variants that are clustered in different cities. If this is true, then it may be more appropriate to draw from the lower technology level. When TFK use six-digit technological classifications to select the control patents, they find substantially less evidence of knowledge spillovers at either the state or metropolitan area level.

There is tension, however, about the selection of control patents—it is important to make the control groups as accurate as possible, but one can narrow the groups from which the controls are drawn too much. Indeed, there is a sense in which every patent is deemed a unique contribution, and yet, one cannot obviously define a peer group within each patent. This tension is especially salient for patent counterfactuals because

¹⁹ The selection of control citations versus other techniques to control for local economic activity may lead to understatement of agglomeration for innovation. In an extreme example, the need for spatial proximity for knowledge spillovers may be so high as to have a whole technological field cluster into a very narrow space. In this setting, looking for deviations in spatial concentration between actual and control citations from the same field would underrepresent how important the localized knowledge flows are for the field.

the three-digit level of the system has 400–500 technologies, depending upon the period studied, while the six-digit level typically has more than 150,000 technologies. Thus, the narrowness of the control groups differs remarkably, and there are unfortunately no midpoints in the hierarchy. In a recent and provocative study, [Murata et al. \(2014\)](#) use sophisticated sensitivity techniques that allow for both levels of controls to illustrate the differences between JTH and TFK. Their work finds significant evidence for localization, suggesting that TFK may have sliced too thinly.

A third branch of inquiry seeks to relax the fixed geographic boundaries used in the original studies, similar to the broader movement toward the continuous-distance metrics described earlier. Since knowledge spillovers tend to be highly localized within metropolitan areas, states and metro areas may not be the appropriate geographies for studying them in many applications. Given the rapid distance decay in knowledge spillovers identified in these studies, researchers using labor market boundaries (such as MSAs) or administrative boundaries (such as states) run the risk of underestimating the importance of knowledge spillovers in the location of innovative activity. [Murata et al. \(2014\)](#) and [Buzard et al. \(2015\)](#) used distance-based approaches and found substantial evidence supporting the localization of patents and patent citations.

Going forward, we hope that researchers will continue to use these techniques. One methodological point is particularly worth emphasizing. Studies today often use the original JTH empirical strategy of drawing a single control citation for each real observation. With today's computing power, there is no reason to stop with a single counterfactual, but instead, researchers should draw a distribution of counterfactuals similar to the [Duranton and Overman \(2005\)](#) technique. [Murata et al. \(2014\)](#) provide a comprehensive depiction. Drawing 100 counterfactuals improves upon a single draw by creating a more accurate baseline and, even more important, by providing the basis for determining if a deviation is statistically significant.²⁰

Other studies find evidence of patent localization without using citations. The [Kerr and Kominers \(2015\)](#) study depicted earlier uses the [Duranton and Overman \(2005\)](#) techniques with the spatial location of patents themselves versus citations. This study then relates tightly packed clusters to those technologies displaying very short spillover ranges. [Lin \(2014\)](#) looks for evidence of knowledge spillovers in patent interferences, which are administrative proceedings to determine which applicant is entitled to the patent when multiple applications are submitted for the same invention. The basic idea is that inventors involved in an interference are likely to share certain knowledge, so patent interferences may offer evidence of knowledge spillovers among inventors. If localized

²⁰ Similarly, patents often list more than one inventor. While the location of the first inventor listed on the patent has typically been used in the past, there is little reason with today's tools to not factor in the locations of all inventors. This critique also applies to technologies, as most patents list multiple technology codes, but historically, only the first one is considered.

knowledge spillovers are important, we should see that inventors in close geographic proximity should be disproportionately involved in interferences. Lin (2014) finds that patent interferences are more likely to be observed between inventors located close to one another as opposed to those located farther apart—evidence that common knowledge inputs among independent inventors are highly localized.

Buzard et al. (2015) described how the geographic concentration of R&D labs can be used to determine more appropriate geographic boundaries in which knowledge spillovers are most likely to occur. They introduced a multiscale core cluster method to measure the localization of R&D labs across continuous space. Among others, they found a cluster of R&D labs centered in Cambridge, MA, and another in Silicon Valley. On the whole, corporate R&D labs are significantly more spatially concentrated than manufacturing employment is. Next, they assign patents and citations to these clusters and find evidence of a significant home bias in patent citations (excluding self-citations) in most clusters. Buzard et al. (2015) found that patent citations are over 12 times more likely to come from the San Jose, CA, cluster and more than six times more likely to come from the Cambridge cluster as from their respective control patents chosen to match the geographic concentration of technologically related activities. This finding provides not only evidence of localized knowledge spillovers in patent citations but also much stronger evidence than reported in prior studies.

As a whole, the patent-based evidence provides confirmation of the localized nature of knowledge flows. This evidence relates to the extreme attenuation of knowledge spillovers noted earlier in this study in Figures 6.3 and 6.4 (e.g., Rosenthal and Strange, 2001, 2003; Arzaghi and Henderson, 2008). Several other studies have found that knowledge spillovers dissipate rapidly with distance (e.g., Conley et al., 2003; Moretti, 2004b,c; Audretsch and Feldman, 1996; Adams and Jaffe, 1996).²¹ While there is much more detail that we need to learn, the general conclusion of this work on localized knowledge flows for innovation appears sound.

An important topic for future work is to evaluate how patent-based estimates of knowledge flow distances compare with other estimates. JTH suggest that since not all ideas are patented, their findings represent a lower bound on the incidence of information flows among inventors. It is also likely that patent-based studies underestimate the decay rates of knowledge flows. The transmission of tacit knowledge likely requires tighter spatial connections, for example, as evidenced by the many seminars and conferences that accompany papers and journal articles. We see the very rapid Arzaghi and Henderson (2008) decay rates as evidence in this regard and hope for ongoing inquiry in this vein.

²¹ Other nonspatial aspects of distance may play a role. Using input–output connections to measure economic distance among manufacturing establishments, Moretti (2004c) also found evidence consistent with the view that human capital spillovers decay with economic distance.

A second important topic is to link local knowledge stocks into patenting outcomes (and further into productivity and local growth). While evidence based on strong correlations exists in this regard, causal identification remains elusive. For example, [Carlino et al. \(2007\)](#) found that of the factors they considered, local human capital has the strongest correlation with patenting rates per capita in US cities. A 1% increase in the adult population with a college degree is associated with an increase in the local patenting rate by about 1%. The focus of [Carlino et al. \(2007\)](#) is on the connection of local employment density and patenting, however, and thus, the authors do not work to establish a causal relationship between human capital and the patenting rate. An extension of their paper would be to use the spatial distribution of land-grant universities as an instrument for local college share.

6.4.3.4 Empirical evidence on localized knowledge spillovers: other approaches

A number of approaches provide evidence on knowledge spillovers without using patents or wage regressions. Some studies consider how differences in education levels across cities translate into differences in firm productivity. The hypothesis is that firms situated in cities with high human capital levels will be able to produce more output using the same level of inputs compared with similar firms located in cities with low human capital levels. [Moretti \(2004c\)](#) looked at the growth in the productivity of manufacturing plants during the 1980s and found that, on average, human capital spillovers account for a meager 0.1% increase in output per year or about \$10,000 per year.

[Lychagin et al. \(2010\)](#) assessed the relative contributions of localized knowledge spillovers on firm productivity using Compustat data for the period 1980–2000. Lacking data on the location of a firm's R&D labs, [Lychagin et al. \(2010\)](#) used USPTO patent data to create a spatial distribution of the locations of a firm's inventors, which is used as a proxy for the firm's R&D. They found a positive correlation between this R&D proxy and productivity growth conditioning on technology and product-market spillovers. They used predetermined variables and the [Arellano and Bond \(1990\)](#) techniques to argue a causal relationship.

A second group of studies consider the rise of the “skilled city.” [Glaeser et al. \(1995\)](#) found that postwar income per capita grew faster in cities with high initial levels of human capital. A 1-year increase in median years of schooling in 1960 increased subsequent income growth by almost 3%. [Glaeser and Saiz \(2004\)](#) found that a one standard deviation increase in a metropolitan area's initial college share results in a 2.5% increase in its size. Similarly, [Shapiro \(2006\)](#) found that from 1940 to 1990, a 10% increase in a metropolitan area's share of college-educated residents (e.g., from 20% to 22%) raised employment growth by approximately 2%.

Beyond the traditional agglomeration literature, several recent papers develop interesting identification approaches to investigate knowledge spillovers among scientists, often in unique historical settings. [Moser \(2011\)](#) used an exogenous increase in patenting

rates for chemicals to examine the effects this increase had on geographic concentration and the diffusion of ideas. She found evidence of a spatial concentration among prize-winning inventors at four World's Fairs in the nineteenth century and early twentieth century, although innovative activity subsequently became less localized after patenting rates increased. [Ham and Weinberg \(2014\)](#) find that scientists in chemistry, medicine, and physics who are proximate to Nobel laureates show a "small, but meaningful" increase in the probability of starting their own Nobel research agenda, suggestive of horizontal knowledge spillovers among individuals of comparable stature. [Menon \(2009\)](#) found that prolific inventors have a positive and lagged effect on the patenting of inventors who are less active in the same city.

While these studies find benefits from proximity to prominent scientists, [Azoulay et al. \(2010\)](#) and [Waldinger \(2012\)](#) challenged the existence of localized knowledge spillovers originating from celebrated academics. [Waldinger \(2012\)](#) dealt with endogeneity by using the firing of scientists by the Nazis as a source of exogenous variation in the peer group of scientists remaining in Germany. He found that the publishing activity of the scientists in physics, chemistry, and mathematics in all German universities whose departments suffered losses during the period from 1925 to 1938 did not decline compared with that of other scientists. [Azoulay et al. \(2010\)](#) used the sudden and unexpected death of superstar collaborators as a source of exogenous variation. While they found that the death of a superstar scientist reduces coauthor productivity, the declines in output are similar regardless of the geographic distance among collaborators.

In sum, much of the extra evidence beyond wages and patents is also consistent with the existence of localized knowledge spillovers, although some studies find null results. A central limitation of the studies reviewed is that there is very little insight into how knowledge is transmitted among individuals living in close geographic proximity. Presumably, this occurs through both professional and social networks, but this has not been confirmed. [Henderson \(2007\)](#), in particular, posed questions about how these networks form, how members are accepted, and what the influence of these phenomena on spatial outcomes is.

While our depiction has mostly presented spillovers operating outside of market-based interactions, geographic proximity may facilitate the exchange of knowledge through contractual and market-based channels. Knowledge can spread in a city as workers move across firms, and this is not a spillover to the extent that these workers are compensated for the knowledge they bring to their new firms. Of course, even with competitive wages, there could remain uncompensated knowledge spillovers originating with these newly hired workers and benefiting the firm's existing workers. Employers use tactics ranging from employment law to stock options (e.g., [Pakes and Nitzan, 1983](#)) to limit this mobility and knowledge diffusion. Later, we depict some studies that suggest that worker mobility over firms may be an important part of the local knowledge effects discussed.

6.4.4 Marshallian equivalence

Duranton and Puga (2004) pointed out that the empirical identification of the various microfoundations (sharing, matching, and learning) can be problematic, since they all are related to the scale of activity at a given location (Marshallian equivalence). The Rosenthal and Strange (2001) and Ellison et al. (2010) studies mentioned earlier attempt to comparatively assess the forces to overcome this issue. A calibration exercise, or quantitative approach, is another way to identify the outcomes of a particular mechanism. For example, Papageorgiou (2013) documented a number of observed features relating to occupational switching, moving patterns, and wages in large cities. Using these features, he designed a model in which larger cities offer workers better matches because they have more occupations. In the calibrated framework, Papageorgiou (2013) found that better matching in large cities accounts for roughly 40% of the observed wage premium. Although the magnitude of the matching externality reported needs to be confirmed, quantitative methods are an underutilized tool in regional analysis.

6.4.5 City size and industrial composition

We next describe work on the link between innovation and the sizes of cities and their industrial compositions. This relates to the previously mentioned formal theories (e.g., thicker markets are more likely in larger cities), but there is also some distinction. The scale of a city in which an invention is found relates to the degree of spillovers in invention, but it is also influenced by factors such as the degree to which internal and external R&D of large firms is consolidated into a single location versus dispersed over plants. How the forces all empirically add up is intrinsically interesting too, especially as a guide for theory models.

Empirically, the rates of innovation generally increase across much of city size distribution, but this pattern does not hold true among the largest cities. Carlino et al. (2007) demonstrated this inverted-U shape to city size and patenting intensity, calculating the inflection point to be a city approximately the size of Austin, TX. Separate from aggregate city size, the density of activity in cities predicts innovation rates. Carlino et al. (2007) quantified that a doubling in the employment density of a city raises patents per capita by about 20%. This pattern is also nonmonotonic, with patenting intensity maximized at a density level on par with that of Baltimore or Philadelphia.²²

Multiple pieces in the urban literature connect innovation to city sizes or densities from a theoretical perspective. The classic theories of Marshall (1890) and Jacobs (1969) link larger cities to the greater development of knowledge because of increased

²² Albouy (2009) described the uneven impact of federal taxation over cities based upon their prices and sizes, given that taxes are applied to nominal wages and not real wages. Albouy described the impact this structure has on agglomeration returns. Agglomeration for innovation occurs most often in large and expensive cities, so Albouy's tax implications apply equally if not more for these areas.

specialization and diversity, respectively. In these formulations, exogenous spillovers monotonically provide a greater foundation for knowledge development in larger cities. This monotonicity is true in the Glaeser (1999) and Helsley and Strange (2002) models discussed earlier, with, for example, congestion for access to a fixed factor ultimately being used to limit city size.

By contrast, Helsley and Strange (2004) depicted how purposeful knowledge barter in cities can create a nonmonotonic relationship to city size and knowledge work. The nonmonotonic role of city size comes in their model because of the greater ease with which agents can evade punishment in large cities if they withhold knowledge that is only *ex post* verifiable in these endogenous exchanges. Berliant et al. (2006) modeled more purposeful search and matching for knowledge exchange that is governed by the city's traits. Their work describes inefficiencies that can occur through endogenous migration choices to cities for knowledge exchange. Further, modeling that captures the nuances of innovation across the city size distribution is to be encouraged.

Holding the size of local areas constant, a large body of work discusses the role of city diversity for innovation and the identification of "new work." This concept dates to Jacobs (1969), who articulated how innovation often exploits the novel recombination of past ideas and tools. Cities that possess a more diverse industrial distribution may be a more fertile soil for these activities. Duranton and Puga (2001) formalized the nursery cities argument. Lin (2011) looked at which cities are the most creative, in that they generate "new work," measured by jobs that did not exist a decade earlier. Lin found that 5–8% of US workers are engaged in new work and that the percentage is higher in cities with a higher-than-average density of college graduates and a more diverse set of industries.

Beyond diversity of the industrial base, recent studies have also considered local firm size distributions and entrepreneurship levels. Agrawal et al. (2014) explore the optimal firm size distribution for innovation in cities. Their work emphasizes the need for both a large incumbent firm and many start-ups in the local area. The large incumbent firm is important for the production of potential spin-off firms, while the local cluster of start-ups provides the capacity for the spin-off to take root. This pattern relates to the themes about internal and external innovation that are discussed in the next section. In a related study, Glaeser et al. (2015) depicts how US cities built up around historical mines developed industrial structures that were dominated by large incumbent firms that were not conducive to these types of dynamics. There is tremendous scope for more research in this regard.²³

²³ Broader studies of local entrepreneurship rates and clusters include Glaeser et al. (1992, 2010), Glaeser and Kerr (2009), and Delgado et al. (2010). It is important to recognize that the agglomeration of entrepreneurship for innovation and growth-oriented firms does not apply more broadly, where studies instead document a local bias of entrepreneurship (e.g., Michelacci and Silva, 2007).

Going forward, there is an unresolved tension about market sizes of cities and innovation. At one extreme, a necessary foundation for agglomeration is the efficient trade of goods over distances. Detroit can innovate in cars for a worldwide market, and Hollywood can do the same for movies, because of the trade that follows. This logic would suggest that city size does not influence innovation because of market size forces (i.e., larger markets create greater incentives to innovate), at least in the United States today. Yet, there are clear deviations from this frictionless starting point. Especially if property rights across places are not secure or are differentially enforced (e.g., developing economies, early US history, and sectors with rapid imitation), local market sizes may be what is relevant for inventors. Thomas Edison's aggressive litigation to support his own property rights was one of the forces that pushed the film industry to the West Coast. More broadly, greater attention needs to be paid to which market sizes are important for firms in their innovation choices and how this is changing over time.

6.4.6 "Natural advantages" for innovation

Traditional discussions of natural advantages focus on geographic features such as harbors and coal mines. For innovation, "natural advantages" can also form in terms of institutions, universities, and similar relatively fixed factors that are required for innovation. These factors are unevenly distributed over places and will continue to be so for some time horizon into the future, and clusters of innovation can form around them. We group these determinants separately from the more generic education and skills noted in the work previously mentioned.

A very visible factor in this regard is the location of research universities and academic R&D (as well as military facilities). An agglomeration of innovation can form around these points, as firms seek to benefit from spillovers out of the university (e.g., [Jaffe, 1989](#); [Anselin et al., 1997](#)). Universities are key players not only in creating new knowledge through the basic research produced by their faculties but also in supplying a pool of knowledge workers on which innovative activity depends. Silicon Valley and Boston became important centers for innovation in part because of their proximity to Stanford University and MIT (e.g., [Saxenian, 1994](#); [Lee and Nicholas, 2012](#)).

There is general empirical evidence that R&D at local universities is important for firms' innovative activity. [Audretsch and Feldman \(1996\)](#) and [Anselin et al. \(1997\)](#) found localized knowledge spillovers from university R&D to commercial innovation by private firms, even after controlling for the location of industrial R&D. [Andersson et al. \(2009\)](#) found evidence that the expansion of the number of university-based researchers in a local labor market is positively associated with an increase in the number of patents granted in that area. [Agrawal and Cockburn \(2003\)](#) argued that local academic R&D is likely to be more productive, in terms of its contribution to additional patents, in the presence of a large research-intensive firm located nearby—the anchor tenant hypothesis.

Taking this effect into account, they reported a significant positive correlation between local patents and academic publications in the fields of medical imaging, neural networks, and signal processing. However, [Carlino et al. \(2007\)](#) found that R&D at local universities has only modest effects on local innovative activity. They found that a 10% increase in R&D intensity at local universities is associated with less than 1% increase in patent intensity.

Recent efforts pursue causal identification of the role of universities through shocks in their activity levels. [Kantor and Whalley \(2014a\)](#) use interactions of lagged university endowments and stock market fluctuations to generate exogenous changes in local university activity. They find that universities have significant spillovers to their surrounding local areas, especially when they are engaged in research efforts and that the long-term composition of local activity may come to reflect the university's traits and spillovers. In a similar manner, [Hausman \(2012\)](#) quantified how US universities stimulate nearby economic activity using the interaction of a national shock to the spread of innovation from universities—the Bayh–Dole Act of 1980—with the initial academic strengths of universities and variations across universities in federal research funding. Hausman found a connection to the long-run growth of the areas surrounding the universities, with particular complementarity evident for large establishments.

Related to these knowledge flows is the specific knowledge contained within scientists and inventors who are also unevenly distributed. [Zucker et al. \(1998\)](#) provided one of the earliest expressions of this phenomenon in the extreme concentration of early biotechnology activity around the locations of star scientists in the nascent field. On a more general level, studies such as [Breschi and Lissoni \(2009\)](#) and [Singh and Agrawal \(2011\)](#) point to substantial concentration of knowledge exchanges that are embedded in the workers moving across firms, most frequently at local distances. [Breschi and Lissoni \(2009\)](#) found that inventor mobility accounts for most of what would be perceived as knowledge flows in their sample. The concentration of these workers has many antecedents similar to the standard labor pooling Marshallian economies (e.g., specialized skills and general investments in knowledge).

Agglomeration of innovation in the United States and some other countries also has another specific factor influencing its spatial location. The United States is very dependent upon immigrants for its innovation workforce, and these workers have general ties to certain US cities (e.g., the Chinese population in San Francisco). As a consequence, adjustments in national immigration admissions differentially influence the rates of innovation in local areas of the country (e.g., [Hunt and Gauthier-Loiselle, 2010](#); [Kerr and Lincoln, 2010](#)). [Kerr \(2010b\)](#) described how the contributions of Chinese and Indian scientists and engineers to US technology formation increased dramatically in the 1990s. At the same time, these ethnic inventors became more spatially concentrated across US cities. The combination of these two factors reversed long-term declines in the overall agglomeration of US inventors evident in the 1970s and 1980s. Similar effects

are discussed in Europe, with a particular link to fostering local diversity (e.g., [Ozgen et al., 2011](#); [Parrotta et al., 2014](#); [Nathan, 2015](#)).

Looking across this work, fixed factors and “natural advantages” play a clear role in the formation of clusters. However, the extent of spatial concentration in innovation is too much to be explained only by these exogenous spatial differences in natural advantages ([Buzard and Carlino, 2013](#); [Murata et al., 2014](#)). More research is particularly needed on the long-term dynamics of these local spillovers. The study of nineteenth-century agricultural experiment stations by [Kantor and Whalley \(2014b\)](#) is an important step in this regard.

6.5. ADDITIONAL ISSUES ON INNOVATION AND AGGLOMERATION

In this section, we discuss some additional, interesting topics related to the agglomeration of innovation and entrepreneurship. These topics often reach into adjacent fields.

6.5.1 Sustaining mechanisms

There are several additional factors that are frequently observed in the operation of successful clusters for innovation. These factors are often linked to the persistence and growth of clusters, rather than the start of clusters, and thus, we consider them separately from the previously mentioned origination theories. In most cases, these factors have the intuitive feel of the cluster feeding back upon itself.

The most natural expression of this sustaining role is localized spin-offs from innovative incumbent firms. [Klepper \(2010\)](#) documented this process in detail for the two very well-known examples of Silicon Valley and Detroit. Fairchild Semiconductor holds a central place in Silicon Valley lore for its pioneering work on transistors and integrated circuit manufacturing. Fairchild was itself a spin-off from Shockley Semiconductor Laboratory, and various reports estimate that some 400 or more ventures subsequently spun out of Fairchild, including Intel and Advanced Micro Devices. [Klepper \(2010\)](#) described how a similar process occurred in the development of Detroit. The [Duranton \(2007\)](#) model of cluster migration to where new innovations occur can be related to this process.

Several parallel strands to these case studies exist in recent work. [Gompers et al. \(2005\)](#) quantified that the spawning creation of VC-backed entrepreneurs is most aided by the presence of local publicly listed companies that were themselves VC-backed earlier. [Chatterji \(2009\)](#) found that spawned firms perform better than new entrants in the medical devices industry, but this is more due to nontechnical transfers than technical knowledge.

Entrepreneurial finance is also thought to play a role in the agglomeration of innovation. We noted financial intermediaries earlier as a shared input, but many features have this feedback mechanism. Studies find spatial proximity to be important between young companies and their financiers, including banks ([Petersen and Rajan, 1994](#)), angel

investors (Kerr et al., 2014b), and VC firms (Chen et al., 2010). Proximity can help with the screening of these ventures, monitoring of the portfolio, transferring of value-added advice, and similar activities. Proximity also allows the formation of reputations and repeated interactions that are useful in situations with asymmetrical information and moral hazard. As the markets can also be thin—e.g., the number of very qualified VC investors is limited—the formation of a critical mass of these investors can create a sustaining force for a cluster. Angel investing—e.g., the Google millionaires funding the next generation of Silicon Valley start-ups—is a particularly strong expression of this feedback loop.

Finally, differences may exist in the cultures of cities for entrepreneurship and innovation. Landier (2005) provided a formal model of an entrepreneurial culture, which is quite different across countries and cities, and Manso (2011) depicted the need for tolerance for failure in innovation and compensation practices. It is often noted that the legal or reputational costs of failure in many places are so high that entrepreneurs get only one shot, while in Silicon Valley, an early failure merely makes an entrepreneur “experienced.” Qualitative discussions at the city level date back to at least Chinitz (1961), and Saxenian (1994) described strong cultural differences between Boston and Silicon Valley in her arguments for why semiconductors grew at a more rapid pace in the latter. Florida (2005) also depicted large differences across cities in their organization and cultural embrace of innovation. Recent contributions include Hofstede (2001) and Falck et al. (2009).

6.5.2 Policy choices

Local and national policies and regulations play an important role in the agglomeration of innovation. At the local level, Singh and Marx (2013) showed the continued importance of political boundaries even after using continuous-distance techniques to model spillovers.

The most prominent example in the United States is state-level differences in noncompete and nondisclosure clauses, which limit the ability of workers to start a new venture that could compete with their past employer. Many states allow such employment contracts, with California being an important exception. Employers may use these clauses to protect proprietary knowledge from leaking to rival firms. While Fallick et al. (2006) found mixed evidence about the importance of these clauses, Marx et al. (2009) found sharper evidence of a decreased mobility of inventors specifically when the clauses are feasible. Bozkaya and Kerr (2014) more broadly show how rigid employment law can hinder the development of innovative sectors that rely on rapid labor turnover.

State-level bankruptcy laws also appear to matter (e.g., Berkowitz and White, 2004), although this impact appears to be determined through the competing forces of greater willingness to take on risk for innovation versus greater creditor caution in lending with

stronger debtor rights. In a recent contribution, [Cerqueiro et al. \(2014\)](#) find that stronger debtor protection decreases the number of patents produced by small firms and makes them less explorative, especially in financially dependent sectors, consistent with a greater wariness on the part of creditors.

Beyond specific policies such as these two examples, general local regulations and investments play important roles. Much of our discussion connects agglomeration of innovation with local human capital levels, and education investments (both K–12 and tertiary) play an important role in this population base. Likewise, external infrastructure can play an important role for innovation and entrepreneurship. Zoning laws play a clear role at the neighborhood level and perhaps higher.

Many policymakers suggest that they want their cities to become “the next Silicon Valley.” Policymakers often view the success of areas such as Silicon Valley as a recipe for local economic development and growth. This chapter has repeatedly emphasized the substantial agglomeration of activity related to innovation. This agglomeration, of course, means that not every city, or even the majority of them, can become a central hub of innovation, such as Austin, TX. Moreover, [Wilson \(2009\)](#) found that efforts by US states to expand R&D tax credits to lure innovative efforts succeeded but mainly by attracting R&D efforts away from neighboring states (“beggar thy neighbor” effects). For these reasons and others, innovation policy may require coordination across regions or national priorities.

More generally, the ecosystems that underlie clusters of innovation are difficult to create, and there is a very questionable record of targeted government interventions to do so ([Lerner, 2009](#); [Chatterji et al., 2014](#)). As [Duranton \(2011\)](#) emphasizes, the coordination problems and market failures are difficult for policymakers to solve, and the benefits derived may be too small to justify the costs. Perhaps, the best advice is to focus on getting the general conditions correct—e.g., functioning local infrastructure, efficient labor regulations, and good schools. These efforts to lower the costs of entry into entrepreneurship and innovative activities may be the most powerful and make good sense regardless. Other chapters in this volume more systematically take up the roles of urban policy and placemaking efforts.

6.5.3 Connections across clusters

Recent research explores the special connections that exist across clusters of innovation. The backdrop for this work is the set of early economic models that focus on a worldwide technology frontier (e.g., [Mankiw et al., 1992](#); Heckscher–Ohlin–Vanek trade models) or technology diffusion processes that are governed only by distance and time (e.g., S-curve diffusions in concentric rings). Empirical work shows, however, that global R&D spillovers and knowledge transfers are stronger when countries share a common language or ethnic bond, even conditional on spatial distances (e.g., [Jaffe and Trajtenberg, 1999](#); [Keller, 2004](#)). This line of research is being further pushed down

to the subnational level to consider special relationships across global cities and their impacts on the heterogeneous transmission of knowledge.

This research has been most frequently connected to immigrant inventors and innovators in an advanced country and their potential interactions with their home countries, although other networks are quite likely too (e.g., alumni and professional networks). The particular focus on immigrants is due to (1) their importance for technology development in the United States as noted previously and (2) the desire by labor and development economists to understand whether this global migration is a net benefit or cost to the sending country (i.e., the brain drain vs. brain gain debate). Despite originating in adjacent fields, this work increasingly speaks to connections across cities for innovation purposes.

A prominent voice in this area is Saxenian, who discusses in many settings the special relationships that exist between Silicon Valley and rising Asian tiger economies in countries such as Taiwan, Korea, and Bangalore, India. This is effectively city-to-city flows given the extreme spatial concentration of innovation in emerging economies. [Saxenian et al. \(2002\)](#) surveyed immigrant scientists and engineers in Silicon Valley to understand the linkages they maintain with their home countries. These surveys, while very unrepresentative for all immigrants even within Silicon Valley, offer some sense of the relative forms of these contributions. Eighty-two percent of the Chinese and Indian immigrant workers in the science, technology, engineering, and mathematics disciplines report exchanging technical information with their respective nations, roughly 50% of the immigrants report aiding the development of contracts or business relationships between the United States and their home countries, and 18% report investing in overseas business partnerships. While the magnitudes of the figures should be treated with heavy caution, as the sample comes from immigrant groups in Silicon Valley who are seeking to maintain global ties (e.g., The Indus Entrepreneurs), the relative likelihood of the various channels reported is interesting.

Moving beyond case studies and surveys, empirical work using patent citation data suggests that immigrant scientists and engineers in the United States do aid technology transfer to their home countries, albeit at a magnitude that is significantly smaller than what the showcase examples suggest. [Kerr \(2008\)](#) emphasized that these ethnic transmission channels are particularly powerful in the first 5–7 years after a new technology is developed. [Agrawal et al. \(2011\)](#) found that the Indian diaspora in the United States aids its home country the most with the development of big, highly cited inventions but that the diaspora is not as helpful in the production of average inventions as a larger domestic base of inventors.²⁴

²⁴ Related work explores these connections in the specific forms of enhanced trade flows, foreign direct investments by multinational companies, and outsourcing contract placement. These ethnic-based placements are particularly strong in industries and fields related to innovation. [Kerr \(2013\)](#) provided a review of this work and the particular role of diaspora connections across clusters.

While these studies have been important for establishing the connections of innovation clusters across countries, they have noticeable gaps. Perhaps, the most prominent gap is a precise articulation of what form of migration matters for the connections. There are many different varieties: permanent migration to the new economy, return migration back to the home countries, regular travel across places, and so forth. [Saxenian \(2006, 2008\)](#) particularly emphasized the latter with her colorful moniker Global Argonauts, but we have very little data that allow the systematic quantification of these features and their relative importance. Data development in this regard is most warranted, and [Hovhannisyan and Keller \(2010\)](#) provided a recent example with respect to business travel.

Second, these studies frequently have less to offer as to how the traits of places shape these connections. [Nanda and Khanna \(2012\)](#) provided a rare example of this form of study. They survey entrepreneurs in India who were members of the National Association of Software and Services Companies trade association. They find that entrepreneurs who had lived outside of India previously are more likely to activate overseas connections when living outside of the prominent software hubs. In this sense, their work pointed to the social networks as substituting for domestic institutions and capabilities when they are very weak in India.

It is hoped that more research refines our understanding of how global agglomerations for innovation relate to each other and how the traits of cities in advanced countries and developing nations modulate these connections. Given the increasing globalization of innovation, these factors will increasingly shape tomorrow's economic geography of innovation.

6.5.4 Innovation in entrepreneurial clusters versus large firms

This chapter has mostly focused on the development, organization, and interactions of clusters of small entrepreneurial firms. Yet, this intense local environment of start-ups is not the only way that innovation can be organized, and many examples of large firms and R&D laboratories abound. In this section, we review several pieces that consider this organizational choice. The studies that we cite are just the tip of the iceberg in regard to large company innovation.

What factors might favor a large, isolated R&D laboratory over a connected cluster of small firms? The discussion in [Section 6.4](#) identifies several initial factors. The larger laboratory can be preferred when the gains from modularization and experimentation are small, when firm investment in worker human capital is central, when the firm is sufficiently large to bring typically shared resources in-house (e.g., general counsel lawyers), and so on. Moreover, to the extent that a branch of innovation requires large fixed costs, the bigger entity is more likely. It is important to note that the fixed costs relate to the innovation (e.g., basic R&D) and not to complementary assets (e.g., manufacturing and distribution capabilities). As the biotech–pharma world illustrates, a cluster of smaller start-ups can connect on the innovation side to large companies that hold complementary

assets if an effective market for ideas exists (e.g., [Gans et al., 2002](#)). More broadly, researchers since Schumpeter note the longer investment horizons of large firms, their incentives to improve the markups on their products, the surplus resources they can devote to the innovative efforts, and so on. Very clearly, the likelihood of engaging in R&D or patenting increases with firm size (e.g., [Cohen and Klepper, 1996a](#); [Akcigit and Kerr, 2010](#)).

On the other hand, extensive work in the management literature also emphasizes limits for innovations at large companies. This has often been expressed in the exploration versus exploitation distinction highlighted earlier. Large companies are frequently accused of having overreaching bureaucracies that limit innovation. Even if the hierarchies of the company are appropriately sized, larger companies may naturally favor forms of innovation that enhance their current products, spending less time and resources on exploratory work. It is also frequently discussed how separate, independent firms can help resolve incentive and compensation constraints; for example, while an entrepreneur's equity can become worth billions of dollars with a breakthrough discovery, large firms face political constraints on whether they could pay an in-house innovator more than the firm's CEO. Because they have resources and do not need to appeal to external markets, large companies can also struggle with the termination of poorly performing projects. This termination is essential for experimentation to be effective and the generation of sufficient returns to innovative efforts (e.g., [Kerr et al., 2014a](#)).

Several theory models consider explicitly the trade-offs between these alternative landscapes, beyond those noted earlier (e.g., modularity and worker training). [Gromb and Scharfstein \(2002\)](#) modeled the trade-off between the high-powered incentives of independent entrepreneurs and the better retention of knowledge about employees in large firms. If external markets cannot fully discern the reason why past projects failed, large firms can derive advantages from their ongoing interactions and monitoring of workers. The authors show that multiple equilibriums can arise and even coexist. In one setting, entrepreneurship is weak, and so, the market for failed entrepreneurs is quite thin, making internal labor markets and innovation within large firms particularly valuable. In a high entrepreneurship equilibrium, the robust labor market and high incentives encourage innovation outside of large firms.

[Hellmann and Perotti \(2011\)](#) considered a second trade-off. In their model, innovative ideas are initially incomplete and must be circulated to provide complementary pieces. Although circulating ideas is essential, innovators face the risks that their ideas will be stolen by those seeing them. Large markets of firms provide innovators with better prospects for finding the best matches compared with efforts to circulate ideas among a single, large company. On the other hand, the internal sharing of the idea within a large company limits the risk of expropriation compared with exposing ideas through markets. Similar to [Gromb and Scharfstein \(2002\)](#), this model can lead to the coexistence of multiple organizational forms related to innovation (internal ventures, spinouts, and start-ups), with the possibility in the [Hellmann and Perotti \(2011\)](#) model of these forms mutually reinforcing each other.

Markets for idea generation work best when the cost of idea generation is low. [Hellmann and Perotti \(2011\)](#) further discussed the literature in this area.

Empirically, [Agrawal et al. \(2010\)](#) quantified how the insular nature of “company towns” affects the types of innovations undertaken locally. Large, isolated plants are more likely to take on lines of innovation that exploit the firm’s past work and reflect less on the newer technologies being created. This work connects with the studies of firm size and innovation choices (e.g., [Rosen, 1991](#); [Akcigit and Kerr, 2010](#)) and the [Duranton and Puga \(2001\)](#) model of nursery cities. Further, development of these cross correlations of city types, firm sizes, and innovation choices will greatly enhance our understanding of these issues.

6.5.5 Multiunit firm locations and agglomeration clusters

A significant body of work in the management literature touches on the location choices of multiunit firms. In an important sense, this work connects to the industrial organization issues identified earlier. This research often uses a foreign direct investment lens, since the overseas operations of multinational firms are feasible to identify. [Alcacer and Chung \(2007, 2014\)](#) provided examples of work that specifically relate these placement decisions to local agglomeration benefits, and [Keller and Yeaple \(2013\)](#) described knowledge flows in particular. These papers also outline the longer earlier literature. [Choudhury \(2014\)](#) and [Choudhury and Khanna \(2014\)](#) flip the lens around and demonstrate in India how the movement of resources and labor in multiunit and multinational firms affects the innovation of the local areas in which these firms operate.

As greater data on plant-level operations have become available, researchers have also considered the optimal placement of domestic facilities with respect to agglomeration. [Alcacer and Delgado \(2013\)](#) contrasted intrafirm linkages that multiunit firms have with the localized external agglomeration economies available in their areas. They found that intrafirm linkages matter a lot and that empirical estimates of agglomeration economies might be overstated as a consequence of not sufficiently modeling them. [Tecu \(2012\)](#) considered how multiunit firms place their industrial innovation. She too found that internal linkages among the plants of a firm matter a lot. Her estimates suggest that the elasticity of R&D productivity with respect to the firm’s production workers is almost as large as the elasticity with respect to the host city’s total patents. These internal–external studies remain relatively rare, yet the existing work suggests great promise for future work.

6.6. CONCLUSIONS

[Marshall \(1890\)](#) noted long ago that, in industrial clusters, “[t]he mysteries of the trade become no mysteries; but are as it were in the air, and children may learn them unconsciously” (p. 271). Much progress has been made since Marshall made this famous statement about understanding exactly how agglomeration, knowledge spillovers, and

innovation fit together and what their impact on local and national economies might be. Over the past two decades, research in this field has particularly surged. This has been unquestionably due in large part to the substantial data now available to better study these questions. It also surely has some connection to the growth of the knowledge economy, the fascination with Silicon Valley and related clusters, and similar popular themes. Particularly, encouraging is the extensive set of connections being made from economic geography to adjacent fields such as labor economics, entrepreneurial finance, and business management.

Much, however, remains to be accomplished. With the handful of important exceptions reviewed previously, we still have not opened the black box of how clusters operate. Most of our empirical work has instead been comparisons over places. Better empirical guidance about the microinteractions within clusters with respect to innovation will allow us to differentiate among models and build stronger theoretical frameworks. Fortunately, the data are within reach to do so. Many employer–employee data sets are being linked to information on innovation (e.g., inventor identifiers), and these data sets often allow precise spatial coordinates of firms. This should prove a powerful lever for seeing further inside local areas, and it is likely that the line of work will increasingly draw upon network theory (e.g., [Jackson, 2008](#); [Dempwolf, 2012](#); [Kerr and Kominers, 2015](#)).

We also need better insight into the long-term nature of agglomeration and innovation—the life cycles of innovative places. This is not only true within countries—innovation cores have shifted between Detroit, Boston, Silicon Valley, etc. and will continue to do so—but also true across countries. How does the rise of Bangalore impact Boston? Does the development of global innovation centers in rapidly emerging countries complement or substitute for those in advanced economies? The study of agglomeration and innovation is exciting because we are starting to make some progress at understanding the local and global nature of these phenomena.

Finally (although by no means closing off a comprehensive list), we need to develop better guidance about what local policymakers can do to foster agglomeration and innovation in their cities. This is a big and difficult question, and we are right to be cautious that we do not have all the answers. However, governments have spent billions of dollars on this, and many will continue to fund “be the next Silicon Valley” initiatives. This is true in advanced economies, in nations currently looking to transition from resource-dependent to a knowledge-based economy, in developing countries looking to leapfrog growth stages, and everywhere in between. Economists must continue to provide insight on these critical matters, and ideally, our insights will get sharper faster.

ACKNOWLEDGMENTS

We thank Gilles Duranton, Vernon Henderson, and Will Strange for their editorial guidance and Ufuk Akcigit and seminar participants for their comments. The views expressed in this chapter are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

REFERENCES

- Acemoglu, A., Angrist, J., 2000. How large are human-capital externalities? Evidence from compulsory schooling laws. In: Bernanke, B., Rogoff, K. (Eds.), *NBER Macroeconomic Annuals*. MIT Press, Cambridge, MA, pp. 9–74.
- Acemoglu, A., Akcigit, U., Bloom, N., Kerr, W., 2014. Innovation, reallocation, and growth. NBER Working paper 18993.
- Acs, Z., Audretsch, D., 1988. Innovation in large and small firms: an empirical analysis. *Am. Econ. Rev.* 78, 678–690.
- Acs, Z., Audretsch, D., Feldman, M., 1994. Resource and output trends in the United States since 1870. *Am. Econ. Rev.* 46, 5–23.
- Acs, Z., Anselin, L., Varga, A., 2002. Patents and innovation counts as measures of regional production of new knowledge. *Res. Policy* 3, 1069–1085.
- Adams, J., Jaffe, A., 1996. Bounding the effects of R&D: an investigation using matched establishment–firm data. *RAND J. Econ.* 27 (4), 700–721.
- Aghion, P., Howitt, P., 1992. A model of growth through creative destruction. *Econometrica* 60, 323–351.
- Aghion, P., Akcigit, U., Howitt, P., 2014. What do we learn from Schumpeterian growth theory? NBER Working paper 18824.
- Agrawal, A., Cockburn, I., 2003. The anchor tenant hypothesis: exploring the role of large, local, R&D-intensive firms in regional innovation systems. *Int. J. Ind. Organ.* 21 (9), 1217–1253.
- Agrawal, A., Cockburn, I., Rosell, C., 2010. Not invented here? Innovation in company towns. *J. Urban Econ.* 67 (1), 78–89.
- Agrawal, A., Kapur, D., McHale, J., Oettl, A., 2011. Brain drain or brain bank? The impact of skilled emigration on poor-country innovation. *J. Urban Econ.* 69, 43–55.
- Agrawal, A., Cockburn, I., Galasso, A., Oettl, A., 2014. Why are some regions more innovative than others? The role of firm size diversity. *J. Urban Econ.* 81 (1), 149–165.
- Akcigit, U., Kerr, W., 2010. Growth through heterogeneous innovations. NBER Working paper 16443.
- Albouy, D., 2009. The unequal geographic burden of federal taxation. *J. Polit. Econ.* 117 (4), 635–667.
- Alcacer, J., Chung, W., 2007. Location strategies and knowledge spillovers. *Manag. Sci.* 53 (5), 760–776.
- Alcacer, J., Chung, W., 2014. Location strategies for agglomeration economies. *Strateg. Manag. J.* 35, 1749–1761.
- Alcacer, J., Delgado, M., 2013. Spatial organization of firms and location choices through the value chain. HBS Working paper 13-025.
- Andersson, R., Quigley, J., Wilhelmsson, M., 2009. Higher education, localization and innovation: evidence from a natural experiment. *J. Urban Econ.* 66 (1), 2–15.
- Anselin, L., Varga, A., Zoltan, A., 1997. Local geographic spillovers between university research and high technology innovations. *J. Urban Econ.* 42, 442–448.
- Arellano, M., Bond, S., 1990. Some tests of specification for panel data: monte carlo evidence and an application to employment equations. *Rev. Econ. Stud.* 58 (2), 277–297.
- Arzaghi, M., Henderson, J.V., 2008. Networking off Madison Avenue. *Rev. Econ. Stud.* 75, 1011–1038.
- Audretsch, D., Feldman, M., 1996. R&D spillovers and the geography of innovation and production. *Am. Econ. Rev.* 86, 630–640.
- Audretsch, D., Feldman, M., 2004. Knowledge spillovers and the geography of innovation. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Urban and Regional Economics*, vol. 4. North-Holland, Amsterdam, pp. 2713–2739.
- Azoulay, P., Zivin, J., Wang, J., 2010. Superstar extinction. *Q. J. Econ.* 125 (2), 549–589.
- Bacolod, M., Blum, B., Strange, W., 2010. Elements of skill: traits, intelligences, education, and agglomeration. *J. Reg. Sci.* 50 (1), 245–280.
- Bairoch, P., 1988. *Cities and Economic Development*. University of Chicago Press, Chicago.
- Baldwin, C., Clark, K., 1997. Managing in an age of modularity. *Harv. Bus. Rev.* 75 (September–October), 84–93.
- Baum-Snow, N., Pavan, R., 2013. Inequality and city size. *Rev. Econ. Stat.* 95 (5), 1535–1548.
- Berkowitz, J., White, M., 2004. Bankruptcy and small firms' access to credit. *RAND J. Econ.* 35, 69–84.

- Berliant, M., Reed, R., Wang, P., 2006. Knowledge exchange, matching, and agglomeration. *J. Urban Econ.* 60, 69–95.
- Billings, S., Johnson, E., 2014. Agglomeration Within an Urban Area. Mimeo (University of North Carolina Charlotte).
- Black, D., Henderson, J.V., 1999. A theory of urban growth. *J. Polit. Econ.* 107, 252–284.
- Bozkaya, A., Kerr, W., 2014. Labor regulations and European venture capital. *J. Econ. Manag. Strategy* 23, 776–810.
- Breschi, S., Lissoni, F., 2009. Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows. *J. Econ. Geogr.* 9 (4), 439–468.
- Briant, A., Combes, P.P., Lafourcade, M., 2010. Dots to boxes: do the size and shape of spatial units jeopardize economic geography estimates? *J. Urban Econ.* 67, 287–302.
- Buzard, K., Carlino, G., 2013. The geography of research and development activity in the U.S. In: Giarratani, F., Hewings, G., McCann, P. (Eds.), *Handbook of Economic Geography and Industry Studies*. Edward Elgar, London.
- Buzard, K., Carlino, G., Carr, J., Hunt, R., Smith, T., 2015. Localized knowledge spillovers: evidence from the agglomeration of American R&D labs and patent data. Federal Reserve Bank of Philadelphia Working paper 15-03.
- Capello, R., Lenzi, C., 2014. Spatial heterogeneity in knowledge, innovation, and economic growth nexus: conceptual reflections and empirical evidence. *J. Reg. Sci.* 54 (2), 186–214.
- Carlino, G., Hunt, R., 2009. What explains the quantity and quality of local inventive activity? In: Burtless, G., Pack, J.R. (Eds.), *Brookings-Wharton Papers on Urban Affairs*. Brookings Institution Press, Washington, D.C.
- Carlino, G., Chatterjee, S., Hunt, R., 2007. Urban density and the rate of invention. *J. Urban Econ.* 61 (3), 389–419.
- Carrincazeaux, C., Lunga, Y., Rallet, A., 2001. Proximity and localisation of corporate R&D activities. *Res. Policy* 30, 777–789.
- Cassey, A., Smith, B., 2014. Simulating confidence for the Ellison–Glaeser index. *J. Urban Econ.* 81, 85–103.
- Cerqueiro, G., Hegde, D., Penas, M., Seamans, R., 2014. Debtor rights, credit supply, and innovation. Tilburg Law and Economics Center Discussion Paper No. 2014-011.
- Chatterji, A., 2009. Spawned with a silver spoon? Entrepreneurial performance and innovation in the medical device industry. *Strateg. Manag. J.* 30 (2), 185–206.
- Chatterji, A., Glaeser, E., Kerr, W., 2014. Clusters of entrepreneurship and innovation. In: Lerner, J., Stern, S. (Eds.), *Innovation Policy and the Economy*, vol. 14. University of Chicago Press, Chicago, IL, pp. 129–166.
- Chen, H., Gompers, P., Kovner, A., Lerner, J., 2010. Buy local? The geography of venture capital. *J. Urban Econ.* 67, 90–102.
- Chinitz, B., 1961. Contrasts in agglomeration: New York and Pittsburgh. *Am. Econ. Rev.* 51 (2), 279–289.
- Choudhury, P., 2014. Return migration and geography of innovation in MNEs: a natural experiment of on-the-job learning of knowledge production by local workers reporting to return migrants. HBS Working paper 14-078.
- Choudhury, P., Khanna, T., 2014. The role of firms in fostering within country migration: evidence from a natural experiment in India. HBS Working paper 14-080.
- Christensen, C., 1997. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press, Boston, MA.
- Ciccone, A., Peri, G., 2006. Identifying human-capital externalities: theory with applications. *Rev. Econ. Stud.* 73, 381–412.
- Cohen, W., Klepper, S., 1996a. A reprise of size and R&D. *Econ. J.* 106 (437), 925–951.
- Cohen, W., Klepper, S., 1996b. Firm size and the nature of innovation within industries: the case of process and product R&D. *Rev. Econ. Stat.* 78, 232–243.
- Cohen, W., Nelson, R., Walsh, J., 2000. Protecting their intellectual assets: appropriability conditions and why U.S. manufacturing firms patent (or not). NBER Working paper 7552.

- Combes, P., Duranton, G., 2006. Labour pooling, labour poaching and spatial clustering. *Reg. Sci. Urban Econ.* 36 (1), 1–28.
- Combes, P., Gobillon, L., 2015. The empirics of agglomeration economies. In: Henderson, J.V., Duranton, G., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. North Holland, Amsterdam, pp. 247–348.
- Conley, T., Flyer, F., Tsiang, G., 2003. Spillovers from local market human capital and the spatial distribution of productivity in Malaysia. *Adv. Econ. Anal. Policy* 3 (1), 1–45.
- Davis, D., Dingel, J., 2013. A spatial knowledge economy. NBER Working paper 18188.
- de la Roca, J., Puga, D., 2012. Learning by working in big cities. CEPR Discussion Paper 9243.
- Delgado, M., Porter, M., Stern, S., 2010. Clusters and entrepreneurship. *J. Econ. Geogr.* 10 (4), 495–518.
- Dempwolf, C., 2012. A Network Model of Regional Innovation Clusters and Their Influence on Economic Growth. Ph.D. Dissertation, University of Maryland.
- Diamond, R., 2012. The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000. Mimeo (Graduate School of Business, Stanford University).
- Directory of American Research and Technology, 1998. Organizations Active in Product Development for Business, 23rd ed. R.R. Bowker, New York.
- Duranton, G., 2006. Human capital externalities in cities. In: Arnott, R., McMillen, D. (Eds.), *A Companion to Urban Economics*. Blackwell Publishing, Ltd., Malden, MA, pp. 24–39.
- Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. *Am. Econ. Rev.* 97, 197–221.
- Duranton, G., 2011. California dreamin': the feeble case for cluster policies. *Rev. Econ. Anal.* 3, 3–45.
- Duranton, G., Overman, H., 2005. Testing for localization using micro-geographic data. *Rev. Econ. Stud.* 72, 1077–1106.
- Duranton, G., Puga, D., 2001. Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am. Econ. Rev.* 91, 1454–1477.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Urban and Regional Economics*, vol. 4. North-Holland, Amsterdam, pp. 2063–2117.
- Duranton, G., Puga, D., 2014. The growth of cities. In: Aghion, P., Durlauf, S. (Eds.), *Handbook of Economic Growth*, vol. 2. North-Holland, Amsterdam, pp. 781–853.
- Eaton, J., Eckstein, Z., 1997. Cities and growth: theory and evidence from France and Japan. *Reg. Sci. Urban Econ.* 27, 443–474.
- Eberts, R., McMillen, D., 1999. Agglomeration economies and urban public infrastructure. In: Cheshire, P., Mills, E. (Eds.), *Handbook of Regional and Urban Economics*, vol. 3. Elsevier Sciences B.V., New York, pp. 1455–1495.
- Ellison, G., Glaeser, E., 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. *J. Polit. Econ.* 105, 889–927.
- Ellison, G., Glaeser, E., Kerr, W., 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. *Am. Econ. Rev.* 100 (3), 1195–1213.
- Falck, O., Fritsch, M., Heblich, S., 2009. Bohemians, human capital, and regional economic growth. *Jena Economic Research Papers* 2009–049.
- Fallick, B., Fleischman, C., Rebitzer, J., 2006. Job-hopping in Silicon Valley: some evidence concerning the microfoundations of a high-technology cluster. *Rev. Econ. Stat.* 88 (3), 472–481.
- Feldman, M., 1994. *The Geography of Innovation*. Kluwer Academic, Boston, MA.
- Feldman, M., Audretsch, D., 1999. Innovation in cities: science-based diversity, specialization, and localized competition. *Eur. Econ. Rev.* 43, 409–429.
- Feldman, M., Kogler, D., 2010. Stylized facts in the geography of innovation. In: Hall, B., Rosenberg, N. (Eds.), *Handbook of the Economics of Innovation*, vol. 1. Elsevier, Oxford, pp. 381–410.
- Florida, R., 2005. *Cities and the Creative Class*. Routledge, New York.
- Fornahl, D., Brenner, T., 2009. Geographic concentration of innovative activity in Germany. *Struct. Chang. Econ. Dyn.* 20, 163–182.
- Freedman, M., 2008. Job hopping, earnings dynamics, and industrial agglomeration in the software publishing industry. *J. Urban Econ.* 64 (3), 590–600.
- Gans, J., Hsu, D., Stern, S., 2002. When does start-up innovation spur the gale of creative destruction? *RAND J. Econ.* 33 (4), 571–586.

- Gerlach, H., Ronde, T., Stahl, K., 2009. Labor pooling in R&D intensive industries. *J. Urban Econ.* 65 (1), 99–111.
- Gertner, J., 2012. *The Idea Factory*. The Penguin Press, New York.
- Gibson, R., 1999. The legal infrastructure of high technology industrial districts: Silicon Valley, Route 128, and covenants not to compete. *N. Y. Univ. Law Rev.* 74, 575–629.
- Glaeser, E., 1999. Learning in cities. *J. Urban Econ.* 46 (2), 254–277.
- Glaeser, E., 2008. *Cities, Agglomeration and Spatial Equilibrium*. Oxford University Press, Oxford.
- Glaeser, E., Kerr, W., 2009. Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain? *J. Econ. Manag. Strateg.* 18 (3), 623–663.
- Glaeser, E.L., Maré, D., 2001. Cities and skills. *J. Labor Econ.* 19 (2), 316–342.
- Glaeser, E., Saiz, A., 2004. The rise of the skilled city. *Brookings-Wharton Pap. Urban Aff.* 5, 47–94.
- Glaeser, E., Kallal, H., Scheinkman, J., Shleifer, A., 1992. Growth in cities. *J. Polit. Econ.* 100 (6), 1126–1152.
- Glaeser, E., Scheinkman, J., Shleifer, S., 1995. Economic growth in a cross-section of cities. *J. Monet. Econ.* 36, 117–143.
- Glaeser, E., Kerr, W., Ponzetto, G., 2010. Clusters of entrepreneurship. *J. Urban Econ.* 67 (1), 150–168.
- Glaeser, E., Kerr, S., Kerr, W., 2015. Entrepreneurship and urban growth: an empirical assessment with historical mines. *Rev. Econ. Stat.*, forthcoming.
- Gompers, P., Lerner, J., 2001. The venture capital revolution. *J. Econ. Perspect.* 15 (2), 169–192.
- Gompers, P., Lerner, J., 2006. *The Venture Capital Cycle*, second ed. MIT Press, Boston, MA.
- Gompers, P., Lerner, J., Scharfstein, D., 2005. Entrepreneurial spawning: public corporations and the genesis of new ventures, 1986 to 1999. *J. Financ.* 60 (2), 577–614.
- Griliches, Z., 1979. Issues in assessing the contribution of research and development to productivity growth. *Bell J. Econ.* 10, 92–116.
- Griliches, Z., 1990. Patent statistics as economic indicators: a survey. *J. Econ. Lit.* 28, 1661–1707.
- Gromb, D., Scharfstein, D., 2002. Entrepreneurship in equilibrium. NBER Working paper 9001.
- Grossman, G., Helpman, E., 1991. Quality ladders in the theory of growth. *Rev. Econ. Stud.* 58 (1), 43–61.
- Hagedoorn, J., Cloudt, M., 2003. Measuring innovative performance: is there an advantage in using multiple indicators? *Res. Policy* 32, 1365–1379.
- Hall, B., Jaffe, A., Trajtenberg, M., 2001. The NBER patent citation data file: lessons, insights and methodological tools. NBER Working paper 8498.
- Ham, J., Weinberg, B., 2014. *Geography and Innovation: Evidence from Nobel Laureate*. Mimeo (Economics Department, The Ohio State University).
- Harhoff, D., Narin, F., Scherer, F., Vopel, K., 1999. Citation frequency and the value of patented inventions. *Rev. Econ. Stat.* 81, 511–515.
- Hausman, N., 2012. University innovation, local economic growth, and entrepreneurship. US Census Bureau Center for Economic Studies Paper No. CES-WP- 12-10.
- Hellmann, T., Perotti, E., 2011. The circulation of ideas in firms and markets. *Manag. Sci.* 57 (10), 1813–1826.
- Helsley, R., Strange, W., 1990. Matching and agglomeration economies in a system of cities. *Reg. Sci. Urban Econ.* 20 (2), 189–212.
- Helsley, R., Strange, W., 2002. Innovation and input sharing. *J. Urban Econ.* 51, 25–45.
- Helsley, R., Strange, W., 2004. Knowledge barter in cities. *J. Urban Econ.* 56 (2), 327–345.
- Henderson, J.V., 1974. The size and types of cities. *Am. Econ. Rev.* 61, 640–656.
- Henderson, J.V., 2007. Understanding knowledge spillovers. *Reg. Sci. Urban Econ.* 37 (4), 497–508.
- Hofstede, G., 2001. *Culture and Organizations*. Harper Collins, London.
- Horvath, M., 2001. Imitating Silicon Valley: regional comparisons of innovation activity based on venture capital flows. In: Gambardella, A., Bresnahan, T. (Eds.), *Building High-Tech Clusters: Silicon Valley and Beyond*. Cambridge University Press, Cambridge, MA, pp. 280–330.
- Hovhannisyan, N., Keller, W., 2010. International business travel: an engine of innovation? Center for Economic and Policy Research Discussion Paper No. DP7829.
- Hunt, J., Gauthier-Loiselle, M., 2010. How much does immigration boost innovation? *Am. Econ. J. Macroecon.* 2 (2), 31–56.

- Inoue, H., Nakajima, K., Saito, Y.U., 2013. Localization of collaborations in knowledge creation. The Research Institute of Economy, Trade and Industry Discussion Papers Series No. 13-E-070.
- Jackson, M., 2008. *Social and Economic Networks*. Princeton University Press, Princeton.
- Jacobs, J., 1969. *The Economy of Cities*. Vintage Books, New York.
- Jaffe, A., 1989. Real effects of academic research. *Am. Econ. Rev.* 79, 957–970.
- Jaffe, A., Trajtenberg, M., 1999. International knowledge flows: evidence from patent citations. *Econ. Innov. New Technol.* 8, 105–136.
- Jaffe, A., Trajtenberg, M., Henderson, R., 1993. Geographic localization of knowledge spillovers as evidenced by patent citations. *Q. J. Econ.* 108, 577–598.
- Jaffe, A., Trajtenberg, M., Fogarty, M., 2000. Knowledge spillovers and patent citations: evidence from a survey of inventors. *Am. Econ. Rev.* 90 (2), 215–218.
- Jones, C., 2005. Growth and ideas. In: Aghion, P., Durlauf, S. (Eds.), *Handbook of Economic Growth*, vol. 1B. North-Holland, Amsterdam.
- Kantor, S., Whalley, A., 2014. Knowledge spillovers from research universities: evidence from endowment value shocks. *Rev. Econ. Stat.* 96, 171–188.
- Kantor, S., Whalley, A., 2014b. Research Proximity and Productivity: Long-term Evidence from Agriculture. Mimeo (Economics Department, University of California-Merced).
- Keller, W., 2004. International technology diffusion. *J. Econ. Lit.* 42 (3), 752–782.
- Keller, W., Yeaple, S., 2013. The gravity of knowledge. *Am. Econ. Rev.* 103 (4), 1414–1444.
- Kelly, M., Hageman, A., 1999. Marshallian externalities in innovation. *J. Econ. Growth* 4 (1), 39–54.
- Kerr, W., 2008. Ethnic scientific communities and international technology diffusion. *Rev. Econ. Stat.* 90 (3), 518–537.
- Kerr, W., 2010a. Breakthrough inventions and migrating clusters of innovation. *J. Urban Econ.* 67 (1), 46–60.
- Kerr, W., 2010b. The agglomeration of U.S. ethnic inventors. In: Glaeser, E. (Ed.), *Agglomeration Economics*. University of Chicago Press, Chicago, IL, pp. 237–276.
- Kerr, W., 2013. U.S. high-skilled immigration, innovation, and entrepreneurship: empirical approaches and evidence. NBER Working paper 19377.
- Kerr, W., Kominers, S., 2015. Agglomerative forces and cluster shapes. *Rev. Econ. Stat.*, forthcoming.
- Kerr, W., Lincoln, W., 2010. The supply side of innovation: H-1B visa reforms and U.S. ethnic invention. *J. Labor Econ.* 28 (3), 473–508.
- Kerr, W., Lerner, J., Schoar, A., 2014a. The consequences of entrepreneurial finance: evidence from angel financings. *Rev. Financ. Stud.* 27 (1), 20–55.
- Kerr, W., Nanda, R., Rhodes-Kropf, M., 2014b. Entrepreneurship as experimentation. *J. Econ. Perspect.* 28, 25–48.
- Klepper, S., 2010. The origin and growth of industry clusters: the making of Silicon Valley and Detroit. *J. Urban Econ.* 67, 15–32.
- Kolympiris, C., Kalaitzandonakes, N., Miller, D., 2011. Spatial collocation and venture capital in the US biotechnology industry. *Res. Policy* 40 (9), 1188–1199.
- Krugman, P., 1991. *Geography and Trade*. MIT Press, Cambridge, MA.
- Lamoreaux, N., Levenstein, M., Sokoloff, K., 2004. Financing invention during the second industrial revolution: Cleveland, Ohio, 1870–1920. NBER Working paper 10923.
- Landier, A., 2005. *Entrepreneurship and the Stigma of Failure*. Mimeo (Stern School of Business, New York University).
- Lee, J., Nicholas, T., 2012. The Origins and Development of Silicon Valley. Harvard Business School Case, pp. 813–1098.
- Lerner, J., 2009. *Boulevard of Broken Dreams*. Princeton University Press, Princeton, NJ.
- Lerner, J., Seru, A., 2014. The use and misuse of patent data. In: Paper Presented at the Allied Social Science Association Meeting, January 3–5, 2014, Philadelphia, PA.
- Lin, J., 2011. Technological adaptation, cities, and new work. *Rev. Econ. Stat.* 93 (2), 554–574.
- Lin, J., 2014. “The paper trail of knowledge transfers.” Federal Reserve Bank of Philadelphia Business Review. Second Quarter.
- Lindley, J., Machin, S., 2014. Spatial changes in labour market inequality. *J. Urban Econ.* 85, 121–138.
- Liu, C., 2010. A spatial ecology of structure holes: scientists and communication at a biotechnology firm. *Acad. Manag. Proc.* 1, 1–6.

- Lucas, R.E., 1988. On the mechanics of economic development. *J. Monet. Econ.* 22, 3–42.
- Lychagin, S., Pinkse, J., Slade, M., Van Reenen, J., 2010. Spillovers in space: does geography matter? NBER Working paper No. 16188.
- Magrini, S., 2004. Regional (di)convergence. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Urban and Regional Economics*, vol. 4. North-Holland, Amsterdam, pp. 2741–2796.
- Mankiw, G., Romer, D., Weil, D., 1992. A contribution to the empirics of economic growth. *Q. J. Econ.* 107 (2), 407–437.
- Manso, G., 2011. Motivating innovation. *J. Financ.* 66, 1823–1869.
- March, J., 1991. Exploration and exploitation in organizational learning. *Organ. Sci.* 2 (1), 71–87.
- Marshall, A., 1890. *Principles of Economics*. Macmillan, London.
- Marx, M., Strumsky, D., Fleming, L., 2009. Mobility, skills, and the Michigan non-compete experiment. *Manag. Sci.* 55 (6), 875–889.
- Matouschek, N., Robert-Nicoud, F., 2005. The role of human capital investments in the location decision of firms. *Reg. Sci. Urban Econ.* 35 (5), 570–583.
- Menon, C., 2009. Star and comets: an exploration of the patent universe. London School of Economics SERC Discussion Paper 37.
- Menon, C., 2012. The bright side of MAUP: defining new measures of industrial agglomeration. *Pap. Reg. Sci.* 91 (1), 3–28.
- Michelacci, C., Silva, O., 2007. Why so many local entrepreneurs? *Rev. Econ. Stat.* 89 (4), 615–633.
- Moretti, E., 2004a. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *J. Econ.* 121, 175–212.
- Moretti, E., 2004b. Human capital externalities in cities. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Urban and Regional Economics*, vol. 4. North-Holland, Amsterdam, pp. 2243–2291.
- Moretti, E., 2004c. Workers' education, spillovers and productivity: evidence from plant-level production functions. *Am. Econ. Rev.* 94, 656–690.
- Moretti, E., 2012. *The New Geography of Jobs*. Houghton Mifflin Harcourt, New York.
- Moretti, E., 2013. Real wage inequality. *Am. Econ. J. Appl. Econ.* 5 (1), 65–103.
- Moser, P., 2011. Did patents weaken the localization of innovations? Evidence from World's Fairs. *J. Econ. Hist.* 71 (2), 363–381.
- Murata, Y., Nakajima, R., Okamoto, R., Tamura, R., 2014. Localized knowledge spillovers and patent citations: a distance-based approach. *Rev. Econ. Stat.* 96 (5), 967–985.
- Nanda, R., Khanna, T., 2012. Diasporas and domestic entrepreneurs: evidence from the Indian software industry. *J. Econ. Manag. Strateg.* 19 (4), 991–1012.
- Nathan, M., 2015. Same difference? Minority ethnic inventors, diversity and innovation in the UK. *J. Econ. Geogr.* 15, 129–168.
- OECD, 2005. *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*, third ed. Organization of Economically Developed Countries, Paris.
- Olson, G., Olson, J., 2003. Mitigating the effects of distance on collaborative intellectual work. *Econ. Innov. New Technol.* 12 (1), 27–42.
- Ozgen, C., Nijkamp, P., Poot, J., 2011. Immigration and innovation in European regions. IZA Working paper 5676.
- Pakes, A., Griliches, Z., 1980. Patents and R&D at the firm level: a first report. *Econ. Lett.* 4, 377–381.
- Pakes, A., Nitzan, S., 1983. Optimum contracts for research personnel, research employment, and the establishment of “rival” enterprises. *J. Labor Econ.* 1 (4), 345–365.
- Papageorgiou, T., 2013. Worker sorting and agglomeration economies. Penn State University Working paper.
- Parrotta, P., Pozzoli, D., Pytlíkova, M., 2014. The nexus between labor diversity and firm's innovation. *J. Popul. Econ.* 27, 303–364.
- Petersen, M., Rajan, R., 1994. The benefits of lending relationships: evidence from small business data. *J. Financ.* 49 (1), 3–37.
- Porter, M., 1990. *The Competitive Advantage of Nations*. The Free Press, New York.
- Porter, M., 1998. Clusters and the new economics of competition. *Harv. Bus. Rev.* 76 (November–December), 77–90.
- Rauch, J., 1993. Productivity gains from geographic concentration in cities. *J. Urban Econ.* 34, 380–400.
- Romer, P., 1986. Increasing returns and long-run growth. *J. Polit. Econ.* 94 (5), 1002–1037.

- Romer, P., 1990. Endogenous technical change. *J. Polit. Econ.* 98 (5), S71–S102.
- Rosen, R., 1991. Research and development with asymmetric firm sizes. *RAND J. Econ.* 22 (3), 411–429.
- Rosenthal, S., Strange, W., 2001. The determinants of agglomeration. *J. Urban Econ.* 50, 191–229.
- Rosenthal, S., Strange, W., 2003. Geography, industrial organization, and agglomeration. *Rev. Econ. Stat.* 85 (2), 377–393.
- Rosenthal, S., Strange, W., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam, pp. 2119–2171.
- Rosenthal, S., Strange, W., 2008. The attenuation of human capital spillovers. *J. Urban Econ.* 64 (2), 373–389.
- Rotemberg, J., Saloner, G., 2000. Competition and human capital accumulation: a theory of interregional specialization and trade. *Reg. Sci. Urban Econ.* 30, 373–404.
- Samila, S., Sorenson, O., 2011. Venture capital, entrepreneurship and economic growth. *Rev. Econ. Stat.* 93 (1), 338–349.
- Saxenian, A., 1991. The origins and dynamics of production networks in Silicon Valley. *Res. Policy* 20 (1), 423–437.
- Saxenian, A., 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge, MA.
- Saxenian, A., 2006. *The New Argonauts*. Harvard University Press, Cambridge, MA.
- Saxenian, A., 2008. Transnational communities and the evolution of global production networks: the cases of Taiwan, China and India. *Industry and Innovation* 9, 183–202.
- Saxenian, A., Motoyama, Y., Quan, X., 2002. *Local and Global Networks of Immigrant Professionals in Silicon Valley*. Public Policy Institute of California, San Francisco, CA.
- Scherer, F.M., 1986. *Innovation and Growth*. MIT Press, Cambridge, MA.
- Schumpeter, J., 1934. *The Theory of Economic Development*. Harvard University Press, Cambridge, MA.
- Schumpeter, J., 1939. *Business Cycles: A Theoretical, Historical and Statistical Analysis of the Capitalist Process*, vol. 1. Yale University Press, New Haven, CT.
- Serrano, C., 2010. The dynamics of the transfer and renewal of patents. *RAND J. Econ.* 41 (1), 686–708.
- Shapiro, J., 2006. Smart cities: quality of life, productivity, and the growth effects of human capital. *Rev. Econ. Stat.* 88 (2), 324–335.
- Singh, J., Agrawal, A., 2011. Recruiting for ideas: how firms exploit the prior inventions of new hires. *Manag. Sci.* 57 (1), 129–150.
- Singh, J., Marx, M., 2013. Geographic constraints on knowledge spillovers: political borders vs. spatial proximity. *Manag. Sci.* 59 (9), 2056–2078.
- Solow, R., 1957. Technical change and the aggregate production function. *Rev. Econ. Stat.* 39, 312–320.
- Strange, W., Hejazi, W., Tang, J., 2006. The uncertain city: competitive instability, skills, innovation, and the strategy of agglomeration. *J. Urban Econ.* 59 (3), 331–351.
- Sturgeon, T., 2002. Modular production networks: a new American model of industrial organization. *Ind. Corp. Chang.* 11 (3), 451–496.
- Tecu, I., 2012. *The location of industrial innovation: does manufacturing matter?* Ph.D. Thesis, Brown University.
- Thompson, P., 2006. Patent citations and the geography of knowledge spillovers: evidence from inventor- and examiner-added citations. *Rev. Econ. Stat.* 88 (2), 383–388.
- Thompson, P., Fox-Kean, M., 2005. Patent citations and the geography of knowledge spillovers: a reassessment. *Am. Econ. Rev.* 95 (1), 450–460.
- Usher, A., 1929. *A History of Mechanical Inventions*. Harvard University Press, Cambridge, MA.
- Waldinger, F., 2012. Peer effects in science: evidence from the dismissal of scientists in Nazi Germany. *Rev. Econ. Stud.* 79, 838–861.
- Wang, Z., 2014. *Location Choice at Labor Force Entry and New Estimates of Selection, Growth, and Level Effects from U.S. Census Data*. Mimeo (Fudan University).
- Wheeler, C., 2001. Search, sorting, and urban agglomeration. *J. Labor Econ.* 19, 879–899.
- Wilson, D., 2009. Beggar thy neighbor? The in-state, out-of-state, and aggregate effects of R&D tax credits. *Rev. Econ. Stat.* 91 (2), 431–436.
- Zucker, L., Darby, M., Brewer, M., 1998. Intellectual human capital and the birth of U.S. biotechnology enterprises. *Am. Econ. Rev.* 88 (1), 290–306.

CHAPTER 7

Cities and the Environment

Matthew E. Kahn*, **Randall Walsh†**

*Department of Economics, UCLA and NBER and IZA, Los Angeles, CA, USA

†Department of Economics, University of Pittsburgh and NBER, Pittsburgh, PA, USA

Contents

7.1. Introduction	406
7.2. Incorporating Local and Global Environmental Externalities into Locational Equilibrium Models	409
7.2.1 Emerging within city environmental trends	413
7.2.1.1 <i>Improving air quality</i>	413
7.2.1.2 <i>Preservation of open space</i>	417
7.2.1.3 <i>Superfund, brownfield, and industrial site remediation</i>	419
7.2.2 Improving quality of life in the center city	421
7.3. Global Externalities Exacerbated by the Intrametro Area Locational Choice of Households and Firms	423
7.3.1 Measuring the causal effect of suburbanization on household energy consumption	424
7.3.2 Suburbanization and carbon politics	426
7.4. Environmental Amenities in a System of Cities	427
7.4.1 A cross city household locational choice model	428
7.4.2 Modeling cross city differences in the local public good supply	430
7.4.3 Firm locational demand and local amenities	431
7.4.4 The evolution of cross city differences in environmental amenities	433
7.4.4.1 <i>The case of Pittsburgh</i>	433
7.4.4.2 <i>Air pollution</i>	435
7.4.4.3 <i>Land use regulation/zoning</i>	436
7.4.4.4 <i>Climate</i>	437
7.4.4.5 <i>Resilience to climate change shocks</i>	438
7.4.5 The rise of superstar green cities	440
7.4.6 Cross city local public good valuation challenges and opportunities	442
7.5. The Urban Building Stock's Energy Consumption	445
7.5.1 Pollution externalities associated with electricity consumption	446
7.5.2 The building stock's energy consumption: a bottom-up approach	447
7.5.3 Weak price incentives for conservation	449
7.5.4 The commercial real estate sector	450
7.5.4.1 <i>Commercial real estate energy efficiency and human capital</i>	451
7.5.4.2 <i>The market for green buildings</i>	452
7.5.4.3 <i>Challenges in increasing the urban real estate's energy efficiency</i>	454
7.5.4.4 <i>The renter/owner split incentive problem in cities</i>	455
7.5.5 Carbon pricing and the building stock's energy efficiency	456

7.6. Conclusion	457
Acknowledgment	458
References	458

Abstract

This chapter surveys recent literature examining the relationship between environmental amenities and urban growth. In this survey, we focus on the role of both exogenous attributes such as climate and coastal access and endogenous attributes such as local air pollution and green space. A city's greenness is a function of its natural beauty and is an emergent property of the types of households and firms that locate within its borders and the types of local and national regulations enacted by voters.

We explore four main issues related to sustainability and environmental quality in cities. First, we introduce a household locational choice model to highlight the role that environmental amenities play in shaping where households locate within a city. We then analyze how ongoing suburbanization affects the carbon footprint of cities. Third, we explore how the system of cities is affected by urban environmental amenity dynamics and we explore the causes of these dynamics. Fourth, we review the recent literature on the private costs and benefits of investing in "green" buildings. Throughout this survey, we pay careful attention to empirical research approaches and highlight what are open research questions. While much of the literature focuses on cities in the developed world, we anticipate that similar issues will be of increased interest in developing nation's cities.

Keywords

Urban economics, Cities, Environment, Carbon policy, Equilibrium sorting, Climate change, Green infrastructure

JEL Classification Codes

H4, Q5, O3, R1

7.1. INTRODUCTION

Both within cities and across cities, nonmarket amenities are an important determinant of where people choose to live and work. While we now spend more time online as we shop on Amazon, socialize on Facebook and Twitter, stream movies from Netflix, and telecommute from home, our physical location continues to define what place-based attributes we enjoy and the environmental conditions to which we are exposed.

Certain locational attributes such as rivers and ocean ports that were once key determinants of local productivity now play a central role as amenities offering unique views as well as recreational and climate amenities. Today, the Ritz-Carlton in Battery Park on the south tip of Manhattan is selling apartments for over \$1300 a square foot. This price premium reflects the desirability of waterfront real estate in modern Manhattan. A majority of the US population and income are clustered in counties within 80 km of a coast or the Great Lakes ([Rappaport and Sachs, 2003](#)). This concentration of

economic activity along the coasts and major rivers is a function of both the historical persistence of early settlement patterns and the key role that transportation access played in determining the locations of early waves of industrial activity (Bleakley and Lin, 2012).

The persistence of these coastal patterns today is in no small part supported by the environmental amenities conveyed by proximity to rivers and coasts and the climate amenities enjoyed by those living on or near the coast. In coastal cities such as Los Angeles, richer people consistently live close to the beach, and thus, there has been less suburbanization of income than in noncoastal metropolitan areas (Lee and Lin, 2013).

Today, nonmarket amenities such as clean air, green spaces, temperate climate, and safe streets help fuel the modern consumer city (Glaeser et al., 2000; Rappaport, 2009). Consumer preferences for these amenities draw wealthier individuals to attractive locations, a trend that is reinforced by employers (much freer of resource- and transportation-related locational constraints than were their predecessors) who choose to locate in high-amenity locations as a way to be more attractive to potential employees. At the same time that cities such as London and San Francisco are experiencing sharp increases in their home prices, there are other cities such as Las Vegas and Phoenix whose population levels continue to grow. Even though these places are not blessed with naturally high levels of amenities, millions of people are expressing a desire to live in such cities. Home prices and wages adjust in the spatial equilibrium and the diverse population and firms sort across space. While some middle-class households will be disappointed that they cannot afford coastal housing, there will be other households with a taste for large new housing who are quite willing to live in an affordable city such as Las Vegas.

At any point in time, a geographic area's environmental attributes are a function of exogenous geography and determined by the (intended and unintended) choices made by the set of households, voters, and industries that cluster within the city. Households and firms trade off the private benefits of a specific location's attributes versus what they must pay in rents to locate there. Given that higher-quality neighborhoods and cities will require a rent premium means that such areas self-select a subset of households and firms to locate there. Expensive high-amenity areas such as Berkeley, California, and Portland, Oregon, attract highly educated individuals. The Tiebout sorting of like-minded progressive environmentalists attracts local firms who cater to such consumers (i.e., organic farmers' markets). The net effect is that green cities emerge due to both selection and treatment effects. Cities such as San Francisco feature few dirty heavy industries. Instead, the "golden goose" for these high-quality-of-life cities tends to be high-tech companies who, in turn, employ workers who demand high-quality local amenities. Such worker/voters support local leaders who enact policies that further enhance local quality of life, and this process feeds on itself as a type of social multiplier effect (Sieg et al., 2004; Bayer et al., 2007).

This causal link between exogenous amenities (i.e., San Francisco's unique coastal beauty) and endogenous amenities (San Francisco becoming a center of green living)

is one of the main themes of this chapter. These synergies pose a challenge for empirical research seeking to measure the marginal willingness to pay for individual local public goods such as air quality improvements. As one local public good improves (perhaps due to local deindustrialization), this will induce a resorting of heterogeneous households and firms such that several aspects of local quality of life will be affected (Kuminoff et al., 2013). Our survey investigates different approaches for addressing this issue.

We explore four main issues related to sustainability and environmental quality in cities. First, we introduce a household locational choice model to highlight the role that environmental amenities play in shaping where households locate within a city. We discuss how the four major recent changes in the US cities, improvements in air quality, rising demand for open space, brownfield remediation, and improvements in center city quality of life influence where and how densely different income groups live within metro areas. Second, we analyze how ongoing suburbanization affects the carbon footprint of cities. Whether this association between suburbanization and carbon production represents a causal effect remains an open question. We discuss various identification strategies for convincingly answering this question.

Third, we explore the causes and consequences of urban environmental amenity dynamics. A city's production of pollution depends on its demographics, industrial structure shifts, and the local and national policies that are adopted. As local public goods shift over time (i.e., Pittsburgh's air quality improves), we seek to understand how such amenity dynamics affect the cross city spatial equilibrium. Unlike standard hedonic analyses, we discuss the cross city sorting of both households and firms and the role that environmental amenities play in this locational decision. We explore how new local and federal policies (i.e., local zoning and differential enforcement of the Clean Air Act) affect the spatial equilibrium, and we study how the emerging new risk of climate change and the resulting increase in summer temperature and sea level rise for coastal cities affects the system of cities. Any change in local public goods introduces distributional effects. We explore the economic incidence of changes in quality of life in different cities and their implications for the well-being of rich, middle-class, and poor urbanites.

While much of the empirical quality of life literature has focused on the United States, we argue that these same issues arise in developing countries as economic development increases the demand for nonmarket amenities. This section ends by discussing open research questions set in the developing world.

Finally, we review the recent literature on the private costs and benefits of investing in "green buildings." In the developing world today, billions of people are moving to cities. Such cities are constructing new residential and commercial real estate that will last for decades. Given that buildings are major consumers of electricity, it is important to understand how urban real estate developers, investors, and real estate tenants demand electricity because these decisions have key implications for greenhouse gas production and thus the severity of future climate change.

7.2. INCORPORATING LOCAL AND GLOBAL ENVIRONMENTAL EXTERNALITIES INTO LOCATIONAL EQUILIBRIUM MODELS

In this section, we present an equilibrium locational choice model to study the role that location-specific environmental amenities play in determining who lives in a given area. The model melds ideas from earlier work including [Ellickson \(1971\)](#), [Stull \(1974\)](#), [Epple and Platt \(1998\)](#), [Epple and Sieg \(1999\)](#), [Sieg et al. \(2004\)](#), and [Banzhaf and Walsh \(2008\)](#).

Assume there exists a continuum of households of measure N that are characterized by their income y , the distribution of which is given by $f(y)$, which has continuous support over the interval $[y_l, y_h]$. In cases where income heterogeneity is of second-order concern, we will often assume that $[y_l, y_h] = \bar{Y}$ such that all households share the same income level. Location choices are discretized and each household must choose to live in one of J discrete locations within the metropolitan area,¹ indexed by $j \in \{1, \dots, J\}$. Conditional on location choice, households then choose their optimal level of land to consume.²

Household preferences are represented by the indirect utility function $V(y, P_j, Z_j)$, where P_j is the price of a unit of housing at location j and $Z_j = [Z_1, \dots, Z_k]$ is a vector of amenities available uniquely to households that choose to live at location j . This vector consists of local public goods that are and are not related to the environment. It includes exogenous environmental locational attributes such as distance to beaches and climate and endogenous environmental amenities such as pollution and protected open space. The elements of Z also include access to employment opportunities and nonenvironmental amenities such as crime and school quality.³ Finally, each location within the metropolitan area is characterized by a continuous housing supply function $S_j(P)$.⁴ This is a

¹ Note that in some models, locations are modeled as continuously varying along certain amenity attributes, the canonical example of this approach being the basic monocentric city model that describes continuously varying locations in terms of their distance from the central business district.

² Alternatively, one could assume that land is used as one input in the production of housing with, for instance, a constant returns to scale technology. However, while tractable, this approach adds an extra notational burden without changing the general intuition that the model provides relative to the role of prices and incomes in determining population density.

³ The specification of $V(\cdot)$ also implicitly assumes the inclusion of a numeraire whose price is normalized to 1. For tractability, $V(\cdot)$ is assumed to be continuous with bounded first derivatives that satisfy $V_y > 0$, $V_Z > 0$, and $V_P < 0$. For simplicity, we assume that indirect utility function is specified such that the associated land demand function is independent of amenity levels Z and is given by $D(P, y)$, which is also continuous with bounded first derivatives $D_y > 0$ and $D_P < 0$. Demand is assumed to be strictly positive and bounded from above.

⁴ The supply function describes what portion of a location's available land L_j will be committed to residential use at a given price. We assume that there exist residual competing demands for land at each location and that there is one special competing use (i.e., agricultural production) for which there is completely elastic demand for land at any price at or below the reservation price $\bar{P}_A$. Thus, the supply of land for residential use at a given location is equal to zero for any price at or below $\bar{P}_A$. At prices above $\bar{P}_A$, residential land supply is increasing to an upper bound of L_j .

very general specification, which embeds the basic monocentric city model, theoretical sorting models such as those considered by [Epple and Platt \(1998\)](#) and [Banzhaf and Walsh \(2008\)](#), and the empirical framework pioneered by [Epple and Sieg \(1999\)](#).

In equilibrium, the sorting of income types across locations will depend on the income elasticities for land and locational amenities. Equilibrium in the model is defined by a vector of allocation functions $\phi_j(y)$, which identify the proportion of each income-type living in each location and a vector of J location-specific housing prices, which satisfy the locational equilibrium conditions:

$$V(y, P_j, Z_j) \geq V(y, P_k, Z_k) \forall k, y \forall j, \text{ where } \phi_j(y) > 0 \quad (\text{i})$$

$$N \int_{y_l}^{y_h} D(P_j, y) \phi_j(y) f(y) dy = S_j(P_j) \forall j \quad (\text{ii})$$

$$\sum_{j=1}^J \phi_j(y) = 1 \forall y \quad (\text{iii})$$

Condition [i](#) requires that in equilibrium, no household would be strictly better off choosing a different location. Condition [ii](#) requires that the housing market clear in each location. And condition [iii](#) requires that each household is assigned to a specific location.⁵

As specified, the model is still quite general. Further simplicity can be gained by assuming that the vector of amenities Z_j can be collapsed into an index such that $V(y, P_j, Z_j) = V(y, P_j, G(Z_j))$ and $\frac{\partial V}{\partial G} > 0 \forall y$. Under this assumption, locations are vertically differentiated with all households agreeing on the relative ranking of “quality” across locations. This assumption when combined with equilibrium condition [i](#) requires that housing prices be strictly increasing in the amenity index G .⁶

⁵ Note that the discreteness of locations can be easily relaxed by instead indexing location by a continuous variable j , that is, distance to the CBD, and then adjusting conditions [ii](#) and [iii](#) as follows. First, the functions that are indexed by j become functions of j (i.e., $\phi_j(y)$ becomes $\phi(y, j)$). Second, in condition [ii](#), $S_j(P_j)$ must be replaced by $s_j(P_j)l(j)$, where $s_j(P_j)$ is the supply of housing per unit land at continuous location j and $l(j)$ is the amount of land per unit of j available at location j (if the “city” is a one unit wide strip of land emanating out of the CBD and j indexes distance from the CBD, then this collapses to $s_j(P_j)$). Finally, in condition [iii](#), the sum is replaced by an integral over the entire support of j , $\int_{j \in 0}^{\infty} \phi(y, j) dj = 1$.

⁶ To offer stronger predictions about the sorting of individuals by income, the model as presented here limits household heterogeneity to the single dimension of income. [Epple and Platt \(1998\)](#) incorporated heterogeneity in tastes for the overall public good level G into a similar model. They showed that under a modified single-crossing concept that considers both a taste parameter and income, stratification occurs with all individuals of a given taste level sorting by income and all individuals with a given level of income sorting based on the strength of their preferences for public goods. Empirical models have generalized this framework to incorporate heterogeneity in tastes for specific attributes and specific locations by incorporating random coefficients on the elements of the G -index and adding an idiosyncratic location parameter, which is typically assumed to have a logit distribution. These analyses then build on the canonical random utility framework developed by [Berry et al. \(1995\)](#). For an overview of this literature, see [Kuminoff et al. \(2013\)](#).

As specified, the model is still too general to make predictions regarding income sorting. In general, sorting will depend on the relationship between income and the marginal (indirect) utilities of G and P (V_G and V_P). To facilitate a characterization of the equilibrium sorting of households across communities, we require further restrictions on preferences. A common assumption made in these models is that preferences satisfy the “single-crossing” property. This condition requires that the slope of an indirect indifference curve in the (G, P) plane, $-\frac{V_G}{V_P}$, be monotonic in income. Intuitively, this assumption implies that willingness to pay for the amenity bundle is either increasing $\left(\text{when } \frac{\partial -\frac{V_G}{V_P}}{\partial y} > 0\right)$ or decreasing $\left(\text{when } \frac{\partial -\frac{V_G}{V_P}}{\partial y} < 0\right)$ in income.

Focusing on the increasing case, single crossing guarantees that equilibrium can be characterized by an ordering of locations that is increasing in both P and G . For each pair of “neighboring” locations (as sorted by this ranking), there will exist a set of boundary households (uniquely identified by income level) that are indifferent between the two locations. Households whose income is below the boundary location will prefer the lower-ordered location, and those whose income is above the boundary income will prefer the higher-ordered location. This leads to perfect income stratification of households across locations based on their level of the amenity index.

The basic model as specified has implications for how spatial heterogeneity in locational attributes affects the spatial distribution of households and the income distribution across the city. To illustrate the basic machinery of the model, consider the case where households care about just two locational amenities, access to employment locations and environmental quality. Suppose, for instance, that employment is at a single factory located at the CBD and that the factory’s pollution mostly lands on residential areas closer to the city center, while areas further from the city center are exposed to less of the factory’s pollution. This differential pollution exposure implies that residential locations vary in two dimensions (commute times and environmental quality), both of which are a function of distance to the CBD and under our index assumption work through the G function. Prices will now increase with distance to the CBD x if

$$\frac{dG}{dx} = \frac{\partial G}{\partial \text{Commute costs}} \frac{d\text{Commute costs}}{dx} + \frac{\partial G}{\partial \text{Pollution}} \frac{d\text{Pollution}}{dx} > 0 \quad (7.1)$$

In a continuous expression of the problem, the actual rent gradient at a given distance from the CBD is given by

$$\frac{dP}{dx} = -\frac{V_g}{V_P} \frac{dG}{dx} \quad (7.2)$$

where V_g and V_P are evaluated at the income of individuals who choose to live distance x from the CBD.

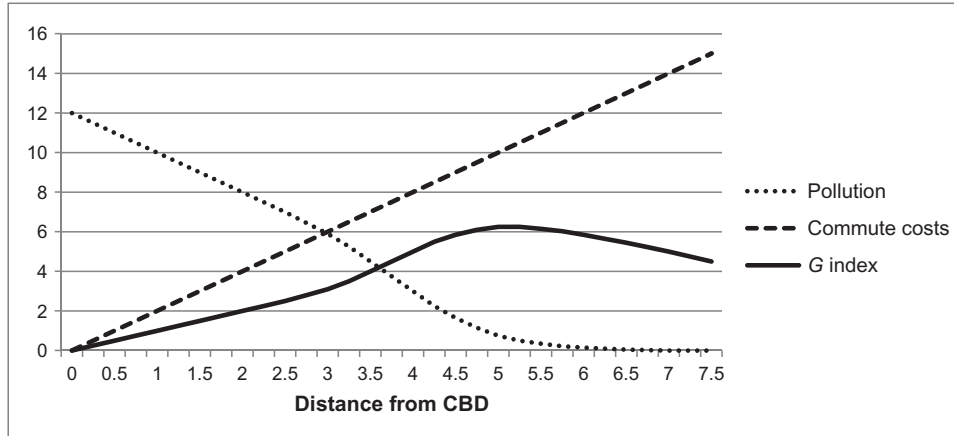


Figure 7.1 The distribution of amenities within a city.

If we assume that pollution levels drop relatively quickly beyond some specified distance from the CBD, then the overall amenity index and its two component parts will be as presented in Figure 7.1, with the most desirable location being just beyond 5 miles from the CBD.⁷ Our model also predicts that prices will be highest in this location and that the richest individuals in the community will locate at this point, with incomes dropping as we move either into the CBD or toward the urban fringe and the lowest-income individuals locating at the CBD. Predictions regarding population density are more nuanced. In equilibrium, prices will be highest at locations where amenity levels (G-index) are highest. *Ceteris paribus*, higher prices lead to lower lot sizes. Thus, at the intensive margin and working through the price channel, higher levels of environmental quality (and amenities in general) serve to increase the density of development. However, at the extensive margin, high-income households tend to sort into the high-amenity locations. *Ceteris paribus*, these high-income households will demand relatively larger residential lots. Thus, there is an income channel through which environmental amenities can effectuate lower housing densities.

This dichotomy between price and income effects is particularly important for understanding the potential impact of zoning restrictions on density. In high-amenity locations, high potential rents create a strong incentive for developers to build up (generate more square footage of living space per unit of land input). By magnifying the relative importance of the price effect, the potential for building up increases the probability that the net

⁷ If the city center is a major employment center, then this will create a pollution “hot spot” downtown as pollution will be created as a by-product of transportation, industrial activity, and individual choices such as litter and smoking.

effect of amenities will be to increase density. However, it is precisely these high-environmental-amenity locations that tend to experience restrictive zoning regulations designed to prevent dense development.

The shape of the income distribution also plays an important role in determining the relative density outcomes in high *G* locations. If there is limited income heterogeneity, then the income sorting channel will be attenuated and price effects will dominate, leading to densification around high-amenity locations. Conversely, in the face of large degrees of income heterogeneity, if the superrich have a high willingness to pay for large homes in high-amenity areas, then we would expect to see large lots and relatively low-density development in high-amenity areas—as the income sorting channel dominates the price channel.

7.2.1 Emerging within city environmental trends

We now examine several environmental dynamics playing out across urban areas and explore how these supply-side changes affect the spatial distribution of economic activity within cities. An overarching recent trend in the urban environment is the improvement in central city quality of life on both environmental and nonenvironmental dimensions. This trend has important implications for the long-run suburbanization trend. We first consider the literature on three specific environmental components of this change (air quality, open space, and brownfield cleanup) before returning to the issue of central city quality of life in general.

7.2.1.1 Improving air quality

The stylized example presented in [Figure 7.1](#) highlights the potential role that air pollution could play in shaping the spatial structure of cities. Historically, the combination of urbanization, industrialization, and heavy reliance on coal in both manufacturing and residential uses led to extreme pollution problems ([Barreca et al., 2014](#)). The “London Fog” made news in the United States as early as 1879 ([Stradling and Thorsheim, 1999](#)). Eighty years later, the Great Smog of 1952 was responsible for the death of as many as 12,000 Londoners ([Clay and Troesken, 2010](#)). Thus, local air pollution has been a major component of the spatial amenity landscape. Perhaps, nowhere are the trade-offs between commute time and air pollution better illustrated than in the East End neighborhoods of late nineteenth-century Pittsburgh. By the late 1800s, the city’s industrial success had led to such severe air pollution problems that it was famously referred to by James Parton as “hell with the lid taken off” ([Tarr, 2003](#)). The air pollution created a situation where “society was forced to flee to the freer air of the East End” ([Nevin, 1888](#)). This environmental haven, separated from the mill-laden rivers by surrounding hillsides, became the location of choice for both the city’s elite and the nation’s elite. By 1900, residents of this neighborhood, located five miles from downtown Pittsburgh, controlled as much as 60% of America’s productive assets ([Skrabec, 2010](#)).

Over the last few decades, there has been a sharp reduction in the US urban air pollution. [Smith \(2012\)](#) documented that the US air pollution (as measured by ambient PM10) has fallen by roughly 50% between the early 1960s and 2008. Similar progress is documented in Germany and Spain. In spite of these improvements, there continues to be strong empirical support for the role that air pollution variation plays in determining the spatial structure within cities. A voluminous set of hedonic studies demonstrate that air quality capitalizes into housing prices suggesting that it belongs in the locational equilibrium model's amenity vector.⁸ Further, direct reduced form evidence regarding the role of air pollution in household location decisions is provided by [Banzhaf and Walsh \(2008\)](#). These authors first showed that the locational equilibrium model in general predicts increases in both density and income when amenity levels increase at a single location. They then tested these density and income predictions *vis-à-vis* air pollution by evaluating the impact of changes in the scale and existence of sources of toxic air pollution that occurred in California between 1990 and 2000.

Their work broadly supports both the density and the income predictions of the model. On density, they found that removing exposure to polluting facilities increased population density on the order of 5%, while adding new exposure led to a drop in population density of similar magnitudes. This density result is likely driven by two related factors. First, working through the price channel, reductions in pollution provide incentives for developers to develop more intensely and households to consume less square footage. Second, removal of polluting facilities provides an opportunity to redevelop the land into residential use. This process is reenforced by the newly increased land prices. In terms of the spatial distribution of households of various income levels, new exposure to toxic air pollution was associated with a drop in a neighborhood's average annual household income of \$1000 or more. Conversely, the loss of exposure to polluting facilities was estimated to lead to concomitant increases in income.

While this research provides reduced form evidence of the potential importance of air pollution in the locational equilibrium framework, such studies do not explicitly model how different households respond to exogenous changes in air pollution. A recent residential sorting literature has sought to test how the spatial distribution of households within a city is affected by regulation-induced changes in the spatial distribution of pollution. Do richer people now move to areas that have experienced a quality of life improvement? A handful of studies have attempted to estimate locational equilibrium models of household location that directly incorporate spatial heterogeneity in air quality.

⁸ [Smith and Huang \(1995\)](#) summarized 37 studies that used hedonic property analysis to evaluate the housing price effects of air pollution. [Zabel and Kiel \(2000\)](#) and [Kim et al. \(2003\)](#) presented additional hedonic studies of air pollution valuation.

The first such efforts are reported in [Sieg et al. \(2004\)](#). These authors estimated a locational equilibrium model for the greater Los Angeles metropolitan area that incorporates spatial heterogeneity in ozone concentrations. They then used the estimated model to compute a counterfactual equilibrium, which describes the predicted sorting of households of different income levels across 92 different neighborhoods located in five different Southern California counties, under the assumption that ozone levels were what they would have been (as estimated by the Environmental Protection Agency) if the Clean Air Act Amendments of 1990 (CAAA) had never occurred.

To highlight the potential importance of accounting for equilibrium sorting when evaluating nonmarginal changes in locational amenities, [Table 7.1](#) reproduces Sieg et al.'s estimates of the county-level impacts of the CAAA. Consider, for instance, the average household that would have located in Los Angeles County in a baseline no-CAAA state of the world. The first three columns of the table report the impact of the CAAA on ozone levels, prices, and the G-index in their baseline county (Los Angeles). On average, implementation of the CAAA was predicted to provide a 9.40% improvement in Los Angeles County's ozone pollution and drop average housing prices by 0.20% (as we discuss below, this price drop arises from the fact that Los Angeles County saw small improvements in air quality relative to overall regional improvements). The 9.40% ozone improvement was associated with an overall 2.30% increase in average G-index levels. However, in assessing the impact of the CAAA on individual households, it is necessary to account for the fact that households move when the spatial distribution of amenities changes. The last two columns of the table account for these movements and report the change in price and G-index predicted to be actually experienced by households in the post-CAAA locational equilibrium as opposed to the baseline no-CAAA equilibrium (i.e., accounting for the relocation of those households who choose different locations under the two different two locational equilibriums). In general, under the post-CAAA equilibrium, households tended to sort out of their initial Los Angeles County neighborhoods and into neighborhoods that saw slightly smaller amenity increases (2.00% vs. 2.30%) and slightly larger price decreases (−0.60% vs. −0.20%). In some instances, the impact of sorting can be quite large. For instance, this analysis predicts that, on average, individuals who

Table 7.1 [Sieg et al. \(2004\)](#) results

County	Initial location			Postsorting	
	Pct. Δ ozone (improvement) (%)	Pct. Δ price (%)	Pct. Δ G index (%)	Pct. Δ price (%)	Pct. Δ G index (%)
Los Angeles	9.40	−0.20	2.30	−0.60	2.00
Ventura	9.40	−2.60	0.40	−1.20	1.60
Orange	10.60	−1.60	1.10	1.00	1.50
San Bernardino	14.30	2.90	5.40	2.80	5.30
Riverside	18.30	5.90	7.60	6.30	8.00

initially live in Ventura County will experience a public good change that is actually four times as large as that experienced at the average household location in Ventura County.

These differences can be extremely important when considering the distribution of welfare from a given policy. For instance, in this example, the partial equilibrium estimate of the average Ventura County household's willingness to pay for these air quality improvements is only \$21 per year, while the general equilibrium estimate, which accounts for relocation, is \$539 per year. [Table 7.1](#) also highlights the importance of relative improvements in evaluating the price effects of environmental improvements. While all counties experienced overall air quality improvements, because those that experienced lower levels of improvement became relatively less attractive, they actually are predicted to experience price decreases in the new equilibrium.⁹ It is important to note that this research assumed that Los Angeles was a closed economy and that the improvement in local air quality did not accelerate population growth of the region as a whole. Future research might examine intrametropolitan area and intermetropolitan area choice within a unified framework. In an open economy, it is likely to be the case that even those areas with the lowest relative gains in air quality will still experience accelerated economic growth.

Two additional recent studies have also incorporated air pollution in a locational equilibrium model. [Tra \(2010\)](#) utilized a random utility model (RUM) approach, which incorporates horizontal differentiation to analyze the impact of air pollution changes in the LA metro area. The RUM framework uses a random coefficients framework that allows individual households to place different weights on the various elements of the amenity index and incorporates idiosyncratic tastes for specific locations. Tra's results are generally consistent with those of [Sieg et al. \(2004\)](#). [Bayer et al. \(2011\)](#) incorporated ozone pollution along with crime rates and racial composition into a dynamic version of the random utility approach, which they estimated using data from the San Francisco Bay area. They find that on average, households would be willing to pay \$295 for a 10% decrease in the number of days that ozone levels exceeded the state's hourly standard.

Applying such a structural approach to studying the valuation of air pollution in cities in the developing world would be quite valuable. In the developing world, cities in China and India feature extremely high levels of air pollution, and within these cities, there is substantial variation in air pollution. [Zheng and Kahn \(2008\)](#) estimated hedonic real estate price regressions using the price per square foot of new residential apartments built across Beijing from 2004 to 2006. Controlling for a large number of attributes of the apartment and the neighborhood, they estimated that a one microgram per cubic meter increase in PM10 is associated with a 0.5% reduction in local home prices.

⁹ While not reported in the [Sieg et al. \(2004\)](#) paper, these relative changes in amenity level can also be expected to be associated with changes in the distribution of income groups across the counties, with the average income of households in the counties that experienced the largest improvements increasing.

One potentially fruitful area for future research is on the spatial distribution of air pollution within cities. Today, regulators are paying greater attention to small particulates such as PM_{2.5}. Atmospheric chemistry studies have documented that “hot spots” exist close to major highways in cities such as Los Angeles (see [Hu et al., 2009](#)). If government regulations reduce such pollution, an important economic incidence issue arises. As particulate problems fall near freeways, will renters in the area benefit or will they be priced out of their previously affordable neighborhoods? If richer people move in, will the type of gentrification documented by [Sieg et al. \(2004\)](#) repeat itself so that the main winner from the local public good improvement are the local land owners?

Some cities continue to have rent control and rent stabilization programs. For example, there are 15 cities in California who have rent control and these include Los Angeles and San Francisco.¹⁰ Incumbents in such programs are less likely to be priced out of gentrifying neighborhoods. In addition, developers in major cities are often asked to set aside a given fraction such as 20% of units in new residential towers to be sold to lower-income households. In this sense, activist local government housing policy interacts with environmental dynamics such that a subset of lower-income households enjoy the gains of local amenity improvements without facing sharp increases in rents.

7.2.1.2 Preservation of open space

Throughout the United States, many local governments have introduced publicly financed efforts to purchase open space development rights. The intent of such legislation is to preserve land for recreational and aesthetic purposes ([Kotchen and Powers, 2006](#); [Banzhaf et al., 2010](#)). Such land set-asides affect the spatial patterns of economic development within a metropolitan area.

Several researchers have considered the role that open space protection can play in determining urban spatial structure. Most of this work has been theoretical in nature and largely builds directly on monocentric city formalizations of locational equilibrium. [Wu and Plantinga \(2003\)](#) incorporated open space into a monocentric city model and found that land protection at the urban fringe can lead to “leap frog” development as open space amenities increase the desirability of land on the far side of the newly protected open space from the city center. [Turner \(2005\)](#) considered open space in a dynamic version of the model that incorporates both urban and suburban locations. He found that in the presence of open space amenities, more remote areas are developed before more central areas. Following on arguments made by [Brueckner \(2000\)](#) about potential market failures in the provision of open space, [Bento et al. \(2006\)](#) developed a computational monocentric city model that incorporates the amenity value of open space at the urban fringe. Under this formulation, the market can lead to inefficiently

¹⁰ <http://www.dca.ca.gov/publications/landlordbook/appendix2.shtml>.

large urban areas as private developers fail to account for the decrease in utility experienced by extant city dwellers due to increased distance to open space at the fringe.

On the empirical front, two papers have considered the issue of open space protection and/or land regulation in a locational equilibrium framework. [Walsh \(2007\)](#) extended the empirical approach developed by [Epple and Sieg \(1999\)](#) to structurally estimate a locational equilibrium model of open space and housing markets in Wake County, North Carolina. The model incorporates multiple employment centers and two types of open space, publicly owned and privately held. Simulated versions of the model are used to undertake counterfactual exercises that incorporate the endogenous nature of privately held open space—modeled essentially as a residual land use. A key finding from this work is that public protection of open space in a given location can lead to an overall decrease in a neighborhood's overall (public + private) provision of open space due to the acceleration of the conversion of privately held open space, which is attracted by the amenity created by the new public land.

A second notable empirical analysis is the work of [Cheshire and Sheppard \(2006\)](#). They used hedonic methods to estimate implicit prices for “amenities produced by the planning system” and then embedded these estimates in a modified monocentric city model that is calibrated to data for the city of Reading in the United Kingdom. The fitted model is then used to calculate the gross monetized value of planning by calculating the change in household expenditure functions needed to hold utility constant under no planning counterfactual outcome—assuming that without planning laws, there would be no open space and that all industrial activity would be equally distributed across the landscape. Working within the framework of the traditional urban model allows Cheshire and Sheppard to implicitly incorporate general equilibrium adjustments in their analysis. Their results suggest that there is a net loss from planning activities that may be as high as 3.9% of annual incomes—with the largest positive benefits from these activities being associated with the provision of accessible open space.

In related work, [Turner et al. \(2014\)](#) use an innovative reduced form approach to evaluate the impact of land use regulation on the price of undeveloped land. They distinguish between three different channels: “own lot” effects, which measure the price effect associated with restrictions that regulations place on the lot itself; “external effects,” which capture the price impacts that derive from restrictions on proximate land uses; and supply effects, which stem from the impact that land use regulations have on equilibrium prices through their attenuating effect on the supply of developed land. They find that a one standard deviation increase in land use regulation decreases land value by about one-third. They attribute the bulk of this effect to own and external lot effects with a 2–3% decrease in developable area contributing the residual.

Finally, we note that a large hedonic literature provides support for the claim that access to open space amenities will lead to higher housing prices. [McConnell and Walls \(2005\)](#) reviewed approximately 40 published articles that use hedonic methods

to assess the capitalization of open space into housing prices. They document capitalization of a wide variety of open space amenities including general open space, parks, and natural areas; greenbelts; wetlands; urban/suburban forest preserves; and agricultural lands.

7.2.1.3 Superfund, brownfield, and industrial site remediation

The legacy of more than a century of manufacturing activity is that many US cities feature large numbers of past industrial sites. These areas suffer from toxic releases and severe localized damage to the land and nearby water (Sigman, 2001). The most toxic of these sites have been placed on the National Priorities List (NPL) and have been targeted for cleanup by the Environmental Protection Agency through the Federal Superfund program. In 1982, an initial 400 sites were placed on the NPL. Today, there are more than 1300 sites on the NPL (US GAO, 2013). These sites are however only the tip of the iceberg. The US Government Accountability Office estimates that there are additionally between 450,000 and 1 million brownfield sites in the United States. While not as toxic as NPL sites, redevelopment or reuse of these sites is complicated by the presence or potential presence of hazardous wastes and can sit abandoned or unused (US GAO, 2005). A large literature documents that property values are negatively impacted by proximity to NPL sites (see Kiel, 1995; Kiel and Williams, 2007).¹¹

As noxious industrial sites are cleaned up, what happens to urban form and the spatial distribution of real estate prices? Two recent studies are of particular interest. Greenstone and Gallagher (2008) examined how tract-level median household values respond to the cleanup of NPL sites. They compared housing price growth in areas surrounding the first 400 hazardous waste sites added to the NPL and thus slated for Superfund cleanup to housing price changes in the areas surrounding the 290 sites that narrowly missed qualifying for the NPL. Greenstone and Gallagher (2008) implemented an instrumental variable approach that embeds a type of regression discontinuity design. The discontinuity arises from the fact that, while all 690 of these sites were scored with an index number representing their level of priority for cleanup, EPA subsequently set an arbitrary score cutoff for listing based on the agency's projection that it could afford to clean 400 sites. Their main outcome equation relates changes in census tract home prices over time to whether the site was designated as an NPL site:

$$\Delta \text{Median census tract home price} = \alpha * X + \beta * \text{NPL site} + \varepsilon \quad (7.3)$$

Greenstone and Gallagher (2008) instrumented for the NPL site listing dummy using a dummy that equals one if the site's hazardous ranking score exceeded 28.5 at the baseline.

Based on this research design, Greenstone and Gallagher (2008) concluded that Superfund listing has little to no effect on local median housing value changes. More

¹¹ A summary of these findings can be found in Sigman and Stafford (2011).

recently, Gamper-Rabindran and Timmins (2013) revisited the Greenstone and Gallagher results and found a positive impact of cleanup on prices. The main reason for this difference in findings is that, while Greenstone and Gallagher evaluated the impact of a site being listed for cleanup, Gamper-Rabindran and Timmins focused on the deletion from the NPL, which occurs after cleanup is actually completed. Thus, Greenstone and Gallagher were measuring the impact of providing information that a site has been designated as an NPL site and may be cleaned up in the future, while Gamper-Rabindran and Timmins were measuring the impact of the actual cleanup. As of August 2014, there were 1318 sites listed as NPL sites and 383 deleted sites on the Superfund list.¹² These data suggest that not all listed sites are cleaned up and this raises an expectation issue, namely, if and when in the future will the cleanup actually occur. Home buyers are likely to have different expectations over these random variables and this will affect their bidding for housing near such sites.

Gamper-Rabindran and Timmins also documented that the use of tract-level data can mask important within-tract variation in exposure. They showed that a cleanup has a larger effect in the lower portions of the within-tract housing price distribution and used house-level data to show that cheaper houses within an exposed census tract tend to be located closer to the NPL site.

Less toxic brownfield sites have also been studied. Three recent studies are of particular note. Currie et al. (forthcoming) geocoded housing sales data from five states (Texas, New Jersey, Pennsylvania, Michigan, and Florida) covering the years 1998–2005. They used data from the Toxics Release Inventory to identify facilities that release toxic emissions and they used microdata from the Longitudinal Business Database to identify the births and deaths of these toxic plants. Their paper measures the housing market and health impacts of 1600 openings and closings of industrial plants that emit toxic pollutants. They found that housing values within one mile decrease by 1.5% when plants open and increase by 1.5% when plants close.

Second, Taylor et al. (2012) studied 105 sites in the Minneapolis metropolitan area, the majority of which are located in the urban core. A major innovation of this work is that they studied the difference between clean industrial sites and toxic sites and noted the high level of spatial correlation between specific contaminated sites and commercial and industrial sites in general. They argued that it is important to control for this potentially confounding factor. They documented that proximity to clean commercial sites (within 0.3 miles) lowers property values by roughly 4.5–5.5%. Environmental contamination increases the negative impacts by approximately 2.5–3.0%. They found that once the site is remediated, the contamination effect disappears.

Third, Haninger et al. (2014) evaluate the US EPA's Brownfields Program, which awards grants for the redevelopment of brownfields. They utilize a variety of

¹² See <http://www.epa.gov/superfund/sites/npl/>.

identification strategies to evaluate a set of 327 brownfield sites that were nominated for cleanup—197 of which were awarded grants for cleanup and 130 were not. They non-parametrically identify a treatment threshold of 2040 m. Estimates of the price effect from cleanup range from 5% to 32%.

A review of the recent literature argues for the presence of price effects associated with close proximity to brownfields and polluted sites. These price effects suggest households are sorting in response to these disamenities. However, from the perspective of the locational equilibrium framework, one would also like to see more direct tests of sorting behavior. A summary of the extant literature is provided in [Banzhaf and McCormick \(2006\)](#) and [Noonan \(2005\)](#). The majority of this literature, going back to [United Church of Christ \(1987\)](#), documents correlations between hazardous facilities and the presence of low-income/minority groups.

However, few of these studies explore the sorting behavior of individuals—instead viewing siting decisions from an environmental justice perspective that focuses on correlations between race/disadvantaged status and pollution. The handful of studies that have looked at demographic changes related to citing or removal of toxic waste facilities have found mixed results.¹³ It remains an open question of whether urban political leaders choose to locate noxious facilities in minority residential communities or whether poor and minority households choose to live close to such facilities because of the housing price discount.

The recent hedonic research documenting the price appreciation in localized areas where noxious sites are cleaned highlights that gentrification will take place in those neighborhoods. The economic incidence is such that land owners in those areas will gain the most. Such hedonic studies do not recover welfare estimates nor do they document changes in population density over time. As a noxious site located in prime piece of center city real estate is remediated, it is possible that the population density in the area will increase. In this sense, the reclamation of center city land can contribute to the center city resurgence documented in many US cities.

7.2.2 Improving quality of life in the center city

In recent years, there have been sharp improvements in center city quality of life brought about by several synergistic trends including declines in center city crime ([Schwartz et al., 2003](#); [Levitt, 2004](#)), improvements in urban air pollution (discussed above; see [Kahn and Schwartz, 2008](#); [Wolff, 2014](#)), multibillion dollar investments in rail transit systems that transport people to the city center ([Baum-Snow and Kahn, 2005](#); [Kahn, 2007](#)), major investments in cleaning the nation's downtown rivers and waterways ([Olmstead, 2010](#)), and building of bike lanes and urban parks such as the Rose Kennedy Greenway ([Tajima, 2003](#)).

¹³ See, for instance, [Been \(1994\)](#), [Been and Gupta \(1997\)](#), [Wolverton \(2009\)](#), and [Cameron and McConnaha \(2006\)](#).

At the same time, there have been emerging demographic trends such as more women working in the labor force and thus delaying marriage and having fewer children (Goldin, 2006). Such households have less of a need to live in the school-focused suburbs. At the industrial level with manufacturing shrinking in its employment share, more people are working in the high-tech service sector and many of these industries gain from downtown locations. Such industries disproportionately hire highly educated people (Moretti, 2012). All of these trends further strengthen the central business district.

Downtown environmental progress is likely to have played a key role in causing the center city revival. While it is very difficult to parse out the individual effects of specific factors, together, these trends have strengthened center cities and have encouraged private developers to make multibillion dollar investments in new downtown real estate projects. The net effect of these investments is to lead more and more people (especially young adults and older adults whose kids are out of the house) to be willing to live downtown. In high-amenity center cities such as New York City and Paris, the rich have chosen to live in the center city (Brueckner et al., 1999; Brueckner and Rosenthal, 2009).

This center city resurgence has been studied in detail (Glaeser and Shapiro, 2003; Glaeser and Gottlieb, 2006) and seems likely to continue. More and more mayors are prioritizing reinventing their downtowns with a focus on beauty and promoting tourism around their waterfronts (see Carlini and Saiz, 2008). Examples range from Pittsburgh, to Chicago, to Shanghai, to Seoul. As center city quality of life improves, center city property values rise. Compare Manhattan's real estate prices in the bleak 1970s versus in Mayor Mike Bloomberg's Manhattan of 2012. By raising center city property values, property tax revenue rises and urban leaders can finance more of their pet projects. It is no accident that New York City mayors focused on enforcing dog poop laws, providing bike lanes, and enforcing antismoking ordinances. At a time when crime is falling sharply, time outdoors is a complementary activity to clean air and safe streets.

As urban quality of life in center cities has improved, it has triggered gentrification in the center cities, and this in turn triggers private sector investments in upgrading buildings and offering better restaurants and retail shopping (Waldfoegel, 2008). The geographic patterns of upscale supermarkets with organic offerings, such as Whole Foods, are not surprising (see Meltzer and Schuetz, 2012).

Whether this demand for downtown living leads millions of more people to live downtown depends on the elasticity of downtown housing supply. Such studies as Glaeser et al. (2006) and in this handbook Gyourko and Molloy (2015) highlight the challenges in building more housing downtown. Kahn (2011) used data from cities across California and argued that more liberal cities are less likely to issue new housing permits. In their study of Massachusetts cities, Glaeser and Ward (2009) concluded that cities with more manufacturing and more minorities in 1940 are also associated with smaller

minimum lot sizes. If environmental quality rises in the center city but the housing supply is highly inelastic, then such demand will simply translate into higher downtown prices and the rich living downtown. This proposition merits more research in urbanizing nations around the world. [Brueckner and Sridhar \(2012\)](#) documented that building height limits in India's cities are contributing to suburban growth. They documented that cities with higher floor to area ratios are more compact.

7.3. GLOBAL EXTERNALITIES EXACERBATED BY THE INTRAMETRO AREA LOCATIONAL CHOICE OF HOUSEHOLDS AND FIRMS

The equilibrium sorting and hedonic models discussed above focus on the demand for location-specific attributes such as air pollution, a short commute, or good school access. Households have the right incentives to trade off such attributes against the market price they must pay for homes located in such communities. However, equilibrium sorting models typically ignore how a specific household's locational choice has consequences for pollution production. For example, if a household chooses to live close to public transit, such a household may be more likely to use public transit more and drive less ([Baum-Snow and Kahn, 2005](#)). In aggregate, such locational choices reduce a city's carbon footprint and help to reduce climate change risk even though each individual household is small in terms of total greenhouse gas production.

Unlike localized environmental goods, changes in greenhouse gas emissions associated with expanded urban footprints do not have any direct effect on the relative attractiveness of different urban locations, so this type of extension to the model essentially boils down to a measurement challenge in which researchers attempt to measure the impact of different spatial structures on greenhouse gas emissions.

A standard claim in the popular media is that suburbanization has a large causal effect on increasing driving and household electricity consumption. Such a result emerges from the standard monocentric model. Suburban homes are far from the CBD's jobs and far from public transit. This leads such households to drive more miles. Since land prices fall with respect to distance from the city center, suburbanites live in larger homes that use more electricity for basic home functions such as air-conditioning and leisure. Such suburban households live further from public transit and are more likely to work in the suburbs. Together, these facts suggest that they will drive more and consume more electricity than if they lived in the center city. These activities will contribute to extra greenhouse gas production if the vehicle is fueled using gasoline and if the home's electricity is generated by coal and natural gas.

From a welfare and policy perspective, this discussion hinges on a failure to price the externalities associated with energy consumption and production. For example, [Parry and Small \(2005\)](#) estimated that internalizing gasoline consumption externalities in the

United States would require a \$1 increase in the gasoline tax.¹⁴ In the presence of such a tax, suburbanites would be more likely to buy more fuel-efficient vehicles, and the carbon externality associated with suburbanization would diminish. This example highlights that the social costs of suburbanization literature are really focused on a second-best world in which households and firms are spreading out without facing the social costs of their actions. A critical implication is that any empirical research focused on the externality consequences of suburbanization must account for the dynamics of the carbon-pricing regime faced by key decision makers.

The greenhouse gas implications of the suburbanization trend depend on the technologies used for generating suburban household electricity demand and for fueling private vehicles. Today, with coal and natural gas providing most of our power and the bulk of private vehicles still being run on gasoline, there is a large carbon impact. In the future, if more households have solar panels fueling electric vehicles, then the carbon consequences of the suburbanization of US homes and jobs could be sharply reduced. In the absence of carbon pricing in the United States, it remains an open question how quickly the research frontier will advance so that the widespread diffusion of solar homes and electric cars takes place.

7.3.1 Measuring the causal effect of suburbanization on household energy consumption

The typical research design seeking to test the hypothesis that suburbanization causes greater household GHG emissions is based on running an OLS regression model of Equation (7.4). In such an equation, the key explanatory variables are such observables as the household's distance from the city center and the population density in the area where the household lives:

$$\text{GHG}_{ijk} = \sum_q \gamma_q X_i^q + \sum_a \theta_a V_j^a + \sum_c \beta_c Z_k^c + \varepsilon_k \quad (7.4)$$

In this regression, the dependent variable is the level of annual household GHG emissions produced by household i living in metropolitan area j in tract k . This variable can be formed with data on the household's annual gasoline consumption and information on the household's annual home fossil fuel consumption such as on natural gas and oil and information on the home's annual electricity consumption and the local electric utility's carbon emissions factor (see Glaeser and Kahn, 2010). X_i^q refers to the value of individual characteristic q for household i . It can include standard household attributes such as the household's income, the head's age, and the household's size. Z_k^c refers to the value of characteristic c in tract k . For example, this may include a household's census tract's population density and the tract centroid's distance to

¹⁴ Davis (2011) and Muller et al. (2011) quantified the externalities associated with coal-fired power plants.

the central business district. The final set of variables, denoted by V_j^a , refers to attribute a of MSA j .

OLS estimates of this equation raise the issue of self-selection.¹⁵ Since households are not randomly assigned to live in neighborhoods, valid concerns can be raised about whether a select subset of people choose to live in the suburbs versus the center city. At the start of this section, we presented a residential locational choice model. That model did not explicitly include an intensive margin of how much do people travel once they choose a location within a metro area. Suppose that households who enjoy driving are more likely to locate in the suburbs further from jobs and shopping centers. The econometrician who ignores this self-selection will be likely to recover biased estimates based on Equation (7.4) because the distance variable is correlated with unobserved attributes of the household. Researchers who rely on OLS have to argue that other factors such as local school quality and housing demand are the key determinants of locational choice and that these choices are independent of the transportation use.

Does living at high density close to the city center induce individuals to consume less energy (i.e., a “treatment effect”)? Or instead, has a select subgroup of the population chosen to live in center cities? This selection effect may be due to unobserved population heterogeneity such that there are people who are predisposed to living the “low-carbon” lifestyle and thus have less demand for residential space and prefer using public transit and walking to driving.¹⁶

Applied economists have recognized this challenge and have tried to make progress using a longitudinal research design. Many urbanists have claimed that people walk and exercise more when they live in center cities. [Eid et al. \(2008\)](#) used a geocoded version of the National Longitudinal Survey of Youth 1979. These data track each individual’s residential address, weight, and other personal characteristics over time. They reported that 79% of these people move address at least once during their 6-year study period. By first differencing their data, they removed individual fixed effects and focused on measuring the association between local neighborhood attributes and a person’s body mass index (BMI). This association can be measured because some households move from the center

¹⁵ Several recent studies have attempted to quantify geography-related variation in carbon intensities. [Glaser and Kahn \(2010\)](#), [Holian and Kahn \(2013\)](#), and [Lee and Lee \(2014\)](#) provided estimates of the relationship between urban population density and the total greenhouse gas emissions from transportation. [Zheng et al. \(2011\)](#) quantified the carbon footprint effects across China’s cities. [VandeWeghe and Kennedy \(2007\)](#) compared per capita auto- and building-related GHG emissions across census tracts within greater Toronto, Canada. Between the most carbon-intensive (typically suburban) and the least carbon-intensive (typically urban) tracts, they found that GHG emissions vary by more than a factor of 4.

¹⁶ The causes of such environmentalism remain an open research question. The peer effect literature and research on parental transmission of culture and values offer possible pathways for examining this issue ([Becker, 1976](#); [Bowles, 1998](#); [Iannaccone, 1998](#)).

city to the suburbs and vice versa.¹⁷ They test and reject the hypothesis that suburbanization contributes to rising obesity levels.

A randomized field experiment design would offer a cleaner way to disentangle selection from treatment effects. Suppose that public housing is scattered across different neighborhoods in a city. Those households who sign up for a public housing lottery would be randomly assigned to their place of residence. A researcher focusing on this subpopulation could study the transportation patterns of these individuals to establish the causal effect of urban form on car use. The researcher would face the challenge of establishing whether results based on this subpopulation are likely to generalize to other subsets of the population.¹⁸

The bottom line in terms of this area of research is that while a handful of papers have been written that highlight potential strategies for identifying the causal link between urban structure and household energy consumption, this issue remains an open question and merits future work both in developed and in developing nations' cities.

7.3.2 Suburbanization and carbon politics

Households who live a high-carbon lifestyle are likely to be aware of this fact and thus to oppose policies such as carbon taxes that would raise electricity and gasoline prices. For example, Cragg et al. (2013) documented that US congressional representative voting on the American Clean Energy and Security Act in 2009 was positively correlated with their district's per capita carbon emissions. Holian and Kahn (2014) document that California voters who live in the suburbs are less likely to support the state's landmark AB32 legislation. California's AB32 commits California to major greenhouse gas reductions by the year 2020 and 2050. It includes new regulations including a cap-and-trade program for carbon dioxide.¹⁹ In the year 2010, California voters had the opportunity to repeal AB32 by voting in favor of Proposition 23. These voting data provide researchers with the opportunity to studying geographic patterns in the precinct level data. Holian and Kahn (2014) document that suburban geographic areas were more likely to vote against cap and trade even after controlling for the fact that more liberal voters live in the center city. This finding is of interest because it highlights that suburbanites are self-interested

¹⁷ The authors do not model why people are moving across geographic areas, and thus, they are implicitly assuming that the migration decision's determinants are not correlated with unobserved determinants of BMI changes.

¹⁸ As an example, the "Move to Opportunity" (MTO) program of the US Department of Housing and Urban Development (HUD) introduced a randomized lottery in which low-income households who "won" the lottery receive vouchers to move to low-poverty areas. The survey data include household vehicle access. This experimental design would allow researchers to study whether the probability that a household owns a vehicle increases when it moves to areas further from public transit and further from the city center. The natural control group in this case is other MTO-eligible families who did not win the lottery (Ludwig et al., 2011).

¹⁹ <http://www.arb.ca.gov/cc/ab32/ab32.htm>.

voters who believe that their operating expenses will rise in the presence of a carbon tax. Center city voters recognize that they live in smaller apartments and drive less and use public transit more. The aggregate effect of these choices is that they face a lower carbon bill if a carbon tax is enacted. The voting patterns support this spatial hypothesis.

A majority of US metropolitan area jobs are located in the suburbs. [Glaeser and Kahn \(2001\)](#) used geocoded zip code-level employment data to document the empirical distribution of jobs across cities and industries. As road networks have improved and center city land has increased in price, jobs have suburbanized. Improvements in information technology have allowed firms to fragment so that they reduce their demand for downtown real estate and only retain workers who require face-to-face interaction with other firms to remain there ([Rossi-Hansberg and Sarte, 2009](#)).

Households who live and work in the suburbs are unlikely to use public transit. Aware of this fact, they are less likely to support public policies that improve public transit or to use this public transit. This fact helps to explain how over 16 major cities have made major investments in rail transit systems over the last few decades, but only Boston and Washington DC have seen significant ridership of these systems ([Baum-Snow and Kahn, 2005](#)).

Throughout the urbanizing world, people are decentralizing as new roads are built and as household income rises. While environmental economists have stressed that the introduction of a carbon tax is likely to offer greater long-run benefits relative to its upfront costs, this group of researchers has tended to ignore the spatial distribution of voters and its implications for the economic incidence of new taxes. Whether the worldwide suburbanization trend affects the likelihood of a global carbon treaty would appear to be an important future research topic.

7.4. ENVIRONMENTAL AMENITIES IN A SYSTEM OF CITIES

In this section, we survey ongoing research examining how environmental amenities shape cross city competition for workers and firms. We examine the causes of differential environmental improvements in some cities (the evolving supply of pollution), and we seek to understand its consequences for the sorting of heterogeneous workers (who differ with respect to human capital) and the types of firms who hire them (i.e., Facebook vs. service firms).

An enormous cross-sectional cross city compensating differentials literature has used the core Rosen/Roback modeling structure to estimate hedonic wage and real estate impacts of nonmarket goods (for a general survey, see [Gyourko et al., 1999](#)). The classic Rosen/Roback model of cross city quality of life implicitly assumes that spatially tied attributes are exogenously determined and fixed over time. The leading papers in the cross city nonmarket quality of life literature have been static studies seeking to estimate the cross-sectional hedonic equilibrium at a point in time ([Graves and Linneman, 1979](#); [Blomquist et al., 1988](#); [Gyourko and Tracy, 1991](#); [Albouy, 2008](#)). Such estimates are

quite useful for learning about the marginal migrant's implicit price index for nonmarket goods and can be used to construct a Laspeyres price index for necessary expenditure to consume a fixed nonmarket local amenity bundle such as climate attributes (Cragg and Kahn, 1999; Costa and Kahn, 2003). This approach is less informative in the case where households face migration costs in moving across cities and when there is significant heterogeneity with respect to the preferences for local public goods. In this case, the hedonic gradient represents the marginal mobile person's willingness to trade off money for local public goods, and such estimates may reveal little about the preferences of those far from the margin (Bayer et al., 2009).²⁰

7.4.1 A cross city household locational choice model

This section modifies the locational equilibrium model presented in Section 7.3 to motivate the discussion below of environmental quality and cross city sorting by heterogeneous households and firms. As presented here, the model is a modification of Bayer et al. (2009) cross city locational choice model. At a point in time, a household is located in a specific metropolitan area $j \in J$ and must choose whether to stay in its current location or move to another location. Households differ with respect to their human capital endowments and can foresee both the incomes and housing prices available to them at all potential locations. Households recognize that if they move, they will incur moving costs (which are heterogeneous; we discuss them below) but that it will enjoy a vector of location-specific attributes G . As with the intracity model presented above, some of these attributes are environmental attributes such as climate, proximity to coasts, and local environmental quality. Finally, we assume that each household has an idiosyncratic preference shock (error term) associated with each possible location ϵ_{ij} . For simplicity, we assume a linear indirect utility function, and thus, the utility that household i that initially is located in community j will receive if it chooses to locate in community k is given by Equation (7.5):

$$V_i^{jk} = \beta * \text{Income}_{ik} - \gamma * \text{Price}_k + \Gamma(D_i, \nu_i) * G_k - \delta * \text{Moving costs}_i^{jk} + \epsilon_{ik} \quad (7.5)$$

In Equation (7.5), moving costs are zero if $j = k$. $\Gamma(\cdot)$ captures both systematic heterogeneity in taste for amenities (based on a vector of demographic characteristics, D_i) and

²⁰ For example, suppose that a Chicago resident who currently earns \$100,000 could earn \$140,000 if he moves to San Francisco. Suppose that housing units are all the same size and that this Chicago resident currently pays \$20,000 in rent but would pay \$120,000 for a housing unit in San Francisco. Abstracting from federal and local taxes, this person sacrifices \$60,000 in private consumption if he moves from Chicago to San Francisco. If we observe this person make this move, then a lower bound on his willingness to pay for San Francisco's amenities over Chicago's amenities is \$60,000. Note that this example assumes that migration costs are zero and abstracts from potential adjustments in the level of housing consumption. Suppose that due to family reasons, this person would face a \$250,000 migration cost for leaving Chicago. An econometrician who does not observe this cost would see the person remain in Chicago and would infer that this person does not value San Francisco's amenities.

a random coefficients component driven by the idiosyncratic shock ν_i . Inclusion of D_i in the Γ function allows the model to reflect the fact that, on average, demographics such as education level impact tastes for environmental quality (Kahn, 2002).²¹

Assuming a logit error term collapses the implicitly high-dimensional problem of identifying the probability of household i , initially located in j , choosing location k as its final destination into the familiar logit probability expression presented in Equation (7.6),

$$\text{Probability}_i^{jk} = \frac{e^{\beta * \text{Income}_{ik} - \gamma * \text{Price}_k + \Gamma(D_i, \nu_i) * G_k - \delta * \text{Moving costs}_i^{jk}}}{\sum_{l \in J} e^{\beta * \text{Income}_{il} - \gamma * \text{Price}_l + \Gamma(D_i, \nu_i) * G_l - \delta * \text{Moving costs}_i^{jl}}} \quad (7.6)$$

Several points regarding empirical implementation of Equation (7.6) merit discussion. First, there is a fundamental asymmetry between the decision maker and the econometrician. The econometrician must impute what this household will earn in each local labor market. While it is easy to impute the rental price of what each household will have to pay for housing in cities such as Boston and Houston, a fundamental self-selection issue arises when using the wages for college graduates who already live in Houston as an estimate of the wage a person now living in Chicago would earn if he moved to Houston. To avoid self-selection concerns, the researcher must assume either that there are no unobserved skills or that the factor price for these skills is equal across cities (Heckman and Scheinkman, 1987).

A key innovation in this model of location choice is the introduction of migration costs (also see Kennan and Walker, 2011). The introduction of migration costs means that migration is an investment and that a forward-looking migrant should also form expectations of how his or her income and the amenities of the area will evolve over time. Incorporating fixed costs is particularly important in the modern economy, where there is an aging population who has built up location-specific social capital (Glaeser et al., 2002). In this case, hedonic differences in prices may sharply understate the marginal valuation of amenities depending on who is the marginal household (Bayer et al., 2009).

Bayer et al. (2009) estimated a version of this model to recover estimates of the marginal willingness to pay for air quality improvements using data on migration while allowing for positive migration costs. They argued that standard hedonic methods will underestimate the marginal willingness to pay for clean air and other location amenities because there are inframarginal households who value such amenities but face large migration costs to moving, and thus, the hedonic gradient contains biased information about household's marginal valuations. In their application, they documented that a

²¹ As presented here, preferences do not exhibit the single-crossing property discussed in Section 7.3. Incorporating income into the Γ function is one way to address this issue.

major bias arises because households are estimated to have high moving costs associated with leaving their state and/or region of birth. A majority of US citizens were born in areas with relatively high air pollution levels. Bayer et al. demonstrated that failure to account for a household's birth-related attachments to these relatively polluted areas imparts a large downward bias in estimates of the willingness to pay for clean air.

In the presence of migration costs, forward-looking households will consider both the amenities of a location today and their expectations of future amenity levels in potential destination cities (see [Bishop and Murphy, 2011](#); [Bishop, 2012](#)). [Bishop \(2012\)](#) assumed that such locational attributes as crime and air pollution follow an AR(1) process and used city-specific estimates of this process to predict future amenity values. In cities where polluted areas are expected to improve over time, today's amenity value understates future progress. If, for instance, the AR(1) process is mean reverting and economic agents anticipate that relatively polluted locations will improve toward the mean over time, then the econometrician who estimates a static model and does not incorporate these dynamics will underestimate the willingness to pay for such an amenity. This bias arises because the econometrician observes people moving to locations with high levels of pollution today and therefore infers that individuals do not prioritize pollution avoidance heavily in their location decisions. When, in fact, economic decision makers are basing their decisions on the lower levels of pollution that they expect to experience in these locations over longer-run time horizons.

7.4.2 Modeling cross city differences in the local public good supply

Environmental and urban economists have not been explicit about how the G is produced and who is doing this "production." Implicit in this is the view that local public goods are either exogenous or being produced through some unintended by-product of economic activity.

In papers such as [Kahn \(1999\)](#) or [Levitt \(2004\)](#), researchers adopt a reduced form approach and model a location's local public good vector by simply regressing measures of amenities (such as air pollution or crime) on a set of observables:

$$G_{jt} = f(X_{jt}, Z_t) \quad (7.7)$$

In this equation, the G is the city's local public good vector, X is a vector of city-specific demographics and industries located within the borders, and Z is a vector of national policies that have been implemented. This reduced form equation representing the supply of local public goods, including many environmental attributes, highlights a key difference between urban economics and industrial organization.

Starting with [Rosen \(1974, 2002\)](#), urban economists have thought of locations as differentiated products whose attributes are in a characteristics space. In this sense, there is a close link between urban and industrial organization economics. In the modern IO literature, profit product sellers choose what to bundle into a product such as a car or an

airplane ride in order to collect more revenue (Berry et al., 1995; Blonigen et al., 2013). A key difference between environmental and urban economics and industrial organization is that the suppliers of real estate do not consciously choose all of the local public goods bundled into a specific property. Instead, such local public goods as clean air and water and what industry and population locate within a geographic area and the daily activities of such economic actors at their specific location are an emergent property of regulation.

Diamond (2012) introduced a straightforward reduced form endogenous amenity modeling technique in which she assumed that endogenous city-specific amenities are an increasing function of the city's share of college graduates. The correlation between college graduates and amenities is likely to reflect both selection and treatment effects. College educated individuals are likely to be richer and thus are willing to pay more to live in nicer areas (selection). The clustering of such college educated in a small geographic area may also have a treatment effect in that their local purchasing power is likely to attract niche retail stores and restaurants (Waldfoegel, 2008). This group's propensity to vote and to follow local politics may discipline local politicians to address quality of life concerns (Moretti, 2004). Given the positive correlation between education and environmentalism, this group is also likely to prioritize green issues (Kahn, 2002).

One promising area for future research relates to how local governments respond to both contemporaneously high levels of local public goods (the G vector) and expectations of future increases in public good levels. In recent work, Brueckner and Neumark (2014) argue that in those areas with better natural amenities (a high G), public sector workers earn higher wages. Their core logic is that households inelastically demand such areas and local officials can tax more (for redistribution purposes) without people moving away. In contrast in cities such as Detroit, if rich people move out of the center city, then the local government has less tax revenue to spend on local public services and such a decline in G may have selective effects on who remains in Detroit and on the probability that outsiders move to this center city.

7.4.3 Firm locational demand and local amenities

In the Roback (1982) model, there is no comparative advantage among heterogeneous firms such that different firms earn greater profits at one location versus another because firms do not differ. The cross city hedonic quality of life literature has developed independently of the empirical firm locational choice literature that explores the role of local agglomeration effects, transportation costs, and access to input markets that affects locational choice (Dumais et al., 2002; Rosenthal and Strange, 2004; Ellison et al., 2010). A second firm literature has explored how differences across space in environmental regulation, labor regulation, and energy prices affect the geography of where different industries cluster (Carlton, 1983; Henderson, 1996; Holmes, 1998; Becker and Henderson, 2000; Greenstone, 2002; Kahn and Mansur, 2013).

Today, firms such as Facebook, Google, Amazon, and Microsoft are locating in high-quality-of-life cities. Nascent agglomerations such as Santa Monica's Silicon Beach are emerging. High-tech firms face a coordination issue that they often have incentives to locate near each other. High quality-of-life areas help to solve the coordination problem and ex post help such firms who seek cross firm learning and labor pooling to retain workers in the local labor market.

For either factories or corporate headquarters, we can write down a locational choice problem where firms choose the location that maximizes their profits. Similar to the Bayer, Keohane, and Timmins model of household locational choice, the core locational choice probability model can be written as

$$\text{Probability}_i^{jk} = \frac{e^{\beta * \text{Agglomeration}_{ik} - \gamma * \text{Factor price}_{ik} - \delta * \text{Moving costs}_i^{jk}}}{\sum_{l \in J} e^{\beta * \text{Agglomeration}_{il} - \gamma * \text{Factor price}_{il} - \delta * \text{Moving costs}_i^{jl}}} \quad (7.8)$$

Further flexibility can be obtained by allowing the coefficients in Equation (7.8) to vary by industry k and firm i . Under such a specification, β , for instance, could be replaced by $B_k(C_i, \epsilon_i)$ where C_i is a vector of firm-level characteristics and ϵ_i is a firm-specific random component.

A key difference between the firm locational choice problem and the household problem is the desire of some firms to locate near other firms in the same industry or complementary industries. The firm agglomeration literature has devoted little attention to the role that place-based amenities (the G_j) play in determining where firms locate. For the subset of high-tech firms that seek high-skilled workers, they should recognize that where they locate is an investment. They must form an expectation of what will be future centers of agglomeration, and the high-quality-of-life cities such as Santa Monica can solve this coordination problem. For example, for high-tech firms, the agglomeration term in Equation (7.8) may be a function of the local public good G -index.

This discrete choice approach captures the sorting by heterogeneous firms. Those firms that are land-intensive will move to areas with cheaper land. Those that are labor-intensive will tend to avoid prounion states (Holmes, 1998). Those that are high emitters will avoid nonattainment counties under the Clean Air Act (Becker and Henderson, 2000; Greenstone, 2002). Those firms that are energy-intensive will tend to concentrate in places with low electricity prices (Kahn and Mansur, 2013). Such a probabilistic approach but can be aggregated up to predict how local state policies influence the sorting of industries across space. California is a state with high land prices and high electricity prices and is prounion. It is no surprise that California has deindustrialized over the last 40 years. This reduction in the scale of manufacturing lowers pollution levels and thus feeds back into a type of virtuous cycle as these amenity improvements can serve to increase the supply of high human capital labor that tends to migrate toward high-amenity locations.

7.4.4 The evolution of cross city differences in environmental amenities

In this section, we discuss four comparative statics with respect to how changes in the spatial distribution of urban environmental amenities affect the spatial distribution of heterogeneous households and firms, manufacturing dynamics, air quality dynamics, land use zoning, and climate amenities. Such environmental amenity dynamics affect the cross city locational choices of households and firms. We motivate the discussion with a discussion of the role environmental amenities have played in the evolution of the city of Pittsburgh's evolution from a polluted industrial center to a high-quality-of-life center of postindustrial commerce.

7.4.4.1 The case of Pittsburgh

Consider the case of Pittsburgh and the evolving role that natural resources have played in its existence over the last 250 years. European settlement in the region began in the 1740s with the establishment of an English trading post at the point of land where the confluence of the Allegheny and Monongahela Rivers form the Ohio River. Early in the region's history, it was the key role that these rivers played in transportation that gave Pittsburgh a natural advantage. The strategic importance of controlling these natural transportation corridors led first the French and then the British to establish forts at the confluence. Over the next 50 years, the city's population slowly grew up around Fort Pitt, its economy relying mainly on trading and the development of a boat building industry—both predicated on the vital role that the city's rivers played in moving people and goods to the west.

Beginning in the 1800s, Pittsburgh's natural transportation endowment was matched with a second critical natural resource endowment as an engine of growth. Throughout the 1800s and early 1900s, the cost of transporting coal was a key determinant of where iron (and later steel) production was located. The ready availability of local coal supplies combined with proximate sources of iron ore gave the Pittsburgh region a critical natural advantage in iron production and led to the region become a national leader in first iron and later steel production. By the close of the nineteenth century, these natural advantages multiplied as manufacturers realized that they could further reduce costs by eliminating the shipping of semifinished products and integrated production at the location of the coal fields. Thus, for more than 200 years (well into the second half of the twentieth century), the city of Pittsburgh was an economic juggernaut fueled by natural resource-based production advantages.²² By the year 1950, Pittsburgh's population had peaked at 676,000. After 150 years as a center for iron and steel production, the city had the third-largest concentration of corporate headquarters in the United States—built almost completely on heavy manufacturing concerns.

²² See Isard (1948) for a discussion of the evolving role of resource location in iron and steel production.

This economic success came with environmental and health costs. As the population density increased along the banks of the city's three rivers, residents experienced increased rates of infectious disease risk because of increasing concentrations of raw sewage being dumped in these rivers as people lived at high population density. Urban public health research has documented the urban death premium (Haines, 2001). Major investments in water treatment and sewage sharply reduced the urban mortality risk, but new threats to urban quality of life emerged. As trading and boat building gave way to iron and steel, an unintended consequence of the greatly increased scale of this dirty activity was large amounts of local pollution that damaged what could have been a beautiful city. In the words of one visitor to the city in 1846,

*Everybody who has heard of Pittsburgh, knows that it is a city of perpetual smoke and looks as if it was built on a descent into a bottomless pit. But the locality is eminently beautiful, in the confluence of two rivers, with bold hills everywhere in the background, richly wooded and verdant.*²³

Of course, this pollution carried health implications as well. Ongoing epidemiology research documents the health impacts of particulate matter from living in an area where coal was being burned (Barreca et al., 2013).

As is well known, by the early 1970s, Pittsburgh's steel economy, predicated on the natural advantages it enjoyed due to transportation and cheap coal, came crashing down as the US economy began its seismic shift away from heavy manufacturing. In the short run, the loss in manufacturing jobs was devastating to the region's economy. However, as documented by Kahn (1999) and Chay and Greenstone (2003), a silver lining of Rust Belt decline (starting in the 1970s and accelerated by the deep 1981 recession) has been a reduction of air and water pollution levels in Rust Belt cities such as Pittsburgh.

Pittsburgh of today has successfully reinvented itself, taking advantage of the reductions in industrial pollution that occurred with the loss of the steel industry to transition from a city where natural resources are used as inputs to production processes to an economy where these resources are packaged as amenities, which make the city an integral recruitment tool for the medical, high-tech institutions of higher education that now characterize the city's economy. Quoting from recruitment materials on Google's Pittsburgh recruitment website,

Back in the day, this was a steel town, known for its smokestacks and smelters. Today, it's a technology hotbed, a model for cities trying to transition from an industrial past to a knowledge-based future. Home to Carnegie Mellon and the University of Pittsburgh, the Burgh produces some of the finest engineering talent in the world. It's also one of the most livable places in the U.S. Don't take our word for it, though. Ask Forbes and the Economist.

²³ Case (1846).

7.4.4.2 Air pollution

In Equation (7.7), we noted that environmental and urban economists take a reduced form approach in modeling a city's G at a point in time. In the case of air pollution, a standard accounting approach is to consider scale, composition, and technique effects. Consider the emissions generated by different industries.²⁴ To simplify the problem, assume there are a clean industry and a dirty industry. At a point in time, the total emissions in this city can be expressed as

$$\text{Total emissions}_{jt} = \text{output}_{jt} * \text{emissions factor}_{ct} + \text{output}_{dt} * \text{emissions factor}_{dt} \quad (7.9)$$

where the c subscript indicates the clean industry and the d subscript stands for the dirty industry.

A growing city can experience reduced pollution if the composition of economic activity is shifting from dirty output to clean output and if the emissions factor for the dirty industry is significantly higher than for the clean industry. Kahn (1999) and Chay and Greenstone (2003) documented the particulate progress observed in the US Rust Belt as heavy steel producers reduced production in the face of heavy international competition. Kahn (2003) showed similar results for eastern European cities after the fall of communism. Similar trends are now playing out in many of China's richer coastal eastern cities (Zheng and Kahn, 2013; Zheng et al., 2014a,b). Pope et al. (1992) presented an early natural experiment documenting the particulate reductions achieved when a steel plant went on strike.

Over the last 40 years, US big cities have experienced an industrial composition shift as the share of jobs in manufacturing has decreased and the share working in the service sector has increased. There have been many causes of the deindustrialization of US major cities including high wages and the rise of international trade (Autor et al., 2013) and differential enforcement of Clean Air Act regulation (Henderson, 1996; Kahn, 1997; Becker and Henderson, 2000; Greenstone, 2002). As major cities feature fewer manufacturing jobs, this reduces the likelihood that middle-class households live in these cities.

This discussion has focused on the United States but the same issues arise in China. As center city land prices rise, the opportunity cost for having land-intensive manufacturing remain in the center cities has increased. In many Chinese cities, urban mayors are eager to reclaim manufacturing land to sell to real estate developers (see Zheng et al., 2014a,b). An open question concerns whether such developers and local governments have strong incentives to remediate the pollution damage (such as localized toxic emissions) caused by

²⁴ To simplify this emissions inventory discussion, we focus on the emissions from industry. We recognize that the transportation sector and the household sector are also major contributors to local air pollution. Their emissions can also be studied by examining scale, composition, and technique effects (see Kahn and Schwartz, 2008).

manufacturing. In the United States, the Superfund and Brownfields cleanups are intended to achieve this goal.

Richer coastal cities such as Shanghai are tightening environmental regulations and this provides a further incentive for manufacturing to migrate west to provinces featuring cheaper land and laxer regulation. This industrial migration from eastern to western China raises the possibility of future environmental progress in China's richer eastern cities at the same time that there is pollution degradation in China's western cities. The extent of this pollution transfer hinges on the emissions control technology built into the new factories opening in western China (Zheng et al., 2014a,b).

7.4.4.3 Land use regulation/zoning

A growing set of empirical studies have documented that coastal high-quality-of-life cities are more likely to engage in land use restrictions than the average city. Such cities as Boston, San Francisco, and New York City have been documented to engage in more land use restrictions (Glaeser et al., 2005, 2006; Glaeser and Ward, 2009; Kahn, 2011). Such local limits to growth are intended to preserve local quality of life, but by limiting housing supply in specific high-amenity areas, they are likely to drive up home prices in such areas and deflect middle-class people to other geographic areas featuring more elastic housing supply such as Phoenix. This in turn has implications for both access to local public goods and the carbon emissions of the residential sector (Mangum, 2014).

Glaeser and Kahn (2010) argued that the large difference in the air-conditioning carbon footprint between San Francisco and other cities such as Houston means that, by downsizing the amount of available housing, aggressive local land use zoning in coastal cities displaces people to higher-carbon areas. This suggests that an unintended consequence of local attempts to preserve coastal quality of life is to exacerbate the global challenge of climate change. Mangum (2014) builds on this research by developing a dynamic model of housing demand that he uses to simulate how the residential sector's production of greenhouse gas emissions is affected by various counterfactual policies. He finds that imposing stricter land use regulations in high-carbon-output cities would decrease the aggregate amount of carbon output by about 1.7% (2.7% in new construction), again mostly through decreasing the housing consumed per person and secondarily by moving population to low-carbon cities. His findings build on the research agenda presented in Gaigné et al. (2012) that stresses the importance of considering general equilibrium effects when evaluating the impact of local policies related to carbon footprints.

To quote Gaigné et al. (2012),

There is a wide consensus among international institutions and national governments in favor of compact (i.e. densely populated) cities as a way to improve the ecological performance of the transport system. Indeed, when both the intercity and intra-urban distributions of activities are given, a higher population density makes cities more environmentally friendly because the average commuting length is reduced. However, when we account for the possible relocation of

activities within and between cities in response to a higher population density, the latter may cease to hold. Indeed, an increasing-density policy affects prices, wages and land rents, which in turn incentivizes firms and households to change place. This reshapes the urban system in a way that may generate a higher level of pollution. Thus, although an increase in compactness is environmentally desirable when locations are given, compactness may not be environmentally-friendly when one accounts for the general equilibrium effects generated by such a policy.

This quote highlights a tension in the applied economics literature. Most papers in the field experiments and regression discontinuity literature adopt a partial equilibrium approach as they focus on cleanly identifying a single parameter of interest. This quote highlights the importance of embedding such estimates within a general equilibrium framework.

7.4.4.4 Climate

A large amount of cross city hedonic research focuses on climate valuation. While climate may appear to be a static attribute, increased access to air-conditioning and rising household incomes have both lead to interesting dynamics with respect to the demand for climate. Based both on the geographic migration of the population over time and on hedonic pricing, one robust finding is the rising demand for warm winter areas. Hedonic research by [Cragg and Kahn \(1999\)](#) and [Costa and Kahn \(2003\)](#) documents the rising implicit price of warm winters in cross city hedonic real estate regressions. Using decennial census data, they documented that all else equal that the implicit price on winter temperature has been rising over time in hedonic real estate regressions. With the exception of coastal California and Oregon and Washington, warm winter states tend to have high summer temperatures. This bundling of “good” and “bad” amenities has been offset by the widespread penetration of air-conditioning access ([Oi, 1996](#)). [Barreca et al. \(2013\)](#) documented that over the twentieth century, there has been a sharp reduction in the mortality rate caused by extremely hot days, and they attribute this trend to the diffusion of air-conditioning.

As air-conditioning has become cheaper and of higher quality over time, warm winter and warm summer states have become more hospitable locations and more economic activity has moved there. The overall trends in population locational choice are presented in [Figure 7.2](#). In the years 1900 and 2010, we report the cumulative percentage of all people in the United States as a function of average February temperature. Consider the median person. In 1900, the median American lived in a state whose February average temperature was 30 °F. In 2010, the median American lived in a state whose temperature was 37 °F. Note that we use the same temperature distribution for both years. The introduction of a cheap market product that offsets summer heat has sharply changed the spatial distribution of economic activity. This example highlights how market products influence climate demand and population locational choice. Future research might explore how the diffusion of air-conditioning in less-developed country cities affects the spatial distribution of economic activity in these nations.

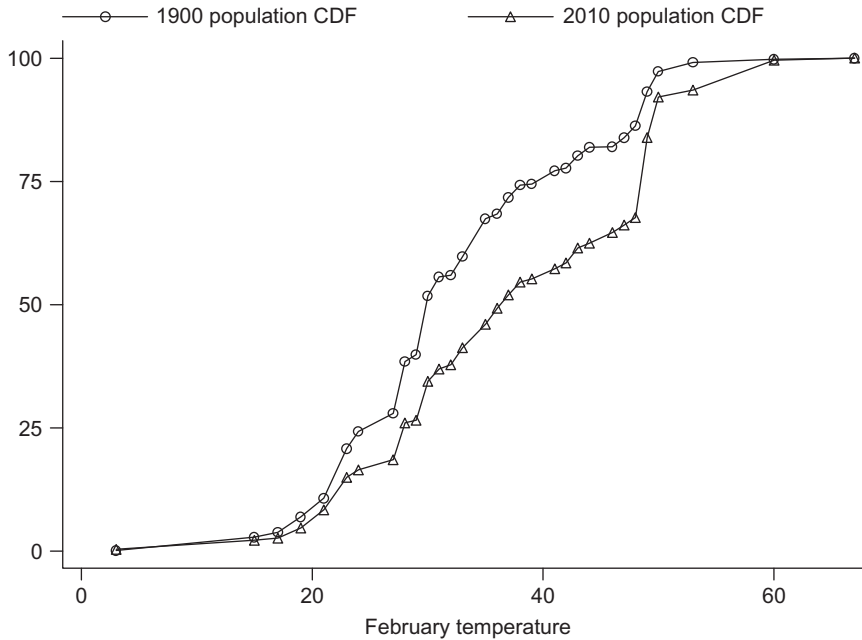


Figure 7.2 The spatial distribution of the United States population in 1900 and 2010.

7.4.4.5 Resilience to climate change shocks

While there are many unknowns about climate change's impacts on different urban areas, climate scientists do agree that mean temperature and the variance of possible outcomes regarding temperature, rainfall, and storms are likely to increase over time as global greenhouse gas emissions continue to rise. Many climate models predict significant changes in average summer and winter climate conditions. For example, in the Southwest, cities such as Las Vegas are expected to become much hotter during summer months. [Albouy et al. \(2013\)](#) used cross city hedonic regressions at a point in time to estimate the marginal valuations of climate. They combined these estimates with predictions from climate change models of how climate conditions are expected to change over time to predict how the climate amenity value of different cities will change due to climate change.²⁵ In addition to shifting mean temperature and rainfall, climate scientists have tried to be clear about their uncertainty about model accuracy and the possibility of fat tail events such as super storms. [Weitzman \(2009\)](#) explored the implications of "fat tails" events becoming more likely and having incredibly high costs if they do take place. This research line has not incorporated spatial features into the analysis. In this

²⁵ Such research implicitly assumes that there will be no future technological progress directed toward reducing the discomfort of summer heat.

section, we focus on an emerging new research agenda examining how climate change is likely to affect the spatial distribution of economic activity across diverse cities.

In an open system of cities, households and firms can migrate. This option provides a type of insurance policy as the new news about climate change's actual impacts unfold. As we discussed earlier, migration costs will inhibit such adaptation. Land prices will change in response to these shifts in migration so that the marginal household and firm are indifferent across locations (Rosen, 2002). Incumbent real estate owners will bear the incidence of new news related to how specific cities are coping with climate change (see Buntun and Kahn, 2014). Those households who face high migration costs of leaving at-risk cities (either due to their demographics or due to idiosyncratic locational capital that they have built up) will be more likely to suffer because they either are unable to move from at-risk places or will lose valuable capital in doing so.

Around the world, a large number of people and billions of dollars of real estate are located close to coastal waterways. Such coastal living offers both easy access to water amenities, beautiful views, and a temperate climate. Such coastal living puts many urbanites around the world at risk from sea level rise and natural disasters.

The impact of Hurricanes Andrew on Miami and Katrina and Sandy perhaps offers a preview of the challenges that cities will face if climate change leads to both sea level rise and more severe and intense storms. These natural disasters highlight an emerging urban research area. How will individual cities and their current urbanites be affected by climate change? Kahn (2010) argued that migration offers urbanites the option to vote with their feet if their current city's quality of life declines due to climate change. Such potential exit incentivizes urban politicians to devote effort to mitigating emerging new risks that different cities will face.

At-risk cities have a variety of strategies to protect themselves from anticipated risks. New Orleans is making enormous investments in new sea walls in response to Hurricane Katrina (Kates et al., 2006). Part of the challenge that older coastal cities face is that their existing infrastructure, such as the New York City subway, is over 100 years old and was built at a time when the emerging risks could not be imagined.

An unintended consequence of public investment in coastal armoring to fend off sea level rise is to crowd out private self-protection. More people are likely to move to a coastal city or remain there if they feel safe (Boustan et al., 2012). Kousky et al. (2006) presented a multiple equilibrium model. If an at-risk city such as New Orleans uses public funds to invest in sea walls, then private investors will be more likely to build hotels because their investment is safer. In this equilibrium, the exogenously risky city experiences economic growth because of the synergy between public and private investments. A second equilibrium in this case would be if the government does not expect that people will move to the risky city and thus the government does not build the sea walls. Anticipating that there will be no sea walls, the hotel investor does not invest in the hotel. In this "libertarian" equilibrium, less economic activity locates in the risky place. From an

ex post regret standpoint, a key unknown concerns the quality of these sea walls. If private investors and migrants are overconfident about the quality of the government-provided sea walls, then a tragedy could result. This example bears a strong similarity to [Peltzman's \(1975\)](#) famous unintended consequence example of mandatory seat belts (a safety regulation) crowding out private effort.

For geographic areas that suffer natural disaster shocks, will their long-run growth be affected? The recent macro urban research focused on urban growth after major bombings would say “no.” [Davis and Weinstein \(2002\)](#) studied Japan's urban population dynamics after World War II bombing and found that cities quickly converge back to their population regression line in the aftermath of a shock (i.e., bombings or natural disasters). Unlike with war bombing, coastal cities face the challenge of the future expectation that they will suffer future shocks of similar types of disasters (i.e., New Orleans will suffer from a future Hurricane Katrina). Such expectations may displace investment to less risky areas. In this sense, climate change risk may lower economic growth in risky areas while increasing investment in other less risky geographic areas that are close substitutes. [Hsiang and Jina \(2014\)](#) document the long-run effects of cyclones on national economic growth. They document large and persistent effects of such place-based shocks.

This section has focused on the impact of climate change on temperature and sea level rise. Another major challenge that climate scientists focus on is drought. In 2014, much of the Southwest is experiencing extreme drought. To economists, an obvious adaptation strategy is to allow water prices to rise in order to trigger both demand- and supply-side responses and to direct technological change to develop new strategies for augmenting our water supply such as water desalinization. Water regulators in the west have been slow to adopt this strategy. They often face complex objectives as they trade off several criteria including equity and efficiency ([Timmins, 2002](#)). Initial property rights disputes further inhibit the efficient allocation of water ([Libecap, 2009](#)). Whether climate change will nudge such nonprofit maximizers to more efficiently allocate a key scarce resource would be a valuable future research topic.

7.4.5 The rise of superstar green cities

Many cities are transitioning from being producer cities to becoming consumer cities ([Glaeser et al., 2001](#)). In the past, coastal cities were major hubs of production and transportation. This suggests that coastal cities and cities connected to rivers featured greater concentrations of both industrial activity and population and this scaled up pollution. In 2014, coastal and river cities in the United States now tend to have better exogenous amenities and high levels of environmental quality (the endogenous amenity). In this section, we start on the supply side to sketch some of the relevant factors responsible for the pollution progress, but then, we focus on the resulting implications for the spatial equilibrium and the economic incidence of uneven amenity improvements within a system of cities.

Today, the United States features some extremely expensive “green cities” such as San Francisco, New York City, Seattle, Portland, Boston, and Chicago, Los Angeles, and San Diego. None of these cities have a major manufacturing center. These cities all feature limited land supply both due to their topography and due to their land use regulations (Glaeser et al., 2006; Saiz, 2010; Kahn, 2011). Such “superstar cities” feature extremely high home prices and a growing share of international investors purchasing properties there to play rather than to work (Gyourko et al., 2013).

A type of social multiplier or environmental multiplier emerges. If an exogenously beautiful area such as Berkeley attracts individuals to live there, then such individuals (both through voting on public goods and taxes and through choices in private markets, i.e., tofu demand) create endogenous local attributes that may further enhance the local amenity vector (Waldfoegel, 2008). This multiplier is likely to feed on itself as like-minded people increasingly move to the area.

Standard political economy logic posits that people vote their economic self-interest. This issue arises in predicting whether specific cities will support or oppose new environmental regulations such as limits on coal-fired power plants. In areas such as West Virginia, whose economy relies on coal extraction and cheap energy, workers have incentives to vote against environmental regulation. In contrast, in high-tech cities such as San Francisco, the workers are likely to already be environmentalists who are inclined to support environmental regulation, and they will anticipate that these regulations will not endanger their employment prospects. This divergence in state environmental policy leads to even more sorting by households and firms, and the net result is that coastal cities become green area hubs; this attracts workers and firms who want this lifestyle; and then, as voters, they vote for policies to further reenforce this lifestyle. Thus, it is not surprising that a red state/blue state divide on greenhouse gas policies has opened up (see Cragg et al., 2013) and that the party composition of region’s congressional delegation can even affect the enforcement actions of federal regulators (Innes and Mitra, 2014). In this sense, the urban system of cities has implications for the political economy of externality mitigation legislation. We believe this connection between urban growth and economic comparative advantage and political voting outcomes is an important future research topic.

At the same time that coastal high-human-capital liberal cities such as Boston and San Francisco engage in strict zoning, there are other metropolitan areas such as Dallas offering more affordable, larger new homes. In a diverse population, people are free to Tiebout sort to pursue their own vision of the good life. Those who enjoy indoor leisure in larger homes will be more likely to choose to live in affordable Texas. In a population where people differ with respect to their tastes for local public goods, the subset of lower-income households with a strong preference for coastal big cities will be priced out of these markets. Such individuals can either live in a distant suburb of that metropolitan area or live in another city. Their loss in consumer surplus merits further research as it speaks to the cost of rising income inequality.

One way to allow more of the middle class to access these elite cities would be to build more housing in them. Critics of such upzoning have countered that cities such as Paris would stop being Paris if too much tall building construction took place. An open question in this literature relates to how much of the desirability of these superstar cities is due to their housing supply regulation? Does such regulation create such desirability by making it exclusive or by preserving the city's charm? To disentangle these effects would require a structural model of the demand and supply for endogenous city attributes. Is it the case that the demand to live in Paris increases because potential entrants know that housing supply is limited and thus that their Parisian neighbors will also be rich? In a diverse city, how do we aggregate public opinion to determine how much a city's "character" and "charm" have been diminished by new construction? Returning to Equation (7.7) where we discussed the supply of a city's local public goods, how much would a San Francisco's quality of life be diminished by introducing higher-rise buildings?

7.4.6 Cross city local public good valuation challenges and opportunities

The previous section highlighted how environmental shifts impact the system of cities through shifting the locational choice of heterogeneous households and firms. In this section, we discuss how such dynamics affect an econometrician's ability to recover marginal valuations of household willingness to pay for local environmental goods such as clean air. Researchers seek to recover the population's distribution of willingness to pay for such nonmarket environmental goods because this is a key input in conducting cost-benefit analysis of the past and future environmental regulations.

Researchers in the valuation literature use several different revealed preference techniques to infer marginal willingness to pay for local public goods. The most common technique continues to be to estimate cross-sectional hedonic wage and rental regressions with the goal of recovering what a standardized worker would earn in each local labor market and what would be the price of a standardized housing unit in each city. The hedonic literature has sought to decompose this differential into the exact observable parts of the bundle using multivariate regression analysis by estimating regressions of the form

$$\text{rent}_{ijk} = \sum_q \gamma_q X_i^q + \sum_c \beta_c Z_k^c + \varepsilon_k \quad (7.10)$$

In this regression, structure controls (X) are included as well as metropolitan area level variables (Z) measuring local public goods and bads such as climate and air pollution. OLS is used to estimate these regressions at a point in time. An obvious challenge that emerges is omitted variable bias. These regressions have a "kitchen sink" feel. In this age of the field experiment, few reviewers are comfortable with the assumption that the

error term in such a regression (which represents all attributes of the home observed by the renter but not observed by the econometrician) is uncorrelated with the explanatory variables.

In response to this concern, leading papers have introduced panel data approaches that incorporate fixed effects in order to sweep out geographically fixed attributes. Consider the seminal paper by [Chay and Greenstone \(2005\)](#) who studied the value of particulate reductions on county home prices. They exploited the fact that counties differ with respect to their assignment to attainment or nonattainment status under the Clean Air Act and the rule for assigning counties to this status has a fixed threshold (i.e., if your county's baseline pollution is less than the regulatory cutoff, then the county faces little regulation as it would be assigned to attainment status). They examined home price dynamics in counties assigned to the Clean Air Act's nonattainment status versus a control set of counties that just missed being assigned to face more severe regulation. As they documented in their first-stage regressions, air quality improved in those counties assigned to Clean Air Act treatment. They used a county's assignment to nonattainment status as an instrumental variable for changes in TSP.

For an instrumental variable to recover consistent estimates of the second stage, it must also pass an exclusion restriction:

$$\text{Second stage regression: } \Delta \text{Home price} = a + B * \Delta \text{TSP} + U \quad (7.11)$$

$$\text{First stage: } \Delta \text{TSP} = \alpha + \gamma * 1(\text{TSP in 1974} > \text{threshold TSP level}) + V \quad (7.12)$$

In Equation (7.12), $1(\text{TSP in 1974} > \text{threshold})$ is a dummy variable that equals one if a county's ambient particulate level in 1974 exceeded the threshold such that the county was assigned to nonattainment status under the Clean Air Act. The standard exclusion restriction in instrumental variable estimation requires that the error term in Equation (7.11) be uncorrelated with the instrumental variable.

Residential sorting and firm sorting in response to the exogenous amenity changes are likely to take place so that this key assumption does not hold. The error term (U) in Equation (7.11) represents all unobserved attributes of the city that are changing over time and capitalized into local home price changes. If the unobserved determinants of housing price changes (which would include the housing stock's quality, local school quality, and restaurant quality) are improving over time by more in cities assigned to high-regulation counties than in low-regulation counties, then the exclusion restriction is violated. Why might this be the case?

Implicit in the Chay and Greenstone argument is a *ceteris paribus* assumption that the regulation's impact on home prices solely works through its impact on ambient air quality. But throughout this survey, we have stressed that an improvement in one amenity such as clean air triggers behavioral responses by both households and firms. Recall, for instance, the work of [Banzhaf and Walsh \(2008\)](#), which documents household sorting in

response to changes in exposure to toxic air pollution. A similar dynamic is likely to play out across US counties as counties are assigned to high and low regulation levels.

As regulation reduces air pollution in a nonattainment county, this will directly lead to air quality progress. In the medium term, it will also reduce air pollution by reducing the probability that dirty footloose manufacturing plants remain in the county. They will be more likely to move to less regulated attainment counties (Kahn, 1997; Becker and Henderson, 2000; Greenstone, 2002). The exit of such firms will reduce both air pollution and other pollution margins such as water pollution and toxic releases. As overall environmental quality improves in these counties, more educated people are more likely to move in to the county and this will tend to raise home prices because local public schools will improve, peer effects will improve, and local retail and shopping quality will improve (Waldfoegel, 2008). In a nutshell, as air pollution sharply improves due to the regulation, all else is not equal.²⁶ The air quality regulation has indirect effects on other local amenities, and home price dynamics reflect this entire bundle's dynamics.²⁷

We posit that exogenous improvements in air quality in cities such as Pittsburgh will attract richer more educated people to move to these improving cities.²⁸ They will move

²⁶ Consider a distinctive example from Falck et al. (2011). Using data from Germany's past, they documented that endogenous cultural amenities affect the spatial distribution of the highly educated. Their study is based on a natural experiment that exploits the fact that baroque opera houses were built as part of local ruler's competition for status among their peers. They estimated cross region growth regression and found that areas that were close to these opera houses attracted the high skilled to live there and this agglomeration of the skilled contributed to further local economic growth.

²⁷ This same issue arises in cross-sectional spatial regression discontinuity studies. Black (1999) studied real estate price variation for homes on different sides of a school attendance boundary. Such physically close homes share the same crime and pollution attributes and the same accessibility to stores and commuting options. She interpreted the difference in the average prices at the boundary to reflect the parents' willingness to pay for better schools. Implicit in such a research design is the assumption that unobservable determinants of home prices are smooth across the school boundary. But, this statistical assumption masks a strong implicit economic assumption that is unlikely to hold in cases where there is a major jump at a physical boundary in local public goods. Consider the following model of household sorting. Households differ only with respect to their income. All households gain utility from private consumption, school quality, and Jacuzzi ownership. Assume that household utility features a complementarity between school quality and owning a Jacuzzi. The econometrician observes the price of each home and the local school quality but does not observe whether the home has a Jacuzzi. In this simple example, rich people will live in the best school quality areas and will install a Jacuzzi. Their nearby poorer neighbors will live in a worse school in a home without a Jacuzzi. The econometrician conducting the regression discontinuity study will overstate the value of school quality because the jump in the school quality at the boundary *causes* the homeowner to install the Jacuzzi. The homes located in the high-school-quality district sell for a price premium because they have better schools and they have a Jacuzzi.

²⁸ Kahn (2007) presented another example of this type of social multiplier effect based on his analysis of the medium-term effects of new rail transit lines. In studying the impact of the fast Red Line that connects Tufts University and Harvard and MIT to downtown Boston, he documented that this subway's local impact on nearby community income and home prices grew over time. He posited that as time passed, new restaurants and retail and increased investment in the local housing supply all contributed to gentrification near the transit stops but that this process took time to unfold.

in and renovate homes (i.e., install Jacuzzis) and improve neighborhoods (perhaps, their kids are better role models at school), and these attributes are not seen by the econometrician. The net effect is that the IV design presented in Equations (7.11) and (7.12) is likely to overstate the partial effect of improvements in particulate matter on home prices because all else is not held constant. The improvement in air quality triggers a sorting of specific types of people and retail that has its own impact on home prices (see [Kuminoff et al., 2013](#); [Kuminoff and Pope, 2014](#)). For example, if homes are now more valuable in Pittsburgh because of the air quality improvement and thus, the city collects more money in property tax revenue and offers better schools, this secondary effect will also be capitalized. This example highlights that researchers using panel methods for inferring marginal valuations for amenity improvements must distinguish between partial effects (i.e., the marginal effect of air quality improvement caused by the regulation holding all else equal) and total effects (i.e., the total change in home prices caused by the initially more stringent regulation).

This example suggests that a valuable research program for studying the value of local environmental amenities in the presence of household and firm sorting will involve the following component parts. First, researchers should identify plausibly exogenous supply-side causes of increases in environmental quality and estimate their direct effects on local environmental quality.²⁹ Second, in the spirit of the approaches considered in [Kuminoff et al. \(2013\)](#), researchers should explicitly model how heterogeneous households and firms respond to amenity improvements. Third, analysts need to investigate the process through which changes in household and firm composition impact other dimensions of local quality of life. For example, if college graduates now move to a city because its air quality improves, how much does local school quality improve because there is greater property tax revenue collected and a better local peer group? By combining these pieces of information, researchers should seek to separately identify both the direct partial equilibrium effect and the indirect equilibrium responses caused by the amenity improvement.

7.5. THE URBAN BUILDING STOCK'S ENERGY CONSUMPTION

In this section, we study how urban growth contributes to greenhouse gas production by exploring how the geography and the physical attributes of the extant urban real estate

²⁹ The developing world's cities offer a number of natural experiments. For example, [Cesur et al. \(2013\)](#) documented the spatial rollout of natural gas access across Turkey's cities. They demonstrated how this cleaner fuel substitutes for coal and thus improves local air quality and reduces infant mortality. [Greenstone and Hanna \(2011\)](#) documented how the phase-in of vehicle control regulations in India's cities has reduced local air pollution and infant mortality. [Zheng et al. \(2014a,b\)](#) documented the changing industrial patterns of China's dirty industry as it moves away from the coasts. All of these cases share the same structure such that there is a large change in the spatial distribution of pollution allowing researchers to estimate sorting models to recover estimates of the willingness to pay for environmental public goods.

stock affect aggregate electricity demand. We focus on five key links between energy economics and urban economics. First, from the standard monocentric model, buildings closer to the city center are likely to be taller and thus less likely to feature a single tenant living or working there. As we discuss below, in such cases, a split incentive problem arises. Second, due to the durability of the capital stock, older buildings are likely to be closer to the city center. Because energy regulations tend to focus on new construction, this suggests that there will be important spatial differences in energy intensity. Third, geographic areas differ with respect to climate conditions, local electricity pricing, energy efficiency regulation, and incentives for conservation. A recent literature has sought to quantify the importance of these factors. Fourth, the types of households and firms that choose to locate in a metropolitan area may influence the demand for solar homes and LEED-certified buildings. Environmental ideology, in addition to economic incentives such as minimizing operating expenses, may play a key role in determining the demand for green buildings.

7.5.1 Pollution externalities associated with electricity consumption

In 2012, residential and commercial buildings were responsible for 74% of total electricity consumption in the United States.³⁰ Given that electricity continues to be generated using fossil fuels (e.g., coal is the fuel source for 70% of India's power and more than 70% in China), urban real estate is a major producer of global greenhouse gas emissions. As nations such as China erect hundreds of new housing and commercial towers that will last for decades, it becomes of even greater importance to study the investment decisions in energy efficiency made by real estate developers, building owners, and their tenants. In the absence of a global carbon pricing, nations are likely to continue to rely on dirty fossil fuels such as coal and natural gas for power generation. Such cheap fossil fuels exacerbate the climate change externality.

Coal-fired power plants are major producers of a large vector of local pollutants that can cause significant harm to health and aesthetics (Davis, 2011). The economic magnitude of these localized damages hinges on how many people live downwind from these power plants and their individual willingness to pay to not be exposed to pollution. Studies based on data from the United States and from China have estimated large Pigouvian social costs associated with the use of fossil fuel-fired power plants (Zhou et al., 2006; Muller and Mendelsohn, 2007; Muller et al., 2011). In the United States today, fewer coal-fired power plants are being built and some are being shut down. The composition shift from coal- to natural gas-fired power plants should reduce both the local and the global greenhouse gas externalities associated with electricity consumption. This in turn suggests that, if major reductions can be achieved in the carbon intensity of electricity, the

³⁰ http://www.eia.gov/totalenergy/data/monthly/pdf/flow/primary_energy.pdf, http://buildingsdatabook.eren.doe.gov/docs/xls_pdf/1.1.9.pdf.

Pigouvian benefits of green buildings (relative to conventional buildings) will shrink as the carbon externality from electricity consumption shrinks.

7.5.2 The building stock's energy consumption: a bottom-up approach

At any point in time, the urban real estate stock is a mixture of residential, commercial, and industrial buildings of different vintages and sizes. In the past, empirical work aimed at understanding the role of these different factors in determining energy consumption was limited by a reliance on highly aggregated data such as per capita energy consumption collected at the nation/year level. In this new big data era, researchers have increased access to highly disaggregated data from the electric utilities that supply the power. These detailed data sets provide consumption information by customer account at monthly, and sometimes even hourly, frequencies. Explaining cross-sectional and temporal variation in residential and commercial electricity consumption has become an active research topic. By merging data from the electric utilities with structural characteristics taken from local appraiser's offices or from databases such as the CoStar data, researchers can observe a building's attributes, its location, and its energy consumption. Given that buildings differ along a number of attributes, researchers are now able to decompose what portion of a building's energy consumption is due to such factors as the physical attributes of the building, its location's regulatory structure, energy prices, and local climate conditions.

It is useful to consider a simple model of a city's energy consumption. In terms of residential energy consumption, assume that there are four key factors that determine an individual household's energy consumption: characteristics of the household members themselves (number, age structure, income, and preferences), the characteristics of the structure in which they live (size, vintage, quality, efficiency investments, etc.), the climate in which the housing unit is located, and the price paid for energy. For the industrial sector, we aggregate each industrial sector to the level of a representative firm. Each industrial sector's total electricity consumption is a function of its scale of activity in a given city, baseline energy intensity, climate, and energy prices.

Under these assumptions, the total energy consumption of the residential sector for city j at time t can be written as

$$KWH_j^t = N_j * \int_{h \in H} KWH(\text{household}_h, \text{structure}_h, \text{climate}_j, \text{price}_j) f_j(h) dh \quad (7.13)$$

In the right-hand side of Equation (7.13), H represents the continuum of household types, the distribution of which in a particular city j is given by $f_j(\cdot)$. $KWH(\cdot)$ is a function that maps the characteristics of household members and the structure in which they live—each of which varies by household type h —along with the climate and energy price vectors experienced in city j into average energy demand. N_j is the total number of households in city j .

Similarly, we can represent the total industrial consumption in city j as

$$KWH_j^i = \sum_{i \in I} KWH(\text{scale}_{ij}, \text{efficiency}_{ij}, \text{climate}_j \text{ price}_j) \quad (7.14)$$

where I represents the set of industries in the economy, scale_{ij} captures the scale of industry in city j , and efficiency_{ij} represents the average energy efficiency of firms in industry i that are located in city j .

The total greenhouse gas emissions associated with this consumption equal the sum of each of these sector's electricity consumption and then multiplied by the city j 's electric power region's emissions factor. eGRID data from the US EPA allow one to calculate the tons of carbon dioxide per megawatt of power by region by year (see <http://www.epa.gov/cleanenergy/energy-resources/egrid/>). A standard number is that the social cost of carbon dioxide is \$35 per ton. These numbers can be used to translate electricity consumption into a bottom-line social cost of carbon dioxide estimate.³¹

This accounting algebra highlights several microeconomic channels through which a city's evolving building stock influences overall city energy consumption dynamics. For example, in the very first issue of the *Journal of Urban Economics*, Harrison and Kain (1974) called attention to the importance of durable capital as a defining characteristic of cities. Over time, newer homes tend to be larger than older homes due to overall income growth. At any point in time, a city's building stock will be a mixture of different vintages. In cities with no new construction such as Detroit, buildings built more than 40 years ago can be a large share of the total stock. In recent years, such studies as Glaeser and Gyourko (2005) and Brueckner and Rosenthal (2009) have examined the role that housing durability plays in determining a city's urban form.

From an analysis perspective, it is critical to understand, and control for, the interplay among these different channels (building stock, climate, energy prices, urban geography, etc.) when evaluating the impact of new policies. For example, if Detroit has no new housing being constructed, does it, at baseline, have a much higher electricity consumption per square foot for its average housing stock than a similar city where old homes are being scrapped and new homes are being built? If so, then, these differences must be accounted for when making cross location policy comparisons.

Over time, energy efficiency codes for new buildings have been tightened (see <http://www.energycodes.gov/regulations>). Given the durability of the building stock, the average energy intensity may only be affected after several years have passed. Recent research has investigated the effectiveness of past building code regulation. Jacobsen and Kotchen

³¹ A second negative externality from electricity consumption is local air pollution. The eGRID data can be used to calculate the sulfur dioxide and nitrogen oxide emissions and mercury emissions associated with local power consumption. These emissions may drift far from the origin city to geographic locations downwind of the power plants (see Bayer et al., 2009). In this sense, the consumption of electricity does not necessarily have local impacts on the city because power is generated in the region not in the locale.

(2013) collected data for roughly 2000 Gainesville, Florida, homes built at different dates. They compared the energy consumption of buildings built just before and after the regulatory changes (introduced in 2001) and found that the regulation is associated with a 4% reduction in electricity consumption and a 6% reduction in natural gas.

The third term in Equations (7.13) and (7.14) is local climate conditions. Both residential real estate and commercial real estate will consume more electricity in more humid hotter summer areas. Researchers have estimated the relationship between summer cooling degree days using both cross-sectional and panel approaches (Glaeser and Kahn, 2010; Miller et al., 2008; Aroonruengsawat and Auffhammer, 2011).

The basic microeconomic model here is a Becker household production function. A household seeks comfort and higher outdoor temperature reduces comfort. Greater expenditures on air-conditioning offset the outdoor temperature. The marginal cost of reducing the indoor temperature by one degree is a function of the efficiency of the air-conditioning unit and the price of electricity. The rational household sets the thermostat so that the marginal benefits of temperature reduction just equal the marginal cost of temperature reduction. In cooler cities such as San Francisco, such comfort can be produced using fewer market inputs, and in fact, many households do not own a central air conditioner. The widespread diffusion of air-conditioning over the twentieth century has allowed US households to enjoy the winter warmth of cities such as Las Vegas and Phoenix without suffering from the summer heat (Oi, 1996; Barreca et al., 2013). In this sense, market products can offset parts of a tied spatial local bundle. California is one of the rare locations in the world to feature temperate winters and summers.

7.5.3 Weak price incentives for conservation

For a typical single-family homeowner, their annual electricity bill is typically around \$1200 per year.³² The marginal cost in terms of effort of achieving a 15% reduction in electricity consumption might exceed the cost savings given current electricity prices. Wolak (2011) presented a bounded rationality model featuring a cost of action such that residential homeowners are only willing to take costly actions (i.e., turning off the lights) if the financial savings in lower electricity bills covers the cost of taking such an action. His analysis suggests in areas with higher electricity prices that the residential sector will be more energy-efficient because there is “more money on the table.” In addition to the impact of the price levels per se, households are confronted with a complicated electricity bill structure. Many electric utilities charge households on an increasing block tariff structure such that the price schedule resembles a staircase with a rising marginal price that jumps when consumption exceeds specific thresholds. For example, in Southern California, electricity consumers face a rate schedule with five steps with the highest price being over 30 cents/kWh. Ito (2014) employs a regression discontinuity design and

³² http://www.eia.gov/electricity/sales_revenue_price/pdf/table5_a.pdf.

demonstrates that Southern California electricity consumers respond to average prices not marginal prices. A notable feature of his research design is that he compares the consumption of nearby homes that are located in adjacent electric utility districts. This means that there is within neighbor variation in the price of electricity. It exploits this and estimates an average price elasticity of consumption of roughly -0.12 . This suggests that residential consumers are not very price-responsive. Recognizing this fact, utilities have sought innovative tools for encouraging conservation. Some have appealed to public conservation campaigns. [Reiss and White \(2008\)](#) presented evidence that these can be effective in the short run. A more novel approach is to distribute Home Energy Reports, which are customized reports educating households on how their consumption of electricity compares to their neighbors. Such peer pressure has been claimed to be an effective social incentive for encouraging conservation. [Allcott \(2011a,b\)](#) evaluation of these reports' impact indicates that they reduce consumption by 2%. These various research designs highlight that there is still the open question of how to make energy efficiency a more salient feature for households to consider. Higher electricity prices would clearly provide an incentive for households to invest more effort studying this issue. International evidence would be valuable here.

New technologies such as smart meters allow households to observe their electricity consumption every 15 min and to see how different devices in the home use different amounts of electricity. In the past, households received a bundled monthly bill telling them their aggregate consumption and total expenditure. With the falling cost of installing smart meters, the cost for electricity consumers of being educated about how different actions such as turning on the lights or running the air conditioner is falling. Such educated consumers should be more sophisticated shoppers for durables that reduce their operating costs and in aggregate improve overall energy efficiency. Field experiments, perhaps randomizing who receives a smart meter, testing this claim would be very useful.

7.5.4 The commercial real estate sector

In many cities, the commercial real estate sector (i.e., office buildings, malls, stores, and restaurants) represents a large share of the total real estate stock, but there has been surprisingly little research on this sector's economic behavior. Until recently, the lack of data has slowed research. The DOE's Commercial Buildings Energy Consumption Survey (CBECS) was last conducted in 2003. In some cases, research partnerships with electric utilities and with private major companies have allowed researchers access to unique microdata that have allowed for progress.

[Kahn et al. \(2014\)](#) partner with a major western electric utility who provided billing account data for every commercial account in their service area. This research team matched these data to the CoStar data. CoStar sells data for buildings that have transacted in the previous 10 years. By matching the two data sets, the researchers observed each

building's monthly electricity consumption and had access to a large set of building attributes including its real estate quality, size, year built, and the tenant roster. They used these data to estimate regressions such as those reported in Equation (7.15):

$$\ln y_i = \gamma \cdot X_i + \sum_{n=1}^k \varphi_p \cdot T_i^n + \varepsilon_i \quad (7.15)$$

In this equation, the dependent variable y_i is the natural logarithm of average daily electricity consumption per square foot. X_i is a vector of the structural characteristics of building i , including building size, vintage, and quality. To control for the impact of occupants on building energy consumption, we also include T_i , a vector of variables measuring the percentage of the building that is occupied by each industry n , based on their SIC classification. ε_i is an error term, assumed to be i.i.d.

Kahn et al. (2014) report that newer vintages of commercial real estate consume more electricity than older vintages and class "A" higher-quality real estate consumes more electricity than lower-quality real estate. They interpret these results to mean that a building's quality to its users and electricity inputs are complements. Given that building quality is likely to be a normal good, this suggests that rising incomes in cities around the world will lead to increased commercial building electricity consumption. They also find that tenant contract terms matter. When tenants pay their own electricity bills through triple net leases, they consume less electricity. Government tenants, who presumably face soft budget constraints, consume more electricity.

7.5.4.1 Commercial real estate energy efficiency and human capital

A key difference between residential single-family homes and commercial entities is that the former are small utility maximizers, while the latter are large profit maximizers. In contrast to the large number of residential households, each of whom spends relatively little on electricity bills, commercial electricity consumers are fewer in number and have much larger electricity bills. Kahn and Kok (2014) study Walmart's electricity consumption across its 220 California stores using monthly data provided by the company covering the years 2006–2012. The company provided additional data on each Walmart's store's climate conditions, year built, and the size and purpose of the store. Kahn and Kok use as a control group a set of other major retail stores that are independently managed but of the same size and age as these Walmart stores. In comparing the electricity consumption across these two sets, an interesting pattern emerges. The Walmart stores feature almost no cross store variation in their electricity consumption per square foot. Using their management systems, Walmart has standardized efficiency gains in electricity consumption. A human capital explanation for this is that Walmart's size is such that it is cost-effective for it to invest in high-quality managers who focus on energy efficiency and then apply the cost-saving innovations that they identify to each of its stores. This claim

mirrors an argument explored by Bloom et al. (2010) who found that higher-quality managers of UK manufacturing plants use less electricity in producing output. This research agenda suggests that there is an inverse relationship between corporate management quality and X -inefficiency. Given that electricity consumption contributes to the global climate change externality, this hypothesis and the potential for economies of scale in achieving energy efficiency merit future research.

7.5.4.2 The market for green buildings

Urban residential and commercial structures are differentiated products. One attribute of such products is their energy efficiency. Standard economic logic suggests that when the price of electricity is high or is expected to rise (perhaps because a carbon tax is expected in the future), there will be greater demand for “green buildings.” An emerging vehicle demand literature has documented that rising gasoline prices are associated with rising demand for fuel-efficient hybrid vehicles (Li et al., 2009; Klier and Linn, 2010; Beresteanu and Li, 2011). There are other factors such as personal ideology that also correlate with demand for distinctive green cars such as the Toyota Prius (Kahn, 2007). Gas prices are highly salient in that car drivers refuel on a weekly basis. In contrast, many households never see their monthly electricity bill because they pay with automatic billing. Sexton (2014) uses data from South Carolina and documents that when households sign up for automatic billing payments, their electricity consumption increases by 4% and commercial consumers increase their consumption by 8%. He attributes these effects to the absence of salient bills as these consumers no longer are confronted with what they are actually paying for electricity.

While fuel economy is the defining attribute of a “green car,” in the case of “green buildings,” there continues to be an open debate about what objective attributes signal a building’s “greenness.” The presence of solar panels is one possible definition. In the United States, the government has played an active role in being a market maker. The ENERGY STAR program, jointly sponsored by the US Environmental Protection Agency and the US Department of Energy, is intended to identify and promote energy-efficient products, appliances, and buildings. In a parallel effort, the US Green Building Council, a private nonprofit organization, has developed the LEED green building rating system to encourage the “adoption of sustainable green building and development practices.”³³ The inventory of certified green commercial space in the United States has increased dramatically since the introduction of rating schemes (Kok et al., 2011).

³³ The LEED label requires sustainability performance in areas beyond energy use, and the requirements for certification of LEED buildings are substantially more complex than those for the award of an ENERGY STAR rating. The certification process for homes measures six distinct components of sustainability: sustainable sites, water efficiency, materials and resources, indoor environmental quality, innovation, and energy performance. Additional points can be obtained for proximity to public transport and for awareness and education.

A growing empirical hedonic real estate pricing literature has sought to estimate the market capitalization of such green attributes using hedonic regressions of the form

$$\log(\text{price}_{ijt}) = B_1 * X_{jt} + \delta_t + \pi_j + B_2 * \text{Green}_{it} + \varepsilon_{ijt} \quad (7.16)$$

In this regression, the researcher controls for housing structure observables and includes interactions of time- and geographic-fixed effects (e.g., zip code by year fixed effects). Controlling for these variables, the researcher tests whether B_2 is positive. One recent study investigated the capitalization of solar panels into the price of homes in San Diego County and Sacramento County in California. Solar panels are a source of pride for environmentalist households and lower the operating costs of owning a home. [Dastrup et al. \(2012\)](#) found that all else equal that solar homes sell for a 3% price premium. A plausible concern is that the presence of solar panels is correlated with other unobserved measures of a structure's quality. Using their panel data on home sales, they observed some solar homes that sold before the solar system was installed. They exploited this within home variation to establish that homes that in the future will install solar panels do not sell for a price premium before such an installation.

In another California hedonic real estate study, [Kahn and Kok \(2014\)](#) study the capitalization of energy star and LEED homes in California. The control groups are homes in the same zip code and on the same street as the energy-efficient homes sold at the same point in time. Kahn and Kok report that ENERGY STAR homes sell for a 3% price premium, but LEED homes do not sell for a price premium. The open question here pertains to what this label premium represents in terms of the present value of saved operating expenses and in terms of better indoor air quality and home performance.

A homeowner who wants solar panels faces a make-versus-buy decision. She can either buy a home that already has them or purchase a home and have the system installed. This option places restrictions on how high the solar price premium for resold homes can be. Over time due to international competition, the price of solar panels is declining. An emerging literature is studying this adoption decision. State and national subsidies have certainly played a role in encouraging the adoption of solar panels ([Bollinger and Gillingham, 2012](#)). [Kahn and Vaughn \(2009\)](#) documented the clustering of LEED buildings in liberal/environmentalist states such as California. An open topic concerns the diffusion of solar technology to cities in the developing world. Such cities often suffer from grid reliability challenges and solar would allow households to be independent.

For the commercial real estate market, a series of papers that study investor and tenant demand for "green" office space in the United States show that buildings with an ENERGY STAR label—indicating that a building belongs to the top 25% of the most energy-efficient buildings—or an LEED label have rents that are 2–3% higher as compared to regular office buildings. Transaction prices for energy-efficient office buildings are higher by 13–16%. Further analyses show that the cross-sectional variation in these premiums has a strong relation to real energy consumption, indicating that tenants and

investors in the commercial property sector capitalize energy savings in their investment decisions (Eichholtz et al., 2010, 2013).

There has also been recent research on the supply side of the green commercial buildings market focused on the motivations of real estate owners and developers to compete on who “is the most green.” The LEED certification has different levels of building greenness such as silver, gold, and platinum, and the threshold of points needed to cross to a higher level of greenness can be achieved through relatively minor investments (Noonan et al., 2013; Matisoff et al., 2014). An open question is whether such a “green” arms race actually contributes to overall sustainability goals such as reducing greenhouse gas emissions. The LEED criteria include many indicators of sustainability (such as using recycled materials), while the ENERGY STAR criteria are more narrowly focused on energy consumption.³⁴

There are several possible next steps for this “value of green buildings” literature. Note that hedonic regressions such as Equation (7.16) do not control for the actual energy and water consumption of the building. Suppose that in estimating Equation (7.16), the key explanatory variable is a dummy that equals one if the building is LEED-certified. If continuous indicators of operating expenses could be observed, would the discrete certification dummy still be capitalized? Given asymmetrical information about a building’s actual operating costs, does the green building dummy act as a coarse partition. Or do green buildings perform better on other criteria such as offering higher indoor air quality and hence raising worker productivity and quality of life? This literature could also explore whether green certification is more valuable in areas with higher average electricity prices. By including such an interaction term, scholars could study whether the demand for green buildings is mostly tied to minimizing operating costs. Such research could mimic Ito’s (2014) spatial regression discontinuity design to observe the pricing of nearby buildings located in different electric utility zones.

7.5.4.3 Challenges in increasing the urban real estate's energy efficiency

Given the durability of the urban real estate stock, the vast majority of buildings are not new. An ongoing effort seeks to increase the energy efficiency of this stock, and many have argued that the equilibrium is unlikely to be a pareto optimum because of information asymmetries.

Consider a car buyer. The Internet offers easy to access information on any make and model’s miles per gallon. Marketing experts have argued that this should be expressed as gallons per mile to help buyers calculate their annual operating expenses (Allcott, 2011a,b). In the case of apartments, there is no analogous energy labeling. Potential renters are likely to be unaware of what their monthly electricity bills will be in a building that they have never lived in before. This information asymmetry poses a sustainability

³⁴ <http://www.usgbc.org/Docs/Archive/General/Docs1095.pdf>.

challenge because if potential renters had full information about each apartment's energy efficiency, then the inefficient apartments would rent for less (especially in areas featuring high electricity prices) and this would incentivize the owners of these apartments to invest in greater energy efficiency.

In recent years, California has passed new regulations (AB1103) requiring the owner or operator to disclose the benchmarking data and rating to a prospective buyer, lessee of the entire building, or lender that would finance the entire building based on a schedule of compliance established by the State Energy Resources Conservation and Development Commission.

Recent research has examined the role of energy labels using data from around the world. The EU energy label seems to be quite effective in resolving the information asymmetry in understanding the energy efficiency of dwellings: [Brounen and Kok \(2011\)](#) estimated hedonic pricing gradients for recently sold homes in the Netherlands and documented that homes receiving an "A" label in terms of energy efficiency sell for a 10% price premium. It is relevant to note that the Netherlands has chosen a color-coded easy to understand educational labeling system where homes are partitioned into those that receive an "A," "B," "C," etc. This report card system is similar to how North Carolina reports grades for local school district's school quality ([Kane et al., 2006](#)). In the case of ENERGY STAR labeling, [Houde \(2014\)](#) uses a structural model to document that consumers respond to labels in different ways such that some greatly value such labels while others do not. The causes of these different responses merit future research.

Conversely, dwellings that are labeled as inefficient transact for substantial discounts relative to otherwise comparable, standard homes. Similar evidence has been documented for Singapore: [Deng et al. \(2012\)](#) found that homes labeled under the government-designed Green Mark scheme sell for a 4–6% price premium. In China today, no one trusted set of standards for ranking buildings exists ([Zheng et al., 2012](#)). The emergence of trusted energy efficiency standards is an important example of the public goods that developing country governments can cheaply provide in furthering urban environmental sustainability.

Electric utilities are becoming increasingly aware of the "power of big data" and are actively working with academics to devise strategies for encouraging energy efficiency. In states such as California where utilities must comply with a renewable portfolio standard (defined as the ratio of green power purchased divided by total power generated), one cost-effective strategy is to invest in energy efficiency and hence shrink the denominator. This regulatory pressure has increased the willingness of utilities to introduce field experiments to learn about what strategies are effective.

7.5.4.4 The renter/owner split incentive problem in cities

As predicted by the monocentric model, center cities are more likely to feature multi-family apartment buildings and multitenant commercial buildings than will be found in

the distant suburbs. Such tall buildings offer significant economies of scale with respect to land consumption, but they raise issues of how to design proper energy efficiency incentives to encourage efficient investment by both the owner of the building and his or her tenants (see [Levinson and Niemann, 2004](#); [Maruejols and Young, 2011](#)).

Based on the year 2000 census, 65% of metropolitan area housing units with five or more units were located within 10 miles of the CBD, while 42% of the metropolitan area single-family housing stock was located within 10 miles of the CBD. A “split incentive” literature has documented the energy efficiency challenge in multifamily housing ([Levinson and Niemann, 2004](#)). If the owner of the apartment complex pays the energy bills, then the renter faces a zero marginal cost for energy consumption. Since renters anticipate that they will not live in the building for a long period, they are unlikely to update energy-intensive durables such as the refrigerator or other appliances that they cannot easily take with them to their next home. If the renter pays the energy bills, then the owner has little incentive to invest in energy-efficient investments such as energy-efficient windows or an energy-efficient air-conditioning system. For evidence from the commercial real estate sector, see [Kahn et al. \(2014\)](#).

This would appear to be a key example where a randomized field experiment research design would yield valuable new knowledge. Suppose that an REIT was willing to select at random which of its buildings in its portfolio that it would be willing to offer net leases versus triple net leases. This randomization of rental contract terms would allow the researcher to study how the building manager and the tenant change their behavior in the face of the new incentives. By studying the dynamics of such buildings’ electricity consumption, the researcher could estimate whether incentives matter for overall electricity consumption. The nonexperimental evidence reported in [Kahn et al. \(2014\)](#) and [Levinson and Niemann \(2004\)](#) suggests that they do, but a field experiment design would be more definitive.

7.5.5 Carbon pricing and the building stock's energy efficiency

At this point in time, the global free-rider problem has precluded the adoption of the global carbon tax. Such a credible policy would raise local electricity prices and incentivize developers and owners of existing buildings to invest in energy efficiency. In many developing nations today, a huge wave of urbanization is taking place. The investments made now will last for decades. If the investors anticipate that carbon pricing will become a reality, then they are more likely to invest in energy efficiency now. Energy efficiency investments in new capital in the developing world’s cities remain an important future research topic. Research on endogenous innovation (see [Acemoglu and Linn, 2004](#)) highlights an optimistic hypothesis. If there is a growing global demand for “green buildings,” then this creates strong market incentives for engineers to focus their efforts in creating such new varieties. The empirical induced innovation literature has

consistently reported evidence of improvements in energy efficiency of new products such as air conditioners shortly after energy prices rise (Newell et al., 1999).

7.6. CONCLUSION

In this chapter, we have explored the connection between cities and the environment, highlighting the bidirectional nature of this relationship. One set of the reviewed literature examines the role that environmental amenities play in determining the economic geography of where diverse households and firms locate both within cities and across cities. A second set of relevant literature explores how a city's economic geography and the composition of its durable capital impact the environment—focusing largely on energy consumption and carbon emissions.

We used the long-run history of Pittsburgh to illustrate our sweeping theme that the relationship between cities and their environment has changed over the last century in a systematic way. This change has had a profound impact on where and how individuals and firms choose to locate. These shifts in the geography of economic activity have in turn influenced how cities impact the environment.

In the past, as highlighted by the case of Pittsburgh, the comparative advantage of cities was largely determined by access to transportation and productive resource that arose out of the natural environment. As these cities grew into the economic engines that characterize the industrial age, unchecked externalities severely impeded quality of life for their residents. Over 40 years ago, Nordhaus and Tobin (1972) wrote

Many of the negative "externalities" of economic growth are connected with urbanization and congestion. The secular advances recorded in NNP figures have accompanied a vast migration from rural agriculture to urban industry. Without this occupational and residential revolution we could not have enjoyed the fruits of technological progress. But some portion of the higher earnings of urban residents may simply be compensation for the disamenities of urban life and work. If so we should not count as a gain of welfare the full increments of NNP that result from moving a man from farm or small town to city. The persistent association of higher wages with higher population densities offers one method of estimating the costs of urban life as they are valued by people making residential and occupational decisions.

Due to a confluence of trends, many cities have enjoyed a sharp improvement in their nonmarket quality of life with many of these gains concentrated along environmental margins. As a result, the requirement for big cities to provide “combat pay” has attenuated as nonmarket quality of life has increased and is a complement of the modern consumer city.

Air quality, water quality, and energy efficiency have become an emergent property of the industry, buildings, regulations, voters, and households who locate within a city's borders. Progress along these dimensions has been seen in cities such as Chicago, London, and New York City. Researchers see the start of similar trends in China. As demonstrated

in work surveyed by [Zheng and Kahn \(2013\)](#), China's emerging system of cities is wrestling with many of these same issues. Some of China's cities are choosing to specialize in heavy industry and this is raising pollution challenges. At the same time, rich eastern Chinese cities are deindustrializing and pollution is declining. As the urban middle class grows in cities around the world, more and more citizens will demand nonmarket quality of life and environmental progress.

ACKNOWLEDGMENT

We thank Devin Bunten, Brandon Fuller, Todd Sinai, and the handbook's editors for useful comments.

REFERENCES

- Acemoglu, D., Linn, J., 2004. Market size in innovation: theory and evidence from the pharmaceutical industry. *Q. J. Econ.* 119 (3), 1049–1090.
- Albouy, D., 2008. Are Big Cities Bad Places to Live? Estimating Quality of Life Across Metropolitan Areas: National Bureau of Economic Research Working paper 14472.
- Albouy, D., Graf, W., Kellogg, R., Wolff, H., 2013. Climate Amenities, Climate Change, and American Quality of Life: NBER Working paper No. 18925. National Bureau of Economic Research.
- Allcott, H., 2011a. Social norms and energy conservation. *J. Public Econ.* 95 (9), 1082–1095.
- Allcott, H., 2011b. Consumers' perceptions and misperceptions of energy costs. *Am. Econ. Rev.* 101 (3), 98–104.
- Aroonruengsawat, A., Auffhammer, M., 2011. Impacts of climate change on residential electricity consumption: evidence from billing data. In: Libecap, G., Steckel, R.H. (Eds.), *The Economics of Climate Change: Adaptations Past and Present*. University of Chicago Press, pp. 311–342.
- Autor, D.H., Dorn, D., Hanson, G.H., 2013. The China syndrome: local labor market effects of import competition in the United States. *Am. Econ. Rev.* 103, 2121–2168.
- Banzhaf, H.S., McCormick, E., 2006. Moving Beyond Cleanup: Identifying the Crucibles of Environmental Gentrification. National Center for Environmental Assessment, US Environmental Protection Agency.
- Banzhaf, H.S., Walsh, R., 2008. Do people vote with their feet? An empirical test of Tiebout's mechanism. *Am. Econ. Rev.* 98 (3), 843–863.
- Banzhaf, H.S., Oates, W.E., Sanchirico, J.N., 2010. Success and design of local referenda for land conservation. *J. Policy Anal. Manag.* 29 (4), 769–798.
- Barreca, A., Clay, K., Deschenes, O., Greenstone, M., Shapiro, J.S., 2013. Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the 20th Century: NBER Working paper No. 18692. National Bureau of Economic Research.
- Barreca, A., Clay, K., Tarr, J., 2014. Coal, Smoke, and Death: Bituminous Coal and American Home Heating: NBER Working paper No. 19881. National Bureau of Economic Research.
- Baum-Snow, N., Kahn, M.E., 2005. Effects of Urban Rail Transit Expansions: Evidence from Sixteen Cities, 1970–2000. *Brookings-Wharton Papers on Urban Affairs*, pp. 147–206.
- Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. *J. Polit. Econ.* 115 (4), 588–638.
- Bayer, P., Keohane, N., Timmins, C., 2009. Migration and hedonic valuation: the case of air quality. *J. Environ. Econ. Manag.* 58 (1), 1–14.
- Bayer, P., McMillan, R., Murphy, A., Timmins, C., 2011. A Dynamic Model of Demand for Houses and Neighborhoods: NBER Working paper No. 17250. National Bureau of Economic Research.
- Becker, G.S., 1976. Altruism, egoism, and genetic fitness: economics and sociobiology. *J. Econ. Lit.* 14, 817–826.

- Becker, R., Henderson, J.V., 2000. Effects of air quality regulations on polluting industries. *J. Polit. Econ.* 108 (2), 379–421.
- Been, V., 1994. Locally undesirable land uses in minority neighborhoods: disproportionate siting or market dynamics? *Yale Law J.* 103, 1383–1422.
- Been, V., Gupta, F., 1997. Coming to the nuisance or going to the barrios—a longitudinal analysis of environmental justice claims. *Ecol. LQ* 24, 1.
- Bento, A.M., Franco, S.F., Kaffine, D., 2006. The efficiency and distributional impacts of alternative anti-sprawl policies. *J. Urban Econ.* 59 (1), 121–141.
- Beresteanu, A., Li, S., 2011. Gasoline prices, government support and the demand for hybrid vehicles in the United States. *Int. Econ. Rev.* 52 (1), 161–182.
- Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. *Econometrica* 63 (4), 841–890.
- Bishop, K.C., 2012. A Dynamic Model of Location Valuation and Hedonic Valuation. ASU Working paper.
- Bishop, K.C., Murphy, A.D., 2011. Estimating the willingness to pay to avoid violent crime: a dynamic approach. *Am. Econ. Rev.* 101 (3), 625–629.
- Black, S.E., 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 114 (2), 577–599.
- Bleakley, H., Lin, J., 2012. Portage and path dependence. *Q. J. Econ.* 127, 587–644.
- Blomquist, G.C., Berger, M.C., Hoehn, J.P., 1988. New estimates of quality of life in urban areas. *Am. Econ. Rev.* 78 (1), 89–107.
- Blonigen, B.A., Knittel, C.R., Soderbery, A., 2013. Keeping It Fresh: Strategic Product Redesigns and Welfare: NBER Working paper No. 18997. National Bureau of Economic Research.
- Bloom, N., Genakos, C., Martin, R., Sadun, R., 2010. Modern management: good for the environment or just hot air? *Econ. J.* 120 (544), 551–572.
- Bollinger, B., Gillingham, K., 2012. Peer effects in the diffusion of solar photovoltaic panels. *Mark. Sci.* 31 (6), 900–912.
- Boustan, L.P., Kahn, M.E., Rhode, P.W., 2012. Moving to higher ground: migration response to natural disasters in the early twentieth century. *Am. Econ. Rev.* 102 (3), 238–244.
- Bowles, S., 1998. Endogenous preferences: the cultural consequences of markets and other economic institutions. *J. Econ. Lit.* 36, 75–111.
- Brounen, D., Kok, N., 2011. On the economics of energy efficiency in the housing market. *J. Environ. Econ. Manag.* 62, 166–179.
- Brueckner, J.K., 2000. Urban sprawl: diagnosis and remedies. *Int. Reg. Sci. Rev.* 23 (2), 160–171.
- Brueckner, J.K., Neumark, D., 2014. Beaches, sunshine, and public sector pay: theory and evidence on amenities and rent extraction by government workers. *Am. Econ. J. Econ. Policy* 6 (2), 198–230.
- Brueckner, J.K., Rosenthal, S.S., 2009. Gentrification and neighborhood housing cycles: will America's future downtowns be rich? *Rev. Econ. Stat.* 91 (4), 725–743.
- Brueckner, J.K., Sridhar, K.S., 2012. Measuring welfare gains from relaxation of land-use restrictions: the case of India's building-height limits. *Reg. Sci. Urban Econ.* 42 (6), 1061–1067.
- Brueckner, J.K., Thisse, J.-F., Zenou, Y., 1999. Why is central Paris rich and downtown Detroit Poor? An amenity-based theory. *Eur. Econ. Rev.* 43 (1), 91–107.
- Bunten, D., Kahn, M.E., 2014. The Impact of Emerging Climate Risk on Urban Real Estate Price Dynamics: NBER Working paper #20018.
- Cameron, T.A., McConnaha, I.T., 2006. Evidence of environmental migration. *Land Econ.* 82 (2), 273–290.
- Carlino, G.A., Saiz, A., 2008. City Beautiful: IZA Discussion Papers No. 3778. Institute for the Study of Labor (IZA).
- Carlton, D.W., 1983. The location and employment choices of new firms: an econometric model with discrete and continuous endogenous variables. *Rev. Econ. Stat.* 65, 440–449.
- Case, L.J.B., 1846. Notes from a Journal No III. Ladies' Reposit. 14, 181–183.
- Cesur, R., Tekin, E., Ulker, A., 2013. Air Pollution and Infant Mortality: Evidence from the Expansion of Natural Gas Infrastructure: NBER Working paper 18736.

- Chay, K.Y., Greenstone, M., 2003. The impact of air pollution on infant mortality: evidence from geographic variation in pollution shocks induced by a recession. *Q. J. Econ.* 118 (3), 1121–1167.
- Chay, K.Y., Greenstone, M., 2005. Does air quality matter? Evidence from the housing market. *J. Polit. Econ.* 113 (2), 376–424.
- Cheshire, P., Sheppard, S., 2006. The welfare economics of land use planning. *J. Urban Econ.* 52 (2), 242–269.
- Clay, K., Troesken, W., 2010. Did Frederick Brodie Discover the World's First Environmental Kuznets Curve? Coal Smoke and the Rise and Fall of the London Fog: NBER Working paper No. 15669. National Bureau of Economic Research.
- Costa, D.L., Kahn, M.E., 2003. The rising price of nonmarket goods. *Am. Econ. Rev.* 93 (2), 227–232.
- Cragg, M.I., Kahn, M.E., 1999. Climate consumption and climate pricing from 1940 to 1990. *Reg. Sci. Urban Econ.* 29 (4), 519–539.
- Cragg, M.I., Zhou, Y., Gurney, K., Kahn, M.E., 2013. Carbon geography: the political economy of congressional support for legislation intended to mitigate greenhouse gas production. *Econ. Inq.* 51 (2), 1640–1650.
- Currie, J., Davis, L., Greenstone, M., Walker, R., forthcoming. Do housing prices reflect environmental health risks? Evidence from more than 1600 toxic plant openings and closings. *Am. Econ. Rev.*
- Dastrup, S.R., Zivin, J.G., Costa, D.L., Kahn, M.E., 2012. Understanding the Solar Home price premium: electricity generation and “Green” social status. *Eur. Econ. Rev.* 56 (5), 961–973.
- Davis, L.W., 2011. The effect of power plants on local housing prices and rents: evidence from restricted census microdata. *Rev. Econ. Stat.* 93 (4), 1391–1402.
- Davis, D., Weinstein, D., 2002. Bones, bombs and break points: the geography of economic activity. *Am. Econ. Rev.* 92 (5), 1269–1289.
- Deng, Y., Li, Z., Quigley, J.M., 2012. Economic returns to energy-efficient investments in the housing market: evidence from Singapore. *Reg. Sci. Urban Econ.* 42 (3), 506–515.
- Diamond, R., 2012. The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980–2000. Job Market Paper, Harvard University, December.
- Dumais, G., Ellison, G., Glaeser, E.L., 2002. Geographic concentration as a dynamic process. *Rev. Econ. Stat.* 84 (2), 193–204.
- Eichholtz, P.M.A., Kok, N., Quigley, J.M., 2010. Doing well by doing good: green office buildings. *Am. Econ. Rev.* 100 (5), 2494–2511.
- Eichholtz, P.M.A., Kok, N., Quigley, J.M., 2013. The economics of green building. *Rev. Econ. Stat.* 95 (1), 50–63.
- Eid, J., Overman, H.G., Puga, D., Turner, M.A., 2008. Fat city: questioning the relationship between urban sprawl and obesity. *J. Urban Econ.* 63 (2), 385–404.
- Ellickson, B., 1971. Jurisdictional fragmentation and residential choice. *Am. Econ. Rev. Pap. Proc.* 61, 334–339.
- Ellison, G., Glaeser, E.L., Kerr, W., 2010. What causes industry agglomeration? Evidence from coagglomeration patterns. *Am. Econ. Rev.* 100, 1195–1213.
- Epplé, D., Platt, G., 1998. Equilibrium and local redistribution in an urban economy when households differ in both preferences and income. *J. Urban Econ.* 43 (1), 23–51.
- Epplé, D., Sieg, H., 1999. Estimating equilibrium models of local jurisdictions. *J. Polit. Econ.* 107 (4), 645.
- Falck, O., Fritsch, M., Heblich, S., 2011. The phantom of the opera: cultural amenities, human capital, and regional economic growth. *Labour Econ.* 18 (6), 755–766.
- Gaigné, C., Riou, S., Thisse, J.-F., 2012. Are compact cities environmentally friendly? *J. Urban Econ.* 72 (2), 123–136. Elsevier.
- Gamper-Rabindran, S., Timmins, C., 2013. Does cleanup of hazardous waste sites raise housing values? Evidence of spatially localized benefits. *J. Environ. Econ. Manag.* 65 (3), 345–360.
- Glaeser, E.L., Gottlieb, J.D., 2006. Urban resurgence and the consumer city. *Urban Stud.* 43 (8), 1275–1299.
- Glaeser, E.L., Gyourko, J., 2005. Urban decline and durable housing. *J. Polit. Econ.* 113 (2), 345–375.
- Glaeser, E.L., Gyourko, J., Saks, R., 2005. Why is Manhattan so expensive? Regulation and the rise in housing prices. *J. Law Econ.* 48 (2), 331–369.
- Glaeser, E.L., Gyourko, J., Saks, R.E., 2006. Urban growth and housing supply. *J. Econ. Geogr.* 6 (1), 71–89.

- Glaeser, E.L., Kahn, M.E., 2001. Decentralized Employment and the Transformation of the American City. No. w8117. National Bureau of Economic Research.
- Glaeser, E.L., Kahn, M.E., 2010. The greenness of cities: carbon dioxide emissions and urban development. *J. Urban Econ.* 67 (3), 404–418.
- Glaeser, E.L., Shapiro, J.M., 2003. Urban growth in the 1990s: is city living back? *J. Reg. Sci.* 43 (1), 139–165.
- Glaeser, E.L., Ward, B.A., 2009. The causes and consequences of land use regulation: evidence from Greater Boston. *J. Urban Econ.* 65 (3), 265–278.
- Glaeser, E.L., Kolko, J., Saiz, A., 2001. Consumer city. *J. Econ. Geogr.* 1 (1), 27–50.
- Glaeser, E.L., Laibson, D., Sacerdote, B., 2002. An economic approach to social capital. *Econ. J.* 112 (483), F437–F458.
- Goldin, C., 2006. The quiet revolution that transformed women's employment, education and family. *Am. Econ. Rev.* 96 (2), 1–21.
- Graves, P.E., Linneman, P.D., 1979. Household migration: theoretical and empirical results. *J. Urban Econ.* 6 (3), 383–404.
- Greenstone, M., 2002. The impacts of environmental regulations on industrial activity: evidence from the 1970 and 1977 clean air act amendments and the census of manufactures. *J. Polit. Econ.* 110 (6), 1175–1219 (University of Chicago Press).
- Greenstone, M., Gallagher, J., 2008. Does hazardous waste matter? Evidence from the housing market and the superfund program. *Q. J. Econ.* 123 (3), 951–1003.
- Greenstone, M., Hanna, R., 2011. Environmental Regulations, Air and Water Pollution, and Infant Mortality in India: MIT Working paper CEEPR WP 2011–014.
- Gyourko, J., Molloy, R., 2015. Regulation and Housing Supply. *Handbook of Regional and Urban Economics* (Chapter 7).
- Gyourko, J., Tracy, J., 1991. The structure of local public finance and the quality of life. *J. Polit. Econ.* 91 (4), 774–806.
- Gyourko, J., Kahn, M.E., Tracy, J., 1999. Quality of life and environmental comparisons. In: Cheshire, P., Mills, E.S. (Eds.), *Handbook of Regional and Urban Economics*, vol. 3. North-Holland, Amsterdam, pp. 1413–1454.
- Gyourko, J., Mayer, C., Sinai, T., 2013. Superstar cities. *Am. Econ. J. Econ. Policy* 5 (4), 167–199.
- Haines, M.R., 2001. The urban mortality transition in the United States, 1800–1940. *Ann. Démogr. Historique* (1), 33–64. Berlin.
- Haninger, K., Ma, L., Timmins, C., 2014. The Value of Brownfield Remediation: NBER Working paper 20296.
- Harrison, D., Kain, J.F., 1974. Cumulative urban growth and urban density functions. *J. Urban Econ.* 1 (1), 61–98.
- Heckman, J., Scheinkman, J., 1987. The importance of bundling in a Gorman–Lancaster model of earnings. *Rev. Econ. Stud.* 54 (2), 243–255.
- Henderson, J.V., 1996. The effect of air quality regulation. *Am. Econ. Rev.* 86 (4), 789–813.
- Holian, M.J., Kahn, M.E., 2013. The Rise of the Low Carbon Consumer City: NBER Working paper No. 18735. National Bureau of Economic Research.
- Holian, M.J., Kahn, M.E., 2014. Household Demand for Low Carbon Public Policies: Evidence from California: NBER Working paper No. 19965. National Bureau of Economic Research.
- Holmes, T.J., 1998. The effect of state policies on the location of manufacturing: evidence from state borders. *J. Polit. Econ.* 106 (4), 667–705.
- Houde, S., 2014. How Consumers Respond to Environmental Certification and the Value of Energy Information, University of Maryland Working paper.
- Hsiang, S., Jina, A., 2014. The Causal Effect of Environmental Catastrophe on Long-Run Economic Growth: Evidence from 6,700 Cyclones: NBER Working paper #20352.s.
- Hu, S., Fruin, S., Kozawa, K., Mara, S., Paulson, S.E., Winer, A.M., 2009. A wide area of air pollutant impact downwind of a freeway during pre-sunrise hours. *Atmos. Environ.* 43 (16), 2541–2549.
- Iannaccone, L.R., 1998. Introduction to the economics of religion. *J. Econ. Lit.* 36, 1465–1495.

- Innes, R., Mitra, A., 2014. Parties, politics and regulation: do Republican Congressmen reduce local enforcement of Clean Air laws? *Econ. Inq.* **forthcoming**.
- Isard, W., 1948. Some locational factors in the iron and steel industry since the early nineteenth century. *J. Polit. Econ.* 56, 203.
- Ito, K., 2014. Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. *Am. Econ. Rev.* 104 (2), 537–563.
- Jacobsen, G.D., Kotchen, M.J., 2013. Are building codes effective at saving energy? Evidence from residential billing data in Florida. *Rev. Econ. Stat.* 95 (1), 34–49.
- Kahn, M.E., 1997. Particulate pollution trends in the United States. *Reg. Sci. Urban Econ.* 27 (1), 87–107.
- Kahn, M.E., 1999. The silver lining of rust belt manufacturing decline. *J. Urban Econ.* 46 (3), 360–376.
- Kahn, M.E., 2002. Demographic change and the demand for environmental regulation. *J. Policy Anal. Manag.* 21 (1), 45–62.
- Kahn, M.E., 2003. New evidence on Eastern Europe's pollution progress. *Top. Econ. Anal. Policy* 3 (1), 1100.
- Kahn, M.E., 2007. Gentrification trends in new transit-oriented communities: evidence from 14 cities that expanded and built rail transit systems. *Real Estate Econ.* 35 (2), 155–182.
- Kahn, M.E., 2010. *Climatopolis: How Our Cities Will Thrive in Our Hotter Future*. Basic Books, New Jersey.
- Kahn, M.E., 2011. Do liberal cities limit new housing development? Evidence from California. *J. Urban Econ.* 69 (2), 223–228.
- Kahn, M.E., Kok, N., 2014. Big-Box Retailers and Urban Carbon Emissions: The Case of Wal-Mart: NBER Working paper No. 19912. National Bureau of Economic Research.
- Kahn, M.E., Mansur, E.T., 2013. Do local energy prices and regulation affect the geographic concentration of employment? *J. Public Econ.* 101, 105–114.
- Kahn, M.E., Schwartz, J., 2008. Urban air pollution progress despite sprawl: the 'greening' of the vehicle fleet. *J. Urban Econ.* 63 (3), 775–787.
- Kahn, M.E., Vaughn, R.K., 2009. Green market geography: the spatial clustering of hybrid vehicles and LEED registered buildings. *BE J. Econ. Anal. Policy* 9 (2).
- Kahn, M.E., Kok, N., Quigley, J., 2014. Carbon emissions from the commercial building sector: the role of climate, quality, and incentives. *J. Public Econ.* 113C, 1–12.
- Kane, T.J., Riegg, S.K., Staiger, D.O., 2006. School quality, neighborhoods, and housing prices. *Am. Law Econ. Rev.* 8 (2), 183–212.
- Kates, R.W., Colten, C.E., Laska, S., Leatherman, S.P., 2006. Reconstruction of New Orleans after Hurricane Katrina: a research perspective. *Proc. Natl. Acad. Sci. U.S.A.* 103 (40), 14653–14660.
- Kennan, J., Walker, J.R., 2011. The effect of expected income on individual migration decisions. *Econometrica* 79 (1), 211–251.
- Kiel, K.A., 1995. Measuring the impact of the discovery and cleaning of identified hazardous waste sites on house values. *Land Econ.* 71, 428–435.
- Kiel, K.A., Williams, M., 2007. The impact of Superfund sites on local property values: are all sites the same? *J. Urban Econ.* 61 (1), 170–192.
- Kim, C.W., Phipps, T.T., Anselin, L., 2003. Measuring the benefits of air quality improvement: a spatial hedonic approach. *J. Environ. Econ. Manage.* 45 (1), 24–39.
- Klier, T., Linn, J., 2010. The price of gasoline and new vehicle fuel economy: evidence from monthly sales data. *Am. Econ. J. Econ. Policy* 2 (3), 134–153.
- Kok, N., McGraw, M., Quigley, J.M., 2011. The diffusion of energy efficiency in building. *Am. Econ. Rev. Pap. Proc.* 101 (3), 77–82.
- Kotchen, M.J., Powers, S.M., 2006. Explaining the appearance and success of voter referenda for open-space conservation. *J. Environ. Econ. Manage.* 52 (1), 373–390.
- Kousky, C., Luttmer, E.F.P., Zeckhauser, R.J., 2006. Private investment and government protection. *J. Risk Uncertain.* 33 (1–2), 73–100.
- Kuminoff, N.V., Pope, J.C., 2014. Do 'Capitalization Effects' for public goods reveal the public's willingness to pay? *Int. Econ. Rev.* 55, 1227–1250.
- Kuminoff, N.V., Kerry, S.V., Timmins, C., 2013. The new economics of equilibrium sorting and policy evaluation using housing markets. *J. Econ. Lit.* 51 (4), 1007–1062.

- Lee, S., Lee, B., 2014. The influence of urban form on GHG emissions in the US household sector. *Energy Policy* 68, 534–549.
- Lee, S., Lin, J., 2013. Natural Amenities, Neighborhood Dynamics, and Persistence in the Spatial Distribution of Income: FRB of Philadelphia Working paper No. 13–48. Federal Reserve Bank of Philadelphia.
- Levinson, A., Niemann, S., 2004. Energy use by apartment tenants when landlords pay for utilities. *Resour. Energy Econ.* 26 (1), 51–75.
- Levitt, S.D., 2004. Understanding why crime fell in the 1990s: four factors that explain the decline and six that do not. *J. Econ. Perspect.* 18, 163–190.
- Li, S., Timmins, C., von Haefen, R.H., 2009. How do gasoline prices affect fleet fuel economy? *Am. Econ. J. Econ. Policy* 1 (2), 113–137.
- Libecap, G.D., 2009. Chinatown revisited: Owens Valley and Los Angeles—bargaining costs and fairness perceptions of the first major water rights exchange. *J. Law Econ. Org.* 25 (2), 311–338.
- Ludwig, J., Sanbonmatsu, L., Gennetian, L., Adam, E., Duncan, G.J., Katz, L.F., Kessler, R.C., et al., 2011. Neighborhoods, obesity, and diabetes—a randomized social experiment. *N. Engl. J. Med.* 365 (16), 1509–1519.
- Mangum, K., 2014. The Global Effects of Housing Policy, Georgia State University, Working paper.
- Maruejols, L., Young, D., 2011. Split incentives and energy efficiency in Canadian multi-family dwellings. *Energy Policy* 39 (6), 3655–3668.
- Matisoff, D.C., Noonan, D.S., Mazzolini, A.M., 2014. Performance or marketing benefits? The case of LEED certification. *Environ. Sci. Technol.* 48 (3), 2001–2007.
- McConnell, V., Walls, M.A., 2005. The Value of Open Space: Evidence from Studies of Nonmarket Benefits. Resources for the Future, Washington, DC.
- Meltzer, R., Schuetz, J., 2012. Bodegas or bagel shops? Neighborhood differences in retail and household services. *Econ. Dev. Q.* 26 (1), 73–94.
- Miller, N.L., Hayhoe, K., Jin, J., Auffhammer, M., 2008. Climate, extreme heat, and electricity demand in California. *J. Appl. Meteorol. Climatol.* 47 (6), 1834–1844.
- Moretti, E., 2004. Human capital externalities in cities. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Holland, pp. 2243–2291.
- Moretti, E., 2012. *The New Geography of Jobs*. Houghton Mifflin Harcourt, New Jersey.
- Muller, N.Z., Mendelsohn, R., 2007. Measuring the damages of air pollution in the United States. *J. Environ. Econ. Manag.* 54 (1), 1–14.
- Muller, N.Z., Mendelsohn, R., Nordhaus, W., 2011. Environmental accounting for pollution in the United States economy. *Am. Econ. Rev.* 101 (5), 1649–1675.
- Nevin, A.M., 1888. *The Social Mirror: A Character Sketch of the Women of Pittsburg and Vicinity During the First Century of the County's Existence*. TW Nevin, New Jersey.
- Newell, R., Jaffe, A., Stavins, R., 1999. The induced innovation hypothesis and energy saving technological change. *Q. J. Econ.* 114 (3), 941–975.
- Noonan, D.S., 2005. Neighbours, barriers and urban environments: are things 'different on the other side of the tracks'? *Urban Stud.* 42 (10), 1817–1835.
- Noonan, D.S., Hsieh, L.-H.C., Matisoff, D., 2013. Spatial effects in energy-efficient residential HVAC technology adoption. *Environ. Behav.* 45 (4), 476–503.
- Nordhaus, W.D., Tobin, J., 1972. Is growth obsolete? In: *Economic Research: Retrospect and Prospect Vol 5: Economic Growth*, pp. 1–80. NBER, New York.
- Oi, W.Y., 1996. The welfare implications of invention. In: *The Economics of New Goods*. University of Chicago Press, Chicago, IL, pp. 109–142.
- Olmstead, S.M., 2010. The economics of water quality. *Rev. Environ. Econ. Policy* 4 (1), 44–62.
- Parry, I.W.H., Small, K.A., 2005. Does Britain or the United States have the right gasoline tax? *Am. Econ. Rev.* 95 (4), 1276–1289.
- Peltzman, S., 1975. The effects of automobile safety regulation. *J. Polit. Econ.* 83, 677–725.
- Pope III, C., Arden, J.S., Ransom, M.R., 1992. Daily mortality and PM10 pollution in Utah Valley. *Arch. Environ. Health Int. J.* 47 (3), 211–217.
- Rappaport, J., 2009. The increasing importance of quality of life. *J. Econ. Geogr.* 9 (6), 779–804.
- Rappaport, J., Sachs, J.D., 2003. The United States as a Coastal Nation. *J. Econ. Growth* 8 (1), 5–46.

- Reiss, P.C., White, M.W., 2008. What changes energy consumption? Prices and public pressures. *RAND J. Econ.* 39 (3), 636–663.
- Roback, J., 1982. Wages, rents, and the quality of life. *J. Polit. Econ.* 90 (6), 1257–1278.
- Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. *J. Polit. Econ.* 82, 34–55.
- Rosen, S., 2002. Markets and diversity. *Am. Econ. Rev.* 92 (1), 1–15.
- Rosenthal, S.S., Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. *Handbook Region. Urban Econ.* 4, 2119–2171.
- Rossi-Hansberg, E., Sarte, P.-D., 2009. Firm fragmentation and urban patterns. *Int. Econ. Rev.* 50 (1), 143–186.
- Saiz, A., 2010. The geographic determinants of housing supply. *Q. J. Econ.* 125 (3), 1253–1296.
- Schwartz, A.E., Susin, S., Voicu, I., 2003. Has falling crime driven New York City's real estate boom? *J. Hous. Res.* 14 (1), 101–136.
- Sexton, S., 2014. Automatic bill payment and salience effects: evidence from electricity consumption. *Rev. Econ. Stat.* Forthcoming.
- Sieg, H., Kerry, S.V., Spencer, B.H., Walsh, R., 2004. Estimating the general equilibrium benefits of large changes in spatially delineated public goods. *Int. Econ. Rev.* 45 (4), 1047–1077.
- Sigman, H., 2001. The pace of progress at superfund sites: policy goals and interest group influence. *J. Law Econ.* 44 (1), 315–343.
- Sigman, H., Stafford, S., 2011. Management of hazardous waste and contaminated land. *Annu. Rev. Resour. Econ.* 3 (1), 255–275.
- Skrabec, Q.R., 2010. *The World's Richest Neighborhood: How Pittsburgh's East Enders Forged American Industry*. Algora Publishing, New Jersey.
- Smith, V.K., 2012. Reflections—in search of crosswalks between macroeconomics and environmental economics. *Rev. Environ. Econ. Policy* 6 (2), 298–317.
- Smith, V.K., Huang, J.C., 1995. Can markets value air quality? A meta-analysis of hedonic property value models. *J. Polit. Econ.* 103, 209–227.
- Stradling, D., Thorsheim, P., 1999. The smoke of great cities: British and American efforts to control air pollution, 1860–1914. *Environ. Hist.* 4, 6–31.
- Stull, W.J., 1974. Land use and zoning in an urban economy. *Am. Econ. Rev.* 64 (3), 337–347.
- Tajima, K., 2003. New estimates of the demand for urban green space: implications for valuing the environmental benefits of Boston's big dig project. *J. Urban Affairs* 25 (5), 641–655.
- Tarr, J.A. (Ed.), 2003. *Devastation and Renewal: An Environmental History of Pittsburgh and Its Region*. University of Pittsburgh Press, New Jersey.
- Taylor, L., Liu, X., Phaneuf, D., 2012. Disentangling the Property Value Impacts of Environmental Contamination from Locally Undesirable Land Uses, Working paper.
- Timmins, C., 2002. Measuring the dynamic efficiency costs of regulators' preferences: municipal water utilities in the arid west. *Econometrica* 70 (2), 603–629.
- Tra, C.I., 2010. A discrete choice equilibrium approach to valuing large environmental changes. *J. Public Econ.* 94 (1), 183–196.
- Turner, M.A., 2005. Landscape preferences and patterns of residential development. *J. Urban Econ.* 57 (1), 19–54.
- Turner, M.A., Haughwout, A., van der Klaauw, W., 2014. Land use regulation and welfare. *Econometrica* 82, 1341–1403.
- United Church of Christ. Commission for Racial Justice, 1987. Toxic wastes and race in the United States: a national report on the racial and socio-economic characteristics of communities with hazardous waste sites. Public Data Access.
- US GAO, 2005. Report to Congressional Requesters, Brownfield Redevelopment: Stakeholders Report That EPA's Program Helps to Redevelop Sites, but Additional Measures Could Complement Agency Efforts.
- US GAO, 2013. Report to Congressional Requesters, Superfund: EPA Should Take Steps to Improve Its Management of Alternative to Placing Sites on the National Priorities List.
- VandeWeghe, J.R., Kennedy, C., 2007. A spatial analysis of residential greenhouse gas emissions in the Toronto census metropolitan area. *J. Ind. Ecol.* 11 (2), 133–144.

- Waldfoegel, J., 2008. The median voter and the median consumer: local private goods and population composition. *J. Urban Econ.* 63 (2), 567–582.
- Walsh, R., 2007. Endogenous open space amenities in a locational equilibrium. *J. Urban Econ.* 61 (2), 319–344.
- Weitzman, M.L., 2009. On modeling and interpreting the economics of catastrophic climate change. *Rev. Econ. Stat.* 91 (1), 1–19.
- Wolak, F.A., 2011. Do residential customers respond to hourly prices? Evidence from a dynamic pricing experiment. *Am. Econ. Rev.* 101 (3), 83–87.
- Wolff, H., 2014. Keep your clunker in the suburb: low-emission zones and adoption of green vehicles. *Econ. J.* 124 (578), F481–F512.
- Wolverton, A., 2009. Effects of socio-economic and input-related factors on polluting plants' location decisions. *BE J. Econ. Anal. Policy* 9, 1.
- Wu, J., Plantinga, A.J., 2003. The influence of public open space on urban spatial structure. *J. Environ. Econ. Manag.* 46 (2), 288–309.
- Zabel, J.E., Kiel, K.A., 2000. Estimating the demand for air quality in four US cities. *Land Econ.* 76, 174–194.
- Zheng, S., Kahn, M.E., 2008. Land and residential property markets in a booming economy: new evidence from Beijing. *J. Urban Econ.* 63, 743–757.
- Zheng, S., Kahn, M.E., 2013. Understanding China's urban pollution dynamics. *J. Econ. Lit.* 51 (3), 731–772.
- Zheng, S., Wang, R., Glaeser, E.L., Kahn, M.E., 2011. The greenness of China: household carbon dioxide emissions and urban development. *J. Econ. Geogr.* 11 (5), 761–792.
- Zheng, S., Jing, W., Kahn, M.E., Deng, Y., 2012. The nascent market for “green” real estate in Beijing. *Eur. Econ. Rev.* 56 (5), 974–984.
- Zheng, S., Cao, J., Kahn, M.E., Sun, C., 2014a. Real estate valuation and cross-boundary air pollution externalities: evidence from Chinese cities. *J. Real Estate Financ. Econ.* 48 (3), 398–414.
- Zheng, S., Sun, C., Qi, Y., Kahn, M.E., 2014b. The evolving geography of China's industrial production: implications for pollution dynamics and urban quality of life. *J. Econ. Surv.* 28 (4), 709–724.
- Zhou, Y., Levy, J.I., Evans, J.S., Hammitt, J.K., 2006. The influence of geographic location on population exposure to emissions from power plants throughout China. *Environ. Int.* 32 (3), 365–373.

This page intentionally left blank

CHAPTER 8

Urban Land Use

Gilles Duranton^{*,†}, Diego Puga^{†,‡}

^{*}Wharton School, University of Pennsylvania, Philadelphia, PA, USA

[†]CEPR, London, UK

[‡]Centro de Estudios Monetarios y Financieros (CEMFI), Madrid, Spain

Contents

8.1. Introduction	468
8.2. Modeling Urban Land Use: The Monocentric Model	472
8.2.1 The Marshallian approach	474
8.2.2 The bid-rent approach	475
8.2.3 The dual approach	477
8.2.4 Quantitative predictions	482
8.3. Extending the Monocentric Model	483
8.3.1 Other elements of travel costs and looking beyond accessibility	484
8.3.2 Heterogeneous residents, properties, and timing	486
8.3.3 Durable housing	497
8.4. Agglomeration and Commercial Land Use: Modeling Polycentric Cities	503
8.5. Land Use Regulation	515
8.6. Empirical Price and Development Gradients	522
8.7. Patterns of Residential Sorting Within Cities	530
8.8. Patterns of Residential Land Development	537
8.9. Employment Decentralization and Patterns of Business Location Changes Within Cities	544
8.10. Conclusion	551
Acknowledgments	553
References	553

Abstract

We provide an integrated treatment of the theoretical literature on urban land use inspired by the monocentric model, including extensions that deal with multiple endogenous business centers, various dimensions of heterogeneity, and durable housing. After presenting the theory and distilling its key empirical implications, we critically review the empirical literature on differences in prices and development across urban locations, patterns of location choices of heterogeneous households in cities, sprawl and residential decentralization, and employment decentralization.

Keywords

Land use, Urban structure

JEL Classification Code

R14

8.1. INTRODUCTION

In this chapter, we provide an integrated treatment of the theoretical literature on urban land use inspired by the monocentric model, including extensions that deal with multiple endogenous business centers, various dimensions of heterogeneity, and durable housing. After presenting the theory and distilling its key empirical implications, we critically review the empirical literature on differences in prices and development across urban locations, patterns of location choices of heterogeneous households in cities, sprawl and residential decentralization, and employment decentralization.

Urban land use is of fundamental importance. Most obviously, it is at the heart of extremely large allocation decisions made by firms and households. On the residential side, American households devote about a quarter of their consumption expenditure to housing, and the value of the residential housing stock may represent up to 2 years of gross national product. Where development occurs and at what intensity is arguably a first-order determinant of the efficiency of these large allocations. Households also engage in a variety of activities that take place in different locations: they work, they sleep, they play, they go to school, they shop, they visit friends, they go to the dentist, etc.¹ To conduct these activities in different locations, people must travel between them. As a result, land use and transport are intimately connected. American households spend between 5% and 10% of their time awake traveling, and the median household devotes 18% of its budget to transport, most of which goes to road transport.²

Beyond this, urban land use is a fundamental determinant of the physical world that surrounds urban dwellers, the majority of the world population. Urban land use determines how the various locations urban dwellers go to or would like to go to are organized and connected with each other. Hence, not only does land use affect the immense resources devoted to housing, commercial property, open space, and transport, it also potentially affects the labor market and the markets for the products we purchase. Land use also arguably affects the ability of firms to produce. In turn, these broader effects of land use may have serious implications for prosperity and equity.

Figure 8.1 depicts the distribution of land across various uses in Paris. The map at the top of the figure classifies land across five uses on a disk with radius 30 km centered on Notre Dame, the conventional center of Paris. We can immediately see that the patterns of land use are quite complex. The next two plots summarize the information by classifying land use by distance from Notre Dame, with the northern half of the map

¹ These travel categories correspond closely to the categories in the US National Household Transportation Survey.

² See Combes et al. (2014) for the sources for the housing figures and Couture et al. (2012) for transport. The figures reported here are for the United States, but the magnitudes for other developed and many developing countries are similar.

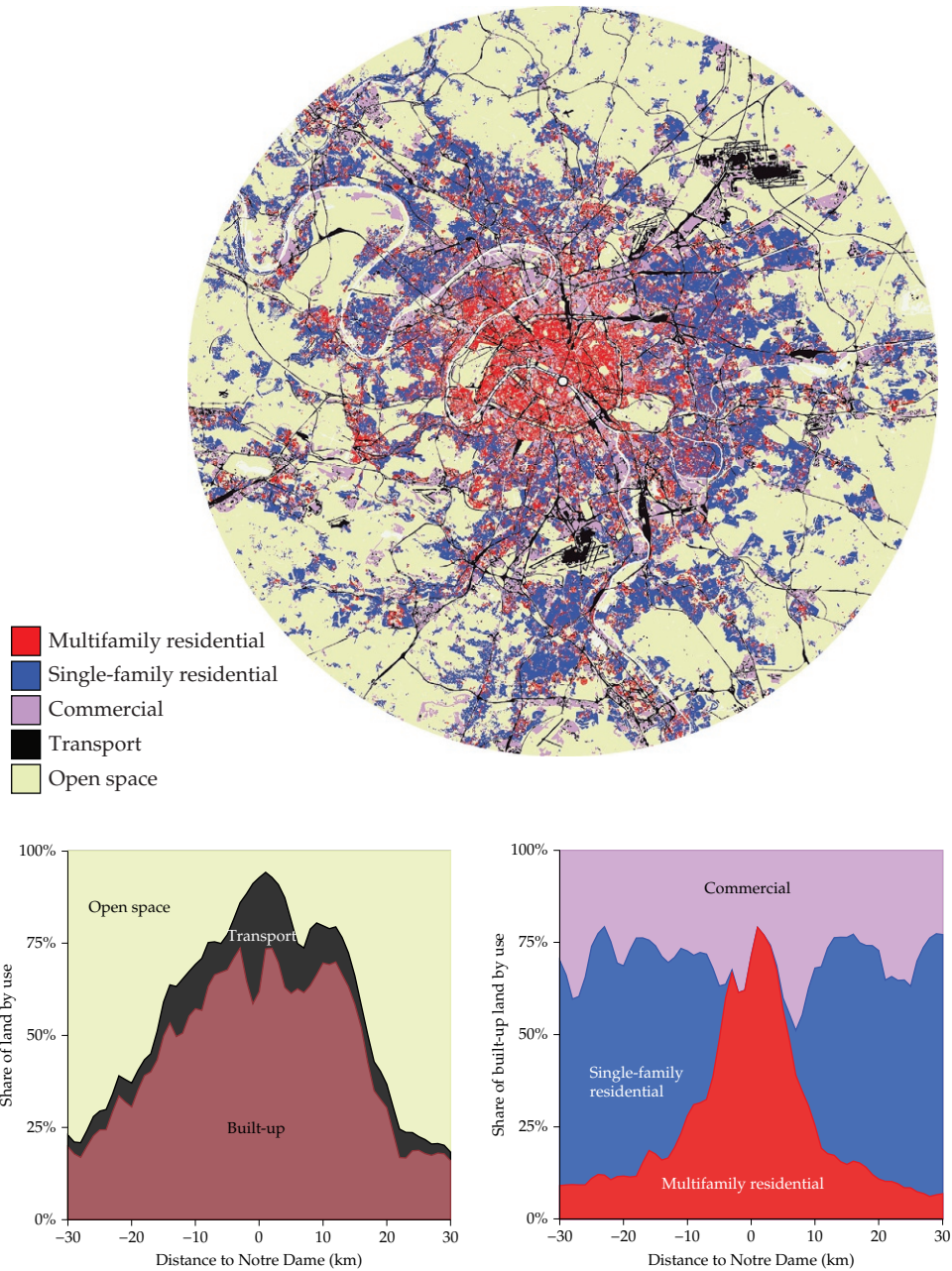


Figure 8.1 Distribution of land across uses in Paris.

plotted on the positive side of the horizontal axis and the southern half plotted on the negative side.³ The first of these two plots splits all land between open space, land used for transport infrastructure, and built-up land. The last plot further divides the built-up land category between multifamily residential, single-family residential, and commercial uses. Both plots show some very clear gradients. In particular, as we look further away from the center (Notre Dame), the percentage of land that is built up falls, with more land being open space instead. The intensity of residential development also falls very clearly with distance to the center, with multifamily buildings giving way very quickly to single-family homes. The distribution of built-up land between residential and commercial uses does not show much variation by comparison, but we do see two peaks of commercial land (pointing downward, since commercial is plotted at the top) at the sides of a central area with more mixed use. It is also worth noting how much space is taken up by transport infrastructure, particularly close to the center, a very graphic illustration of how closely tied transport is to land use issues in cities. The rest of the chapter will help the reader to understand both the complexity and the order that appear in [Figure 8.1](#).

Before proceeding any further, we will draw some intellectual boundaries for this chapter and justify its organization further. Since everything is located somewhere, land use potentially touches on a large number of topics. At minimum, it could certainly overlap greatly with all the other chapters in this volume. To retain a finite agenda, we think of urban land use as covering mainly the following issues: (a) the differences in land and property prices across locations, (b) the patterns of location choices by types and sub-groups of users, (c) the patterns of land conversion across uses, and (d) the patterns of residential and business location changes within cities.

To explore these four sets of issues, we first present an integrated summary of theoretical developments on urban land use before turning to the empirical work on the aforementioned issues. A first reason for using this structure instead of providing a different model for each empirical question is that the theory that underlies the issues listed above is unified. There is no point repeating it several times. Furthermore, the economic analysis of land use first saw some important theoretical developments with empirical

³ Since they transform spatial data from two dimensions (latitude and longitude) to just one dimension (distance to the center), these plots are more directly comparable with economic models of land use. As we shall see below, typically these models either represent cities as a segment on the real line (e.g., [Ogawa and Fujita, 1980](#)) or represent them as a disk that is circularly symmetric by assumption (e.g., [Lucas and Rossi-Hansberg, 2002](#)) so that all points at the same distance from the center of the disk have the same land use. It is worth emphasizing that, in addition to the market forces that are the focus of economic models of land use, the allocation of land in Paris to the different uses depicted in [Figure 8.1](#) also reflects public sector land use decisions and important restrictions and regulations. Figure 8.1 is constructed based on the dataset Mode d'occupation du sol 2012 from l'Institut d'aménagement et d'urbanisme de la Région d'Île-de-France.

work lagging behind or developing independently.⁴ We endeavor to reconnect empirical work more strongly with theory both by making sure that we highlight the empirical content of the models as we describe them and by trying to tie empirical work to land use models as strongly as possible (or by highlighting the weakness of those links in some cases). Another reason for presenting the theory in a self-contained manner is that it is relevant not only to the issues explored here, but also to many issues explored in other chapters in this volume such as regulation, neighborhood dynamics, and transport, to list just a few.

Following a long and well-established economic literature, our starting point is that accessibility determines land and housing prices at different locations. However, the patterns of accessibility are also affected by the location choices of firms and workers, which are determined by prices. Hence, the land use problem is in essence a hard equilibrium problem with many feedbacks. The literature first solved it by restricting accessibility to be solely about access to jobs and by treating the location of these jobs as exogenous within a simple geography and with frictionless markets. This is the monocentric model that we explore in [Section 8.2](#).

While the simplest version of the monocentric model may be viewed as a reasonable first-order description of many cities and delivers a number of plausible predictions, it remains extremely crude. Even if we are willing to restrict production in cities to take place in a centralized area, the model does not include a number of fundamental urban features. In particular, city dwellers are highly heterogeneous in their incomes, demographics, and preferences. The study of the heterogeneity of urban residents is interesting in itself since, beyond making predictions about prices and the intensity of development, we also expect good models of land use to offer insights into who lives where. The heterogeneity of residents, coupled with that of the housing stock, also implies that land and property markets may not be as frictionless as assumed in the simplest land use models. In addition, the basic model is static in nature, but properties are long-lived and we cannot expect land use in cities to adjust immediately to any shock. This creates further frictions. We explore all these issues in [Section 8.3](#).

But perhaps the most obvious criticism of the monocentric model is that cities have become less and less monocentric. The main problem with the standard approach is not that it cannot accommodate more realistic employment distributions. It can. What the standard approach cannot do easily is allow the distribution of jobs to be endogenous, interacting with the distribution of residents. Much modeling effort has been devoted to this problem since the late 1970s. Residents face a trade-off between accessibility and land and property prices. Businesses benefit from proximity to other businesses

⁴ We discuss below the empirical work of [Clark \(1951\)](#), an exception that appeared before the main framework was established. See also the introduction in [Quigley \(1998\)](#) for a brief intellectual history of the early work on urban land use.

because of agglomeration economies but, if they cluster, they must pay higher land prices and also compensate their workers for longer commutes with higher wages. [Section 8.4](#) provides a tractable model of land use in cities under endogenous business location dealing with these complex issues and summarizes other efforts at modeling secondary centers and job decentralization.

No work on urban land use would be complete without a discussion of government intervention. Land and the properties erected on it are usually highly regulated. We explore and discuss the possible reasons for these regulations and their possible effects in [Section 8.5](#).

Our treatment of the theoretical literature gives a dominant role to the accessibility of jobs. While clearly important, job accessibility is not the sole determinant of how land is used and how properties are valued. First, commuting is only one aspect of urban travel. Thus, accessibility should be broadly understood to include proximity to shops, school, amenities, etc. Second, other aspects of location, such as heterogeneity and neighborhood interactions, matter greatly. This said, we believe focusing on accessibility is warranted because it seems uniquely important in shaping cities at a broader scale.

To be useful and become more than a speculation, a conceptualization must confront the empirical reality. This is what the last four sections of our chapter aim to do. [Section 8.6](#) examines the empirical literature that assesses the gradient predictions of the simplest models of urban land use. [Section 8.7](#) then turns to the empirical location patterns of heterogeneous city residents. [Section 8.8](#) looks at recent patterns of residential land development. Finally, [Section 8.9](#) examines changes in business location within cities.

8.2. MODELING URBAN LAND USE: THE MONOCENTRIC MODEL

Land use is one of the oldest topics of economic analysis. [Ricardo \(1817\)](#) and [von Thünen \(1826\)](#), two of the pioneers of the discipline, offered early insights. Ricardo observed that the rent differential between two parcels of land should be equal to the difference in the revenue derived from these two parcels, all else being equal. von Thünen discovered the same idea independently and embedded it in a simple model of farming to show that, on a flat featureless plain, crops that are more costly to transport should be located closer to the center of the village while crops that are cheaper to transport should be cultivated in outer rings.⁵ As will become clear below, the insights of Ricardo and Thünen still form the basis of much of our understanding of land use issues today.

The modern approach to urban land use has its origins in the work of [Alonso \(1964\)](#), [Mills \(1967\)](#), and [Muth \(1969\)](#), who first encapsulated city transport, land use, and

⁵ von Thünen's work may have been the first ever fully specified formal economic model. He is also credited as being one of the fathers of marginalism ([Samuelson, 1983](#); [Fujita, 2012](#)).

population issues into what is known as the monocentric city model. In this section, we first offer a formal exposition of this model. Then, we turn to a number of extensions in Sections 8.3 and 8.4. Our review of these extensions is highly selective because this model has been thoroughly investigated. A full review would take us well beyond the scope of this chapter.

In this section, we expand on the exposition of the monocentric city model we introduced in earlier work (Duranton and Puga, 2014).⁶ Consider a linear monocentric city. Land covered by the city is endogenously determined and can be represented by a segment on the positive real line. Production and consumption of a numéraire good take place at a single point $x = 0$, the central business district (CBD). In addition to the numéraire good, individuals living in the city consume housing. The production of housing from land in the city is described below. For now, it is nonetheless important to keep in mind that land and the housing that sits on it are allocated competitively to the highest bidder at each location. Preferences can be represented by a utility function $u(h, z)$ derived from individual consumption of housing, h , and of the numéraire, z . This function is increasing in both its arguments and is strictly quasi concave.

Commuting costs increase linearly with distance to the CBD, so a resident living at distance x from the CBD incurs a commuting cost τx . This leaves $w - \tau x$ for expenditure on housing and the numéraire. If we denote by $P(x)$ the rental price of housing at a distance x from the CBD, this resident's budget constraint is thus $w - \tau x = P(x) h + z$.

All residents in the city are identical in income and preferences and are freely mobile within the city, and hence must derive a common level of utility $\underline{u}$ from the heterogeneous combinations of housing and the numéraire they consume.

Relative to the standard consumer problem studied in introductory economics, there are two main differences. First, residents must choose the location of their residence as well as allocate their disposable income optimally between housing and the numéraire. The price of housing, and thus the budget constraint they face, varies with their location choice. Without loss of generality, we can think of each resident solving for a standard budget allocation between housing and the numéraire at each location and then choosing the location that offers the highest utility. Second, the price of housing at each location is unknown to the analyst and needs to be computed as part of the equilibrium.

This first version of the monocentric model makes six simplifying assumptions. The first three concern the linearity of cities, the restriction of urban travel to commutes, and a particular specification for commuting costs paid in the numéraire and increasing proportionately with distance. These assumptions are for expositional purposes. Most of the

⁶ For further reading, Brueckner (1987) offers a remarkably clear and intuitive graphical presentation of the monocentric model. A more detailed presentation can be found in the classic work of Fujita (1989). See also the appendix in Zenou (2009), where the Marshallian and bid-rent approaches are further developed and compared.

results we derive below readily generalize to two-dimensional cities, other reasons for travel, and other specifications for commuting costs, including commuting costs paid in time or affecting utility. We highlight the few cases where the results depend on these simplifying assumptions. The other three assumptions made here are that of a static model, an exogenous geography of jobs, and homogeneous residents. The next two sections explore extensions that attempt to relax these assumptions. Finally, we also consider that accessibility is the only factor that determines land use. We discuss below how other determinants of land use can be considered together with accessibility.

8.2.1 The Marshallian approach

There are several ways to solve this model. Since they all shed a different light on its mechanics, we describe them in turn. The first path is the Marshallian approach (also sometimes referred to as the indirect approach). This approach solves the individual budget allocation between housing and the numéraire at each location and then obtains house prices by ensuring that, with each consumer allocating optimally his or her disposable income, utility is equalized across locations in the city.

Maximizing utility $u(h, z)$ with respect to h and z subject to $w - \tau x = P(x) h + z$ is equivalent to maximizing $u(h, w - \tau x - P(x) h)$ with respect to h . The first-order condition of this problem yields a unique Marshallian demand for housing at each location $h(x)$ defined implicitly by $\frac{\partial u(\cdot)}{\partial h} - \frac{\partial u(\cdot)}{\partial z} P(x) = 0$ or, equivalently,

$$P(x) = \frac{\frac{\partial u(\cdot)}{\partial h}}{\frac{\partial u(\cdot)}{\partial z}}. \quad (8.1)$$

This expression is of course the standard first-order condition for utility maximization by residents stating that the marginal utility of more housing per amount spent should be equal to the marginal utility of the numéraire. Using the budget constraint again, we can recover the Marshallian demand for the numéraire $z(x)$ as a function of the Marshallian demand for housing $z(x) = w - \tau x - P(x) h(x)$.

In equilibrium, given that all individuals have the same income and are freely mobile, they must obtain the same level of utility $\underline{u}$ defined above (we return later to how it is determined). Thus,

$$u(h(x), w - \tau x - P(x) h(x)) = \underline{u}. \quad (8.2)$$

Totally differentiating Equation (8.2) with respect to x yields

$$\frac{\partial u(h, z)}{\partial h} \frac{\partial h(x)}{\partial x} - \frac{\partial u(h, z)}{\partial z} P(x) \frac{\partial h(x)}{\partial x} - \frac{\partial u(h, z)}{\partial z} \left(\tau + h(x) \frac{dP(x)}{dx} \right) = 0. \quad (8.3)$$

By Equation (8.1), the first two terms in Equation (8.3) cancel out (this is just the envelope theorem at work), which implies

$$\frac{dP(x)}{dx} = -\frac{\tau}{h(x)} < 0 . \quad (8.4)$$

Equation (8.4) is known as the Alonso–Muth condition.⁷ It states that, at the residential equilibrium, if a resident moves marginally away from the CBD, the cost of his or her current housing consumption falls just as much as his or her commuting costs increase. Thus, the price of housing decreases with distance to the CBD. This housing price gradient is the first of a series of gradients that occur in the monocentric model.

The main drawback of the Marshallian approach is that it gets to the solution in a roundabout way. It solves first for the consumer program in a location before recovering the price of housing at this location through the residential equilibrium condition. Then, knowing the price of housing, it returns to the choice of consumption before solving for the optimal location. The main advantage of the Marshallian approach is to make clear that the price of housing at each location is endogenous and emerges within the model.

8.2.2 The bid-rent approach

The Alonso–Muth condition can be derived more directly using the so-called bid-rent approach (also known as the direct approach). We define the bid-rent function for housing $\Psi(x, \underline{u})$ as the maximum price a resident is willing to pay for housing at distance x from the CBD while enjoying utility $\underline{u}$ and satisfying the budget constraint:

$$\Psi(x, \underline{u}) \equiv \max_{h(x), z(x)} \{P(x) | u(h, z) = \underline{u}, w - \tau x = P(x)h(x) + z(x)\} . \quad (8.5)$$

To reduce the problem to a single constraint instead of two constraints, one can solve the budget constraint for $P(x)$ and substitute this into the original program to obtain

$$\Psi(x, \underline{u}) = \max_{h(x), z(x)} \left\{ \frac{w - \tau x - z(x)}{h(x)} \mid u(h, z) = \underline{u} \right\} . \quad (8.6)$$

We can then turn the program into an unconstrained maximization problem by replacing $z(x)$ with the restricted Hicksian demand for the numéraire $z(h(x), \underline{u})$:

⁷ Although it is also sometimes referred to as the spatial equilibrium condition, we prefer the Alonso–Muth label and reserve spatial equilibrium for the equalization condition (8.2). As made clear below, the Alonso–Muth condition will occur in models where there is no spatial condition involving a full equalization like (8.2). We also note that spatial equilibrium conditions are also present in many other models, including settings that involve amenity or wage differences across locations.

$$\Psi(x, \underline{u}) = \max_{h(x)} \left\{ \frac{w - \tau x - z(h(x), \underline{u})}{h(x)} \right\}. \quad (8.7)$$

Now the Alonso–Muth condition can be obtained by direct application of the envelope theorem to Equation (8.7):

$$\frac{dP(x)}{dx} = \frac{d\Psi(x, \underline{u})}{dx} \Big|_{h(x)=h(\Psi(x, \underline{u}), \underline{u})} = -\frac{\tau}{h(x)} < 0. \quad (8.8)$$

The consumption of housing when the individual pays the bid-rent price can be obtained from the first-order condition for the program of Equation (8.7):

$$\frac{\partial z(h(x), \underline{u})}{\partial h(x)} h(x) + w - \tau x - z(h(x), \underline{u}) = 0. \quad (8.9)$$

Note that this can be rewritten as

$$\frac{\partial z(h(x), \underline{u})}{\partial h(x)} = -\frac{w - \tau x - z(h(x), \underline{u})}{h(x)}, \quad (8.10)$$

where the left-hand side is the slope of the indifference curve $u(h, z) = \underline{u}$ and the right-hand side is the slope of the budget constraint. Panel (a) in Figure 8.2 illustrates the approach. The vertical axis represents the consumption of numéraire $z(x)$ and the horizontal axis that of housing $h(x)$. The line $z(x) = w - \tau x - P(x)h(x)$ is the budget constraint, which has vertical intercept $w - \tau x$ and slope $-P(x)$. The bid-rent or maximum price that the consumer can pay for housing while staying within the budget constraint and achieving utility $\underline{u}$ is found by pivoting the budget constraint around its vertical intercept until it is tangent to the indifference curve $u(h, z) = \underline{u}$.⁸

While less standard, the bid-rent approach is a more direct way for solving the monocentric model. It also highlights the bidding nature that prevails for the allocation of housing (and land) at each location.

To make the bid-rent approach more concrete, assume, for instance, that $u(h, z) = h^\alpha z^{1-\alpha}$, with $0 < \alpha < 1$. Then, $z(h(x), \underline{u}) = h(x)^{-\alpha/(1-\alpha)} \underline{u}^{1/(1-\alpha)}$. In this particular case, Equation (8.7) can be written as

$$\Psi(x, \underline{u}) = \max_{h(x)} \left\{ \frac{w - \tau x - h(x)^{-\frac{\alpha}{1-\alpha}} \underline{u}^{\frac{1}{1-\alpha}}}{h(x)} \right\}. \quad (8.11)$$

Solving for $h(x)$ in Equation (8.11) and reinserting the solution into it finally yields

⁸ Note that this is different from the standard consumer problem solved when deriving the expenditure function, which keeps the slope of the budget constraint fixed and moves it in parallel until reaching the tangency point with $u(h, z) = \underline{u}$. Here, we instead keep the intercept fixed and rotate the budget constraint to reach the tangency point.

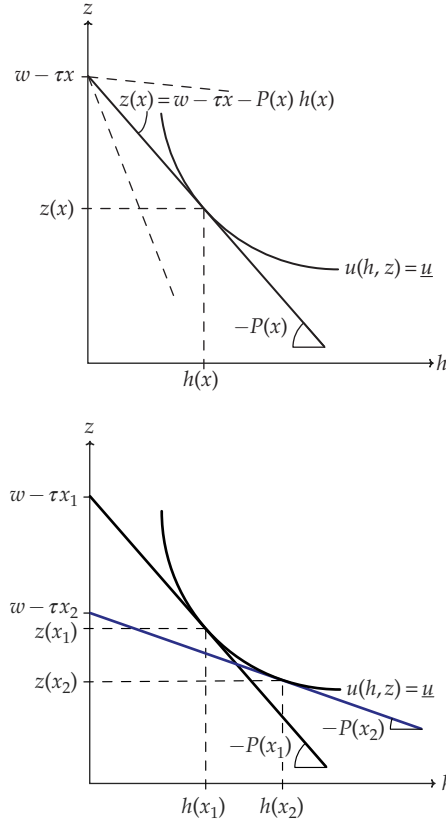


Figure 8.2 A graphical representation of the monocentric model. Panel (a): Deriving housing prices in x . Panel (b): Comparative statics.

$$h(x) = \left(\frac{\underline{u}}{(1-\alpha)^{1-\alpha}(w-\tau x)^{1-\alpha}} \right)^{\frac{1}{\alpha}} \quad \text{and} \quad P(x) = \Psi(x, \underline{u}) = \frac{\alpha(w-\tau x)}{h(x)} = \alpha(1-\alpha)^{\frac{1-\alpha}{\alpha}} \left(\frac{w-\tau x}{\underline{u}} \right)^{\frac{1}{\alpha}}. \quad (8.12)$$

Note that the expenditure on housing $P(x)h(x)$ is a constant share α of the wage net of the commuting cost in the second expression. This is unsurprising given the Cobb–Douglas nature of the utility function assumed in this example. Note also that the housing price gradient is convex.

8.2.3 The dual approach

For the third path to the solution, we use a dual representation of the utility function $\nu(P(x), w - \tau x)$, where $\frac{\partial \nu}{\partial P(x)} < 0$, and $\frac{\partial \nu}{\partial (w - \tau x)} > 0$. The residential equilibrium can be restated as

$$v(P(x), w - \tau x) = \underline{u} . \quad (8.13)$$

By the definition of the expenditure function $e(P(x), u)$,

$$e(P(x), v(P(x), w - \tau x)) = w - \tau x . \quad (8.14)$$

Substituting Equation (8.13) into (8.14) and totally differentiating with respect to x yields

$$\frac{\partial e(P(x), \underline{u})}{\partial P(x)} \frac{dP(x)}{dx} = -\tau , \quad (8.15)$$

which implies the Alonso–Muth condition immediately after using Shephard’s lemma:

$$\frac{dP(x)}{dx} = -\frac{\tau}{\frac{\partial e(P(x), \underline{u})}{\partial P(x)}} = -\frac{\tau}{h(P(x), \underline{u})} < 0 . \quad (8.16)$$

While perhaps less intuitive, the dual approach offers the most direct way to derive the Alonso–Muth condition. Because it makes subsequent derivations easier, we mainly retain the dual approach from now on.⁹

Residents react to the lower price of housing by consuming more of it (i.e., living in larger residences) the farther they live from the CBD. To see this, simply differentiate the Hicksian demand for housing with respect to x :

$$\frac{\partial h(P(x), \underline{u})}{\partial x} = \frac{\partial h(P(x), \underline{u})}{\partial P(x)} \frac{dP(x)}{dx} \geq 0 . \quad (8.17)$$

Equation (8.17) introduces a second gradient: the consumption of housing increases with the distance to the CBD. Note this is a pure substitution effect, since utility is being held constant at $\underline{u}$. Equation (8.17) together with Equation (8.16) immediately implies that the price of housing is convex in distance to the CBD, $\frac{d^2 P(x)}{dx^2} > 0$: house prices do not need to fall as fast as commuting costs increase with distance to the CBD to keep city residents indifferent, since they enjoy having a larger house. This convexity of the housing price gradient already appeared in the example above. This is a generic property of the model, not an artifact of functional forms used to specify preferences.¹⁰

⁹ The use of the dual approach in urban economics was originally introduced by Solow (1973) and was subsequently used by Polinsky and Shavell (1975, 1976), Henderson (1977), and Kanemoto (1980). Duranton and Puga (2014) also use the dual approach in their derivation of the monocentric model but totally differentiate Equation (8.13) and invoke Roy’s identity to derive the Alonso–Muth condition (8.16) instead of using the expenditure function and Shephard’s lemma.

¹⁰ The housing price gradient may be concave only if commuting costs are sufficiently convex in distance instead of being linear as assumed here. With commuting costs being $\tau(x)$ instead of τx , the Alonso–Muth condition has the marginal cost of commuting $\frac{d\tau(x)}{dx}$ instead of τ in the numerator. Hence, only when the commuting function is convex enough can the housing price gradient be concave. Commuting costs highly convex in distance seem unlikely.

Panel (b) in Figure 8.2 illustrates these results. It considers two locations x_1 and $x_2 > x_1$. The price of housing in x_1 , $P(x_1)$, is given by the slope (in absolute value) of the budget constraint with intercept $w - \tau x_1$ that is tangent to the indifference curve $u(h, z) = \underline{u}$. The point where the indifference curve and the budget constraint are tangent allows us to read the consumption of numéraire $z(x_1)$ on the vertical axis and the consumption of housing $h(x_1)$ on the horizontal axis. For the resident in x_2 , the price of housing, $P(x_2)$, and the consumption of numéraire, $z(x_2)$, and housing, $h(x_2)$, can be read in the same way using this time a budget constraint with intercept $w - \tau x_2$. This lower budget constraint must be flatter—that is, involve a lower price of housing—for it to be tangent to the indifference curve. Then it is also immediate that the consumption of housing is higher in x_2 relative to x_1 , whereas the consumption of the numéraire is lower.

We now turn from housing to the land it is built on. To supply housing, a perfectly competitive construction industry uses land and capital under constant returns to scale to produce an amount $f(x)$ of housing floor space per unit of land at a distance x from the CBD. The production function for housing is increasing in both its arguments and is strictly quasi concave. The rental price of land, denoted $R(x)$, varies across the city. The rental price of capital is constant and endogenously given, so we omit it as an argument of the unit cost function in construction $c(R(x))$. The zero-profit condition for the construction sector can then be written as $P(x) = c(R(x))$. Totally differentiating this expression with respect to x yields

$$\frac{dP(x)}{dx} = \frac{\partial c(R(x))}{\partial R(x)} \frac{dR(x)}{dx}, \quad (8.18)$$

which implies

$$\frac{dR(x)}{dx} = \frac{dP(x)}{dx} \frac{1}{\frac{\partial c(R(x))}{\partial R(x)}} = \frac{dP(x)}{dx} f(x) < 0, \quad (8.19)$$

where the simplification follows from Shephard's lemma.¹¹ Thus, the reduction in the price of housing as one moves away from the CBD is reflected in a reduction in the price of land. The construction industry then reacts to lower land prices by building with a lower capital-to-land ratio further away from the CBD. Put differently, there are two other gradients here as one moves away from the CBD: declining land prices and

¹¹ Note that the relationship between the land gradient price and the housing price gradient can alternatively be written as $\frac{d \log R(x)/dx}{d \log P(x)/dx} = \frac{f(x)P(x)}{R(x)}$. Thus, the land price decline is many times the house price decline. In particular, the ratio of the percentage decline in the land price to the percentage decline in the housing price equals the ratio of the price level of housing and the price level of the land the housing is built on.

declining capital intensity in housing (i.e., both larger gardens and properties with fewer stories).

Land is built if the rent $R(x)$ it can fetch in residential use is at least as high as the rent $\underline{R}$ it can fetch in the best alternative use (e.g., agriculture). The edge of the city is thus located at a distance $\bar{x}$ from the CBD such that

$$R(\bar{x}) = \underline{R} . \quad (8.20)$$

The physical extent of the city must also be sufficient to hold its population N :

$$N = \int_0^{\bar{x}} n(x) dx , \quad (8.21)$$

where $n(x)$ denotes population density at a distance x from the CBD. Using Equations (8.16) and (8.19), we can express population density as

$$n(x) = \frac{f(x)}{h(x)} = \frac{\frac{dR(x)}{dx} / \frac{dP(x)}{dx}}{-\tau / \frac{dP(x)}{dx}} = -\frac{1}{\tau} \frac{dR(x)}{dx} . \quad (8.22)$$

By Equation (8.19) and the fact that capital intensity declines with distance to the CBD, $\frac{df(x)}{dx} < 0$, density also declines with distance to the CBD, $\frac{dn(x)}{dx} < 0$. This gradient of declining density of residents as one moves farther from the CBD is the fifth gradient predicted by this model. As made clear by Equation (8.22), it is a direct implication of two other gradients already discussed: the increase in housing consumption and the decline in the capital intensity of development as one moves farther from the CBD.

Substituting expression (8.22) for $n(x)$ into Equation (8.21), solving the integral, and using Equation (8.20) yields $N = \frac{R(0) - \underline{R}}{\tau}$. This implies a very simple expression for land rent at the CBD ($x = 0$):

$$R(0) = \underline{R} + \tau N . \quad (8.23)$$

Valuing the residential equilibrium condition $v(P(x), w - \tau x) = \underline{u}$ at $x = 0$ and using Equation (8.23), we can write the price of housing at the CBD as $P(0) = c(\underline{R} + \tau N)$. The residential equilibrium condition holds for any location in the city, so valuing it at an arbitrary x and at $x = 0$, and using the previous expression for $P(0)$ yields

$$\begin{aligned} v(P(x), w - \tau x) &= \underline{u} = v(P(0), w) \\ &= v(c(\underline{R} + \tau N), w) . \end{aligned} \quad (8.24)$$

This can be inverted to solve for house prices $P(x)$ as a function of x , N , w , τ , and $\underline{R}$. That is the “closed city” version of the monocentric city model, which treats population N as a parameter. The “open city” version allows N to be endogenously determined by

migration across cities to attain a common utility level. We can write the condition of utility equalization across cities as

$$v(c(\underline{R} + \tau N), w) = \underline{u}. \quad (8.25)$$

This spatial equilibrium condition can be inverted to solve for N as a function of $\underline{u}$, w , τ , and $\underline{R}$.

The gradients presented so far compare locations within the city for given parameters. While a full analysis of the comparative statics of the model is beyond our scope here, it is also of interest to describe how the city is affected by changes in these parameters. We focus on the open city case, where we can think of the comparative statics as the consequences of changes that affect a city that is small relative to the entire economy. Then the level of utility $\underline{u}$ is pinned down by what consumers can achieve elsewhere. An increase in the wage in a city increases house prices everywhere in the city: Equation (8.7) immediately implies $\frac{dP(x)}{dw} > 0$. Housing needs to become more expensive to offset higher wages as residents need to retain the same level of utility as elsewhere in the economy, and this is attained through a population increase in the city.

A reduction in transport costs—for instance, from road or transit improvements—also increases house prices everywhere in the city: $\frac{dP(x)}{d\tau} < 0$. This increase in house prices is again the result of individuals moving into the city in response to the utility gain from lower commuting costs. To accommodate this larger population, cities physically expand outward and also experience rising densities. Of these two channels, the monocentric model predicts that the physical expansion of the city is more important. To see this, consider a segment of the city between the CBD and a point x_C , and think of this segment as the historical central city. We can then think of the segment between x_C and the city edge $\bar{x}$ as the suburbs. Let $N_C = \int_0^{x_C} d(x)dx$ denote the (endogenous) population of the central city. Then, using Equations (8.22) and (8.23), we can calculate the share of the population in the central city as

$$\frac{N_C}{N} = \frac{R(0) - R(x_C)}{R(0) - \underline{R}}. \quad (8.26)$$

A reduction in τ increases land rent at any given point beyond the CBD including x_C , but it does not affect land rent $R(0)$ at the CBD (where there is no need to commute and migration keeps utility unchanged) nor land rent at the city edge, which is fixed at $\underline{R}$. Then Equation (8.26) implies that the share of the population in the central city falls when commuting costs are reduced. This implies that improvements in local transport foster the suburbanization of the population.

To sum up, the monocentric model makes a number of important predictions. It is best known for predicting gradients. We saw five of them above. As one moves away from the CBD, housing prices decline, housing consumption increases, land prices

decline, the density of construction declines (i.e., capital intensity in housing declines), and population density declines. These five gradients are potentially measurable and appear intuitively plausible.

8.2.4 Quantitative predictions

Beyond these qualitative predictions, the monocentric model also makes a number of sharp quantitative predictions. First, the Alonso–Muth condition in Equation (8.4) is more precise than simply giving a direction of change. It states that the slope of the housing price gradient is equal to the ratio of the marginal cost of commuting, τ , to the consumption of housing, $h(x)$. More intuitively, as one moves away from the CBD, the marginal decline in house prices should be exactly offset by the marginal increase in the cost of travel. This is observable in principle.

This is not the only quantitative prediction that can be derived from the monocentric model. The second one is apparent in Equation (8.19) and results from optimal investment in housing and zero profit in construction: the ratio of the land price gradient to the housing price gradient, $\frac{dR(x)/dx}{dP(x)/dx}$, should be equal to the amount of housing $f(x)$ (in terms of floor space per unit of land). Put differently, if for the same housing price gradient the land price gradient is twice as steep, there should be twice as much housing.

The third quantitative prediction appears in Equation (8.22). It states that the population density in x , $n(x)$, is equal to minus the ratio of the land price gradient, $\frac{dR(x)}{dx}$, to the marginal cost of commuting, τ . Just like with housing price, land prices adjust to reflect the greater cost of commuting to the CBD but, for land, this relationship needs to be weighted by the number of local residents. This condition is, of course, the land price equivalent of the Alonso–Muth condition (8.4).

While these three quantitative predictions of the monocentric model are “local” in the sense that they apply at each location, the monocentric model also generates two aggregate predictions. The first of these is made clear by Equation (8.23): the differential land rent between the CBD and the edge of the city, $R(0) - \underline{R}$, should be proportional to the city population, N , and to the unit commuting cost, τ . This is a strong quantitative prediction, which can also be potentially applied to the data. Because it concerns city aggregates, this quantitative prediction is more sensitive to our modeling assumptions and, in particular, to the geography that is imposed on cities. For instance, the proportionality of the population to the differential land rent at the center does not carry through to two-dimensional cities.¹² This is because Equation (8.21) becomes $N = \int_0^{\infty} l(x)n(x)dx$, where $l(x)$ is the amount of land at distance x from the center. If the city sits on a homogeneous plane, then $l(x) = 2\pi x$. However, in reality geographical

¹² The model proposed by Combes et al. (2012) implies the proportionality of land rent at the CBD to population in two-dimensional cities. The three key elements that generate this result are an isoelastic demand for housing, constant returns in construction, and multiplicatively separable commuting costs that directly enter the utility function.

constraints and the fact that parcels of land more suitable for construction tend to be built first imply that cities probably lie somewhere between the one-dimensional and the two-dimensional cases (and in some more extreme cases may be below the one-dimensional case).

The second aggregate quantitative prediction states that total commuting costs are equal to the total differential land rent:

$$\text{TCC} \equiv \int_0^{\bar{x}} \tau x n(x) dx = \int_0^{\bar{x}} -\frac{x}{\tau} dR(x) = \int_0^{\bar{x}} (R(x) - \underline{R}) dx \equiv \text{TDLR}, \quad (8.27)$$

where the first equality is obtained using (8.22) and the second requires integrating by parts. This result for the proportionality of city aggregates is less sensitive to the modeling of cities as shown by [Arnott and Stiglitz \(1981\)](#). Different geographies imply a different, but still constant, ratio of total differential land rent to total commuting costs. For instance, with two-dimensional circular cities, the total commuting cost is twice the total differential land rent. The linearity of commuting costs remains nonetheless crucial for this aggregate result.

Finally, it remains to be said that the equilibrium of the monocentric model presented here is efficient. This is an expected outcome given the competitive behavior of residents and housing builders and the absence of externalities. It has, however, been well known since [Mirrlees \(1972\)](#) and [Dixit \(1973\)](#) that a non-Rawlsian planner may want to treat *ex ante* equal residents unequally *ex post*. As shown by [Wildasin \(1986\)](#), this is because the marginal utility of income differs across locations at the symmetric equilibrium. Hence, a utilitarian planner will allocate workers differently from the competitive equilibrium case explored here, where *ex ante* equal residents end up with the same level of utility *ex post* (despite living in different locations). A Rawlsian planner will mimic the market outcome. This can be proven by showing that the market allocation minimizes the total social cost (the sum of the commuting cost, the opportunity cost of land, and cost of numéraire consumption) required to achieve a given level of utility $\underline{u}$. We do not develop the proof here for the sake of brevity, but the interested reader can consult [Fujita and Thisse \(2013, p. 85\)](#).

8.3. EXTENDING THE MONOCENTRIC MODEL

The monocentric model presented above is in many respects a remarkable achievement. It brings together housing, construction, transport, and the choice of location and consumption made by households in a tractable way. It derives many predictions. Many of these predictions regard the existence of several gradients, and casual observation is certainly supportive of these predictions. The monocentric model is also a very elegant construct whose apparent simplicity actually reveals a lot of subtlety. Unsurprisingly, the

work of [Alonso \(1964\)](#), [Mills \(1967\)](#), and [Muth \(1969\)](#) hailed the creation of a new sub-field of economics: urban economics. Although, as the rest of this volume hopefully makes clear, this field has broadened its objects of study to a variety of other issues, most urban economists still recognize the monocentric model as being core to the urban field.

This said, the exposition in [Section 8.2](#) relies on strong simplifying assumptions. Our objective in this section is threefold. First, we aim to review the main extensions of the monocentric models to take stock of the state of the literature. Second, we want to distinguish between the extensions that add to the richness of the monocentric model and work that questions its main results. Our review here puts more emphasis on the latter. Finally, we also emphasize the extensions that are the most relevant for the empirical work which is reviewed from [Section 8.6](#) onward.

8.3.1 Other elements of travel costs and looking beyond accessibility

Our review of extensions to the monocentric model begins with generalizations that do not constitute major theoretical challenges for the baseline model presented in [Section 8.2](#). Rather, they allow us to clarify what the monocentric model is really about and also help highlight a number of issues which are of first-order empirical importance.

The first of these issues regards the specification for commuting costs. Travel is costly, not only in pecuniary terms, but also in time. Transport economists routinely value time spent in privately owned vehicles at half the wage of the traveler ([Small and Verhoef, 2007](#)). To consider the time cost of commuting, we assume that travel costs t units of time per unit of distance in addition to out-of-pocket expenses. We also assume that each resident chooses how much labor ℓ to supply and how much leisure s to consume. Leisure enters the utility function as a positive argument, $u(h, z, s)$. Each resident located at distance x from the CBD now faces both a pecuniary budget constraint, $w \ell(x) - \tau x = P(x) h(x) + z(x)$, and a time budget constraint, $\ell(x) + s(x) + t x = 1$, following the normalization of the total endowment of time of each resident to unity. We can insert the latter budget constraint into the former and obtain

$$w = P(x) h(x) + z(x) + w s(x) + w t x + \tau x = P(x) h(x) + z(x) + w s(x) + T(x), \quad (8.28)$$

where $T(x) \equiv w t x + \tau x$ is the total cost of commuting, and the wage w is also the shadow cost of time. A resident in x maximizes his or her utility $u(h(x), z(x), s(x))$ with respect to his or her consumption of housing, leisure, and other goods subject to the budget constraint (8.28). We can insert the budget constraint into the utility and have this resident maximize $u(h(x), w(1 - s(x)) - P(x) h(x) - T(x), s(x))$ with respect to housing h and leisure s . From there, the Marshallian demands can be easily recovered. In equilibrium, utility must also be equalized across locations:

$$u(h(x), w(1 - s(x)) - P(x) h(x) - T(x), s(x)) = \underline{u}. \quad (8.29)$$

This spatial equilibrium condition is the analogue of Equation (8.2) after introduction of a time cost of commuting as part of an endogenous allocation of time. After simplifications using the first-order conditions with respect to housing and leisure, totally differentiating Equation (8.29) with respect to x to derive the optimal choice of residence yields

$$\frac{dP(x)}{dx} = -\frac{1}{h(x)} \frac{dT(x)}{dx}, \quad (8.30)$$

which is the Alonso–Muth condition that corresponds to Equation (8.4) above with leisure and a time cost of commuting. It is then easy to show that all the other gradients exhibited above are also present here. As argued above, the aggregate properties of the simpler monocentric models are more reliant on the linearity of commuting costs. Note also that similar results with the same five gradients can be obtained by assuming that commuting enters the utility function directly as a negative argument $u(h, z, x)$. Again, solving for the first-order conditions in h , z , and x and totally differentiating the spatial indifference condition analogue to Equation (8.29) yields another version of the Alonso–Muth condition and a negative housing price gradient under mild conditions regarding preferences.¹³

While a more realistic modeling of commuting is more intricate than what the simple monocentric model of Section 8.2 assumes, this does not affect the main properties and predictions of this model. A similar conclusion holds for “accessibility.” The simple monocentric model in Section 8.2 equates accessibility with distance to the CBD. This is restrictive for two reasons. First, because in reality employment is far from being concentrated around a single center, so travel to work patterns are more complex than trips to a common center. We deal with generalizations of this aspect in Section 8.4. Second, much of observed travel takes place for reasons other than commuting. Couture, Duranton and Turner (2012) report that in the largest 100 US metropolitan areas, commutes represented less than a quarter of all trips by privately owned vehicles in 2008. Shopping trips, recreational trips, and trips for other personal/family business are about equally important. Even in a city with a single employment center, the location of retail, entertainment, and family and friends is thus likely to matter a lot as well.

While these considerations are empirically of first-order importance, richer notions of accessibility can be readily incorporated into the monocentric model. Assume, for instance, that there is still a single CBD where all employment is located but there are also many equally spaced retail centers. Residents need to travel to work on a daily basis and go shopping, say, every other day. It would be easy to extend the model in Section 8.2 to account for this. Taken alone, the location of jobs still leads to a negative housing price gradient, while the location of retail implies a sawtooth pattern with a small peak at every retail location. Bringing these two elements together, one can easily see that

¹³ We are grateful to Frédéric Robert-Nicoud for pointing this out to us.

the job accessibility gradient dominates since travel to retail centers is half as frequent as travel to employment and since there are many retail centers compared with just one employment center. The overall gradient is thus negative but flatter before each retail center and steeper after as one moves away from the CBD. Increasing the frequency of shopping trips or the average distance to a retail center sufficiently can lead to a nonmonotonic overall gradient.

These more complex notions of accessibility can be incorporated in a variety of other ways. For instance, [Anas and Moses \(1979\)](#) and, more recently, [Baum-Snow \(2007b\)](#) consider radial commuting highways in two-dimensional cities. This introduces some heterogeneity in the unit price of housing for properties that are located at the same physical distance to the CBD since they will differ in their distance to the radial highway. Should access to the radial highway be limited by a finite number of exits, the gradient could become nonmonotonic again because the travel distance to the CBD for a location close to a radial highway exit may be less than that of another location that is closer to the CBD but farther from this exit.

Nonmonotonic gradients also occur naturally if one considers several modes of travel that may be combined with one another. For instance, residents may walk to a railway station before their ride to the CBD. By the same type of argument as with radial roads, this generates nonmonotonic gradients with local peaks of housing prices at the railway stations.

As discussed below, it is also possible to embed considerations other than accessibility into a monocentric framework. Large metropolitan areas are often divided into several municipalities or districts, which in some countries have considerable latitude regarding taxation and the local public goods that they offer, including education in the United States. The provision of local public goods will differ across districts and affect housing prices and, in turn, patterns of land use (see, e.g., [De Bartolome and Ross, 2003](#)). Local amenities may also have a spatial dimension and will naturally affect housing prices and thus land use and the various gradients (see, e.g., [Brueckner et al., 1999](#)).

Although none of these extensions generate theoretically surprising results, they are useful to make two important points that are empirically highly relevant: accessibility cannot be reduced to the distance to the CBD and accessibility is not the only determinant of housing prices and thus of patterns of land use.

8.3.2 Heterogeneous residents, properties, and timing

The model described in [Section 8.2](#) assumes the existence of a representative resident. Even though residents all end up in different locations, they are ex ante symmetric with similar preferences and the same labor income. From the theoretical standpoint, generating asymmetric outcomes from ex ante symmetric agents is a strength. Having residents in different locations and with different consumption levels of housing and other goods is

the equilibrium result of a market allocation. It is not the direct and somewhat trivial implication of location-specific tastes, for instance, nor a choice made by an urban planner. While exploring a case with representative residents is theoretically important, it remains nonetheless crucial to understand how household heterogeneity affects residential location patterns.

A first way to model household heterogeneity is to consider different groups with, for instance, heterogeneous income levels.¹⁴ Consider, for instance, N_0 poor residents with a wage w_0 and a utility level $\underline{u}_0$ and N_1 rich residents with a wage $w_1 > w_0$ and a utility level $\underline{u}_1 > \underline{u}_0$. If housing is essential, residents from both groups must consume a positive amount of housing. This implies that in some locations poor residents must outbid rich residents. There must also be other locations where the opposite is true. Consider points that separate rich and poor and denote $\tilde{x}$ one such “boundary” point.¹⁵ At $\tilde{x}$, the bid rents of rich and poor must be the same: $P_0(\tilde{x}, \underline{u}_0) = P_1(\tilde{x}, \underline{u}_1)$. Otherwise, if one group were willing to pay strictly more, there would be only one group of residents, rich or poor, living on both sides of this point, which contradicts the definition of $\tilde{x}$ as a boundary point. This equality in bid rent across groups implies that rich residents must then consume more housing than poor resident at $\tilde{x}$ as long as housing is a normal good. That is, we must have $h(P_1(\tilde{x}, \underline{u}_1)) > h(P_0(\tilde{x}, \underline{u}_0))$. If we use the Alonso–Muth condition (8.16), this implies $\frac{dP_1(\tilde{x})}{dx} = -\frac{\tau}{h(P_1(\tilde{x}), \underline{u}_1)} > -\frac{\tau}{h(P_0(\tilde{x}), \underline{u}_0)} = \frac{dP_0(\tilde{x})}{dx}$. Thus, the rent gradient should be steeper for poor residents than for rich residents.

If poor residents have a steeper housing price gradient, in equilibrium they will live closer to the CBD, whereas rich residents should live further away. Hence, when commuting costs are the same for both groups and housing is a normal good, poor residents are predicted to occupy small dwellings close to the CBD, where housing is more expensive. The key driver of this somewhat counterintuitive result is that rich residents are more willing to pay greater commuting costs and live further from the CBD because their higher wage allows them to consume more land.

As discussed in [Section 8.3.1](#), in practice the cost of commuting involves a time cost, and the opportunity cost of time is likely larger for rich residents. Working in the opposite direction is the fact that poor residents are more likely to rely on public transport. With different commuting costs for poor and rich τ_0 and τ_1 , the condition for poor residents to occupy small central dwellings and rich residents to live in larger residences out in the suburbs is

¹⁴ For the sake of brevity, most of our focus here is on income heterogeneity. There are other forms of heterogeneity that matter. Race is of particular salience in the United States.

¹⁵ The reasoning can be extended to intervals where rich and poor coexist (although they do not occur in equilibrium here).

$$\frac{dP_1(\tilde{x})}{dx} = -\frac{\tau_1}{h(P_1(\tilde{x}), \underline{u}_1)} > -\frac{\tau_0}{h(P_0(\tilde{x}), \underline{u}_0)} = \frac{dP_0(\tilde{x})}{dx}. \quad (8.31)$$

The literature sometimes uses this condition directly, implying that rich residents live further out if the ratio of commuting costs per unit distance to housing consumption is lower for them (note the minus sign in front of both sides of the inequality in Equation (8.31)). However, it is also quite frequent to see the condition expressed in terms of elasticities. In this case, the condition for rich residents to live further out is that the income elasticity of commuting costs is smaller than the income elasticity of the demand for housing.¹⁶ Finally, some articles prefer to express the condition in terms of the income elasticity of the demand for land (although residents do not value land per se, and consume it indirectly as an input used in the production of housing). Using Equation (8.19), we can rewrite Equation (8.31) as

$$\frac{dR_1(\tilde{x})}{dx} = \frac{dP_1(\tilde{x})}{dx} f(\tilde{x}) = -\frac{\tau_1 f(\tilde{x})}{h(P_1(\tilde{x}), \underline{u}_1)} > -\frac{\tau_0 f(\tilde{x})}{h(P_0(\tilde{x}), \underline{u}_0)} = \frac{dP_0(\tilde{x})}{dx} f(\tilde{x}) = \frac{dR_0(\tilde{x})}{dx}. \quad (8.32)$$

Note that $\frac{h(P(x), \underline{u})}{f(x)}$ is the demand for land (embedded in housing consumption), calculated as the product of housing floor space, $h(P(x), \underline{u})$, and land input per unit of floor space produced, $\frac{1}{f(x)}$. Hence, rich residents live further out if the ratio of commuting costs per unit distance to land consumption is lower for them. Alternatively, expressed in terms of elasticities, the condition is that the income elasticity of commuting costs must be smaller than the income elasticity of the demand for land.

While we discuss the empirical relevance of this result below, several properties of this extension to multiple demographic groups must be discussed. First, within each demographic group the monocentric model remains exactly as in Section 8.2 and the results are pinned down by utility equalization across locations. Of course, utility must be equalized within groups but not across groups. Second, at any location, residents from different groups will want to consume different amounts of housing. In turn, this implies bid-rent curves of different steepness. Hence, in equilibrium, the group with the steepest bid-rent curve will have the highest willingness to pay close to the CBD, whereas the group with the flattest bid-rent curve will occupy the locations closest to the urban fringe. The overall bid-rent curve will be formed by the upper envelope of the bid-rent curves from the different groups of residents. This should result in further convexity of the city bid-rent curve. The equilibrium with different groups of residents also implies perfect segmentation of these groups across city segments.

¹⁶ In the specific case where commuting costs are proportional to the wage, this condition implies that rich residents remain on the outskirts of the city provided the income elasticity of the demand for housing is above 1 (Becker, 1965).

Another way to model household heterogeneity is to assume a continuous distribution for the characteristic by which households differ. Let us retain income as the key dimension by which households differ. A continuous distribution presents the monocentric model with a much greater challenge since, under a continuum of incomes, we expect a continuum of utilities and we can no longer rely on the residential equalization condition used above: $u(h, z) = \underline{u}$.

For the problem to remain tractable and for the key intuitions to be conveyed easily, we return to the example used above where residents maximize a Cobb–Douglas utility function $u(h, z) = h^\alpha z^{1-\alpha}$ subject to the budget constraint $P(x)h + z = w(x) - \tau x$. The main difference is that wages are now distributed according to the probability distribution function $g(w)$, whereas before they took a discrete number of levels. For simplicity, we disregard construction and assume that there is one unit of housing available per unit of land.¹⁷

Solving the model involves characterizing the functions $P(x)$, $w(x)$, and $h(x)$ that describe the housing price, the wage, and the consumption of housing at each location. Formally, we are solving an assignment problem. Unlike the standard assignment problems, where, for instance, a fixed number of workers are assigned to a fixed number of machines, residents choose how much housing to consume.¹⁸ To keep the exposition straightforward, we follow the heuristic originally proposed by Beckmann (1969).¹⁹ This heuristic should be viewed as the counterpart to the Marshallian approach derived above with heterogeneous residents. In essence, we derive the demand for housing at each location and equate it with the supply of housing. Behrens et al. (2015) propose a more formal derivation of the equilibrium where the assignment function is explicitly modeled.²⁰

With a Cobb–Douglas utility function, solving for the allocation of disposable income between housing and the numéraire readily yields

¹⁷ We could also solve for housing development, but that would introduce yet another equation into the system below.

¹⁸ The standard assignment problem can be traced to Koopmans and Beckmann (1957). A nice presentation can be found in Sattinger (1993).

¹⁹ Beckmann's choice of functional forms differs slightly from ours. His proposed solution contained minor mistakes later corrected by Montesano (1972).

²⁰ Brueckner et al. (2002) propose another heuristic that uses a bid-rent approach. More specifically, they rely on the notion that the bid rent of a given type of resident should be maximized at the equilibrium location. That is, they consider that land is allocated through a first-price auction where landowners are able to extract all the surplus. The rest of the literature explicitly or implicitly views housing and land allocation as a second-price (English) auction or a sealed-bid first-price auction, where the seller is not able to extract all the surplus. See section 13.6 in Chapter 13 in this handbook for more on real-estate auctions.

$$h(x) = \frac{\alpha(w(x) - \tau x)}{P(x)}, \quad (8.33)$$

which was already part of expression (8.12). Solving for the location choice of residents implies another version of the Alonso–Muth condition:

$$\frac{dP(x)}{dx} = -\frac{\tau}{h(x)}. \quad (8.34)$$

Using the same type of argument as above, we expect residents with higher wages to reside further from the CBD. That is, and with a slight abuse of language, we expect positive assortative matching between residents (ordered by increasing income) and endogenously defined parcels (ordered by increasing distance to the CBD). More formally, between locations x and $x + dx$, we will find residents with income between w and $w + \frac{dw(x)}{dx}dx$, where $\frac{dw(x)}{dx}$ describes how wages change across locations.

We can now equate the supply and demand of housing between x and $x + dx$. Following the simplifying assumption made above of disregarding the construction sector and the assumption of a linear city, the supply of housing between x and $x + dx$ is dx . There is a density $g(w(x))$ of residents with income between $w(x)$ and $w(x) + \frac{dw(x)}{dx}dx$. Hence, equating demand and supply leads to $h(x)g(w(x))\frac{dw}{dx}dx = dx$, or after simplifications

$$\frac{dw(x)}{dx} = \frac{1}{h(x)g(w)}. \quad (8.35)$$

Substituting (8.33) into (8.34) and (8.35), we obtain a system of two differential equations for $P(x)$ and $w(x)$. In the case where the distribution of wages is a Pareto distribution, $g(w) = (b+1)\frac{w^{b+1}}{w^b}$, it is easy to verify that the system formed by Equations (8.33)–(8.35) admits a solution of the following type:

$$w(x) = \underline{w} x, \quad P(x) = P(0) x^{-b}, \quad \text{and} \quad h(x) = h(0) x^{b+1}. \quad (8.36)$$

These expressions imply that the housing price gradient and the housing consumption gradients are both power functions of the distance to the CBD. This may not be very surprising since a constant share of disposable income is spent on housing and the distribution of wages is assumed to follow a power law. This expression should also make it clear why with other distributions of wages a closed-form solution will typically be unavailable. It must also be the case that if the housing consumption and housing price gradients are to be power laws, the wage gradient must be linear in x for Equation (8.33) to factorize and simplify.

Relative to the canonical model with a representative resident derived above, it is easy to see that all its key local properties are retained here with heterogeneous residents despite

the absence of utility equalization across locations within the city.²¹ That is, the model still predicts an Alonso–Muth condition for the housing price gradient. It also still predicts gradients of housing consumption, density of residents, etc. A development gradient could also be derived in a more complete model where construction is explicitly modeled.

Note that these gradients differ from the situation with homogeneous residents explored above. With Cobb–Douglas utility, the house price gradient with homogeneous residents given in expression (8.12) depends on the income net of commuting costs $w - \tau x$ elevated to the power $1/\alpha$, the inverse share of housing in consumption. With heterogeneous residents, it is equal to distance elevated to minus the slope parameter b of the Pareto distribution of income. Interestingly, this does not depend on the commuting cost parameter. To explain this surprising result, note first that, for a given resident to be at his or her optimal location, he or she should have no incentive to move further away from the CBD. Hence, the Alonso–Muth condition continues to hold for that person. But then, recall that land is no longer allocated through an indifference condition where residents must be indifferent everywhere. Instead, residents with different levels of income compete to occupy land. More specifically, a resident competes for land with his or her poorer neighbor closer to the CBD and with his or her other, richer, neighbor located on the other side. How much this resident will bid in equilibrium will depend on how much richer he or she is relative to his or her poorer neighbor and how much poorer he or she is relative to his or her richer neighbor. In equilibrium, the shape parameter of the distribution of income solely drives the housing price gradient.

Finally, the ordered sorting of residents by increasing income as one moves away from the CBD is extreme. In a richer model with residents that also differ in their commuting costs, Behrens et al. (2015) show that it is easy to relax this result to obtain some social mixing. In this case, the model predicts only broad trends where income rises only on average as one considers locations further from the CBD.

Assignment models have the great advantage of being able to deal more naturally with the inherent heterogeneity of city residents. This advantage comes at the cost of a much greater technical complexity. Setting up the equilibrium conditions in models less rudimentary than the one considered here is often challenging, and closed-form solutions are available only in specific cases. Since in assignment models, a resident in a given location pays only what the second person with the highest willingness to pay is willing to offer, these models naturally link different submarkets without forcing full equalization. Since Braid (1981), these models have been used to understand price changes in different quality segments of the housing market. Recent contributions include those of Määtänen and Terviö (2014) and Landvoigt et al. (2011). We do not dwell on these models further here because they do not model land use explicitly.

²¹ The more aggregate properties of the monocentric models with homogeneous residents that relate total land rents and total commuting costs do not hold in general in this type of setting (Behrens et al., 2015).

While much of the literature has been concerned with heterogeneous users of land, land parcels and the properties that sit on them are also highly heterogeneous. Starting with land, [Combes et al. \(2012\)](#) report that for vacant parcels that were developed into single-family homes in France in 2008, the mean parcel area is 1100 m², with a large standard deviation of 1200 m². Even after the city where parcels are located and their distance to the center have been controlled for, there remains considerable heterogeneity. Furthermore, after their location and area have been accounted for, parcels also differ in a number of other dimensions, including their shape. The shape of parcels is more difficult to characterize. [Combes et al. \(2012\)](#) suggest an *ad hoc* measure using the ratio of the road frontage of a property to the square root of its area. The mean of this ratio for parcels with newly built single-family homes in France is 0.68. The standard deviation is nearly as large at 0.50, with a first decile at 0.23 and a ninth decile at 1.07. Again this suggests considerable heterogeneity, this time in the shape of parcels.²²

This heterogeneity of parcels matters. According to data used by [Combes et al. \(2012\)](#), land prices per square meter decline fairly sharply beyond a certain size threshold. Similarly, a standard deviation of the frontage to area ratio is valued at 8% of the sale price. When properties are considered instead of only land, the heterogeneity in land parcels is compounded by the heterogeneity of the structure that sits on them. This heterogeneity of structure concerns their size, quality, level of depreciation, and style.

Despite such heterogeneity being widely mentioned to explain why property markets do not work like other asset markets, there has been very little to no work on the determinants and implications of parcel and property heterogeneity for land use in cities.²³ Several articles by Asami and coauthors (see, e.g., [Asami and Ohtaki, 2000](#)) attempt to provide axiomatic metrics for parcel shapes and explore their development implications. There is also a debate about continuous versus discrete models of land use. [Berliant \(1985\)](#) questioned the consistency of continuous land use models as an approximation for large discrete economies since each resident (in continuum) can consume only a zero amount of land in equilibrium. In their responses, [Papageorgiou and Pines \(1990\)](#) and [Asami et al. \(1991\)](#) showed under which conditions appropriately defined urban models with a discrete number of residents approximate classical continuous models. Because even in the discrete models considered by these authors land parcels

²² There is also considerable heterogeneity in residential and job densities. [Anas et al. \(1998\)](#) provide striking evidence of such heterogeneity in Los Angeles. A strongly smoothed three-dimensional representation of employment density makes the city look strongly monocentric with only a few subcenters. Higher levels of resolution reveal instead a highly “jagged” picture. This feature is also apparent when looking at the details of the map of Paris in [Figure 8.1](#).

²³ Instead, there is a long tradition that treats parcel heterogeneity as a nuisance that needs to be conditioned out. This is usually done through hedonic regressions. However, the models presented here suggest that parcel location, parcel area, and their intensity of construction are all determined simultaneously. This points to some obvious endogeneity problems for hedonic regressions.

are determined endogenously without friction, this debate is about the foundations of the monocentric model, not about the implications of parcel heterogeneity.

To explain the heterogeneity of parcels at a given distance from the CBD, natural geography and roads must play a prominent role. Parcels are heterogeneous because the land itself is heterogeneous. In addition, roads and other man-made obstacles need to cut through the land.

One might object that despite this unevenness of land, we may be able to come reasonably close to the optimal parcel size and optimal development for each parcel suggested by the model above. Even if the land is uneven and divided by the roadway, the blocks of parcels are usually large enough to be able accommodate a fairly close approximation to the optimal number of parcels. The heterogeneity of parcels and properties would then be a reflection of heterogeneous residents choosing to live in the same location as, for instance, in [Behrens et al. \(2015\)](#), where heterogeneity in income and commuting costs implies the presence of different residents occupying parcels of different sizes with different amounts of housing living next to each other.²⁴ Note that this approach to parcel heterogeneity is in line with the modeling approach taken in [Section 8.2](#) where parcels arise endogenously following choices made by residents.

There are, however, reasons to believe that the heterogeneity of demand is not the entire explanation for parcel and property heterogeneity. There might be some elements of (supply) exogeneity in the sizes and shapes of parcels. Taken literally, the monocentric model presented above implies that parcels and properties should be modified as incomes grow or commuting costs decline. The debate about the foundations of the monocentric model mentioned above has also stressed the indivisible nature of parcels and housing units ([Ellickson, 1977](#); [Berliant, 1985](#)). In reality, most established neighborhoods often see little change in many years despite changes in income, progress in transport technology, or wild fluctuations in gasoline prices. [Brooks and Lutz \(2012\)](#) provide more direct evidence from the land assembly process and report that assembled parcels trade at an about 40% premium. [Cunningham \(2013\)](#), using different data, finds a lower but still large premium of about 20%. Whether this large premium is due to a holdout problem or to parcels hosting properties at different stages of their life cycle is an open question.²⁵

²⁴ An alternative is to assume that residents have idiosyncratic preferences for locations within the city as in [Anas \(1990\)](#), where this heterogeneity in tastes is modeled using a logit specification. Greater taste heterogeneity relaxes the competition for land close to the CBD and flattens the bid-rent curve. With income heterogeneity it should also weaken the sorting of residents by income. A difficulty with the logit framework is that the utility valuation of a parcel is independent of the valuation of the parcel next to it.

²⁵ See [Strange \(1995\)](#) for a consistent model of the holdout problem. In this model, a developer first makes offers for parcels, which are accepted or rejected by landowners, who do not know how much the developer stands to gain. Landowners can make counteroffers which are accepted or rejected by the developer. There are many possible equilibria. The weakly dominant equilibrium has a number of interesting properties. The first-stage offer is not informative. Small landowners ask for relatively more and often hold out on projects. In many cases, socially profitable projects do not get implemented.

This nonetheless suggests the existence of considerable frictions regardless of from where exactly they stem. Empirically, we would like to know how much of the observed heterogeneity of residents in similar locations is caused by the exogenous supply of heterogeneous parcels and how much is caused by the demand from residents that are heterogeneous in two or more dimensions and endogenously make parcels and properties heterogeneous.²⁶

Although, to our knowledge, the implications of exogenous parcel heterogeneity have not been formally explored, we can form a number of conjectures about their effects. First, parcel heterogeneity will generate some mixing of heterogeneous residents. Second, the frictions that underlie parcel heterogeneity may also be at the root of significant inefficiencies in land use, particularly in areas that were developed long ago—that is, the central part of cities. Parcels may have been of optimal size when the city was first developed. However, given changes in income, transport, or construction technology, parcels that were of optimal size 100 years ago are unlikely to be of optimal size today. They may be instead be grossly suboptimal. It would be important to know how far existing parcels are from unconstrained optimality. Getting an idea of the implied welfare losses would also be important because large deviations from optimal parcel size may only imply small welfare losses if the effects of suboptimal parcel size can be partially offset through several other margins (changing intensity of development, rising housing quality, preservation of historical buildings, etc.). More generally, the theoretical literature (and the empirical literature) has paid much more attention to trends (i.e., gradients) than to variations around those trends (i.e., the heterogeneity between neighboring properties) even though understanding variations around those trends is potentially very informative about the mechanisms that drive land use and its implications.

That both residents and properties are highly heterogeneous is a challenge to the bidding assumptions made above regarding the determination of land prices. The representative monocentric model of [Section 8.2](#) and the assignment model developed above both assume that residents bid competitively for land (or properties) and that the process of allocating land is decided (implicitly) by a second-price auction. Heterogeneity in both supply and demand will make land and property markets thin, and the assumption of competitive bidding may no longer be warranted. The thinness of land and property markets is compounded by the fact that not everyone tries to buy and sell at the same time. This calls into question the static nature of the models used so far.

²⁶ Again, with only one dimension of heterogeneity as in the assignment model described above, extreme ordered sorting is expected in equilibrium. It is only when residents differ in two dimensions of heterogeneity that some mixing will occur. See [Behrens et al. \(2015\)](#) or [Epple and Platt \(1998\)](#) for different modeling approaches. See also [Davis and Dingel \(2013\)](#) for a model of income mixing across cities.

At this stage, we need to recognize that many land and property markets may be better characterized as search markets. In the simple case of a resident seeking to buy a starter home, this resident will first search for a number of properties. When he or she sees one that is “good enough,” he or she will view it as a potential match and make an offer. Sometimes there will be other potential buyers interested in the same property. More often perhaps, this prospective buyer will negotiate on a one-to-one basis with the seller to reach an agreement on the price before going forward with the transaction. This portrayal corresponds closely to the mechanics of the canonical job search model (Mortensen and Pissarides, 1994; Pissarides, 2000). More specifically, after the tags “properties,” “seller,” and “resident” have been replaced by “job,” “employer,” and “job seeker,” respectively, this stylized description of the housing market is the one usually applied to a job search and the formation of standard employment relationships. The main difference is that, for the labor markets, supply and demand are independent and employers can create jobs by investing, whereas for the housing market, sellers of houses are often also buyers.²⁷

The key model that describes the housing market as a two-sided search process is due to Wheaton (1990). Beyond being consistent with the simple depiction of the housing market in the previous paragraph, this model is useful in replicating a number of interesting stylized facts about property markets, such as the prevalence of bargaining and the existence of persistent vacancies with properties coming in and out of the market.

There has been a steady stream of research on searches in the housing market. This literature is discussed in Chapter 13 in this handbook. Because of the close resemblance of a housing search with many other forms of searches in economic life, the lessons of this broader literature are arguably relevant for housing (see Rogerson et al., 2005, for a review), keeping in mind nonetheless the substantive differences between housing and labor such as the buyer–seller problem mentioned above. There is little empirical work on housing searches that closely relates to theory. A first exception is Carrillo (2012), who calibrates a housing search model to infer some of its unobserved parameters. One can also cite recent work by Genesove and Han (2012a,b) or Merlo et al. (2013). A key reason behind the paucity of empirical work is that teasing out the empirical content of search models is notoriously difficult (Postel-Vinay and Robin, 2002). The last issue with this class of model is that, to our knowledge, no one has uncovered the implications of a housing search for land use.²⁸

²⁷ The joint buyer–seller problem is studied theoretically and empirically by Anenberg and Bayer (2013), who suggest that it may amplify housing market fluctuations both in prices and in volumes.

²⁸ For instance, there is only a small fraction of properties available on the market at any point in time. This implies that residents may be able to get the most suitable property on the market at the time of their search but not the overall best property for them. Hence, residential land might be misallocated because of search and relocation frictions. A similar misallocation will also occur with commercial properties. This may then affect productivity. In turn, this may prevent house builders from providing the optimal amount of heterogeneity in housing. More specifically, they will refrain from supplying more “extreme houses” since there may be no buyer for them at the time they try to sell them.

Another class of search models has implications for land use: labor search models. The reason is that depending on where an unemployed resident lives, searching for a job may be more or less costly. In a simple monocentric setting, imagine, for instance, that searching for work involves going to the CBD. Then, the cost of searching for a job increases with the distance to the CBD. This has a direct effect on the search effort of unemployed residents. In turn, this will affect residential patterns in cities and land development. It is also perhaps reasonable to assume that the distance to the CBD affects the efficiency of the job search process and not only its costs.

The first model of labor search with an explicit modeling of land is due to [Wasmer and Zenou \(2002\)](#). Further developments can be found in [Zenou \(2009\)](#). In the simplest model proposed by [Zenou \(2009\)](#), the unemployed locate either close to or far from jobs depending on the fundamental parameters that govern the functioning of the labor market. More specifically, if these parameters imply a labor market that is sufficiently tight, unemployed individuals (or workers with a greater propensity of becoming unemployed) will prefer to live close to the CBD. This proximity to the CBD will lower their search costs, facilitating a more intense search, which will be rewarded with a job more often. Instead, if the labor market is not sufficiently tight and unemployment is high in equilibrium, the incentives to search intensively are muted, and unemployed individuals will prefer to live further from the CBD. In turn, this greater distance to jobs will weaken their incentives to search for jobs.²⁹

To sum up, there are several take-away points that emerge from this discussion of heterogeneity in land use models. First, the exploration of several dimensions of heterogeneity is still in its infancy. This is true in particular for the heterogeneity of parcels and for the assignment of heterogeneous parcels to heterogeneous residents. Progress will be slow because this type of work is often technically challenging and relies on techniques that are new to the field. Next, a key lesson from existing work is that for many first-order questions such as the location choices of different socioeconomic groups, the results are sometimes sensitive to fine details about the assumptions or to some key parameter values. Allowing for heterogeneous residents often modifies the results or even, sometimes, changes their qualitative nature. Recall, for instance, that all the results for the proportionalities of the urban aggregates no longer hold with heterogeneous residents or that the land price gradients no longer depend on commuting costs in simple models of assignments with a continuous distribution of income. Finally, and most importantly, despite the aforementioned changes to secondary results, all the key insights in [Section 8.2](#) still hold with heterogeneous residents. At the individual level, all the key trade-offs remain qualitatively the same, and an appropriately redefined Alonso–Muth condition still

²⁹ Although it may sound from this description that multiple equilibria are possible with a high unemployment-peripheral unemployed configuration versus a low unemployment-central unemployed configuration, this is not the case in the model of [Wasmer and Zenou \(2002\)](#), where the equilibrium is unique.

governs location choices and the gradients for land prices, housing prices, the intensity of development, population density, and parcel sizes.

8.3.3 Durable housing

When we do comparative statics in the monocentric model, we are implicitly letting the city be completely rebuilt from scratch to fit the new conditions. This is often seen as a reasonable simplification because the comparative statics fit well with comparisons of actual cities built under different conditions. However, some important details of the standard monocentric model conflict with reality. For instance, the model predicts that cities will be built contiguously and with building heights monotonically decreasing from the center. In practice, however, we see some centrally located parcels left vacant while others further out are developed, and building heights can both increase and decrease as we travel outward from the CBD. Replacing the assumption that housing is completely malleable with the more realistic assumption that housing is durable helps address these shortcomings. In addition, it provides useful additional insights into how cities react to changing conditions.³⁰

Once we acknowledge the durability of housing, we must take into account that housing developers will recover their investment over an extended period of time. Then it becomes important to consider how they form their expectations about the future evolution of prices. A simple possibility is that developers have myopic expectations, as in [Anas \(1978\)](#), behaving as if current conditions will last forever. Then, a developer who owns a parcel of land at a distance x from the CBD will develop it at time T if and only if

$$\int_T^\infty R_T(x)e^{-rt}dt = \int_T^\infty P_T(x)f(x)e^{-rt}dt - \int_T^\infty rk_T(x)e^{-rt}dt \geq \int_T^\infty \underline{R}_Te^{-rt}dt. \quad (8.37)$$

The return, $R_T(x)$, that the developer expects to obtain from a parcel of land at a distance x from the CBD from time T onward is the difference between the expected present value of the rent from the $f(x)$ units of housing floor space developed on the parcel of land and the expected present value of the cost at a constant interest rate r of the capital $k_T(x)$ used in the development (note that, by Shephard's lemma, $k_T(x) = \frac{\partial c(R_T(x), r)}{\partial r}$, which varies with $R_T(x)$). For land to be developed, this return must be greater than the expected present value of the agricultural rent, $\underline{R}_T$. Owing to myopic expectations, all variables have time subindex T . Integrating Equation (8.37) and simplifying the result implies that, at time T , the edge of the city is still given by the same condition as in the standard static monocentric model:

³⁰ In this section, we review briefly extensions to the monocentric model featuring durable housing. For additional details, see the survey by [Brueckner \(2000\)](#).

$$R_T(\bar{x}) = P_T(x) - rk_T(x) = \underline{R}_T, \quad (8.38)$$

which is the same as Equation (8.20) of the standard model but with a time subindex T . Consequently, a city that grows over time has contiguous development, as in the static monocentric model. The key difference is that the capital intensity of development $k_T(x)$ at each point x reflects the conditions at the time T when the city edge was at this point instead of current conditions. Several situations may occur, with building heights and population density both decreasing (as in the static model), staying constant, or increasing with distance from the CBD. To understand under which conditions these outcomes occur, recall that the optimal intensity of development at the city edge is given by $k_T(\bar{x}) = \frac{\partial c(R_T(\bar{x}), r)}{\partial r} = \frac{\partial c(\underline{R}_T, r)}{\partial r}$ after using Equation (8.38). Hence, unless there is a change in the interest rate or in the agricultural land rent, new development at the urban edge driven by higher wages or lower commuting costs will occur at the same level of capital intensity. A combination of higher wages (for the city to expand) and a lower interest rate will lead to more capital-intensive development at edge of the city. A combination of higher wages (for the city to expand) and higher interest rates will, on the other hand, lead to less capital-intensive new development at the urban edge.

The assumption of myopic foresight, while convenient, is not very satisfactory. For instance, in a city that is growing gradually over time, one would expect developers to take this growth trajectory into account. Developers may be able to predict the future quite well. Taken to the extreme, this implies assuming that developers have perfect foresight. A tractable monocentric model with irreversible development and perfect foresight is that of Capozza and Helsley (1989), who assume for simplicity that dwelling size and floor space per unit of land are both equal to unity: $h_t(x) = f_t(x) = 1$. They also assume that conversion of one unit of land from nonurban use to urban use involves a fixed amount of capital k instead of an endogenously chosen amount $k_t(x)$. Denoting by T the endogenous date at which a parcel located at a distance x from the CBD will be converted to urban use, we can express the expected present value of returns for a developer who owns that parcel as

$$\int_0^T \underline{R}_t e^{-rt} dt + \int_T^\infty P_t(x) e^{-rt} dt - k e^{-rT}. \quad (8.39)$$

The first term in Equation (8.39) is the land rent obtained up until time T while the parcel is still used for agriculture. The second term is the rent obtained from the development at time T onward for the unit of housing built on the land. The third term is the present value of the conversion cost. Note that, unlike in the case of myopic expectations, developers fully anticipate the evolution of rents of agricultural land and housing, $\underline{R}_t$ and $P_t(x)$. The first-order condition for the developer can be found by differentiating Equation (8.39) with respect to T and equating it to zero, implying

$$P_T(x) - rk = \underline{R}_T. \quad (8.40)$$

Note this condition is the same as Equation (8.38) for the case of myopic foresight, so land is still developed when the urban land rent minus the cost of conversion from agricultural to residential use equals the agricultural land rent. However, perfect foresight introduces an important difference: the price of land is no longer proportional to the rent. Instead, the price of land is equal to the maximized value of Equation (8.39). As a result, in a growing city, land beyond the city edge is priced above the expected present value of the agricultural rent, reflecting the anticipation of its conversion to urban use. This model also generates a price gradient with higher prices for undeveloped land as we approach the city edge from outside the developed area.

If we let the structural characteristics of housing change endogenously, the dynamic monocentric model with perfect foresight can generate leapfrog development, where parcels are not developed contiguously starting from the CBD (Fujita, 1982; Wheaton, 1982; Turnbull, 1988). Letting the structural characteristics of housing vary introduces a second choice variable in the expected present value of the developer's return relative to Equation (8.39). This leads to an additional first-order condition for optimal structural characteristics of development in addition to (8.40), which regards the optimal timing of development. Since for any given time T these two first-order conditions can intersect for more than one value of x , it follows that parcels of land separate from each other can be developed simultaneously, while some parcels in between may be left vacant.

In this section, we have so far assumed that developers either have no anticipation of the future and expect current conditions to last forever (the myopic foresight case) or can predict the future precisely (the perfect foresight case). A more realistic case sits in between these two, with developers anticipating the future but realizing that there is uncertainty about the evolution of rents. Capozza and Helsley (1990) explore such a monocentric model with irreversible development under uncertainty. As in Capozza and Helsley (1989), they fix the dwelling size and floor space per unit of land ($h_t(x) = f_t(x) = 1$) so that the only choice variable for the developer is when to convert land from agricultural to urban use. Conversion again involves a constant capital cost k . They focus on an open city within a large urban system where the ongoing level of utility is constant at $\underline{u}$. With every resident consuming one unit of housing built on one unit of land, this implies a constant level of consumption of the numéraire $\underline{z}$. Income w for every resident in the city is assumed to vary. More specifically, it follows a Brownian motion with drift $g > 0$ and variance σ^2 . When income rises in the city, this makes the city more attractive relative to other cities, bringing in more residents, which raises land rents until utility is restored to level $\underline{u}$. When income falls, the city loses population and land rents fall. From Equations (8.16) and (8.18) with $h_t(x) = 1$ and $\frac{\partial c(R_t(x))}{\partial R_t(x)} = 1$, the bid-rent curve is linear: $\frac{dP_t(x)}{dx} = \frac{dR_t(x)}{dx} = -\tau$. Hence,

$$P_t(x) = w - \underline{z} - \tau x . \quad (8.41)$$

Capozza and Helsley (1990) show that the optimal conversion date T for the developer owning a parcel at a distance x from the CBD satisfies

$$P_T(x) - rk = \underline{R}_T + \frac{r - \psi g}{\psi r}, \quad (8.42)$$

where

$$\psi \equiv \frac{\sqrt{g^2 + 2\sigma^2 r} - g}{\sigma^2} \leq \frac{r}{g}. \quad (8.43)$$

Comparison of Equation (8.42), for irreversible development under uncertainty, and Equation (8.40), for irreversible development under perfect foresight, shows that developers require higher urban rents to develop land under uncertainty. This implies that, in a growing city, uncertainty delays urban development. The reason is that a developer who converts land to urban use may be surprised by lower rents than expected and come to regret the conversion. As in the case of perfect foresight, land beyond the city edge is priced above the expected present value of the agricultural rent. This now reflects not just the anticipation of its conversion to urban use as a result of expected urban growth (as was the case with perfect foresight) but also an additional option value of agricultural land arising from the uncertainty about future urban rents. Holding a parcel of agricultural land implies holding an option to convert it to urban use. Urban development is equivalent to exercising that option.

A key feature of land development is the existence of significant lags between the time when a development project is decided on and the time when new floor space can be put on the market. These lags are caused by construction and the time it takes to obtain a building permit. They imply that a developer will face a potentially very different rent relative to the one that prevailed when the project was decided on. In addition, developers also have the option to cut their losses and stop a project should the circumstances become too unfavorable. Bar-Ilan and Strange (1996) extend the model of Capozza and Helsley (1990) to account for these two features. The main result is that development lags reduce the amount of delay. As in Capozza and Helsley (1990), developers still want to delay their investment for fear of lower rents in the future. However, development lags make the opportunity cost of a delay more expensive when rents are high. The option of stopping a project also puts a floor on the returns from a new development. In turn, this makes returns from development convex in rents and leads the value of land development to increase with uncertainty. Since rents are lower further away from the CBD, the same income uncertainty results in greater rent uncertainty in less central locations. In the presence of long development lags, developers may prefer to convert to urban use land that is further away, leaving more central locations undeveloped. This provides a motivation for patterns of leapfrog development.

The assumption that development is irreversible is justified because buildings are long-lived, and once it has been converted to urban use, land very rarely reverts to nonurban use (Burchfield et al., 2006). Nevertheless, buildings deteriorate over time and require periodic maintenance, and while no development is rare, redevelopment is very common. It is therefore important to consider not just developers' decisions about the initial development but also their decisions about redevelopment. This can again be done under different scenarios regarding developers' expectations. Brueckner (1980) studies the case of myopic foresight. The inconsistency of developers' behavior is enhanced by redevelopment because now myopic developers do not just assume that current rent levels will persist forever, they also decide whether to redevelop now while ignoring that they will again be deciding whether to redevelop in the future. Nevertheless, studying redevelopment with myopic developers is a useful exercise because some key conclusions carry over to more sophisticated treatments of developers' expectations. In a growing city, land is still initially developed as in the case of the myopic irreversible development of Anas (1978) when the value of land in urban use minus the conversion costs equals the value of land in nonurban use. The main difference is that there will be periodic redevelopment. Consider, for instance, a city where the agricultural land rent is constant at $\underline{R}$ but wages keep rising over time. Wage growth causes the city to expand outward. However, since developers are myopic, they expect the current state of the city to be the permanent state, so they keep building at the edge of the city with the same constant capital intensity $k_t(x) = \frac{\partial c(R_t(\bar{x}), r)}{\partial r} = \frac{\partial c(\underline{R}, r)}{\partial r}$. At the same time, they may find it optimal to compensate for their past myopic behavior by redeveloping land closer to the CBD at a higher intensity. Thus, close to the city center there will be some tall new buildings standing next to shorter older buildings. If we isolate structures built around the same time, more central buildings are taller than those in the periphery, as in the standard static monocentric model. However, since many generations of building coexist, the overall pattern is one of sawtooth building heights.

Combining perfect foresight and redevelopment greatly complicates the developer's problem. A possible solution is to focus on a city in a stationary state, as in Brueckner (1981) and Arnott et al. (1983). Alternatively, Braid (2001) studies a city in a nonstationary state by using specific functional forms that make the developer's problem at any distance from the CBD a simple transformation of the problem at unit distance.

Housing redevelopment has two important components: the deterioration of existing structures over time and their eventual replacement with new structures. Even in the absence of deterioration, changing conditions could be sufficient to generate redevelopment. Deterioration strengthens the incentives for redevelopment and also raises additional issues—for instance, enriching the analysis of income sorting we considered in Section 8.3.2. A developer may initially build housing at a high quality level to target high-income residents. However, as the passing of time deteriorates the building and

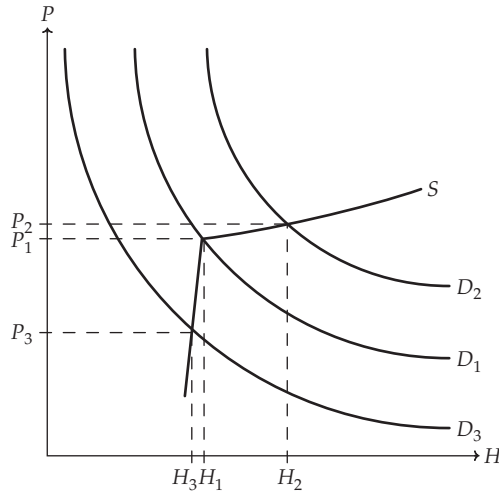


Figure 8.3 The asymmetry between growing and declining cities.

some of its features become obsolete, the effective level of quality it provides falls. This may lead its residents to move to a higher-quality dwelling, while lower-income residents move in to replace them. This filtering process is studied formally by [Sweeney \(1974a,b\)](#) and is thought to be crucial for the provision of housing to low-income groups. [Rosenthal \(2014\)](#) examines the filtering process empirically and shows that it is amplified by dwellings shifting over time from being owner-occupied to being rented, with the filtering process accelerating once units are rented. Combined with the amplifying effect on filtering from a low income elasticity of housing, [Rosenthal \(2014\)](#) concludes that filtering can be quite effective in providing suitable low-income housing. [Brueckner and Rosenthal \(2009\)](#) develop a model where filtering leads to cyclical changes in the location of higher-income households. In their model, dwellings have a fixed time span, during which fixed-size dwellings provide housing services that decline with their age. Thus, residents who wish to consume more housing must do so by residing in newer dwellings. In a growing city, locations close to the CBD are developed first. As they age, high-income households relocate toward the city edge to consume more housing. Eventually, central dwellings reach the end of their life and are replaced by new structures, prompting high-income households to relocate toward the city center away from now ageing suburban dwellings.³¹

One of the main implications of the durability of housing is that there are important asymmetries between growing cities and declining cities. This point, on which [Figure 8.3](#) is based, is made by [Glaeser and Gyourko \(2005\)](#). The figure represents housing supply

³¹ See Chapter 16 in this handbook for more on filtering in the housing market.

and demand in a city, with the price of housing P measured on the vertical axis and the quantity of housing H measured on the horizontal axis. Suppose that initially the city has a housing stock H_1 priced at P_1 per unit of floor space. If the city experiences a positive housing demand shock that moves the demand curve upward from D_1 to D_2 , then, provided that suitable land is available and new construction is not significantly constrained by regulation and zoning, additional housing will be built, increasing the stock from H_1 to H_2 possibly with a small price increase from P_1 to P_2 .

Consider now that the city experiences a negative housing demand shock. This shock moves the demand curve downward from D_1 to D_3 . Then, because housing is durable, existing housing will remain in place. At most, housing may deteriorate owing to lack of maintenance, but this will reduce the housing stock only very slowly. Thus, a negative demand shock will be almost completely reflected in a sharp drop in prices from P_1 to P_3 , with almost no change in quantities other than depreciation.

Putting everything together, we find the key is that housing supply will be kinked at the level of the current housing stock, being relatively elastic above this current level and very inelastic below it. This has several implications. First, as shown in [Figure 8.3](#), positive shocks increase the population more than they increase housing prices, while negative shocks decrease housing prices more than they decrease the population. Second, cities grow more quickly than they decline because the durability of existing housing slows the fall in the population. Third, the abundance of cheap housing priced well below construction costs is a sign of large negative demand shocks that, since the population in declining cities falls slowly, can be taken as a sign of future decline. [Glaeser and Gyourko \(2005\)](#) show that these features hold empirically in the United States.

8.4. AGGLOMERATION AND COMMERCIAL LAND USE: MODELING POLYCENTRIC CITIES

The monocentric model is a remarkable achievement and a very useful stylized representation of cities. However, two aspects of the model are in particular tension with important empirical facts about cities. First, the monocentric city model explains patterns of residential land use and commuting within a city. However, it does not explain why individuals wish to be in a city to start with. If we treat the wage w as a parameter independent of a city's population in Equation (8.24), then $\frac{dv}{dN} < 0$. This implies that any individual prefers to live alone than to live in a city of any size. To explain why cities exist at all, we must introduce agglomeration economies. This is particularly important given the growing evidence about the importance of such agglomeration economies.³² A simple way to incorporate agglomeration economies into the monocentric model is to have the wage depend positively on the city's population: $w = w(N)$, with $\frac{dw}{dN} > 0$. This leads to a trade-off in Equation (8.24) between urban costs ($-\frac{dv}{dP(0)} \frac{dP(0)}{dN} < 0$) and stronger agglomeration

³² They are reviewed in [Chapter 5](#) in this volume.

economies ($\frac{dv}{dw} \frac{dw}{dN} > 0$). For a more detailed exposition of the implications of introducing agglomeration economies in the monocentric city model, including the modeling of systems of cities and the analysis of systematic and stochastic determinants of city growth, we refer the reader to [Duranton and Puga \(2014\)](#) and [Chapter 4](#) in this volume.

A second aspect of the monocentric city model that appears at odds with modern cities is precisely its monocentric structure: in the monocentric model described in [Section 8.2](#), firms do not use any land and locate, by assumption, at a single central point. In reality, firms use land as an input, and the division of land between residential and commercial uses within a city follows complex patterns. In particular, as we discuss in [Section 8.9](#), land use patterns in actual cities are far less extreme than are assumed by the monocentric city model. In 1996, only about 25% of employees in US metropolitan areas worked within 5 km of their CBD ([Glaeser and Kahn, 2001](#)). Increasingly, secondary employment centers have emerged in metropolitan areas, and they have absorbed a growing number of jobs ([Anas et al., 1998](#)).

Extending the monocentric model so that land is used in production is straightforward and simply involves modeling the CBD as a segment or a disk instead of as a point. Having more than one employment center or area also does not change the basic mechanics of the model either, provided the location of secondary centers is exogenously given (see [White, 1976](#), for an early example). A simple way to endogenize the location of secondary centers is to take a central city and its capacity as given and examine the problem of a developer who sets up an edge city, choosing its distance with respect to the central city and its capacity. [Henderson and Mitra \(1996\)](#) examine this problem and highlight the trade-off faced by the edge-city developer: locating the secondary center further away from the central city alleviates competition for land and lowers costs, but it also weakens productivity spillovers between the central city and the edge city. The weakening of spillovers reduces the productivity of the edge city but, by also reducing the productivity of the central city, strengthens the developer's monopsony power.

A more difficult problem is to endogenize the location of both firms and workers throughout the city, with areas emerging endogenously with commercial, residential, or mixed land use as a result of the interactions of the location decisions of all agents directly with each other and through land markets. This problem was first tackled independently by [Ogawa and Fujita \(1980\)](#) and [Imai \(1982\)](#) in a framework where firms benefit from proximity to each other due to communication externalities that decay linearly with distance. [Fujita and Ogawa \(1982\)](#) studied a case with exponential decay instead of linear decay in externalities. This was revived and generalized by [Lucas and Rossi-Hansberg \(2002\)](#). We now describe a simple version of the [Ogawa and Fujita \(1980\)](#) and [Imai \(1982\)](#) framework. For a neat exposition of the [Fujita and Ogawa \(1982\)](#) framework, see [Fujita and Thisse \(2013\)](#).

The city occupies a segment of endogenous length on the real line with one unit of land available at each location x . Denote by $m(x)$ the endogenous density of firms and

by $n(x)$ the endogenous density of residents at location x . In equilibrium, there can be areas with mixed land use ($m(x) > 0$ and $n(x) > 0$), areas with only commercial development ($m(x) > 0$ and $n(x) = 0$), and areas with only residential development ($m(x) = 0$ and $n(x) > 0$).

Agglomeration economies arise owing to spillovers that raise a firm's productivity when its workers are able to interact more closely with other workers in the city. In particular, suppose that, using one unit of labor and λ units of land, each firm produces one unit of output for every unit of communication spillovers involving its workers. Hence its cost function is $(w(x) + \lambda P(x))/A(x)$ where $A(x)$ denotes communication spillovers. In turn, communication spillovers between workers depends on how far their jobs are located, starting at β units of communication spillovers for workers employed at the same location and decreasing at a rate γ per unit of distance between their job locations. The output of a firm choosing to locate at x then depends on the location of all other firms as follows:³³

$$A(x) = \int_{-\infty}^{\infty} (\beta - \gamma|x - y|)m(y)dy. \quad (8.44)$$

Differentiation of Equation (8.44) yields $\frac{dA(x)}{dx} = -\gamma(\int_{-\infty}^x m(y)dy - \int_x^{\infty} m(y)dy)$ and $\frac{d^2A(x)}{dx^2} = -2\gamma m(x)$. This implies that $A(x)$ reaches a global maximum at the point in the city where half the firms are located to its left and half the firms are located to its right. Without loss of generality, assign coordinate $x = 0$ to this point, so that $A(x)$ increases with x for $x < 0$ and decreases with x for $x > 0$. Furthermore, $A(x)$ is a concave function of x in areas wherever there is commercial development ($m(x) > 0$) and a linear function of x wherever there is no commercial development.

Free entry of firms exhausts their profits. The bid-rent function for commercial land $\Phi(x)$ is the maximum price a firm can pay for land at each location x while making zero profit:

$$\Phi(x) = \frac{1}{\lambda}[A(x) - w(x)]. \quad (8.45)$$

In the standard monocentric model each worker commutes from his or her location x to the exogenous CBD located at 0. Now, instead, a worker residing at x chooses the work location that best suits him or her. We maintain the assumption of commuting costs increasing linearly with distance at a rate τ . Let $T(x)$ denote the utility-maximizing job location of a worker as a function of his or her residential location x :

$$T(x) \equiv \arg \max_y \{w(y) - \tau|x - y|\}. \quad (8.46)$$

³³ β is assumed to be large enough that $A(x)$ does not end up being negative.

Thus, for a given residential location, workers choose their job location by trading off wages against commuting costs.

Let us simplify the residential location problem by assuming that all residences have the same unit size. Then, maximizing utility $u(1, z)$ subject to the budget constraint $w(x) - \tau|x - T(x)| = P(x) \times 1 + z(x)$ is equivalent to maximizing consumption of the numéraire $z(x) = w(x) - \tau|x - T(x)| - P(x)$. Suppose all houses are built with one unit of land and a fixed amount of capital. To avoid carrying around additional constants, we set both the constant cost of capital throughout the city and the price of land in agriculture to zero, so that the price of housing and the price of land coincide instead of differing by a constant and so that the price of land at the city edge is zero: $P(x) = R(x)$ and $R(\bar{x}) = 0$. The bid-rent function for housing and for residential land $\Psi(x, \underline{u})$ is the maximum price a resident can pay for housing at each location x while consuming the amount of numéraire $z(\underline{u})$ that allows him or her to enjoy utility $\underline{u}$ and while also satisfying the budget constraint:

$$\Psi(x, \underline{u}) = w(T(x)) - \tau|x - T(x)| - z(\underline{u}) . \quad (8.47)$$

Land will be allocated to the highest bidder. This implies that the rental price of land is given by

$$R(x) = \max(\Phi(x), \Psi(x, \underline{u})) , \quad (8.48)$$

$$R(x) = \Phi(x) \text{ if } m(x) > 0 , \quad (8.49)$$

$$R(x) = \Psi(x, \underline{u}) \text{ if } n(x) > 0 . \quad (8.50)$$

In turn, land use is described by

$$\lambda m(x) + n(x) = 1 \text{ if } R(x) \geq 0 , \quad (8.51)$$

$$m(x) = n(x) = 0 \text{ if } R(x) < 0 . \quad (8.52)$$

Labor market clearing implies

$$\int_X n(x) dx = \int_{T(X)} m(x) dx , \quad (8.53)$$

for every interval X . Finally, we must consider the aggregate population constraint,

$$\int_{-\infty}^{\infty} n(x) dx = N , \quad (8.54)$$

and the aggregate firm constraint,

$$\int_{-\infty}^{\infty} m(x) dx = N . \quad (8.55)$$

Equations (8.44)–(8.55) are the equilibrium conditions of this framework.

The general form of the equilibrium is the following. Patterns of land use are symmetric around the point $x = 0$, which splits the distribution of firms into halves. There is a central area of mixed land use centered at $x = 0$ and extending from $-x_0$ until x_0 , where firms and residences coexist continuously. Beyond this central mixed-use area, and at both sides of it, there are areas of pure commercial land use, extending from $-x_1$ to $-x_0$ and from x_0 to x_1 . Finally, beyond the commercial areas there are areas of pure residential land use extending from $-\bar{x}$ to $-x_1$ and from x_1 to $\bar{x}$. Workers living in the mixed-use area work where they live, whereas workers living in the purely residential area commute to work in the purely commercial area. These patterns of land use can be expressed more formally as

$$m(x) = \begin{cases} \frac{1}{1+\lambda} & x \in [-x_0, x_0] , \\ \frac{1}{\lambda} & x \in [-x_1, -x_0] \cup [x_0, x_1] , \\ 0 & x \in [-\bar{x}, -x_1] \cup [x_1, \bar{x}] , \end{cases} \quad (8.56)$$

$$n(x) = \begin{cases} \frac{1}{1+\lambda} & x \in [-x_0, x_0] , \\ 0 & x \in [-x_1, -x_0] \cup [x_0, x_1] , \\ 1 & x \in [-\bar{x}, -x_1] \cup [x_1, \bar{x}] , \end{cases} \quad (8.57)$$

where the densities in each interval follow from the above description and Equations (8.51) and (8.53). Given that each firm uses λ units of land and each resident uses one unit, the shares of commercial and residential land are $\lambda m(x)$ and $n(x)$, respectively. Panel (a) in Figure 8.4 plots the share of commercial land in the equilibrium we have just described. We now show that such an equilibrium exists and derive the equilibrium values of the communication spillovers $A(x)$, wages $w(x)$, and land prices $R(x)$, as well as the values of x_0 , x_1 , and $\bar{x}$.

Substituting Equation (8.56) into (8.44) yields the equilibrium value of the communication spillovers at each point in the city:

$$A(x) = \begin{cases} \beta N - \gamma \left(\frac{1}{\lambda} x_1^2 - \frac{1}{\lambda(1+\lambda)} x_0^2 + \frac{1}{1+\lambda} x^2 \right), & x \in [-x_0, x_0] , \\ \beta N - \gamma \left(\frac{1}{\lambda} x_1^2 - \frac{2}{\lambda(1+\lambda)} x_0 |x| + \frac{1}{\lambda} x^2 \right), & x \in [-x_1, -x_0] \cup [x_0, x_1] , \\ \beta N - \gamma \left(\frac{2}{\lambda} x_1 - \frac{2}{\lambda(1+\lambda)} x_0 \right) |x|, & x \in [-\bar{x}, -x_1] \cup [x_1, \bar{x}] . \end{cases} \quad (8.58)$$

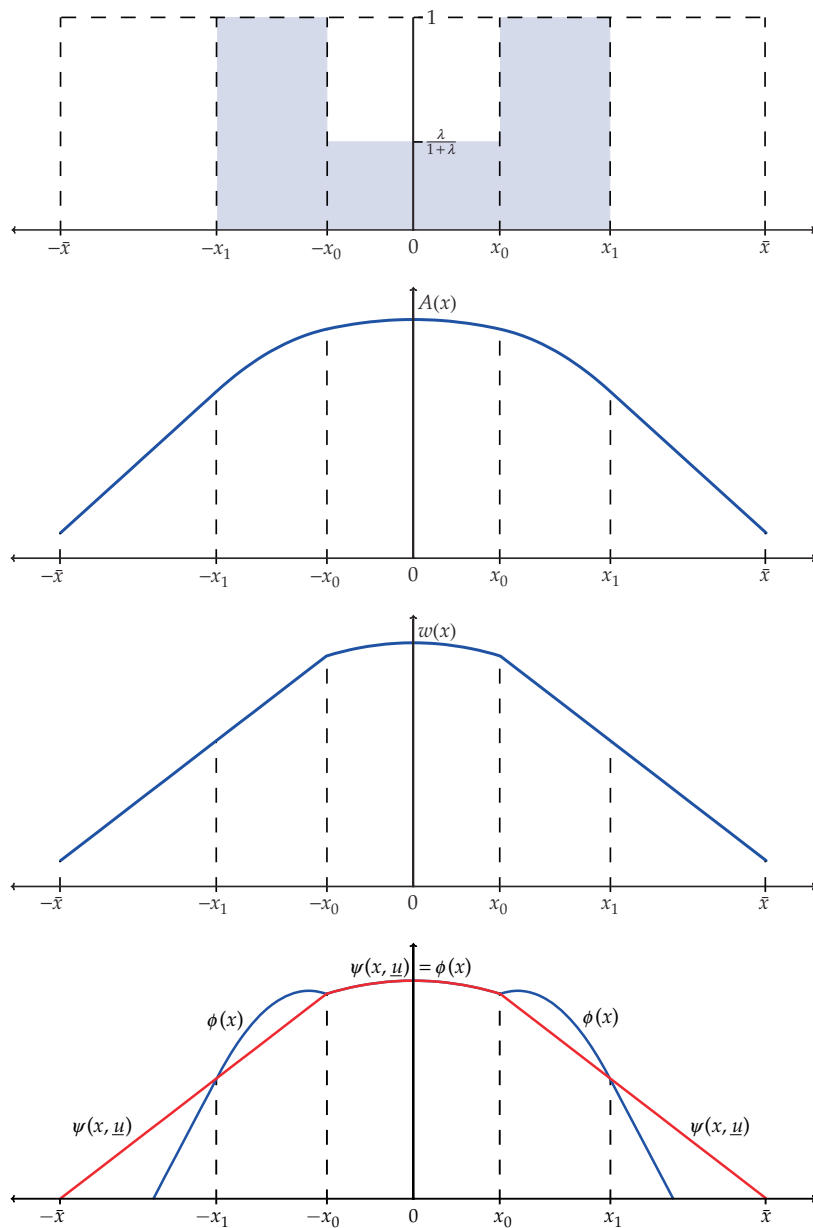


Figure 8.4 Equilibrium land use patterns and gradients in Ogawa and Fujita (1980). Panel (a) shows the share of land in commercial use. Panel (b) shows spillovers. Panel (c) shows wages. Panel (d) shows bid-rent gradients.

Panel (b) in Figure 8.4 plots the equilibrium spillovers $A(x)$. As shown above, $A(x)$ is concave over the interval where firms are located, $x \in [-x_1, x_1]$, with a maximum at the (endogenous) center of the city and it linearly decreases with the distance to the center over the interval where all land is in residential use, $x \in [-\bar{x}, -x_1] \cup [x_1, \bar{x}]$. It also follows immediately from Equation (8.58) that $A(x)$ is continuous in x and so is its first derivative.

Turning to wages, consider two points x and y with firms to which workers are commuting. Then, by the definition of the commuting cost function in Equation (8.46), $w(x) - \tau|T^{-1}(x) - x| \geq w(y) - \tau|T^{-1}(x) - y|$ and $w(y) - \tau|T^{-1}(y) - y| \geq w(x) - \tau|T^{-1}(y) - x|$. These two inequalities together imply $w(x) - w(y) = -\tau(x - y)$. Equivalently, $\frac{dw(x)}{dx} = -\tau$ for any $x > 0$ such that $m(x) > 0$ and $T^{-1}(x) \neq x$ and $\frac{dw(x)}{dx} = \tau$ for any $x < 0$ such that $m(x) > 0$ and $T^{-1}(x) \neq x$. Thus, in any area with firms to which workers are commuting, the wage is a linear function of x decreasing as one moves away from $x = 0$ with a slope equal in absolute value to the commuting costs parameter τ . Consider now firms whose workers instead do not need to commute because their workers reside at the same location—that is, where $T(x) = x$. By Equations (8.48)–(8.50), (8.56), and (8.57), $\Psi(x) = \Phi(x)$ for $x \in [-x_0, x_0]$. Substituting Equations (8.45) and (8.47) and $T(x) = x$ into this equality turns it into $w(x) - z(\underline{u}) = \frac{1}{\lambda}[A(x) - w(x)]$, which we can solve for $w(x)$. Thus, the wage gradient in the city is given by

$$w(x) = \begin{cases} \frac{1}{1+\lambda}A(x) + \frac{\lambda}{1+\lambda}z(\underline{u}), & x \in [-x_0, x_0], \\ w(x_0) - \tau(|x| - x_0), & x \in [-\bar{x}, -x_0] \cup [x_0, \bar{x}]. \end{cases} \quad (8.59)$$

Panel (c) in Figure 8.4 plots the equilibrium wage $w(x)$. Note from Equation (8.59) that $w(x)$ is continuous in x . Also, given that $A(x)$ is a concave function of x for $x \in [-x_1, x_1]$ with a maximum at $x = 0$, it follows from Equation (8.59) that the wage $w(x)$ is a concave function of x for $x \in [-x_0, x_0]$ with a maximum at $x = 0$. Outside this central area of mixed land use, the wage decreases linearly with x as one moves away from the center with a slope equal in absolute value to the commuting costs parameter τ .

Consider next the bid-rent functions for land. On the residential side, substituting Equation (8.59) into (8.47) and using $T(x) = x$ for $x \in [-x_0, x_0]$ yields

$$\Psi(x, \underline{u}) = \begin{cases} \frac{1}{1+\lambda}[A(x) - z(\underline{u})], & x \in [-x_0, x_0], \\ \Psi(x_0, \underline{u}) - \tau(|x| - x_0), & x \in [-\bar{x}, -x_0] \cup [x_0, \bar{x}]. \end{cases} \quad (8.60)$$

Note that $\Psi(x, \underline{u})$ is a continuous function of x . Also, it follows from Equation (8.60) and the concavity of $A(x)$ for $x \in [-x_1, x_1]$ that, like the wage, the bid-rent function for land in residential use $\Psi(x, \underline{u})$ is also a concave function of x for $x \in [-x_0, x_0]$ with a maximum

at $x = 0$. Within the mixed-use area, workers living and working closer to the center obtain a higher wage, but this is exactly offset by a higher price for their residence. Outside this central area of mixed land use, $\Psi(x, \underline{u})$ decreases linearly with x as one moves away from the center with a slope equal in absolute value to the commuting costs parameter τ . Within the purely residential area, workers living closer to the center are able to obtain a higher wage at any given commuting distance from their home, but this higher wage is exactly offset by a higher price for their residence.

Regarding the bid-rent function for land in commercial use, substituting Equation (8.59) into (8.45) results in

$$\Phi(x) = \begin{cases} \frac{1}{1+\lambda} [A(x) - z(\underline{u})], & x \in [-x_0, x_0], \\ \Phi(x_0) + \frac{1}{\lambda} [A(x) - A(x_0)] + \frac{\tau}{\lambda} (|x| - x_0), & x \in [-\bar{x}, -x_0] \cup [x_0, \bar{x}]. \end{cases} \quad (8.61)$$

Note that $\Phi(x)$ is also a continuous function of x . From Equations (8.58) and (8.61), $\Phi(x)$ is a concave function of x for $x \in [-x_0, x_0]$ and also for $x \in [-x_1, -x_0] \cup [x_0, x_1]$, although its slope changes discretely at $-x_0$ and x_0 . Beyond x_1 and $-x_1$, $\Phi(x)$ becomes a linear function of x , maintaining the same slope it has at x_1 and $-x_1$, respectively.

Panel (d) in Figure 8.4 plots the bid-rent functions for land in residential use and in commercial use, $\Psi(x, \underline{u})$ and $\Phi(x)$. For $x \in [-x_0, x_0]$, they both coincide and land is in mixed use. For $x \in [-x_1, -x_0] \cup [x_0, x_1]$, firms bid for land more than residents and land is in purely commercial use. Finally, for $x \in [-\bar{x}, -x_1] \cup [x_1, \bar{x}]$, residents bid for land more than firms and land is in purely residential use.

The edge of the city can be obtained by integrating both sides of (8.51) between $-\bar{x}$ and $\bar{x}$ and using the aggregate constraints (8.54) and (8.55):

$$\bar{x} = \frac{1+\lambda}{2} N. \quad (8.62)$$

For the land use patterns of Equations (8.56) and (8.57) to be consistent with the equilibrium conditions (8.48)–(8.51), we must have

$$R(x) = \Phi(x) = \Psi(x, \underline{u}), \quad x \in [-x_0, x_0], \quad (8.63)$$

$$R(x) = \Phi(x) \geq \Psi(x, \underline{u}), \quad x \in [-x_1, -x_0] \cup [x_0, x_1], \quad (8.64)$$

$$R(x) = \Psi(x, \underline{u}) \geq \Phi(x), \quad x \in [-\bar{x}, -x_1] \cup [x_1, \bar{x}], \quad (8.65)$$

$$R(\bar{x}) = \Psi(\bar{x}, \underline{u}) = 0. \quad (8.66)$$

Equations (8.64) and (8.65) and the continuity of $\Psi(x, \underline{u})$ and $\Phi(x)$ imply $\Phi(x_1) = \Psi(x_1, \underline{u})$. Substituting Equations (8.60) and (8.61) into this equality yields $A(x_0) - A(x_1) = (1 + \lambda)\tau(x_1 - x_0)$. Substituting (8.58) into this gives a first equation linking x_0 and x_1 :

$$\frac{\gamma}{\lambda} \left(x_1^2 - x_0^2 - \frac{2}{1+\lambda} x_0(x_1 - x_0) \right) = (1+\lambda)\tau(x_1 - x_0). \quad (8.67)$$

A second equation linking x_0 and x_1 is obtained by substituting Equations (8.56) and (8.57) into the aggregate firm constraint (8.55):

$$2 \left(\frac{1}{1+\lambda} x_0 + \frac{1}{\lambda} (x_1 - x_0) \right) = N. \quad (8.68)$$

Equations (8.67) and (8.68) have two solutions. The first solution has the entire city under mixed land use: $x_0 = x_1 = \frac{1+\lambda}{2}N = \bar{x}$. For this solution to be an equilibrium, we must make sure that starting from such a configuration, a firm at the edge of the city is not willing to outbid workers for land in order to concentrate its production, thus pushing workers out into a purely residential area. That is, the bid-rent gradient for firms cannot be steeper (have a more negative slope) at $x = x_1$ when $x_0 = x_1 = \bar{x}$. Differentiating Equations (8.60) and (8.61), we can express the condition $\frac{d\Phi(x)}{dx}|_{x=x_1} \geq \frac{d\Psi(x,u)}{dx}|_{x=x_1}$ as $\frac{2\gamma}{\lambda} (x_1 - \frac{1}{1+\lambda}x_0) \leq \tau(1+\lambda)$. Valuing this at $x_0 = x_1 = \bar{x} = \frac{1+\lambda}{2}$, we find the condition for a fully integrated equilibrium becomes $N \leq \frac{\tau(1+\lambda)}{\gamma}$.

The second solution to Equations (8.67) and (8.68) takes the values $x_0 = \frac{\tau}{\gamma}(1+\lambda)^2 - \frac{1+\lambda}{2}N$ and $x_1 = \frac{\tau}{\gamma}(1+\lambda) - \frac{1-\lambda}{2}N$. We must check that this second solution involves values in the admissible range—that is, such that $0 \leq x_0 \leq x_1 \leq \bar{x}$. From the solution itself, we see that $0 \leq x_0$ is equivalent to $\frac{N}{2(1+\lambda)} \leq \frac{\tau}{\gamma}$, and $x_0 \leq x_1$ is equivalent to $\frac{\tau}{\gamma} \leq \frac{N}{1+\lambda}$. Using Equation (8.62), we find $x_1 \leq \bar{x}$ is also equivalent to $\frac{\tau}{\gamma} \leq \frac{N}{1+\lambda}$. In addition, we must again check that land is allocated to the highest bidder, as reflected in conditions (8.63)–(8.66). By inspection of Equations (8.60) and (8.61), we can see that condition (8.63) is satisfied. Since $\Phi(x_0) = \Psi(x_0, u)$ and $\Phi(x_1) = \Psi(x_1, u)$, conditions (8.64) and (8.65) are equivalent to $\frac{d\Phi(x)}{dx}|_{x=x_0^+} \geq \frac{d\Psi(x,u)}{dx}|_{x=x_0^+}$ and $\frac{d\Phi(x)}{dx}|_{x=x_1} \leq \frac{d\Psi(x,u)}{dx}|_{x=x_1}$, respectively. Using Equations (8.58), (8.60), (8.61), we can see that they are both satisfied provided that $\frac{\tau(1+\lambda)}{\gamma} \leq N$. And population has been determined by ensuring that Equation (8.66) is satisfied.

Pulling all of the above together, we can summarize it as follows:

$$x_0 = \begin{cases} 0 & \text{if } \frac{\tau(1+\lambda)}{\gamma} \leq \frac{N}{2}, \\ \frac{\tau}{\gamma}(1+\lambda)^2 - \frac{1+\lambda}{2}N & \text{if } \frac{N}{2} < \frac{\tau(1+\lambda)}{\gamma} < N, \\ \frac{1+\lambda}{2}N = \bar{x} & \text{if } N \leq \frac{\tau(1+\lambda)}{\gamma}. \end{cases} \quad (8.69)$$

$$x_1 = \begin{cases} \frac{\lambda}{2}N & \text{if } \frac{\tau(1+\lambda)}{\gamma} \leq \frac{N}{2}, \\ \frac{\tau}{\gamma}(1+\lambda) - \frac{1-\lambda}{2}N & \text{if } \frac{N}{2} < \frac{\tau(1+\lambda)}{\gamma} < N, \\ \frac{1+\lambda}{2}N = \bar{x} & \text{if } N \leq \frac{\tau(1+\lambda)}{\gamma}. \end{cases} \quad (8.70)$$

The thresholds $\bar{x}$, x_0 , and x_1 , in Equations (8.62), (8.69), and (8.70), respectively, depend on the population. By substituting these three equations, as well as Equations (8.58) and (8.60) into (8.66), we can obtain the equilibrium population N in the city as a function of the level of utility $\underline{u}$ that can be achieved elsewhere.

Before we discuss these equilibria, a final check remains: commuting patterns must be utility maximizing, as defined by Equation (8.46). First, note that $T(x) = x$ for $x \in [-x_0, x_0]$. Workers within the mixed-use area cannot do better by commuting outward since this both lowers the wage and makes them incur commuting costs. They cannot do better by commuting to a job closer to the center if the increase in the wage is not greater than the commuting costs incurred: $-\frac{dw(x)}{dx} \leq \tau$. Using Equations (8.59) and (8.69), we see that this condition is satisfied if $0 \leq x_0$, which must hold. Regarding workers living in the purely residential area $[-\bar{x}, -x_1] \cup [x_1, \bar{x}]$, they will work in the closest purely commercial area, and if they change jobs within this area, by Equation (8.59), the change in the wage they obtain is exactly offset by the change in the commuting costs they incur. For concreteness, the following commuting function is utility maximizing and supports the equilibrium:

$$T(x) = \begin{cases} \frac{x_1(x + x_1) - x_0(\bar{x} + x)}{\bar{x} - x_1}, & x \in [-\bar{x}, -x_1], \\ x, & x \in [-x_0, x_0], \\ \frac{x_1(x - x_1) + x_0(\bar{x} - x)}{\bar{x} - x_1}, & x \in [x_1, \bar{x}]. \end{cases} \quad (8.71)$$

This expression implies that the worker living furthest away in the purely residential area at $\bar{x}$ commutes to the job in the purely commercial area closest to home at x_1 , with workers living more centrally commuting to more centrally located jobs up until the worker living at x_1 who commutes to x_0 . Thus, in Equation (8.71) there is no cross-commuting.

The equilibrium depends on the value of $\frac{\tau(1+\lambda)}{\gamma}$, which increases with the rate τ at which commuting costs increase with distance and with the land requirement of firms λ , and decreases with the rate γ at which productivity spillovers decay. If this combination of parameters is within some intermediate range ($\frac{N}{2} < \frac{\tau(1+\lambda)}{\gamma} < N$), we have the richest equilibrium configuration: there is an area close to the (endogenous) city center where there is mixed land use and where workers reside locally; on each side of this mixed land

use area, there is an area in pure commercial use; and beyond each of these two commercial areas there is an area of pure residential use, from where workers commute to the nearest commercial area.

For firms, being clustered together in a purely commercial area increases their productivity through spillovers (with the gain from proximity being greater the higher the spillover decay γ), but forces them to compensate their workers for commuting costs (which are greater the higher the cost per unit of distance, τ , and the more space that each firm takes up, λ). If $\frac{\tau}{\gamma} < \frac{N}{2(1+\lambda)}$, productivity spillovers dominate commuting costs and the mixed-use area does not exist. In this case, there is a central commercial area surrounded by two residential areas. Figure 8.5 plots the bid-rent functions for land in residential use and in commercial use, $\Psi(x, \underline{u})$ and $\Phi(x)$, in this equilibrium configuration where commercial and residential uses are fully separated. This is like a monocentric model, except that now the configuration of a central area where all firms are located surrounded by land in residential use is an equilibrium outcome instead of an initial assumption. If we let firms produce without using land ($\lambda = 0$), then from Equations (8.69) and (8.70), $x_0 = x_1 = 0$ and we get exactly the monocentric city outcome with all firms clustered at one central point.

Finally, if $N \leq \frac{\tau(1+\lambda)}{\gamma}$, we are at the opposite extreme, and commuting costs dominate productivity spillovers and the pure commercial and residential areas do not exist. In this case, the entire city is under mixed use and every worker lives where they work.

It is worth noting that, unlike in the standard monocentric model presented in Section 8.2, when firms generate spillovers for each other the equilibrium may not be efficient. This is because, in choosing its location within the city, each firm takes into account the spillovers it receives from other firms but not the spillovers it creates for other firms. Thus, to calculate optimum land use patterns, we must replace γ by 2γ in the calculations above. Replacing γ (for the market equilibrium) by 2γ (for the optimum) in Equations (8.69) and (8.70) immediately implies the following. If $\frac{N}{2} < \frac{\tau(1+\lambda)}{\gamma} < N$, the market equilibrium involves a configuration like the one shown in Figure 8.4, but the

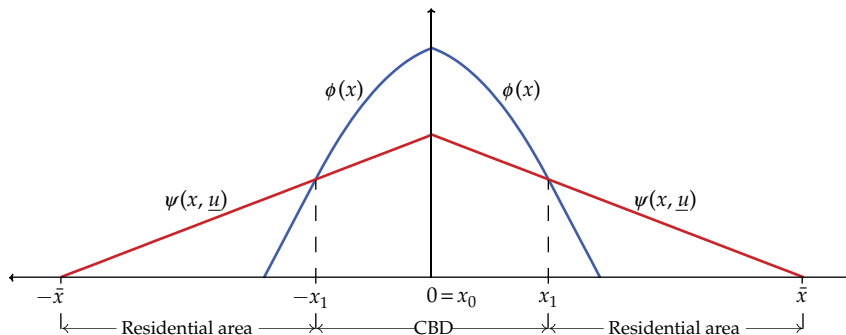


Figure 8.5 The monocentric equilibrium in Ogawa and Fujita (1980).

optimum configuration involves a monocentric city with a purely commercial district from $-\frac{\lambda}{2}N$ to $\frac{\lambda}{2}N$ surrounded by purely residential land. If $N < \frac{\tau(1+\lambda)}{\gamma} < 2N$, the market equilibrium involves mixed commercial and residential use throughout the city, but the optimum configuration involves a pattern like that in [Figure 8.4](#). In both cases, the market delivers commercial development that is too dispersed. Only if $\frac{\tau(1+\lambda)}{\gamma} \leq \frac{N}{2}$ (monocentric configuration) or if $2N \leq \frac{\tau(1+\lambda)}{\gamma}$ (completely mixed configuration) do market equilibrium and optimum coincide.

The above framework can be extended and many of its simplifying assumptions relaxed. For instance, [Fujita and Ogawa \(1982\)](#) replace the linear decay of spillovers by an exponential decay. [Lucas and Rossi-Hansberg \(2002\)](#) also use an exponential functional form for commuting costs and further allow firms to substitute between land and labor and workers to choose their consumption of land. Many of the general features follow, but, without linearity, not only is the framework much more difficult to solve but the complexity of possible equilibria explodes.³⁴

Explicitly modeling the location choice of both firms and workers in a city where they both compete for land is very useful for several reasons. First, it shows that the same method used to solve the monocentric model and its extensions to multiple groups can be applied in a context where firm location is endogenous. To do so, we construct the bid-rent function of land used by firms and, as before, ensure that land is allocated to the highest bidder. Second, the gradients that were present in the monocentric model still apply in this context. However, since multiple production centers can arise, these gradients are no longer necessarily monotonic. For instance, in the richer configuration described above, land rent gradients typically have local peaks both at the city center and at the center of commercial districts, with their relative height depending on parameters. Also, new gradients arise, in particular for wages, which also tend to be higher at denser locations. Third, we see that the monocentric model can be an equilibrium of a model where firms choose their location freely. This arises when the benefits of proximity for firms are strong relative to commuting costs. Finally, we are also able to explore alternative patterns of land use in the city. The richer configuration of the [Ogawa and Fujita \(1980\)](#) framework, while complex, may be a realistic stylized description of many cities: a central area where businesses and homes coexist and most people tend to work locally, with the next ring having a stronger commercial component, and finally an outer ring of residential areas from where people commute longer distances into the city. While the distribution of built-up land between commercial and residential uses shown in [Figure 8.1](#) for Paris is more nuanced than the theoretical outcome in panel (a) in [Figure 8.4](#), one can nevertheless see a very central area in Paris with more mixed use and then further out two peaks of commercial land (pointing downward, since the share of land in commercial use is plotted at the top in the bottom right panel in [Figure 8.1](#)).

³⁴ See also [Helsley \(1990\)](#) and [Ota and Fujita \(1993\)](#).

8.5. LAND USE REGULATION

In most countries and cities, land use is not solely the outcome of the market forces highlighted above. Zoning and other restrictions on land use and property transactions usually play a fundamentally important role in the determination of whether a parcel of land is developed, how it is developed, and who ends up using it. A comprehensive review of zoning and other restrictions on land use is beyond the scope of this chapter. We refer to Chapter 19 in this handbook for more details. See also [Fischel \(2000\)](#) for an introductory conceptual overview of the issues surrounding land use regulations. In developing countries, an important additional aspect of regulation is the insecurity that frequently surrounds land property and housing tenure. See Chapter 21 in this handbook for a review of work on this issue and other particular aspects of urbanization in developing countries. In what follows, we provide only a succinct account to show how land use regulations can be incorporated into the models we have explored so far and what type of issues they raise.

At a broad level, land use regulations limit the type and intensity of land development. They limit the type of development by specializing land use and separating users. They limit the intensity of development by imposing constraints on the size of buildings, limiting the floor-to-area ratio, limiting the fraction of land that can be developed for each parcel, or simply by making the application process for development extremely demanding and lengthy.

The main case for separating users relies on the existence of negative externalities that certain users may impose upon others.³⁵ This argument was first formalized by [Stull \(1974\)](#). He considers a city with land divided between a central commercial area and surrounding residential areas, as in [Figure 8.5](#). Rather than considering this configuration as an equilibrium outcome where firms and residents compete for land everywhere in the city, as we did when constructing [Figure 8.5](#), [Stull \(1974\)](#) considers the allocation of land from the point of view of a city developer who is constrained by such a monocentric configuration and has to decide where to place the limit between commercial and residential uses so as to maximize the aggregate value of land in the city. In the absence of any externalities across land uses, the limit between commercial and residential uses that maximizes aggregate land prices is at the point where the bid-rent curves for commercial and residential land intersect.

Consider now introducing an externality across uses. For instance, commercial activity may generate noise or pollution that reduces the utility of nearby residents. Suppose that this externality affects only the residents' utility and decays as one moves away from

³⁵ This is sometimes referred to as "Euclidian zoning" in the United States following the 1926 Supreme Court case *Euclid v. Ambler*, in which the Euclid community in suburban Cleveland opposed Ambler Realty, a developer of manufacturing space. See [Fischel \(2004\)](#) for an insightful history of zoning in the United States.

commercial land. Then this externality depresses the willingness to pay of residents to live close to the CBD and lowers their bid-rent curve. Firms are not directly affected by the externality so, absent any other changes, the point where the bid-rent curves for commercial and residential land intersect moves outward and commercial activity expands at the expense of residents. However, firms and workers interact not only in the land market but also in the labor market. The expansion of firms and the reduction in the number of residents increases the wage, lowering the bid-rent curve for commercial land and raising the bid-rent curve for residential land. In the end, wages will be higher and commercial land prices lower than in the absence of the externality. Residential land prices may be higher or lower depending on the magnitude of wage changes relative to the externality (they could be lower close to the CBD and higher toward the edge of the city). A city developer achieves a higher aggregate land value by constraining commercial land use, placing the limit between uses at a point where the commercial bid-rent curve is strictly higher than the residential bid-rent curve.

This example may seem of limited interest since most CBDs are no longer dominated by manufacturing in developed countries. However, it is easy to see that the same argument applies to any nonconforming use generating negative externalities in its neighborhood anywhere in a city.

We have just seen an example where zoning is justified on the basis of negative externalities across users. A similar argument can be made on the basis of positive externalities within uses. Consider the model in [Section 8.4](#). We have already seen that in the presence of density externalities across firms the equilibrium may differ from the optimum because firms do not take into account the spillovers they generate for other firms by locating closer to them and they may end up too dispersed in equilibrium. For instance, we saw that if $N < \frac{\tau(1+\lambda)}{\gamma} < 2N$, the market equilibrium involves mixed commercial and residential use throughout the city, but the optimum configuration involves full separation between suburban residential zones and central commercial area.³⁶ Firms do not cluster enough because they do not internalize the effect of their own spillovers on other firms. Through zoning, one could achieve the optimal configuration.

While the sort of externalities examined here provides a strong justification for specializing land use and containing some types of users, some questions remain. First, it is obviously hard to know how much land should be devoted to manufacturing and where it should be located. Second, while the case for separation between dirty manufacturing and residential areas is strong, separating, as many cities do, residential from commercial areas is less obvious. Even the case for separating office space from housing is far from clear-cut (and far from being universally practiced). Third, even if separation is desirable, it is unclear whether this is best achieved by zoning to prevent the externality occurring in the first place or by having nuisance laws that force appropriate compensation *ex post* for

³⁶ The comparison between the equilibrium and the optimum configurations in this framework can be found in [Imai \(1982\)](#). See also [Rossi-Hansberg \(2004\)](#).

negative externalities. The latter can be optimal if the “compensation” is equal to the optimal Pigovian tax associated with the externality that is generated.

Turning to regulations that restrict the intensity of development, they could also be justified on efficiency grounds by a similar type of externality argument. For instance, a minimum lot size regulation in a nice neighborhood will prevent the construction of low-end housing which could affect the aesthetic quality of the neighborhood.³⁷ Restricting the intensity of residential development in certain areas could then be a way to specialize land use even further among subtypes of users in a desirable way. While this argument is logically consistent, it seems hard to believe that the negative externalities associated with mixing different forms of residential development are significant enough to justify the prevalence and importance of these regulations.

A more reasonable possibility is that residents value the (low) density around them. More specifically, assume residents value the open space that surrounds them as in [Turner \(2005\)](#). Although it is hard to fully operationalize the notion of open space empirically, there is little doubt that city residents value it highly.³⁸ In their property descriptions, realtors strongly underscore, when possible, nice views and proximity to parks and nature. They also keenly emphasize privacy using a positive adjective such as “secluded” or “oasis.” This is not anecdotal. Parks and open spaces are quantitatively important in cities. Even a very intensely developed city such as New York boasts that more than a quarter of its land area is devoted to open spaces, parks, and other recreational areas. The data underlying [Figure 8.1](#) indicate that open space represents 14% of the land in the 5 km ring around the center of Paris. The proximity to undeveloped land is highly appealing when residents value open space. Hence, the presence of open space begs new residential development, which of course reduces it. In equilibrium, this is self-defeating and there will be too much development. In turn, this may justify regulations that limit the intensity of development ([Turner, 2005](#)).³⁹ In practice, these regulations can take many forms, such as a maximum share of development for a parcel, maximum height,

³⁷ Restrictions on development impose an upper limit in the vast majority of cases. Although we ignore them here for the sake of brevity, lower limits such as minimum density are also used in some cases. They can find their justifications either as measures countervailing other inefficiencies that make the intensity of development suboptimally low or as measures necessary to achieve a “critical mass” to justify some amenities such as a neighborhood park, convenience stores, or public transport. Note also that preventing development in some locations is an implicit incentive to have it somewhere else as in the case of urban growth boundaries.

³⁸ See, for instance, [Irwin \(2002\)](#) and [Geoghegan \(2002\)](#).

³⁹ [Strange \(1992\)](#) studies this type of feedback effect in a model where density in neighboring city blocks imposes a negative disamenity, which in turn affects housing production decisions. [Turner \(2005\)](#) also provides some interesting results for the dynamics of development when residents have a taste for open space, showing, for instance, that remote locations will be developed before less remote residential areas and provides a consistent explanation for the leapfrogging that is often observed in the development process. A limitation to these results is that suburban development is often the work of nonatomistic developers who may internalize the externalities of the model, at least partially.

or “rights to lights” that allow neighbors to appeal against high-rise developments in large cities.

While a theoretical case can be made to restrict development and keep some land undeveloped in the form of parks and other green spaces, too little is known to provide firm guidance for urban planning. For instance, low density can occur through uniform development at a very low level with, say, one house per hectare over a 100 ha tract. The same level of density in this 100 ha tract can also be the outcome of a high-density development hosting 100 households over a 1 ha parcel surrounded by 99 ha of green space. Extant models and our knowledge of what residents value are not detailed enough for us to be able to deal with those issues at this stage. One might be tempted to use land values as a guide for making decisions. Although, in practice, land value appreciations seem grossly underused by the planning process, we know that in theory land prices provide an unbiased guide to investment in public projects only under restrictive conditions.⁴⁰ In the absence of solid knowledge of those issues, there are of course suspicions that many existing land use regulations might be widely off the mark.

A second set of reasons for controlling the intensity of development is given by possible externalities arising from commuting. This activity has been viewed so far as free of externalities. This is obviously counterfactual. Having more cars on the road slows down traffic. This is a solidly established fact (see, e.g., [Small and Verhoef, 2007](#)), which implies that the unit commuting cost, τ , used in [Section 8.2](#) should be treated not as a constant but as a function of the number of commuters. The main complication is that the number of commuters is not the same everywhere in the city. In a monocentric city, all commuters need to enter the CBD, but only one commuter uses the last segment of the road at the urban fringe. More generally, commuting costs at a given point will depend on the number of commuters that live beyond this point. Hence, land use determines commuting costs, which determine land use. This problem was studied by [Solow \(1972\)](#). Commuting costs at location x depend on the number of city residents that leave beyond x , which we denote N_x so that the equivalent of the Alonso–Muth condition (8.16) will include $\frac{\partial \tau(x, N_x)}{\partial x}$ instead of simply τ in the case where commuting costs are linear with distance. Hence, in equilibrium, $\frac{dP(x)}{dx}$, the gradient of the bid-rent curve at location x , depends on the density at location x and a function of the cumulative of housing density beyond x . Since in equilibrium the price of housing $P(x)$ also determines the quantity of housing consumed by a resident and hence residential density, solving for the allocation of land amounts to solving a nonlinear second-order differential equation, for which closed-forms solutions are available only in particular cases as shown by [Solow \(1972\)](#).

⁴⁰ See [Kanemoto \(1988\)](#) for further discussions. Changes in land prices locally provide a biased measure of welfare changes when mobility is imperfect, when residents are heterogeneous, or when lot size is endogenous.

More generally, road congestion leads to inefficient land use. Zoning has been alleged to provide a solution to this problem. Pines and Sadka (1985) show that an urban planner can implement optimal land use by controlling parcel size (and the intensity of development). Should it turn out to be impossible to control parcel size, Pines and Sadka (1985) show that an urban growth boundary can improve patterns of land use in a second-best world. However, it is unclear why an indirect instrument such as the regulation of lot size and the intensity of development should be used instead of a more direct tool such as a congestion toll. Such a toll is analyzed by Kanemoto (1980), who considers a model in the spirit of Solow (1972) with congestible roads. The danger with indirect instruments is that they often do not include important other margins of adjustment. For instance, the results obtained by Pines and Sadka (1985) and others, who explore the possibility of using land use regulations instead of congestion tolling, were obtained by typically assuming that the number of trips per household and their destination are fixed. Regulations of land use by benevolent and highly sophisticated urban planners may manage to impose the “right number” of residents in each location, but nothing guarantees that they will drive the “right amount.” On the other hand, it is true that existing congestion tolls are subject to difficult acceptance issues. They also form fairly crude instruments: typically a cordon instead of, optimally, a charge at each location that depends on the state of the traffic.⁴¹ Even if congestion tolling is not available, there are other instruments such as parking pricing that could substitute for congestion tolls (Armott and Inci, 2006). Parking pricing seems easier to use, more direct, and more flexible than comprehensive zoning.

In turn, road congestion is raising the issue of the provision of roadway, which has been ignored so far. Roads and parking spaces are quantitatively important in cities. The data underlying Figure 8.1 show that, while, as already mentioned, parks and other public spaces represent around 14% of land within 5 km of the center of Paris, the roadway, parking areas, and other transport infrastructure occupy around 18%. These figures are likely higher in the central parts of American cities, which tend to have wider streets and more land fully dedicated to parking lots.⁴² The use of most roads is not excludable, and as a result, roads are publicly provided. Following the pioneering work of Solow and Vickrey (1971), roads have been explicitly considered in land use models by Kanemoto (1980) and Pines and Sadka (1985). The provision of roads differs from the provision of other local amenities such as parks and green spaces discussed above because roads are not directly enjoyed by residents. Instead, they provide a link between locations, and one road might affect different locations alongside it differently. As just argued, the issue of congestion also looms

⁴¹ See Small and Verhoef (2007) for further discussion of congestion tolling. There has been some recent progress toward more time-dependent pricing for roads, but this is still limited to a handful of roads in the United States and a few cities elsewhere in the world such as Singapore, Stockholm, and London.

⁴² See Manville and Shoup (2003) for further discussion of these issues and the difficulty of providing accurate numbers for them.

very large. Despite these differences with other local public projects, many of our conclusions regarding roads are the same: we know too little to provide good guidance for policy, and land prices will provide an imperfect guide for decisions.⁴³

Although land use regulations are often justified by efficiency motives and curbing externalities, in practice zoning is often motivated by other reasons and is exclusionary in nature.⁴⁴ This may be because of peer effects. For instance, residents of rich suburban areas may want to maintain some exclusivity for their neighborhood, wish to remain among themselves, or want socially selected children to attend local schools. While this is perhaps part of the explanation, a fiscal externality is also likely to be at play.

In many countries, including the United States, a large share of local public goods, including education, is financed through property taxation. This can generate a fiscal externality which can be curbed using exclusionary zoning. To understand this, it is worth going back to Tiebout's (1956) original model of fiscal federalism. In his model, a heterogeneous population will sort into homogeneous jurisdictions that efficiently provide local public goods. In each jurisdiction, local public goods will cater to the tastes and incomes of local residents. To obtain this efficient outcome, a number of stringent conditions must be met, including having residents being able to vote with their feet, the absence of public good spillovers across jurisdictions, and the availability of lump-sum taxes. In practice, lump-sum taxes are not available for a variety of reasons. Instead, local public goods are financed through property taxation. The tax paid on a property is usually roughly proportional to its value. This creates a problem because property taxation makes it possible for poorer resident to free-ride by moving to a rich jurisdiction to consume a high level of public goods but pay only low taxes by consuming a small quantity of

⁴³ Skepticism about land instruments to fund roads was first raised by Mohring (1961) and Solow and Vickrey (1971). After Mohring and Harwitz (1962), transport economics is more upbeat about the use of tolls to implement an optimal provision of roads. See Small and Verhoef (2007) for further discussion and the exposition of a number of "self-financing" theorems for roads.

⁴⁴ For instance, the municipality where one of the two authors of this chapter lives imposes strict requirements on the maximum share of "impervious" (i.e., developed) surface for each parcel. This regulation is ostensibly justified on environmental grounds, but it seems hard to believe that 75–80% of land in a close suburb should remain unbuilt or unpaved to avoid floods and other environmental damage when the central part of the city is much more densely built. More likely, this forces residents to consume a lot of land and thus selects mostly rich residents willing to finance high-quality primary and secondary schools as argued in what follows. The municipality where the other author lives, on the basis of similar environmental concerns, limits the total amount of impervious surface by excluding most parcels from development instead of limiting development within each individual parcel. Unlike the previous example, this reduces the number of detached houses built on large lots, but also raises house prices and selects residents. This alternative regulation preserves large contiguous natural spaces instead of leading to many large private gardens. However, it generates a large price gap between parcels on which development is allowed or not contingent on a local policy decision, which opens the potential for corruption.

housing. That is, a decentralized provision of public goods financed by property taxation induces the poor to chase the rich. To avoid this, the rich impose exclusionary zoning regulations.⁴⁵

Then, exclusionary zoning, to the extent that it restores a Tiebout equilibrium, could promote efficiency, if not equity. There are worries, however, that exclusionary zoning may go beyond implementing a Tiebout equilibrium and may instead maintain land development at inefficiently low levels within the urban fringe. As eloquently illustrated by [Fischel \(2001\)](#), local residents may be incentivized to restrict development locally to maximize property values. In many countries, including the United States, land use regulations are local decisions taken by local officials elected by homeowners. [Fischel \(2001\)](#) notes that although some new property developments might be expected to be beneficial to incumbent property owners, the latter may nonetheless rationally resist those changes because there is a risk that things may not go according to plan.⁴⁶ Risk aversion is heightened by the fact that homeowners usually have most of their assets vested in their house. It is also possible that gains from new property developments are highly unequally distributed, with some residents ending up losing from them. Appropriate compensation schemes are difficult to set up. As a result, the status quo may naturally arise as a political economy equilibrium.

A more direct argument is that restricting housing supply may lead to higher prices. Note that this argument requires reneging on another assumption of the Tiebout model: perfect mobility. If the demand for locations is flat (i.e., residents are indifferent between locations after being appropriately compensated for differences in accessibility), housing must be efficiently provided for property values to be maximized. Put differently, under perfect mobility and in the absence of local preferences, residents have an incentive for optimal land use regulations since this is what maximizes property values. Overly restrictive regulations will entail the inefficient use of capital in housing and reduce values.⁴⁷ If the demand for locations is not perfectly elastic (if, for instance, residents have a preferred location, all else being equal), incumbent residents have an incentive to restrict entry and limit housing supply locally, in effect acting as monopolies. [Ortalo-Magné and Prat \(2014\)](#) and [Hilber and Robert-Nicoud \(2013\)](#) propose some versions of this political economy argument. If it is right, overly restrictive regulations in developed areas would be a powerful force explaining excessive urban sprawl in undeveloped areas. We return to this issue in [Section 8.8](#).

⁴⁵ See [Fischel \(1987\)](#) for more regarding this type of argument.

⁴⁶ [Breton \(1973\)](#) provides an early version of this argument.

⁴⁷ Overly restrictive regulations may also imply negative feedbacks through agglomeration effects as suggested by [Chatterjee and Eyigungor \(2014\)](#).

8.6. EMPIRICAL PRICE AND DEVELOPMENT GRADIENTS

We now turn our attention to empirical work on land use patterns, beginning with estimations of the gradients predicted by the monocentric model. Even before the work of [Alonso \(1964\)](#), [Mills \(1967\)](#), and [Muth \(1969\)](#), research had taken an interest in some of the predictions of the monocentric model. [Clark \(1951\)](#) is usually credited for being the first to show a decline in population density as one moves away from the CBD for a diverse cross section of cities. The popularity of Clark's work is to some extent due to its simplicity and its weak data requirements. Studies in this tradition first decide on a city center before drawing concentric rings around it. They count the population within each ring and regress it on the distance to the center. Unsurprisingly, the population density in most cases decreases smoothly with the distance to the center. This sort of regression usually yields a high R^2 that authors often interpret as strongly supportive of the monocentric model. However, this high R^2 is mostly due to the inherent smoothing associated with the ring approach. Another approach is to consider small areas within a city, such as tracts, and regress tract density on their distance to the center. The R^2 associated with this type of regression is much lower as there are often areas of fairly high density that are located relatively far from the main center.⁴⁸

Under the assumption of a linear relationship between the logarithm of the population density and distance, [Mills \(1972\)](#) showed that the density gradient could be estimated knowing only the population of the main city, its area, and the population of the entire metropolitan area. This “two-point” approach lowers the data requirements even further, perhaps at the cost of highly noisy estimates.

The large literature that followed Clark's work is generally supportive of negative population density gradients. See [McDonald \(1989\)](#) for an early review and [Bertaud and Malpezzi \(2003\)](#) for further evidence from world cities. [Kim \(2007\)](#) documents a gradual flattening of the density gradient in US cities over the twentieth century. Interestingly, exceptions to negative gradients include cities in formerly socialist economies such as Moscow ([Bertaud and Renaud, 1997](#)) and South African cities under apartheid ([Selod and Zenou, 2001](#)), where the market mechanisms at play in the monocentric model was heavily constrained. Even in countries where cities tend to follow a broadly monocentric pattern, several interesting features emerge. First, the density gradient typically becomes weaker far from the center. Second, other variables such as the distance to secondary centers and various geographical landmarks also often have some explanatory power.

⁴⁸ Gradients also tend to become much flatter far from the CBD in American cities. Mechanically, the fit of a weak negative gradient will be poor, with R^2 going to zero as the slope also goes to zero.

There is also work that attempts to measure the gradient of the price of housing as illustrated, for instance, by [Yinger \(1979\)](#) or [Coulson \(1991\)](#).⁴⁹ For reasons to be explained below, this literature has often struggled to provide evidence of negative gradients for the unit price of housing. Perhaps because of the mixed success of the approaches looking at unit housing price gradients, much of the literature has focused instead on land price gradients after [Mills' \(1969\)](#) pioneering effort. The works of [Cheshire and Sheppard \(1995\)](#) and more recently [Ahlfeldt \(2011\)](#) are perhaps the most advanced. This literature is too large to review extensively here. See [McMillen \(2006\)](#) and [McMillen \(2010\)](#) for surveys. Although the findings are generally supportive of a negative land price gradient, [McDonald and Bowman \(1979\)](#) raise some doubts.

In contrast to the abundance of work looking at density, housing price, and land price gradients, little to no work has been devoted to the predicted gradients of the capital intensity of housing development and of housing consumption per household. The work of [McMillen \(2006\)](#) is a lone exception that examines the floor-to-area ratio in metropolitan Chicago and shows a strong declining trend as one moves away from the CBD.⁵⁰ As far as we know, there is no work looking at housing consumption per household in relation to the distance to the center. [Figure 8.1](#) provides a crude but sharp illustration for the case of Paris through the shares of land in residential use occupied by single-family and multifamily buildings at various distances from the city center. Only 3% of land in residential use within 5 km of the center of Paris is occupied by single-family homes. In contrast, single-family homes occupy 50% of residential land between 5 and 10 km from the center, a figure that rises to 79% between 10 and 20 km from the center, and to 87% between 20 and 30 km from the center.

Despite a large body of work that has developed over more than 60 years since the work of [Clark \(1951\)](#), the empirical knowledge accumulated on the monocentric urban model and its extensions remains limited. The first reason is that, until recently, data have been difficult to collect. Assembling data about tracts in cities or property prices often needed to be done manually. As a result, much of the literature is limited in scope and focuses on one particular city or perhaps a small number of cities. To make matters worse, these cities are often acknowledged to have been chosen for particular reasons, including strong priors about how monocentric or multicentric these places might have been. Looking at broad cross sections of cities is needed to avoid such sample selection issues.

⁴⁹ It is important to remember that the monocentric model makes a prediction for the price of housing per unit not for property prices. As one moves away from the CBD, the unit price of housing is expected to fall, but the size of properties is expected to increase. The net effect of these two forces on overall property prices is theoretically ambiguous.

⁵⁰ See also [Clark \(1967\)](#) for early evidence.

It is also important to assess how much cross-city heterogeneity there is within one country. The work of [Combes et al. \(2012\)](#) is of particular interest in that respect. They estimate land price gradients for a broad cross section of cities in France and find a lot of heterogeneity. The elasticity of land prices with respect to distance varies from about -50% for cities in the first decile to essentially zero for some cities in the last decile. Casual observation also suggests tremendous heterogeneity in urban land use across cities of different countries, with commentators often referring to the “American city” as synonymous for urban sprawl with extremely low suburban densities and flat gradients there. Instead, the “Asian city” is often taken as synonymous with extremely high density in the core. The behavior of gradients over time is also of interest as we expect growth and technological progress in the last 200 years to have affected cities in a major way. Unique in the literature, [McMillen \(1996\)](#) uses comprehensive land price data over 150 years in Chicago to document a flattening of the land price gradient. Hopefully, future work will make use of data that are now broadly available (at least in some countries) about property prices, urban land use, and the population to document the evolution of land use and its heterogeneity both within and between countries.

Data availability is not the only issue. The data and approaches that have been used in the literature so far are problematic in several respects. Let us first discuss measurement issues. The first problem concerns the definition of centers or subcenters. Older studies often assign a center in a somewhat arbitrary manner on the basis of history or casual evidence.⁵¹ More recent studies tend to compute a centroid or use a peak of density for residents or, better, employment. While the latter approach is probably sufficient to determine a CBD for some applications, matters are more complicated when the analyst wishes to consider several subcenters. The most convincing approach to detect centers and subcenters was proposed by [McMillen \(2001\)](#). It is inspired by two related features from theory. First, subcenters will be associated with a concentration of jobs. This concentration will attract residents, and as a result, subcenters will affect land and housing prices in nearby locations. Consistent with these two features, the approach developed by [McMillen \(2001\)](#) is in two steps. First, a smooth employment density function is estimated nonparametrically. Candidate subcenters are locations with positive and significant residuals in this employment regression. Then in the second step, a semiparametric rent function is estimated. Subcenters will be those among the candidates identified in the first

⁵¹ In [Figure 8.1](#), we assigned the cathedral of Notre Dame as the center of Paris. Although the French government uses this location to measure all distances to Paris, it is hard to argue that the area around this cathedral constitutes a CBD. The historical business district of Paris is about 4 km to the west and the current main business center of La Defense is 9 km to the west. At the same time, we can see in [Figure 8.1](#) that the fraction of land that is built up and the percentage of residential development accounted for by multifamily dwellings both peak at about the location of Notre Dame, indicating that it may be a reasonable choice for the city center in terms of residential uses. Once again, Paris provides a good example of the complexity of land use allocations in practice.

step of the analysis that provide significant explanatory power for rents. Despite its appeal, this approach is likely to be sensitive to the smoothing parameters since most cities will look fully monocentric with enough smoothing.⁵² The significance of any subcenter in the second-step rent regression will also depend on the physical extent of the study area as the effect of small subcenters becomes more difficult to detect in larger areas.⁵³

Another measurement problem with density gradients is that density is usually measured at the area level in the empirical work that seeks to estimate density gradients, whereas the model makes predictions at the parcel level. This is a worry because a smaller fraction of land may be used for residential purpose further from the CBD. This may explain a negative population density gradient irrespective of what happens to density at the parcel level.⁵⁴ A simple solution here is to rely on data from a land registry or any other source that is able to identify residential parcels. The difficulty is that such data are far less widely available than area population data at the tract level.⁵⁵

While measuring population density involves some unexpected difficulties, measuring unit housing prices is distinctly harder. The chief reason is that we usually observe only the price of a house—that is, the product of the unit price of housing and the number of housing units that this house provides. The standard solution—used, for instance, in [Coulson \(1991\)](#)—is to use house characteristics and introduce them in the regression that estimates the housing price gradient, such as

$$\log P_i(x) = F(x) + X_i\alpha + \epsilon_i, \quad (8.72)$$

where $P_i(x)$ is the price of house i located at distance x from the CBD, $F(x)$ is a function of distance x , X_i is a collection of house characteristics, and ϵ_i is the error term. The mapping from the Alonso–Muth condition (8.4) to the empirical Equation (8.72) which aims to estimate it is not immediate. It is, however, possible to rewrite Equation (8.4) in first difference for a given house relative to another house nearby. After linearizing, we obtain $\Delta \log P(x) \approx \tau \Delta \log x + \Delta \log h$. Approximating h , the number of housing units of a property, by a vector of characteristics and using the same hypothetical house as a comparison

⁵² Recall that the so-called optimal smoothing is based only on a rule of thumb that trades off noise reduction against systematic changes in curvature.

⁵³ See [Redfean \(2007\)](#) for a more local approach that is not subject to this criticism. Another possible criticism is that this procedure is inspired by theory but not directly led by it. This criticism is less convincing because no existing theory of land use is detailed enough to go beyond the broad features used by [McMillen \(2001\)](#). In addition, one might be wary of using one particular theory to define subcenters as more precise definitions will likely be driven by specific features of models.

⁵⁴ Cheaper land further from the CBD may, of course, explain a greater prevalence of nonresidential use in more complex versions of the monocentric model where several types of users could be competing for suburban land. It remains that a negative density gradient at the area level is not the same thing as the negative density gradient at the parcel level predicted by the simplest models.

⁵⁵ The work of [Mieszowski and Smith \(1991\)](#) is an exception using parcel-level data for the city of Houston.

point for all houses yields regression (8.72) for the particular case of a logarithmic specification for the distance effect. Taking a more general functional form for commuting costs in the theoretical model still yields Equation (8.72). Note that when the range of distances from the CBD becomes large, the approximation made in the linearization above which disregards the $\Delta \log x \Delta \log h$ term may not be warranted. The use of hedonics to condition out housing heterogeneity is not innocuous either, and several well-known problems may arise when using this tool. The most important problem is that, as predicted by the theory, house characteristics will be correlated with the distance to the CBD and it is unlikely that the vector of the observed house characteristics used by the analyst is both comprehensive and well specified. As a result, there may be missing house characteristics that are correlated with distance.⁵⁶

Much attention in the literature has been devoted to the choice of the functional form for the distance function $F(x)$ in Equation (8.72) (and in the corresponding equation estimating density gradients). Following Clark (1951), the early literature typically used a dependent variable in log and the level of distance as the explanatory variable. As made clear by the model developed in Section 8.2, the price of land may be a complicated function of distance which depends finely on which assumption is made regarding the utility function, commuting costs, and residents' income. Nothing guarantees that the resulting gradient will be a negative exponential in distance. In general, it will not. An alternative is to use the log of distance instead of its level as explanatory variable.⁵⁷ Combes et al. (2012) argue that the fit is approximately the same for both specifications when estimating land price gradients for French cities. Many authors have used more comprehensive specifications by adding distance terms of higher order. These terms obviously increase the explanatory power of the regressions, sometimes by enough that the authors justify them by invoking specification tests.⁵⁸ McMillen (2010) defends the use of nonparametric or semiparametric estimation methods.

At some level, the usefulness of this debate about functional forms in the estimation of gradients is questionable because it forgets which proposition from theory is tested. An exception is the work of Brueckner (1986). He proposes using switching regression methods to capture discontinuities caused by vintage effects in housing construction such

⁵⁶ As possible alternatives, Epple et al. (2010b) and Combes et al. (2014) develop new methods to reconstruct the amount of housing offered by each property. Epple et al. (2010b) use land and property values as well as land area. Land values are a function of house prices. Then, one can estimate the capital-to-land ratio in housing from housing values per unit of land and retrieve the production function. The quantity of housing can then be recovered. Combes et al. (2014) use the first-order condition for profit maximization and free entry that gives the marginal product of housing with respect to capital (or land) and recover the quantity of housing by integrating.

⁵⁷ While negative exponentials impose a lot of curvature, linear specifications are at the other extreme since, as argued above, theory predicts convex gradients.

⁵⁸ Recall nonetheless that these tests weigh explanatory power against the number of explanatory variables, imposing somewhat arbitrary penalties for the latter.

as those described in [Section 8.3.3](#). Hence, the objective is not to improve the fit of the regression for the sake of it but to capture an empirically important feature that the simplest version of the model misses.

The theory described in [Section 8.2](#) makes a series of predictions under the assumption of a single city center. These concern accessibility, chiefly through the Alonso–Muth condition for land and housing prices and related conditions for the intensity of development, housing consumption, and the density of the population. This assumption of a single center is mostly for convenience and, as argued in [Section 8.4](#), many of the key propositions regarding the importance of accessibility for housing prices and, in turn, land prices, the intensity of development, local population density, and the consumption of housing do not depend on the existence of a single center.

Much of the disagreement in the literature about the monocentric model is about whether cities are monocentric or not. They are not about whether accessibility is valued in accordance with the Alonso–Muth condition. These are two separate questions. The first question is really about an assumption about the geography of cities, whereas the second question is the analytical substance of the model. This said, while the monocentricity of cities in the monocentric model is an assumption, in the models with endogenous job locations reviewed in [Section 8.4](#) it is just one possible outcome. This suggests that cities may be well approximated as monocentric or not depending on the conditions, making it valuable to assess the monocentricity of real cities. Unfortunately, much of the debate here has been misconstrued. The interesting question is not whether cities are monocentric or not. Strictly speaking, most cities are not monocentric and maybe no city is truly monocentric. For instance, [McMillen \(2001\)](#) provides convincing evidence that all the large American cities he looks at contain subcenters, sometimes a sizeable number of them. The prediction of a monotonic decline in land prices, housing prices, and population density in all directions as one moves away from the CBD will always be rejected provided one can work at a sufficiently high level of geographical resolution, be it only because of access to transit stops or major arterial roads. Adding to this, we cannot expect the forces highlighted by the monocentric model to be the only determinants of land and property prices, housing consumption, and population density. As shown by the review of the extensions of the monocentric model in [Section 8.3](#), other factors need to be conditioned out. The world will never be as smooth as the simplest models.

Since monocentricity can always be rejected, the more interesting question is: How monocentric are cities? As argued above, the key difficulty here is that the R^2 of gradient regressions do not provide a good metric by which to measure this, given the granular aspect of population patterns. Reasonable metrics that allow us to measure how monocentric cities are independently of the level of resolution at which population and prices are measured or independently of how the data are smoothed are yet to be developed.

Turning to the second question, about accessibility, it has been only half answered by the literature. We know that indeed distances to the center, subcenters, and other

landmarks nearly always matter greatly to explain house and land prices. There is also a body of literature which has developed a number of accessibility indices that weigh the distances of each location to various subcenters. See [Anas et al. \(1998\)](#) for a review. That accessibility matters is necessary for the accessibility versus price trade-off highlighted by the Alonso–Muth condition to be true. However, this does not guarantee that accessibility is valued as suggested by this condition, where, all else being equal, the difference in property prices between two blocks should be equal to the difference in transport costs involving not only commuting but also all other errands.

As highlighted in [Section 8.2](#), the monocentric model also makes a number of predictions about some aggregate quantities and land rent at the center.⁵⁹ As argued in [Section 8.3](#), these predictions are theoretically less robust than those that rest on the Alonso–Muth condition. Like with the monocentricity question above, the issue is not whether these predictions hold exactly—they will not—but how far we are from them in reality and what accounts for the difference between these simple theoretical predictions and reality. To our knowledge, the predictions about urban aggregates like the proportionality between total land rent and total commuting costs have never been seriously assessed. The proposition that land rent at the center is proportional to city population is assessed by [Combes et al. \(2012\)](#), who provide a measure of urban costs based on equilibrium land prices at the city center in France. Their preferred estimate for the elasticity of the price of land at the city center with respect to the city population is 0.72, and they attribute the difference between 0.72 and unity to greater decentralization in larger cities. The same article also provides results for regressions attempting to explain land price gradients with a variety of city characteristics. The results are not particularly encouraging. None of the explanatory variables that [Combes et al. \(2012\)](#) is robustly significant. Models of heterogeneous residents in the spirit of those explored in [Section 8.3.2](#) predict that the parameters that govern income distribution will also determine land price gradients. [Combes et al. \(2012\)](#) find no evidence of that. Obviously, more work will be needed on this before definitive conclusions are reached.

A recent contribution by [Ahlfeldt et al. \(2012\)](#) estimates structurally a model of internal city structure. The city is a collection of blocks. Although these blocks are taken as given, the intensity of development of each block is endogenous. Residents consume housing, residential amenities, and a composite good. In turn, amenities in a block depend on the fundamentals of this block and amenities in neighboring blocks, which are discounted by a negative exponential function of the distance to the block under consideration. Production requires floor space and labor. As in the model described in [Section 8.4](#), firms benefit from agglomeration economies that are measured by employment density in neighboring blocks discounted by distance just like with the consumption externalities. Productivity is also allowed to vary idiosyncratically across blocks.

⁵⁹ As argued in [Section 8.2](#), it also predicts that the ratio of the land and housing gradients should give the amount of housing. This prediction has not been tested to the best of our knowledge.

When choosing their residence, workers not only consider (endogenously determined) housing prices, local amenities, and the distance to work, but they also receive an idiosyncratic utility shock for each possible commute between blocks. These shocks introduce some heterogeneity in residential and employment choices. Residents choose the block that gives them the highest utility. A key technical difficulty is that the highest utility is given by a potentially extremely hard-to-compute order statistic regarding the distribution of commuting shocks. To keep the problem tractable, [Ahlfeldt et al. \(2012\)](#) use the structure developed in international trade by [Eaton and Kortum \(2002\)](#), which relies on the Fréchet distribution. The main reason for doing so is that the maximum of a sequence of Fréchet-distributed variables is also Fréchet distributed.

An interesting property of the framework of [Ahlfeldt et al. \(2012\)](#) is that it predicts a gravity pattern for commuting flows where the logarithm of the number of commuters between two blocks is expected to be proportional to the travel time between them, after conditioning out features of the origin and destination blocks. Obviously, commuting flows between nearby residential blocks will be nil when there is no employment in these blocks. [Ahlfeldt et al. \(2012\)](#) provide some evidence consistent with this feature for Berlin commuters. The full estimation of their model requires recovering two parameters for commuting costs (the base utility cost of commuting per unit of distance and the idiosyncratic dispersion around it), four agglomeration parameters (the intensity and the spatial decay for both consumption and production externalities), and the productivity and amenity fundamentals of each block. There exists a unique mapping from observed data on commuting, land rents, the intensity of development, employment, and the number of residents for each block and knowledge of key aggregate shares in consumption, production, and construction to the unknown parameters and fundamentals of the model.

The first key result of [Ahlfeldt et al. \(2012\)](#) regards the estimation of rather large but extremely localized agglomeration effects for both production and consumption. They also take advantage of the fact that they are able to observe Berlin at three points in time: before its division and during and after reunification. Assuming unchanging fundamentals for each block in expectation, they are able to replicate with their model the changes in the employment and residential composition of blocks depending on where they are located.

This article is important for several reasons. First, it constitutes a pioneering effort to implement the methods of structural estimation to understand what happens within cities. These methods have been developed in the last 10–20 years in other fields of economics and only now are beginning to be used to analyze land use issues.⁶⁰ Second, the

⁶⁰ See [Chapter 2](#) in this volume on structural estimation in urban economics. Structural methods have been used to explore housing and local public finance issues, but little to none of that work has an explicit spatial dimension. See [Epple et al. \(2010a\)](#) for a model that considers localized amenities within a multijurisdiction framework. Very recently, quantitative versions of the monocentric have been developed to explore a range of applied questions such as the welfare effects of a simple congestion toll ([Brinkman, 2013](#)).

work of Ahlfeldt et al. (2012) allows an economic interpretation of the various gradients observed in cities. Like Combes et al. (2012), it is an attempt to bring empirical work closer to theory, where gradients and other spatial patterns are no longer examined for their own sake but to learn about more fundamental issues. Third, the weaker points of the model or the dimensions where results are less satisfactory become more clearly apparent. For instance, the modeling of commuting behavior in Ahlfeldt et al. (2012) is somewhat *ad hoc* since it is not clear what drives the heterogeneity of bilateral commutes for residents. This is an avenue for future research. Some of the more provocative conclusions of Ahlfeldt et al. (2012) regarding local agglomeration effects in production and consumption will also be examined through other lenses. The robustness of the approach may also be assessed by implementing it for other cities. More generally, the interaction between theory and empirical work leads to their mutual questioning and paves the way for progress in our knowledge.

A few recent contributions notwithstanding, there is still a great paucity of empirical work that takes theory seriously. The main thrust of land use models in the spirit of Alonso (1964), Mills (1967), and Muth (1969) lies in the Alonso–Muth condition that highlights the key trade-off between prices and accessibility. This condition remains to be assessed empirically. The key extensions of the monocentric models explored in Section 8.3 have received scant empirical attention. As just argued, close to nothing is known regarding the effect of income heterogeneity on the various gradients. Similarly, this literature on urban land use has not really taken zoning constraints into account. The work of Cheshire and Sheppard (1995) that integrates planning together with a broader notion of accessibility is a lone exception.⁶¹ As described in Chapter 16 in this handbook, there is a significant literature that deals with the durability and decay of housing, but only little of it is explicitly tied to broader and more formal land use models.⁶²

8.7. PATTERNS OF RESIDENTIAL SORTING WITHIN CITIES

The empirical work discussed in the previous section relates to the framework developed in Section 8.2 which considered homogeneous agents. As argued in Section 8.3.2, many of the predictions of the monocentric model with homogeneous residents carry through to heterogeneous residents. Models with heterogeneous residents also make further important predictions regarding the sorting of residents depending on their observable characteristics.

⁶¹ A key difficulty here is that land use regulations are hard to measure and raise simultaneity problems that have turned out to be extremely hard to deal with. See Chapter 19 in this handbook for further developments on these two issues.

⁶² The work of Brueckner and Rosenthal (2009) is an exception that is discussed further below.

In the simplest case where residents differ only along one dimension (e.g., income), perfect segmentation across income groups is expected. When residents differ along two or more dimensions, only partial sorting is predicted. For instance, a household with high commuting costs may choose to live far from the CBD because it also demands a lot of land. That household may end up being a neighbor of a household with lower commuting costs and a lower demand for land. Income has been of particular interest in the literature because it is a key dimension along which households differ within cities and because it generates an interesting tension. We have seen that, from a theoretical standpoint, the effect of income on distance to the CBD is ambiguous. Higher household income is expected to increase the demand for housing and thus push households further from the CBD. However, since commuting costs are also paid in time, higher income and the associated higher value of time make households want to live closer to the CBD. As shown in [Section 8.3.2](#), richer households will want to live further from the CBD than poorer households if the income elasticity of the demand for housing/land is greater than the income elasticity of commuting costs.

Before we discuss the predictive power of this theoretical proposition, it is important to know what the basic facts are. Looking at the historical record for the United States, [LeRoy and Sonstelie \(1983\)](#) argue that until the second half of the nineteenth century, richer residents were living in more central locations. That pattern reversed afterward with the emergence of the streetcar, followed by the rise of the car. [LeRoy and Sonstelie \(1983\)](#) then argue that some regentrification of central cities occurred after 1970. Their model attributes these changes to the fact that, early on, the streetcar and the automobile were available only to richer households but allowed considerable time savings on commutes. This could have led to the suburbanization of richer households. Eventually, the automobile became available to all households but the poorest ones. This allowed a much broader range of households to decentralize, and rising land prices in the suburbs may have driven richer households to return to central locations. Hence, these broad patterns are consistent with a simple monocentric framework in which the demand for land is relatively inelastic.

More recently, [Lee and Lin \(2013\)](#) have assembled a long-run dataset of neighborhood income and population in major US cities. They confirm that average neighborhood income declined with distance to the CBD in 1880. This income gradient changed sign as early as 1930, and became very steep by 1940. Since 1960 the income gradient has flattened, a process that continues today. These patterns of decentralization and recentralization of wealthier households are much less marked for coastal cities. [Lee and Lin \(2013\)](#) argue that this is due to the better amenities of central locations in coastal cities as stability in neighborhood income strongly correlates with measures of natural amenities.

[Brueckner and Rosenthal \(2009\)](#) also provide detailed evidence for contemporary American cities. Using information at the tract level, they show that there is a tendency for tracts located further from the densest tract to be richer. Importantly, they show that

this tendency is not uniform since tract income flattens and even declines beyond a certain distance from the center. They also show that the income gradient is particularly steep close to the center. The major income gap occurs between central cities and their close suburbs. For large metropolitan areas, the gap between the richest and poorest tracts is about 50%.

This said, this broad characterization is not an absolute norm, far from it. A few American cities have a rich central part. New York is certainly an example. [Glaeser et al. \(2008\)](#) report U-shaped curves plotting incomes as a function of distance to the CBD for New York, Chicago, and Philadelphia. Importantly, as highlighted by [Brueckner et al. \(1999\)](#), European cities tend to have rich cores and poor peripheries, with, again, some exceptions such as Brussels. Adding to this, many cities also exhibit important directionality patterns. In most European cities, western suburbs tend to be richer than eastern suburbs.

It is also important to note that residential areas are far from homogeneous. [Epple and Platt \(1998\)](#) document that in the Boston area median income in the richest municipality was four times as high as that in the poorest municipality. However, nearly 20% of the households in the richest municipality had an income below that of the 20% of richest residents in the poorest municipality. This is indicative of considerable overlap in the income distributions across both municipalities. Hence, while the forces that push toward sorting by income are certainly present, they cannot be overwhelming since spatial sorting by income is far from perfect.

To assess the ability of the monocentric model with heterogeneous residents to explain residential location patterns by income, one needs to estimate both the income elasticity of the demand for housing/land and the income elasticity of the cost of commuting. The first attempt to compare these two elasticities was made by [Wheaton \(1977\)](#), who raised some early skepticism regarding the performance of simple monocentric models along that dimension.

In his work, [Wheaton \(1977\)](#) assumes a utility function with constant elasticity of substitution. He then estimates the coefficients of that utility function using household data collected before the construction of the Bay Area Rapid Transit in the San Francisco metropolitan area. Note that commuting appears both as a disutility and as an expenditure. [Wheaton \(1977\)](#) estimates the income elasticity of the demand for land and the income elasticity of commuting costs to be roughly equal at around 0.25. This suggests that the net effect of income on location within the city stemming from the trade-off highlighted by the monocentric model is likely too small to explain actual residential patterns.

[Glaeser et al. \(2008\)](#) revisit the issue. They first argue that since commuting is mainly paid in time, the income elasticity of commuting costs should be 1 or close to it. Then, they focus their attention on estimating the income elasticity of the demand for land. They first regress the logarithm of parcel area on the logarithm of household income

and estimate an elasticity of the demand for land between 0.1 for single detached houses and 0.3 when some land consumption is ascribed to households living in apartments. Given that their estimate for the income elasticity of the demand for land is well below the unit income elasticity of commuting costs that they assume, they conclude that the monocentric model with heterogeneous residents strongly predicts that richer households should live in the urban cores of US cities. This is obviously counterfactual. As we discuss below, Glaeser et al. (2008) incorporate transit into their analysis to explain why the poor live close to the urban core of American cities. However, before we discuss transit and other alternative explanations, it is worth noting that the conclusions of Glaeser et al. (2008) regarding the relevance of the monocentric model for income sorting are perhaps overly strong. First, their estimate of the income elasticity of the demand for land is likely to be downward biased.⁶³ But more importantly, the income elasticity of commuting costs is likely to be substantially below 1.

While the cost of time is an important component of the cost of commuting, it is not the only one. Consider 1 h of urban driving at a typical speed of 40 km/h. For a car consuming 12.5 l of gasoline per 100 km, this correspond to about \$5.⁶⁴ Adding a depreciation of 12.5 cents per kilometer, the variable cost of the vehicle is \$10 for 1 h.⁶⁵ Consider now a low-skilled worker making \$10 an hour driving this car. This is slightly less than half the median wage. Estimates from transport economics suggest that time in vehicle is valued at about half the wage (Small and Verhoef, 2007). The total cost of driving for this worker is thus \$15 per hour but the cost of time is only a third of this. For this low-skilled worker, the elasticity of the cost of commuting with respect to the wage is only 0.33, which is far from unity. Even for a worker making close to the median wage of \$20 per hour, the wage elasticity of commuting costs is still 0.5. For highly paid workers, workers making close to twice the median wage, \$40 per hour, this elasticity is still 0.66 and maybe not very different from the elasticity of the demand for housing.⁶⁶

⁶³ There are three reasons for this. First, the price of land is missing from their regression of the logarithm of parcel area on the logarithm of household income, and cheaper land in the suburbs is likely to increase demand for land by wealthier households. Second, they focus on instantaneous as opposed to permanent income. When they use education as an instrument for permanent income, the income elasticity of the demand for land rises above 0.5. Third, the consumption of land might be mismeasured if wealthy households choose to live in particularly low density areas with parks and a lot of open space. Considering total land per household instead of lot size raises the income elasticity of the demand for land to close to 0.5.

⁶⁴ We use 2014 prices close to \$4 per gallon. A gasoline consumption of 12.5 l per 100 km is equivalent to about 19 miles per gallon, which is slightly better than the fuel economy of a Ford Taurus.

⁶⁵ This depreciation would apply to a car worth \$25,000 fully depreciating over 200,000 km. This price tag is slightly below that of a Ford Taurus.

⁶⁶ This calculation neglects to consider that this driver may drive a more expensive car. With a car worth \$50,000 and a 20% lower fuel efficiency, we return to an elasticity of 0.55.

Hence, the relevance of the monocentric model with heterogeneous agents to explain patterns of location by income is still very much an open question.⁶⁷ Given a lack of decisive evidence, the literature has proposed a range of complementary explanations. In the spirit of [LeRoy and Sonstelie \(1983\)](#), [Glaeser et al. \(2008\)](#) emphasize the importance of transit in explaining why the poor live close to the urban core of American cities. While the majority of American households can count on a car for each adult, about one-third have only one car and about 10% have no car at all. This suggests that a sizeable minority of American adults must rely on public transport to go to work and do many of their daily errands. Transit is usually present in the core of most American cities, but is often absent from their suburbs. [Glaeser et al. \(2008\)](#) show a strong connection between poverty rates in urban tracts, access to public transport, and closeness to the city center.

While these correlations are interesting, they do not establish any form of causality. [Glaeser et al. \(2008\)](#) buttress the case in favor of public transport by comparing cities with subways and cities without subways. Unsurprisingly in both types of cities, patterns of transit usage differ considerably. Transit usage declines fast with the distance to the center in cities without subways, whereas it increases over the first 5 km and remains on a high plateau for about 10 km in cities with subways. In both types of cities, transit usage is closely negatively mirrored by local median incomes. Given that most subways have been in operation for three generations or more, the difference between subway and non-subway cities has some degree of exogeneity.

[Brueckner et al. \(1999\)](#) focus on the difference between typical American cities, where the poor live in the urban core, and typical European cities, where the core is occupied by the rich, and propose an amenity-based explanation. They develop an extension of the monocentric model with heterogeneous residents who need to commute not only to the CBD but also to enjoy localized amenities. Their main comparison is between Paris and Detroit, and they argue that amenities in Paris are centrally located, whereas, if anywhere, they are located in the suburbs in Detroit. Depending on the assumptions made regarding the demand for land, many equilibrium configurations are possible. When the pull of central amenities is strong enough and the demand for land is not too income elastic, a Paris-type equilibrium is possible, with rich residents located in the core. When the demand for land is not too inelastic, peripheral amenities can induce rich residents to locate in the outer suburbs. Their analysis also allows amenities to be endogenous. They naturally occur where the rich live, and this allows for multiple equilibria since, for instance, the rich can live in the core and generate

⁶⁷ Better estimates of the demand for land would need to solve the three issues raised by [Glaeser et al. \(2008\)](#). Getting better estimates of commuting costs will be difficult since commuting arguably enters the utility function directly. Survey data seem to suggest that commuters have a strong dislike for long commutes but may not want to live exactly where they work ([Krueger et al., 2009](#)).

endogenous amenities such as restaurants which are enough to retain them even though the exogenous amenities such as open space are at the edge of the city.

A first shortcoming of this analysis is that it does not appear to rule out much. Many configurations can be rationalized as an equilibrium outcome that depends on the subtle interplay of features that are hard to observe, such as the presence of amenities and the shape of the demand for land. The model of [Brueckner et al. \(1999\)](#) also remains devoid of explicit empirical evidence. This said, casual evidence nonetheless suggests that there is some explanatory power to this type of explanation. Amenities may also explain an interesting asymmetry observed in many European cities, where, as already mentioned, western suburbs are often much richer than eastern suburbs. This appears to be closely related to winds blowing from the West in those cities. While wind direction no longer constitutes an important issue in these cities, it certainly did at the peak of the industrial era, when a windward location helped the affluent avoid factory emissions and smells.

A third extension of the monocentric model concerns the possible role of the age of the housing stock. [Brueckner and Rosenthal \(2009\)](#) show that, in American cities, controlling for the age of the housing stock explains away the positive income gradient as one moves away from the center. Older constructions, which are located more centrally in American cities, offer lower-quality housing and thus end up being allocated to poorer residents in equilibrium. This is an important element of the filtering mechanism explored in [Section 8.3.3](#). Assessing causality is hard because, to some extent, the state of the housing stock is an endogenous variable. To limit the scope of this problem, [Brueckner and Rosenthal \(2009\)](#) use as an instrument for the age in 2000 of the housing stock in city tracts the same variable in 1980. Using information on the current age distribution of the housing stock and likely reconstruction patterns, they offer 20-year predictions regarding the distribution of the population by income in American cities. These predictions entail further gentrification of urban cores.⁶⁸

While all the explanations examined so far build on the monocentric model, the last one puts much less emphasis on relative locations and the trade-off between accessibility and land prices. It instead relies on public finances and the provision of various public goods in the spirit of [Tiebout \(1956\)](#). The two main ideas are that central cities started engaging in costly social policies and significant redistribution in the 1960s and the 1970s. See, for instance, [Inman's \(1995\)](#) vivid account of these issues in Philadelphia. These policies led well-off residents to flee to suburban municipalities. In turn, poorer urban cores were afflicted by a variety of social ills ranging from crime to failing schools. This led to a further exodus of the middle class from the blight of central cities. This conjecture is consistent with the documented fact that the major income gap is between central cities and surrounding municipalities. This conjecture also provides a natural explanation for the

⁶⁸ See Chapter 16 in this handbook for more on these issues.

differences between Europe, where municipalities often have little fiscal autonomy, and the United States, where education is locally financed and fiscal autonomy is much greater.

There has been quite a lot of work over the years that has thought to evaluate the “flight from blight” conjecture. Among the social ills that have been explored, one may mention crime (Cullen and Levitt, 1999), the degradation of the housing stock (discussed in Chapter 16 in this handbook), racial preferences (Boustan, 2010), and related changes in the school system (Baum-Snow and Lutz, 2011). De Bartolome and Ross (2003) propose a model that embeds aspects of Tiebout fiscal competition into a monocentric framework with two groups of residents, rich and poor. There has been, however, no empirical work to our knowledge that runs a clear horse race between the role of fiscal decentralization and the trade-off between commuting costs and land prices to explain patterns of location by income within cities.

Before closing this section, we need to return to the considerable overlap in income in most areas regardless of how far they are located from the city center. This problem is tackled head on by Epple and Platt (1998) in a nonspatial framework with multiple Tiebout communities. In addition to the usual Tiebout mechanism of catering to the needs to local residents, Epple and Platt (1998) also introduce idiosyncratic preferences for locations. When they quantitatively estimate their model, they find that idiosyncratic preferences explain most of the location decisions of households. This is a disappointing result in that the idiosyncratic preferences introduced by Epple and Platt (1998) are a residual that, in the end, only measures our ignorance. At the same time, this result is consistent with the literature described above starting with Wheaton (1977) which suggests that the difference in the accessibility-land price trade-off for the rich and the poor is unlikely to lead to overwhelming patterns of segmentation.

Bayer and McMillan (2012), using a method developed in Bayer et al. (2011), took the work of Epple and Platt (1998) further by considering explicitly job accessibility and the heterogeneity of the housing stock as determinants of residential choices. Interestingly, they find that the dispersion of jobs and the heterogeneity of the housing stock act as strong brakes on the tendency for households to segregate by race, education, and income. Although they do not explicitly consider the trade-off of the monocentric model and pay minimal attention to the income elasticity of the demand for land, they find that a reduction in commuting costs would lead to large increases in segregation. This is because lower commuting costs make access to jobs a less relevant consideration for households when choosing a residence.

Overall, the evidence indicates a lack of predictive ability of the monocentric model for patterns of household location in cities by income. In itself, this is not a failure of the monocentric model. First, the literature has yet to come up with reliable estimates of the income elasticity of the demand for land and the income elasticity of commuting costs. In cities where the majority of jobs are well outside the urban core, it is also empirically problematic to equate accessibility to jobs (or even general accessibility) to the distance

to the city center. In other words, better work is needed to reduce uncertainty. Second, the best estimates at hand of the two key elasticities for the demand for land and commuting costs are fairly close. This suggests that the monocentric model predicts only weak income effects in location choices. The fact that this matches the empirical reality may not be a huge success, but at least it does not constitute a failure. Third, just like in the attempt to understand geographical patterns of housing prices and population density, the forces highlighted by the monocentric model are arguably not the only ones. This was true for the main housing price versus accessibility trade-off. This is even more so when we consider how this trade-off changes with income. What might be more worrying here is the limited success of the literature at explaining patterns of location choices by income using any explanation. There is good evidence about the effects of the housing stock but disturbingly little regarding public finance. Identifying the effects of local amenities also raises considerable challenges.

8.8. PATTERNS OF RESIDENTIAL LAND DEVELOPMENT

The patterns of residential development in cities have changed substantially over time. These changes are most frequently described as urban sprawl. The term “sprawl” is used with very different meanings, often confusing patterns of residential development characterized by low density and scatteredness with possible causes of such patterns (e.g., car-based commuting) and with possible consequences (e.g., loss of environmental quality). Nevertheless, the most commonly emphasized characteristics of urban sprawl are low-density spread-out development and scattered development (Galster et al., 2001). Figure 8.1 shows just how pervasive is leapfrog development even in an extremely dense city such as Paris. Even as centrally as between 5 and 10 km from the center, 26% of the land remains undeveloped, 18% in the form of urban parks but 8% as unbuilt natural spaces. Between 10 and 20 km from the center, parks take up 15% and unbuilt natural spaces take up 31% of the land.

A first important change in residential patterns has been the gradual displacement of residents away from city centers. A simple measure of the suburbanization of the population compares the population in central cities with the population in the suburban parts of a metropolitan area. Between 1940 and 2000, the share of residents in metropolitan areas of the United States who lived in the central city fell from 56% to 32% (Boustan and Shertzer, 2012). This decline was strongest in the 1940s, but continued over the second half of the twentieth century.

This displacement of the population toward the suburbs has been most often characterized in terms of flatter population density gradients. The work on density gradients beginning with Clark (1951), discussed in Section 8.6, finds that urban population density gradients have declined over time. According to Mills (1972), most US cities have seen their population density gradients decline since 1880. The sharpest decline occurred

during the 1940s, but it has continued at a steady pace since then (Edmonston, 1975). The decline in density gradients has also been observed in many other countries, even if it has been faster and more pronounced in the United States (Mieszkowski and Mills, 1993).

Documenting patterns in the extent to which development is scattered or compact is more difficult. Burchfield et al. (2006) merge data based on high-altitude photographs from 1976 with data based on satellite images from 1992 to create a grid of 8.7 billion 30 m by 30 m cells that tracks the evolution of land use across the United States. These high-resolution data make it possible to observe the amount of open space in the neighborhood of houses in every US city. Thus, they can measure urban sprawl by calculating the amount of undeveloped land in the 1 km neighborhood around the average house in each city. Burchfield et al. (2006) find that residential development in 1992 was no more scattered than in 1976. For the United States as a whole, the proportion of open space in the square kilometer of land surrounding the average house was 42% in 1976 compared with 43% in 1992. While a substantial amount of scattered residential development was built between 1976 and 1992, overall residential development did not become any more biased toward such sprawling areas. On average, areas that were already densely built up in 1976 experienced little change, largely unbuilt areas in the neighborhood of earlier development saw some scattered development, while areas with initial scattered development had the highest rate of new development and became more densely built up in the process. As a result, the total amount of developed land grew substantially, but the proportions of sprawling and compact development remained substantially unchanged.

The nationwide finding that residential development in 1992 is no more scattered than development in 1976 also holds for most individual metropolitan areas. However, this stability over time coexists with large cross-sectional variation: there are very large persistent differences in the extent of sprawl across metropolitan areas. Angel et al. (2012) study the fragmentation of development in a global sample of 120 cities from around the world. In addition to the fragmentation index of Burchfield et al. (2006), they calculate four other measures helping characterize infill and leapfrog development. They find that the proportion of open space in the square kilometer of land surrounding the average development in cities in developed countries was 44% in 1990 and 39% in 2000, figures very similar to those reported by Burchfield et al. (2006) for the United States. Construction in cities from their world sample located in developing countries was more fragmented, with 48% and 43% of open space in the immediate square kilometer in 1990 and 2000, respectively.

Of the two aspects of sprawl we have discussed, falling densities is most easily related to the monocentric model. Lower density is equivalent to a larger amount of developed land for a given population in a city. Brueckner and Fansler (1983) study the comparative statics for the physical size of the city predicted by a closed-city version of the monocentric model. They measure the physical size of the city based on the surface area of urbanized areas in the United States as defined by the 1970 census. These are constructed by aggregating contiguous census blocks with high population density. For data availability

reasons, they focus on 40 small urbanized areas contained within a single county. Their empirical analysis regresses the area of each of these 40 cities on the population, average household income, median agricultural land value per acre for the county, and two alternative proxies for commuting costs (the percentage of households owning one or more cars, which is thought to be higher where commuting costs are low, and the percentage of commuters using urban transit, which is thought to be lower where commuting costs are low). Cities are found to occupy more land when they host a larger population, when average income is higher, and when agricultural land is less valuable. All of these findings conform with the comparative statics of the closed-city version of the monocentric model (the open-city version treats the population as endogenous, but makes similar comparative-statics predictions for income and agricultural land rents). Neither of the two proxies for commuting costs is statistically significant. Several articles have replicated the analysis of [Brueckner and Fansler \(1983\)](#) with data for more cities or other time periods. [McGrath \(2005\)](#) studies large metropolitan areas between 1950 and 1990, while [Paulsen \(2012\)](#), in addition to studying all metropolitan areas regardless of the size for several time periods, also uses data on developed land derived from satellite imagery.

One possible reason why the proxies for commuting costs in [Brueckner and Fansler \(1983\)](#) do not perform well is endogeneity. In a city where residents own more cars, they may be able to travel longer distances more easily, and this can lead to more land being developed. However, it is also possible that the city expands for other reasons and residents buy more cars in response to the larger distances that must be traveled. Most other measures of commuting costs, such as the availability of roads, are subject to similar concerns. [Baum-Snow \(2007a\)](#) addresses these endogeneity concerns while studying another measure of sprawl aiming to capture suburbanization. We saw in [Section 8.2.3](#) that the monocentric model predicts that lower commuting costs are associated with a lower central city population relative to the suburban population.⁶⁹ Baum-Snow regresses the change between 1950 and 1990 in the logarithm of the central city population on a measure of the change in commuting costs, while controlling for a change in the logarithm of the population for the entire metropolitan area and the radius of the central city. His measure of the change in commuting costs is the change in the number of rays of interstate highways that converge toward the central city. Of course, it is possible that the rays of interstate highway going through the city center are as much a consequence as a cause of suburbanization. To tackle this identification issue, [Baum-Snow \(2007a\)](#) uses as an instrument for actual highways crossing cities those that were part of the 1947 interstate highway

⁶⁹ We derived the result in the context of the open-city version of the monocentric model. [Baum-Snow \(2007a\)](#) uses a closed-city version of the monocentric model to motivate his analysis. In the closed-city version, where the population is constant, when a fall in commuting costs flattens the land and house price gradients, each resident consumes more housing and land. This expands the city boundary outward and also (unlike in the open-city version of the model with an endogenous population) reduces density. Suburbanization then follows from the relocation of some former central city residents to the suburbs.

plan. The instrument is relevant because planned rays of interstate highways are a strong predictor of rays that were actually built. At the same time, a reasonable case can be made for exogeneity because the 1947 highway planners were interested in linking US cities together but were not trying to address future commuting patterns. Baum-Snow (2007a) finds that an extra ray of interstate highways leads to a decline in the central city population of about 9%. Interestingly, this instrumental variable estimate is larger than its ordinary least squares counterpart, perhaps because more highways were built in cities that suburbanized less. Glaeser and Kahn (2004) emphasize that the prevalence of cars for commuting has not only reduced commuting costs but has also eliminated the larger fixed cost associated with public transport use, further promoting decentralization.

Baum-Snow's (2007a) finding of a relative decline in the central city population in response to lower commuting costs is consistent with the monocentric model. What is more difficult to reconcile with the monocentric model is the fact that the central city population fell not only in relative terms but also in absolute terms in the United States. Between 1950 and 1990, the population of central cities fell by an average 17%, while the total metropolitan area population rose by 72%. One possible explanation is the simultaneous economy-wide increase in income. In the monocentric city model, it follows from Equation (8.24) that a rise in income affecting all cities equally leaves their populations unchanged. By Equation (8.23), land rent at the CBD is also unchanged. Since land rent at the city edge must still equal the rent paid for land when it is allocated to the best alternative use, if housing is a normal good, the economy-wide increase in income simply makes the house-price gradient flatter. This is consistent with an absolute decline in the central city population and a rise in the suburban population, as residents drift outward to consume more land. Margo (1992) studies the contribution of rising incomes to suburbanization. He first estimates an individual-level regression of the probability of living in a suburban area in 1950 as a function of income while controlling for other variables. He then applies the economy-wide increase in average income to those individual-level estimates and suggests that rising incomes account for about 40% of the suburbanization of US population between 1950 and 1980. Note, however, that this sort of exercise does not include the equilibrium effects in the housing market.

Kopecky and Suen (2010) simultaneously look at the influence of commuting costs and income on suburbanization. They calibrate a monocentric model, extended to allow for two forms of transport (cars and public transport), and perform counterfactual simulations. These suggest that the increasing affordability of cars together with the fall in the fixed and time costs of commuting by car were the major driver of population suburbanization in the United States between 1910 and 1950. However, between 1950 and 1970, rising incomes played a greater role.

In addition to falling commuting costs and rising incomes, a number of different explanations have been proposed for the suburbanization of the population in the United States. Cullen and Levitt (1999) show that increases in crime rates in the city and

decreases in crime rates in the suburbs are associated with more residents leaving the city. They address the endogeneity of changes in crime rates by using as instruments for these lagged changes in the punitiveness of the state criminal justice system.

Boustan (2010) emphasizes another cause of suburbanization, which is specific to the United States: black immigration into northern cities from the South, followed by white flight toward the suburbs. Between 1940 and 1970, 4 million black migrants relocated from the South of the United States. Boustan (2010) estimates that each black arrival was associated with 2.7 white departures from central cities in the North. These may have been motivated by distaste for racial or income diversity. However, it is also possible that white departures had a different motive and, by lowering housing prices, encouraged black immigrants to settle in central cities. To address the endogeneity of black location choices, she uses as an instrument for actual black immigration the variation in local agricultural conditions in southern states to predict black emigration flows and assigns these flows to northern cities on the basis of settlement patterns from an earlier wave of black migration. She also argues that the fact that more than one white resident left for every black arrival indicates that white flight was not just a consequence of black immigration raising house prices. Baum-Snow and Lutz (2011) study the extent to which desegregation in public schools also contributed to suburbanization. They find that while desegregation had a very significant impact on sorting by race within cities, it generated only a small fraction of the decentralization of the urban population between 1960 and 1990.

Besides the decentralization of the residential population, another aspect of urban sprawl is the fragmentation of residential development. Burchfield et al. (2006) study the determinants of urban fragmentation by estimating the relationship between the percentage of open space in the square kilometer around an average new house built between 1976 and 1992 in each metropolitan area of the United States and a host of characteristics for each metropolitan area in 1976. The standard monocentric model can generate scattered development only through larger private open space (a lower capital intensity of development will result in shorter and smaller buildings on larger yards). This kind of development is more likely to take place in cities where commuting costs are lower because it is easier to get around by car. Burchfield et al. (2006) find that development is more scattered in cities built around the automobile (in particular those where streetcars were not widely used in the early twentieth century).

To generate scattered urban expansion in the sense of leapfrog development, leaving some parcels undeveloped while building up other parcels further away, we must either turn to dynamic models as covered in Section 8.3.3 or allow open space to provide some consumption value as in Turner (2005). Regarding dynamic models, we have seen that in the presence of long development lags, greater uncertainty can favor leapfrog development (Bar-Ilan and Strange, 1996). Since developers will likely see future local population growth as more uncertain in cities that have had more ups and downs in population growth rates over previous decades, Burchfield et al. (2006) use as a proxy for uncertainty the

standard deviation of the decennial percentage population growth rates for 1920–1970. As expected, higher uncertainty leads to more sprawl. The value of open space will be greater if open space is likely to remain undeveloped for a long time and if local conditions, such as climate, make being outdoors more enjoyable. [Burchfield et al. \(2006\)](#) show that cities that have been growing more slowly in the past, where undeveloped land is less likely to be converted soon, have more scattered development. This is true even after controlling for contemporaneous population growth rates. Also, cities with a more temperate climate have more scattered development. As we discuss in the following section, employment is often much less centralized than what is assumed in the standard monocentric model. The models of endogenously polycentric cities reviewed in [Section 8.4](#) tell us that more localized agglomeration spillovers will tend to endogenously produce an urban structure with employment centralized in a CBD, as in the monocentric model. Such cities, with strong agglomeration economies, will pay higher wages but also have higher housing and land prices. This will encourage more capital-intensive development. Cities specializing in sectors that, economy-wide, tend to be more concentrated close to city centers indeed tend to have more compact development.

Most models assume a homogeneous landscape. However, natural geography matters for spatial development patterns. [Burchfield et al. \(2006\)](#) examine several aspects of heterogeneity in the natural environment. Mountains are an important aspect of geography. These have been prominent, for instance, in Los Angeles, where the mountains bordering the city have limited further expansion of its sprawling suburbs (a situation described locally as “sprawl hits the wall”). However, in studying the effect on sprawl of mountains more generally, one needs to be careful with two features. First, it is important to focus on mountains in the vicinity of earlier development where they truly act as a barrier to further expansion. Second, one needs to be careful to separate large-scale from small-scale terrain irregularities. Mountains and hills tend to have opposite effects. When an expanding city hits a mountain range, further scattered development in the urban fringe becomes very costly. Thus, high mountains in the urban fringe encourage infilling and lead to increasingly compact residential patterns. On the other hand, small-scale irregularities in the urban fringe have the opposite effect. This is perhaps because rugged terrain encourages scattered development as steep hillsides, where development is more costly, alternate with flat portions, where development is less costly.

Another physical feature with important effects on sprawl is aquifers. Most households in the United States get their water through the nearest municipal or county water supply. However, extending water systems to service new scattered development in the urban fringe requires substantial infrastructure investments, the cost of which is typically borne by developers through connection fees and is ultimately reflected in housing prices. In places where water-yielding aquifers are pervasive, developers can instead sink a well at a fraction of the cost of connecting to the municipal or county water supply. The presence of aquifers is a particularly interesting dimension of underlying heterogeneity in the

physical landscape because of the way it interacts with scale economies: wherever aquifers underlie the urban fringe, household water can be obtained without the large increasing returns associated with public water systems, and this facilitates scattered development.

A final set of determinants of scattered development are of political nature. Two of the main dimensions emphasized in public discussions—competition between municipalities of different sizes and the extent of municipal fragmentation—do not matter for sprawl in practice. Municipal boundaries matter, but for different reasons. [Burchfield et al. \(2006\)](#) find that a disproportionate share of 1976–1992 development happened in unincorporated areas that were close to existing development but just beyond the municipal boundaries at the beginning of the period, where land use regulation was laxer. This development is also more dispersed than that on incorporated land. Finally, one of the common complaints about urban sprawl is that as development spreads, municipal services such as roads, sewers, the police, and fire protection are more expensive. Indeed, when a smaller fraction of local expenditures is financed through transfers from other levels of government as opposed to local taxes, development tends to be less scattered. This suggests that when local taxpayers are held accountable for infrastructure costs, they respond by insisting on more compact patterns of development that require less infrastructure spending.

In their study of fragmentation in a global sample of cities, [Angel et al. \(2012\)](#) find that cities that are smaller and had more buildable land around them have more fragmented development, which is consistent with an initial phase where a city grows in a fragmented way, followed by gradual infilling of undeveloped land as the city continues to grow. Like [Burchfield et al. \(2006\)](#) for the United States, they also find that more geographical constraints lead to more compact development, while aquifers providing well water encourage fragmentation. One of their more interesting findings is that cities in countries with greater car ownership have more compact development, after income differences have been controlled for. Greater car ownership can facilitate commuting over long distances, which would seem to favor leapfrog development. However, public transport relies on stations and bus stops and is likely to encourage development to be concentrated around these. By enabling easy travel between any two points, car ownership can in fact favor infilling. [Angel et al. \(2012\)](#) suggest that their result arises because the latter effect dominates.

If urban sprawl has received so much attention in the popular press, it is mostly because it is presumed to have important consequences. Most prominently, sprawling residential development is presumed to encourage car use. [Bento et al. \(2005\)](#) study the effects of various measures of urban form and centralization on transport mode choices and distance traveled. They find that individual measures of urban form have only a modest impact on the probability of commuting by car and distance traveled. However, when multiple aspects of urban form are considered simultaneously, the effects are much larger. For instance, they calculate that changing Atlanta to look like Boston in terms of

the central location of its population, the shape of the city, and the availability of public transport would reduce the total distance traveled by car by 25%. Glaeser and Kahn (2004) suggest that by making commutes more car dependent, sprawling cities may hurt poorer households, who are much less likely to be able to own a car for every working adult.

In addition to favoring car use, sprawl is also sometimes thought to reduce social interactions (Putnam, 2000) and to encourage unhealthy behavior increasing the prevalence of obesity (Ewing et al., 2003; Giles-Corti et al., 2003; Saelens et al., 2003; Frank et al., 2004). Regarding social interactions, empirical studies show that social interactions are, in fact, greater in suburban areas and in areas with lower residential density (Glaeser and Gottlieb, 2006; Brueckner and Largey, 2008). Regarding obesity, people living in sprawling neighborhoods tend to be heavier than those living in neighborhoods where development is compact and there are plenty of shops and amenities within walking distance. However, using panel data tracking the characteristics of people's residential locations and their weight over time, Eid et al. (2008) show that this correlation between sprawl and obesity is fully accounted for by sorting. Once they take advantage of the panel dimension of their data to control for unobserved propensity to be obese, the correlation between obesity and sprawl vanishes: changes in neighborhood characteristics do not lead to changes in weight. People in sprawling neighborhoods are heavier because the same characteristics that make them obese (e.g., a distaste for walking) make them prefer to live in sprawling neighborhoods (e.g., so that they can get around by car and not walk).

8.9. EMPLOYMENT DECENTRALIZATION AND PATTERNS OF BUSINESS LOCATION CHANGES WITHIN CITIES

Section 8.8 discussed the causes and manifestations of residential decentralization in cities. Decentralization, however, did not only affect patterns of residential locations. It was also followed and accompanied by the decentralization of employment and economic activity.

As already mentioned, Glaeser and Kahn (2001) report that only 24% of jobs in US metropolitan areas in 1996 were located within 3 miles (about 5 km) of the CBD. Consistent with this, Baum-Snow (2014) reports that about 24% of workers in 2000 worked in the central city of one of the largest 100 metropolitan areas in the United States. Second, although residential decentralization has been stronger (Glaeser and Kahn, 2001), the decentralization of jobs has been considerable. Baum-Snow (2014) reports that for the 100 largest US metropolitan areas, the employment share of central cities went from 61% in 1960 to 34% in 2000. Third, the decentralization of jobs started after the decentralization of residents, but the high correlation between employment and residential decentralizations across cities makes it difficult to separate the two. Fourth, since jobs were initially more centralized than residents, they have become even more centralized

in relative terms.⁷⁰ Fifth, although job decentralization affected most sectors, it has been relatively stronger for manufacturing than for services (Baum-Snow, 2014), and less-skilled jobs have decentralized more than high-skilled and managerial jobs (Rossi-Hansberg et al., 2009). Finally, there is also considerable heterogeneity in the patterns of employment decentralization. Some cities such as Los Angeles have become extremely decentralized, whereas others such as New York have been far less affected.⁷¹

While these broad facts are well established, little is known about the details of the spatial patterns of decentralized employment. As reported above, there is evidence of the emergence of new subcenters. Following a pattern initially and vividly described by Garreau (1991) and, more formally, by Henderson and Mitra (1996), a fraction of CBD employment went to subcenters that, as a first approximation, may be viewed as miniature replicas of the main CBD.

Using the subcenter detection method proposed by McMillen (2001), McMillen and Smith (2003) examine the existence of subcenters in large US cities inspired by the model of Fujita and Ogawa (1982) discussed in Section 8.4. For 62 large American metropolitan areas with a mean population of 2.1 million, they detect an average of 2.7–4.4 subcenters per metropolitan area depending on the stringency that they apply to the detection of subcenters. Importantly, they find that cities with more population and a higher travel time index have more subcenters. These two variables alone account for 80% of the variation in the number of subcenters in their sample.⁷²

Urban employment decentralization is not only about the emergence of subcenters. Employment has also decentralized in a “diffuse” manner. In fact, subcenter formation and diffuse employment decentralization should not be seen as a binary dichotomy. Reality is about a continuum ranging from small isolated facilities, to groups of several offices in a strip-mall, to small industrial parks with a couple hundred workers, to full-fledged business subcenters with tens of thousands of employees. Although diffuse employment is hard to define precisely and measure empirically, casual evidence suggests it perhaps accounts for a large part of decentralized employment.

As argued in Section 8.4, firms face a complicated trade-off when they choose to decentralize. The equilibrium layout of the city in Ogawa and Fujita (1980), as captured

⁷⁰ Put differently, the gradients of employment density were initially steeper than those of residential density. Both types of gradients have become much flatter. Since residential density has flattened more, employment gradients have become relatively steeper.

⁷¹ The findings of Glaeser and Kahn (2001) are also suggestive of a role of institutional fragmentation with more decentralization in more fragmented metropolitan areas, a weak role of city age, with more recent cities being more decentralized, and no role for city demographics.

⁷² While these results are interesting, they should be taken as a description of the data but not be given a causal interpretation. The number of subcenters is arguably determined simultaneously with the metropolitan population and with a travel time index. Appropriate identification of causal effects is still missing here.

in Equations (8.69) and (8.70), features three key parameters: the spatial decay of interactions, the intensity of land use in production, and commuting costs. By moving away from central locations, firms lose some agglomeration benefits, the more so the steeper the spatial decay of interactions. At the same time, commercial real estate in central locations is more expensive. Hence, by moving away from the center, firms lower their land costs. This represents a larger saving the higher the land intensity of production. Finally, if firms concentrate in a CBD, they force their workers to commute to this center instead of working locally. To offset those commuting costs, firms must pay higher wages if they are centrally located. This complicated trade-off is summarized by the bid-rent curve for land in commercial use. If agglomeration economies are very local, production is not very land intensive, and commuting costs are not too high, the bid-rent curve for commercial land features a central peak putting it above the bid-rent curve for land in residential use, as in [Figure 8.5](#). If agglomeration economies spread out more easily, production is more land intensive, or commuting is more costly, firms become decentralized, first through the emergence of subcenters as in [Figure 8.4](#) and then through the diffusion of employment throughout the city.

The model of [Ogawa and Fujita \(1980\)](#) features a single production sector. However, its comparative statics are informative of which sectors are more likely to become decentralized: those with less localized agglomeration economies and those that use land less intensively. Thus, manufacturing plants would be among the first to spread out, while, for instance, financial services are likely to remain concentrated in the CBD. Commuting costs are likely to matter more for comparisons across cities and countries than across sectors, with lower commuting costs favoring decentralization. Decentralization raises further accessibility issues when the new location is difficult to access with public transport. In this respect, we expect major differences between North America and the rest of the world, where the car is less prevalent in commutes. Job decentralization, especially in its more diffused forms, is arguably much easier when the vast majority of workers commute by car.

The standard monocentric model characterizes a number of important gradients. However, by assuming a single central location for all firms, it features a single wage for each occupation. As we saw in [Section 8.4](#), once we allow for the endogenous decentralization of jobs, an immediate new feature is the existence of wage gradients. Early evidence consistent with the existence of wage gradients was provided by [Madden \(1985\)](#) and [McMillen and Singell \(1992\)](#). It is summarized in [Anas et al. \(1998\)](#). As noted by [Manning \(2003\)](#), extant work on wage gradients mostly establishes that workers with longer commutes receive higher wages. This is consistent with decentralized firms offering lower wages to their workers and capturing the gains from better access. However, in a monocentric framework, a positive correlation between wages and commuting distance is also consistent with the workers that receive higher wages choosing to live further from the CBD. As shown in [Section 8.3.2](#), this is a standard prediction of the

monocentric model extended to multiple income groups. Finally, as underscored by Manning (2003), a positive correlation between commuting distance and wages is also consistent with spatially dispersed firms having some monopsony power. With this type of labor market friction, employers need to offer higher wages to workers who commute from further away irrespective of their location.⁷³ Manning (2003) provides a range of evidence consistent with the latter feature.

Rather than examine the wage received by workers depending on where they live, evidence for wage gradients should be sought using information about the wages offered by employers depending on how far they are from their workforce and how their workers are distributed in the city. Since the location choices of both employers and workers are likely to be simultaneously determined, this is a challenging problem. Even with a plausible identification strategy, getting significant results will be difficult. The models with decentralized employment we have examined assume that, all else being equal, workers commute to the closest workplace. In reality, workers do not go to the closest job. They commute much more, perhaps a full order of magnitude more than indicated by the work–residence assignment that would minimize aggregate commuting. This fact is usually referred to as “wasteful” or “excess” commuting and was originally evidenced by Hamilton (1982) (see Ma and Banister, 2006, for a more recent survey). The adjective “wasteful” is probably misleading since there are many reasons why workers do not work for the employer closest to their residence. These reasons include preferences for specific local amenities, schools in particular, the fact that labor is far from being a homogeneous commodity, more complex location optimization for dual earners, and the costs associated with changing residence after changing job. Interestingly and despite strong job decentralization, the commuting time of workers has remained fairly stable (Gordon et al., 1991; Levinson and Wu., 2005), a fact dubbed the “commuting time paradox.” Proper integration of these features into a consistent theoretical framework that would form the basis for subsequent empirical work is still missing.

Aside from wage gradients, a second prediction of employment decentralization is a flattening of land and housing price gradients as employment decentralizes. As already mentioned, McMillen (1996) documents a flattening of the land price gradient for Chicago over a long period. Further evidence about declining gradients is discussed in Anas et al. (1998). The issue with declining gradients is that they could also be caused by lower commuting costs in the absence of decentralization.

Another implication of employment decentralization is that greater job dispersion will make it harder for some workers to access jobs. In particular, diffuse job decentralization makes it harder for workers with no car to reach potential employers once these

⁷³ If the labor market for each firm is thin enough, its workers will be widely dispersed across the city. Then, it could even be that firms further away from the CBD need to offer, on average, higher wages because of worse accessibility. This would then imply an inverse wage gradient.

are located far from the CBD. This is the spatial mismatch hypothesis, first proposed by [Kain \(1968\)](#). According to this hypothesis, in the US context, the departure of jobs from central cities made it much harder for minorities to find employment since these minorities have traditionally resided in the more central parts of cities. Whether this increased disconnect between residence and jobs is a major explanatory factor for the plight of minorities in the United States is still an open question. There is no doubt that, as reported above, a decrease in the employment share of central cities from 57% to 47% over 20 years is a major change. These figures probably understate the true extent of the change for minority workers since, as already stated, low-skilled manufacturing jobs decentralized much more than high-skilled jobs, which in many cases grew in central cities ([Gobillon et al., 2007](#); [Rossi-Hansberg et al., 2009](#)). Given that minority residents are overwhelmingly over represented in central cities, their access to jobs diminished, perhaps considerably so.

A key complication in the empirical analysis of the spatial mismatch hypothesis is that a variety of mechanisms are likely to be at play. As noted above, greater distance may reduce both the efficiency of the search for jobs and its intensity. Poor accessibility may also make workers less likely to accept job offers since the wage net of commuting costs is more likely to fall below some reservation level. Finally, minority workers may also suffer from various forms of discrimination in predominantly white suburban areas ranging from customer-driven discrimination in service jobs to redlining and statistical discrimination afflicting workers coming from deprived central areas with high crime rates.⁷⁴

Because of this variety of mechanisms, it is hard to get a sense of the importance of the spatial mismatch hypothesis from the studies that examine individual channels.⁷⁵ Some empirical studies of the spatial mismatch hypothesis have taken a more aggregated approach. In one of the most careful explorations in this literature, [Raphael \(1998\)](#) estimates two sets of regressions for a cross section of neighborhoods in the San Francisco metropolitan area. In the first, he regresses neighborhood labor market participation rates on measures of job accessibility and other neighborhood characteristics as controls. The key finding is that the coefficient on the share of black residents drops after accessibility has been controlled for. Hence, much of the lower employment rate of blacks in San Francisco neighborhoods appears to be accounted for by the poor accessibility of these neighborhoods. To confirm this, [Raphael \(1998\)](#), in the second set of regressions, uses employment rates across neighborhoods, but this time, by race. His decomposition of racial employment rate differentials also suggests an important role for accessibility. Overall, he concludes that about 20% of the employment rate differentials can be attributed to differences in job accessibility. The key identifying assumption is that residential choices are exogenous. To minimize problems, he looks only at young workers, below the age of

⁷⁴ See [Zenou \(2009\)](#) for a theoretical exposition of these various channels.

⁷⁵ The various threads of the literature about specific channels are discussed in [Gobillon et al. \(2007\)](#).

19 years, who disproportionately live with their parents. This, however, may not help much with the main identification challenge if employability of youths is highly correlated with the employability of their parents, who choose their residential location on that basis.

Even if we ignore identification problems, it is unclear what this type of result implies in terms of policy. One might want to bring the residents to the jobs, bring the jobs to the residents, or help central city residents access suburban jobs. While a full discussion of the evidence on those issues would take this chapter too far, note that the literature on place-based policies is often skeptical about bringing jobs to central city residents as argued in Chapter 18 in this handbook. The results from the rare cases where residents from poor neighborhoods were relocated to richer areas are not supportive of large employment benefits either (see, e.g., [Katz et al., 2001](#)). Improving access might be more promising provided major transit investments with poor returns are avoided. In this respect, subsidizing private transport is sometimes suggested ([Glaeser and Kahn, 2004](#)), but this possibility does not receive much interest from policy makers.

Turning to the causes of employment decentralization, several conjectures have been proposed in the literature. As with residential decentralization discussed in [Section 8.8](#), changes in transport obviously loom large ([Anas et al., 1998](#); [Glaeser and Kahn, 2004](#); [Baum-Snow, 2007a](#)). Most extant empirical work focuses on the effect of transport on residential decentralization. This work is discussed in [Section 8.8](#). [Baum-Snow et al. \(2013\)](#) study the effect of transport networks on both the decentralization of the population and the decentralization of specific types of economic activity in China. They use historical transport networks as an exogenous source of variation. They find that railroads and roads had a strong effect on population decentralization. Production in industries with high weight-to-value ratios did not respond to transport improvements, while production in lighter industries did. [Baum-Snow \(2014\)](#) studies residential and employment decentralization for the United States. He uses some plausible exogeneity associated with the construction of US interstate highways to identify the effect of new radial roads on both employment and residential decentralization. He finds that new roads had a much larger effect on residential decentralization and uses a model to back out the importance of agglomeration effects that retain firms in the more central locations of cities.

In addition, the same social ills that have pushed many residents outside central cities in the United States may also have been responsible for the departure of businesses. These ills were briefly discussed in [Section 8.8](#).

There is a factor that is nonetheless uniquely important to firms: advances in communication and computing technologies have made it much easier for firms to separate their activities across various sites. In particular, firms have been able to separate administrative functions from production, and within administrative functions high-end and front-office activities from back-office work. Such separation is expected to occur when the cost of separating activities and the possible loss of agglomeration for the separated activity is more than offset by lower production costs in a cheaper location. A large share of these

separations involves relocating back-office work or production to different countries (Markusen, 2002) and different cities (Duranton and Puga, 2005; Davis and Henderson, 2008; Liao, 2012). In many cases however, close coordination between activities is still needed and production or back-office labor moves to the fringe of the city, while management remains in the CBD.

This hypothesis was first articulated by Ota and Fujita (1993) and later by Rossi-Hansberg et al. (2009). It is also discussed in Anas et al. (1998). Unfortunately, there is no empirical work that goes beyond documenting broad descriptive facts to substantiate the importance of the telecommunication revolution as an explanatory factor for the decentralization of employment.

Unlike what the material in Section 8.8 and in this section may implicitly suggest, cities are not only about change. Much of the recent literature actually suggests there is a lot of persistence in urban patterns. Bleakley and Lin (2012) document the persistence and continuous development of many cities in the United States located along a fall line. These cities initially specialized in the portage of merchandise traveling by boat but remained and prospered even after their main economic justification disappeared. In a different vein, Davis and Weinstein (2008) show that Japanese cities recovered in population terms and redeveloped the same economic specializations after World War II even though many of those cities were heavily bombed.

In an interesting exploration of the evolution of subcenters in Los Angeles, Redfearn (2009) shows that the most important predictor of clusters of decentralized jobs in 2000 in a location is the presence of decentralized jobs in the same location in 1980. This strong stability in the location patterns of firms is surprising against a backdrop of a population growth of more than 25% during the period, ethnic change, and strong deindustrialization. Even more interesting, as noted by Redfearn (2009), old highway networks are more relevant in explaining the current spatial distribution of employment than is the modern highway system.

Brooks and Lutz (2013) present related evidence also for Los Angeles. They examine to what extent streetcars, whose importance culminated in the 1910s and which were completely gone by the early 1960s, still affect land use patterns. They document strong negative and persistent correlations between the distance to old streetcar tracks and contemporaneous population density or property prices. These correlations became stronger over time and hold even after conditioning out contemporary transit and measures of locational quality.

This suggests an important role for durable housing and persistence in zoning patterns in explaining why urban decentralization did not change cities even more than it did. Further indirect evidence supportive of this idea is proposed by Siodla (2014), who documents changes in the city of San Francisco after the fires that followed the 1906 earthquake. These fires destroyed large parts of the city. He shows that, at the border of fires, the density subsequently increased by 40% on the side that burnt down relative to the side that did not burn.

The bottom line of any conclusion on residential and employment decentralization is that much remains to be done. A first set of priorities is to document facts. For employment decentralization, the best account is still arguably the one provided by Glaeser and Kahn (2001). It is nearly 15 years old and much of it relies on county-level information. With lots of more data available at much finer spatial scales, a more precise and up-to-date documentation of the patterns of employment decentralization should be within reach. It is also disconcerting that the overwhelming majority of the little we know about employment decentralization concerns only one country, the United States, which is arguably an outlier.

While the drivers of residential decentralization have begun to be explored, much remains to be done. There is very little on employment decentralization and on the relationship between employment and residential decentralization. While the exploration of some drivers of urban decentralization will benefit from clever identification strategies relying on transport innovations or exogeneity in institutional change, some broader issues will probably need to rely on modeling assumptions to be able to back out key parameters from observable moments of the data. There is an emergent body of work in that direction. One could cite the articles by Ahlfeldt et al. (2012), Baum-Snow (2014), or Brinkman et al. (2012) as examples of the current frontier. A lot of the evidence also points at the persistence created by the durability of commercial and residential properties and the persistence of zoning. A better understanding of how houses and buildings are built, maintained, and torn down and how zoning regulations affect this process are clearly important avenues for future research.⁷⁶

8.10. CONCLUSION

This chapter aimed to selectively review the large literature on urban land use. The theoretical literature is well developed and has provided numerous important insights. It mostly builds on the monocentric structure initially developed by Alonso (1964), Mills (1967), and Muth (1969). Even in the simple version presented in Section 8.2, the monocentric model is an outstanding piece of theory which delivers subtle and non-trivial results from the interplay of fairly straightforward and natural assumptions. Higher housing prices in some locations arise to offset better accessibility. In turn, higher housing prices are reflected in higher land prices, and affect housing consumption, the intensity of housing development, and the number and characteristics of residents locally.

The monocentric model faces three key challenges. The first is heterogeneity of users of land. While, as seen in Section 8.3, much attention has been devoted to broad trends

⁷⁶ A detailed review of the literature on this is beyond our scope here. About the production function for single family homes, we already mentioned Epple et al. (2010b) and Combes et al. (2014). See also Albouy and Ehrlich (2012). For larger buildings and differentials in construction costs, see Glaeser et al. (2005) and Gyourko and Saiz (2006). Finally, see Dye and McMillen (2007) for tear-downs.

such as the average income of residents in relation to the distance to the CBD, much less effort has been devoted to understanding heterogeneity in parcel size, housing, and local residents in different locations. Existing models tend to predict extreme forms of sorting, while actual patterns of land use tend to be smoother. Understanding what drives this local heterogeneity and its implications is still part of the agenda for future research. With much recent progress elsewhere in the discipline in dealing with heterogeneity, this is a task that seems feasible. The second main challenge is to accommodate more fully the durability of housing. Here again, tractability issues loom large, although progress in computational economics is likely to help. The third modeling challenge seems much harder. For more than 30 years urban economists have struggled to accommodate more complex spatial structures that involve the endogenous formation of centers and patterns of mixed land use. As we showed in [Section 8.4](#), these models are both fragile and difficult to manipulate.

Relative to theory, empirical work is less advanced. The first problem has been one of data availability. For a long time, the best empirical research could do was to combine aggregate statistics for urban tracts with maps and attempt to uncover some patterns predicted by theory, usually some sort of gradient. The data environment has changed dramatically in the last 15 years. It went from being extremely data poor to incredibly data rich. Landsat data recently released by the US Geological Survey offer land cover information for the entire world for pixels of 30 m by 30 m since the mid-1970s. This represents more than 500 billion observations for a single cross section.

Better data will certainly help produce better and more informative descriptive work. For many first-order questions, we still lack basic facts. Very little is known about how land is being used in cities. Models often assume that all noncentral land is residential, but commercial land, open space, and roads and parking may use three-quarters or more of the land of contemporary cities. As highlighted in [Sections 8.8](#) and [8.9](#), much of our knowledge of urban decentralization and urban sprawl in the United States derives from a tiny number of articles. There is a wealth of detailed studies about small areas but very little regarding broad cross sections of cities and even less about their evolution. Also, the vast majority of what is known about land use concerns American cities. Much less is known about other cities from the developed world, although they appear to differ greatly from American cities in a number of aspects. Our knowledge of land use in developing cities is even more rudimentary.

However, data availability is not and should not be everything. The second major problem of empirical investigations of land use in cities is that, although they are often inspired by theory, often they are only loosely connected to the models they claim to draw from. For instance, empirical predictions of the monocentric model are examined qualitatively and not quantitatively. Gradients of property prices are then exhibited but without much notion of whether they are quantitatively consistent with the models. We certainly call for a tighter connection between theory and empirical work in future research.

Third, identification concerns are often ignored. To return to the example of property price gradients, alternative explanations for these gradients not based on the accessibility-price trade-off are usually ignored. Hence, although we would like theory to be taken more seriously in empirical work, the limitations of extant theory and the empirical implications of these limitations also need to be better understood and recognized. Hopefully this chapter will be helpful in this respect.

Fourth, land use is often studied in isolation. While land use is of clear interest on its own, it should also be studied in relation to other issues. The most obvious connection is with transport. Although the land use and transport nexus is at the core of the monocentric model, most of the empirical work on land use ignores transport or treats it superficially with low-quality data such as ad hoc measures of congestion. As made clear by Chapter 20 in this handbook and in [Sections 8.8](#) and [8.9](#), recent advances in the transport literature are starting to provide useful insights into the drivers of land use. Land use and changes in land use are also expected to reflect technological change since the location choices of firms and households depend on existing technologies and the availability of substitutes for travel. Aside from some broad trends relating to the importance of the automobile and changes in what firms do and how they are organized, our knowledge here is still extremely sparse. As shown by Chapter 13 in this handbook, urban and real estate research is developing a more detailed and sophisticated understanding of the land and property markets. However, the spatial and land use implications of this better understanding of how sales of land and properties are transacted are still mostly unexplored.

To conclude, while in this chapter we have highlighted the deficiencies of extant theory, we believe that the main priority for future research on urban land use should be empirical work.

ACKNOWLEDGMENTS

We thank Frédéric Robert-Nicoud for his early comments and several novel insights and Frédéric Gilli for his help with the data. We are also grateful to Will Strange, Vernon Henderson, Dan O’Flaherty, and Steve Ross for their detailed comments. Gilles Duranton gratefully acknowledges funding from the Zell Lurie Real Estate Center at Wharton. Diego Puga gratefully acknowledges funding from the European Union’s Seventh Framework Programme under ERC Advanced Grant agreement 269869.

REFERENCES

- Ahlfeldt, G., 2011. If Alonso was right: modeling accessibility and explaining the residential land gradient. *J. Reg. Sci.* 51, 318–338.
- Ahlfeldt, G.M., Redding, S.J., Sturm, D.M., Wolf, N., 2012. The Economics of Density: Evidence from the Berlin Wall. Processed, London School of Economics.
- Albouy, D., Ehrlich, G., 2012. Metropolitan land values and housing productivity. Working paper 18110, National Bureau of Economic Research.
- Alonso, W., 1964. *Location and Land Use; Toward a General Theory of Land Rent*. Harvard University Press, Cambridge, MA.

- Anas, A., 1978. Dynamics of urban residential growth. *J. Urban Econ.* 5, 66–87.
- Anas, A., 1990. Taste heterogeneity and urban spatial structure: the logit model and monocentric theory reconciled. *J. Urban Econ.* 28, 318–335.
- Anas, A., Moses, L.N., 1979. Mode choice, transport structure and urban land use. *J. Urban Econ.* 6, 228–246.
- Anas, A., Arnott, R., Small, K.A., 1998. Urban spatial structure. *J. Econ. Liter.* 36, 1426–1464.
- Anenberg, E., Bayer, P., 2013. Endogenous Sources of Volatility in Housing Markets: The Joint Buyer-Seller Problem. Processed, Duke University.
- Angel, S., Parent, J., Civco, D.L., 2012. The fragmentation of urban landscapes: global evidence of a key attribute of the spatial structure of cities, 1990–2000. *Environ. Urban.* 24, 249–283.
- Arnott, R., Inci, E., 2006. An integrated model of downtown parking and traffic congestion. *J. Urban Econ.* 60, 418–442.
- Arnott, R.J., Stiglitz, J.E., 1981. Aggregate land rents and aggregate transport costs. *Econ. J.* 91, 331–347.
- Arnott, R., Davidson, R., Pines, D., 1983. Housing quality, maintenance and rehabilitation. *Rev. Econ. Stud.* 50, 467–494.
- Asami, Y., Ohtaki, T., 2000. Prediction of the shape of detached houses on residential lots. *Envir. Plann. B* 27, 283–296.
- Asami, Y., Fujita, M., Smith, T.E., 1991. On the foundations of land use theory: discrete versus continuous populations. *Reg. Sci. Urban Econ.* 20, 473–508.
- Bar-Ilan, A., Strange, W.C., 1996. Urban development with lags. *J. Urban Econ.* 39, 87–113.
- Baum-Snow, N., 2007a. Did highways cause suburbanization? *Q. J. Econ.* 122, 775–805.
- Baum-Snow, N., 2007b. Suburbanization and transportation in the monocentric model. *J. Urban Econ.* 62, 405–423.
- Baum-Snow, N., 2014. Urban Transport Expansions, Employment Decentralization, and the Spatial Scope of Agglomeration Economies. Processed, Brown University.
- Baum-Snow, N., Lutz, B.F., 2011. School desegregation, school choice and changes in residential location patterns by race. *Am. Econ. Rev.* 101, 3019–3046.
- Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M.A., Zhang, Q., 2013. Roads, Railroads and Decentralization of Chinese Cities. Processed, Brown University.
- Bayer, P., McMillan, R., 2012. Tiebout sorting and neighborhood stratification. *J. Publ. Econ.* 96, 1129–1143.
- Bayer, P., McMillan, R., Rueben, K., 2011. An Equilibrium Model of Sorting in an Urban Housing Market. Processed, Duke University.
- Becker, G.S., 1965. A theory of the allocation of time. *Econ. J.* 75, 493–517.
- Beckmann, M.J., 1969. Distribution of urban rent and residential density. *J. Econ. Theory* 1, 60–67.
- Behrens, K., Combes, P.P., Duranton, G., Gobillon, L., Robert-Nicoud, F., 2015. Household Sorting in the City. University of Pennsylvania, Wharton School.
- Bento, A., Cropper, M.L., Mobarak, A.M., Vinha, K., 2005. The effects of urban spatial structure on travel demand in the United States. *Rev. Econ. Stat.* 87, 466–478.
- Berliant, M., 1985. Equilibrium models with land: a criticism and an alternative. *Reg. Sci. Urban Econ.* 15, 325–340.
- Bertaud, A., Malpezzi, S., 2003. The Spatial Distribution of Population in 48 World Cities: Implications for Economies in Transition. Processed, The Center for Urban Land Economics Research, University of Wisconsin.
- Bertaud, A., Renaud, B., 1997. Socialist cities without land markets. *J. Urban Econ.* 41, 137–151.
- Bleakley, H., Lin, J., 2012. Portage and path dependence. *Q. J. Econ.* 127, 587–644.
- Boustan, L.P., 2010. Was postwar suburbanization “white flight”? Evidence from the black migration. *Q. J. Econ.* 125, 417–443.
- Boustan, L., Shertzer, A., 2012. Population trends as a counterweight to central city decline, 1950–2000. *Demography* 50, 125–147.
- Braid, R.M., 1981. The short-run comparative statics of a rental housing market. *J. Urban Econ.* 10, 286–310.

- Braid, R.M., 2001. Spatial growth and redevelopment with perfect foresight and durable housing. *J. Urban Econ.* 49, 425–452.
- Breton, A., 1973. Neighborhood selection and zoning. In: Hochman, H. (Ed.), *Issues in Urban Public Economics*. Institut International de Finance Publique, Saarbrücken, pp. 241–251.
- Brinkman, J., 2013. Congestion, Agglomeration, and the Structure of Cities. Processed, Federal Reserve Bank of Philadelphia.
- Brinkman, J., Coen-Pirani, D., Sieg, H., 2012. Estimating a Dynamic Equilibrium Model of Firm Location Choices in an Urban Economy. Processed, Federal Reserve Bank of Philadelphia.
- Brooks, L., Lutz, B., 2012. From Today's City to Tomorrow's City: An Empirical Investigation of Urban Land Assembly. Processed, Trachtenberg School of Public Policy and Public Administration, George Washington University.
- Brooks, L., Lutz, B., 2013. Vestiges of Transit: Urban Persistence at a Micro Scale. Processed, Trachtenberg School of Public Policy and Public Administration, George Washington University.
- Brueckner, J.K., 1980. A vintage model of urban growth. *J. Urban Econ.* 8, 389–402.
- Brueckner, J.K., 1981. A dynamic model of housing production. *J. Urban Econ.* 10, 1–14.
- Brueckner, J.K., 1986. A switching regression analysis of urban population densities. *J. Urban Econ.* 19, 168–198.
- Brueckner, J.K., 1987. The structure of urban equilibria: A unified treatment of the Muth–Mills model. In: *Handbook of Regional and Urban Economics*, vol. 2. Elsevier, North-Holland, pp. 821–845.
- Brueckner, J.K., 2000. Urban growth models with durable housing: An overview. In: Huriot, J.M., Thisse, J.F. (Eds.), *Economics of Cities: Theoretical Perspectives*. Cambridge University Press, Cambridge, pp. 263–289.
- Brueckner, J.K., Fansler, D.A., 1983. The economics of urban sprawl: theory and evidence on the spatial sizes of cities. *Rev. Econ. Stat.* 65, 479–482.
- Brueckner, J.K., Largey, A.G., 2008. Social interaction and urban sprawl. *J. Urban Econ.* 64, 18–34.
- Brueckner, J.K., Rosenthal, S.S., 2009. Gentrification and neighborhood housing cycles: will America's future downtowns be rich? *Rev. Econ. Stat.* 91, 725–743.
- Brueckner, J.K., Thisse, J.F., Zenou, Y., 1999. Why is central Paris rich and downtown Detroit poor? An amenity-based theory. *Eur. Econ. Rev.* 43, 91–107.
- Brueckner, J.K., Thisse, J.F., Zenou, Y., 2002. Local labor markets, job matching, and urban location. *Int. Econ. Rev.* 43, 155–171.
- Burchfield, M., Overman, H.G., Puga, D., Turner, M.A., 2006. Causes of sprawl: a portrait from space. *Q. J. Econ.* 121, 587–633.
- Capozza, D.R., Helsley, R.W., 1989. The fundamentals of land prices and urban growth. *J. Urban Econ.* 26, 295–306.
- Capozza, D.R., Helsley, R.W., 1990. The stochastic city. *J. Urban Econ.* 28, 187–203.
- Carrillo, P.E., 2012. An empirical stationary equilibrium search model of the housing market. *Int. Econ. Rev.* 53, 203–234.
- Chatterjee, S., Eyigungor, B., 2014. Agglomeration Economies, Geography and the Value of Urban Land. Processed, Federal Reserve Bank of Philadelphia.
- Cheshire, P., Sheppard, S., 1995. On the price of land and the value of amenities. *Economica* 62, 247–267.
- Clark, C., 1951. Urban population densities. *J. R. Stat. Assoc. Ser. A* 114, 490–496.
- Clark, C., 1967. *Population Growth and Land Use*. Macmillan, London.
- Combes, P.P., Duranton, G., Gobillon, L., 2012. The Costs of Agglomeration: Land Prices in French Cities. Processed, Wharton School, University of Pennsylvania.
- Combes, P.P., Duranton, G., Gobillon, L., 2014. The Production Function for Housing: Evidence from France. Processed, Wharton School, University of Pennsylvania.
- Coulson, N.E., 1991. Really useful tests of the monocentric city model. *Land Econ.* 67, 299–307.
- Couture, V., Duranton, G., Turner, M.A., 2012. Speed. Processed, University of Toronto.
- Cullen, J.B., Levitt, S.D., 1999. Crime, urban flight, and the consequences for cities. *Rev. Econ. Stat.* 81, 159–169.

- Cunningham, C., 2013. Estimating the Holdout Problem in Land Assembly. Processed, Federal Reserve Bank of Atlanta.
- Davis, D.R., Dingel, J.I., 2013. The Comparative Advantage of Cities. Processed, Columbia University.
- Davis, J.C., Henderson, J.V., 2008. The agglomeration of headquarters. *Reg. Sci. Urban Econ.* 38, 445–460.
- Davis, D.R., Weinstein, D.E., 2008. A search for multiple equilibria in urban industrial structure. *J. Reg. Sci.* 48, 29–65.
- De Bartolome, C.A.M., Ross, S.L., 2003. Equilibria with local governments and commuting: income sorting vs income mixing. *J. Urban Econ.* 54, 1–20.
- Dixit, A., 1973. The optimum factory town. *Bell J. Econ. Manag. Sci.* 4, 637–651.
- Duranton, G., Puga, D., 2005. From sectoral to functional urban specialisation. *J. Urban Econ.* 57, 343–370.
- Duranton, G., Puga, D., 2014. The growth of cities. In: Aghion, P., Durlauf, S.N. (Eds.), *Handbook of Economic Growth*, vol. 2B. Amsterdam, North-Holland, pp. 781–853.
- Dye, R.F., McMillen, D.P., 2007. Teardowns and land values in the Chicago metropolitan area. *J. Urban Econ.* 61, 45–63.
- Eaton, J., Kortum, S., 2002. Technology, geography, and trade. *Econometrica* 70, 1741–1779.
- Edmonston, B., 1975. *Population Distribution in American Cities*. Lexington Books, Lexington.
- Eid, J., Overman, H.G., Puga, D., Turner, M.A., 2008. Fat city: questioning the relationship between urban sprawl and obesity. *J. Urban Econ.* 63, 385–404.
- Ellickson, B., 1977. Economic analysis of urban housing markets: a new approach. Report R-2024-NSFRand Corporation.
- Epplé, D., Gordon, B., Sieg, H., 2010a. Drs. Muth and Mills meet Dr. Tiebout: integrating location-specific amenities into multi-community equilibrium models. *J. Reg. Sci.* 50, 381–400.
- Epplé, D., Gordon, B., Sieg, H., 2010b. A new approach to estimating the production function for housing. *Am. Econ. Rev.* 100, 905–925.
- Epplé, D., Platt, G.J., 1998. Equilibrium and local redistribution in an urban economy when households differ in both preferences and incomes. *J. Urban Econ.* 43, 23–51.
- Ewing, R., Schmid, T., Killingsworth, R., Zlot, A., Raudenbush, S., 2003. Relationship between urban sprawl and physical activity, obesity, and morbidity. *Am. J. Health Prom.* 18, 47–57.
- Fischel, W.A., 1987. *The Economics of Zoning Laws: A Property Rights Approach to American Land Use Controls*. Johns Hopkins University Press, Baltimore, MD.
- Fischel, W.A., 2000. Zoning and land use regulations. In: Boudewijn, B., Geest, G.D. (Eds.), *Encyclopedia of Law and Economics*, vol. 2. Edward Elgar, Cheltenham, pp. 403–442.
- Fischel, W.A., 2001. *The Homevoter Hypothesis*. Harvard University Press, Cambridge, MA.
- Fischel, W.A., 2004. An economic history of zoning and a cure for its exclusionary effects. *Urban Stud.* 41, 317–340.
- Frank, L.D., Andresen, M.A., Schmid, T.L., 2004. Obesity relationships with community design, physical activity, and time spent in cars. *Am. J. Prev. Med.* 27, 87–96.
- Fujita, M., 1982. Spatial patterns of residential development. *J. Urban Econ.* 12, 22–52.
- Fujita, M., 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge University Press, Cambridge, MA.
- Fujita, M., 2012. Thünen and the new economic geography. *Reg. Sci. Urban Econ.* 42, 907–912.
- Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Reg. Sci. Urban Econ.* 12, 161–196.
- Fujita, M., Thisse, J.F., 2013. *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*. Cambridge University Press, Cambridge.
- Galster, G., Hanson, R., Ratcliffe, M.R., Wolman, H., Coleman, S., Freihage, J., 2001. Wrestling sprawl to the ground: defining and measuring an elusive concept. *Hous. Pol. Deb.* 12, 681–717.
- Garreau, J., 1991. *Edge City: Life on the New Frontier*. Doubleday, New York, NY.
- Genesove, D., Han, L., 2012a. Measuring the Thinness of Real Estate Markets. Processed, University of Toronto.
- Genesove, D., Han, L., 2012b. Search and matching in the housing market. *J. Urban Econ.* 72, 31–45.
- Geoghegan, J., 2002. The value of open spaces in residential land use. *Land Use Pol.* 19, 91–98.

- Giles-Corti, B., Macintyre, S., Clarkson, J.P., Pikora, T., Donovan, R.J., 2003. Environmental and lifestyle factors associated with overweight and obesity in Perth, Australia. *Am. J. Health Prom.* 18, 93–102.
- Glaeser, E.L., Gottlieb, J.D., 2006. Urban resurgence and the consumer city. *Urban Stud.* 43, 1275–1299.
- Glaeser, E.L., Gyourko, J., 2005. Urban decline and durable housing. *J. Pol. Econ.* 113, 345–375.
- Glaeser, E.L., Kahn, M., 2001. Decentralized employment and the transformation of the American city. In: *Brookings-Wharton Papers on Urban Affairs*, pp. 1–47.
- Glaeser, E.L., Kahn, M.E., 2004. Sprawl and urban growth. In: Henderson, V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam, pp. 2481–2527.
- Glaeser, E.L., Gyourko, J., Saks, R., 2005. Why is Manhattan so expensive? Regulation and the rise in housing prices. *J. Law Econ.* 48, 331–369.
- Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *J. Urban Econ.* 63, 1–24.
- Gobillon, L., Selod, H., Zenou, Y., 2007. The mechanisms of spatial mismatch. *Urban Stud.* 44, 2401–2427.
- Gordon, P., Richardson, H.W., Jun, M.J., 1991. The commuting paradox: evidence from the top twenty. *J. Am. Plann. Assoc.* 57, 416–420.
- Gyourko, J., Saiz, A., 2006. Construction costs and the supply of housing structure. *J. Reg. Sci.* 46, 661–680.
- Hamilton, B.W., 1982. Wasteful commuting. *J. Pol. Econ.* 90, 1035–1051.
- Helsley, R.W., 1990. Knowledge production in the CBD. *J. Urban Econ.* 28, 391–403.
- Henderson, J.V., 1977. *Economic Theory and the Cities*. Academic Press, New York, NY.
- Henderson, J.V., Mitra, A., 1996. The new urban landscape: developers and edge cities. *Reg. Sci. Urban Econ.* 26, 613–643.
- Hilber, C., Robert-Nicoud, F., 2013. On the origins of land use regulations: theory and evidence from US metro areas. *J. Urban Econ.* 75, 29–43.
- Imai, H., 1982. CBD hypothesis and economies of agglomeration. *J. Econ. Theory* 28, 275–299.
- Inman, R.P., 1995. How to have a fiscal crisis: lessons from Philadelphia. *Am. Econ. Rev.* 85, 378–383.
- Irwin, E.G., 2002. The effects of open space on residential property values. *Land Econ.* 78, 465–480.
- Kain, J.F., 1968. Housing segregation, negro employment, and metropolitan decentralization. *Q. J. Econ.* 82, 175–197.
- Kanemoto, Y., 1980. *Theories of Urban Externalities*. North-Holland, Amsterdam.
- Kanemoto, Y., 1988. Hedonic prices and the benefits of public projects. *Econometrica* 56, 981–989.
- Katz, L.F., Kling, J.R., Liebman, J.B., 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. *Q. J. Econ.* 116, 607–654.
- Kim, S., 2007. Changes in the nature of urban spatial structure in the United States, 1890–2000. *J. Reg. Sci.* 47, 273–287.
- Koopmans, T.C., Beckmann, M., 1957. Assignment problems and the location of economic activities. *Econometrica* 25, 53–76.
- Kopecky, K.A., Suen, R.M.H., 2010. A quantitative analysis of suburbanization and the diffusion of the automobile. *Int. Econ. Rev.* 51, 1003–1037.
- Krueger, A.B., Kahneman, D., Schkade, D., Schwarz, N., Stone, A.A., 2009. National time accounting: The currency of life. In: Krueger, A.B. (Ed.), *Measuring the Subjective Well-Being of Nations: National Accounts of Time Use and Well-Being*. NBER and University of Chicago Press, Chicago, pp. 9–86.
- Landvoigt, T., Piazzesi, M., Schneider, M., 2011. *The Housing Market(s) of San Diego*. Processed, Stanford University.
- Lee, S., Lin, J., 2013. *Natural Amenities, Neighborhood Dynamics, and Persistence in the Spatial Distribution of Income*. Processed, Federal Reserve Bank of Philadelphia.
- LeRoy, S.F., Sonstelie, J., 1983. The effects of urban spatial structure on travel demand in the United States. *J. Urban Econ.* 13, 67–89.
- Levinson, D., Wu, Y., 2005. The rational locator reexamined: are travel times still stable? *Transportation* 32, 187–202.
- Liao, W.C., 2012. Inshoring: the geographic fragmentation of production and inequality. *J. Urban Econ.* 72, 1–16.
- Lucas Jr., R.E., Rossi-Hansberg, E., 2002. On the internal structure of cities. *Econometrica* 70, 1445–1476.
- Ma, K.R., Banister, D., 2006. Excess commuting: a critical review. *Transp. Rev.* 26, 749–767.

- Määtänen, N., Terviö, M., 2014. Income distribution and housing prices: an assignment model approach. *J. Econ. Theory* 151, 381–410.
- Madden, J.F., 1985. Urban wage gradients: empirical evidence. *J. Urban Econ.* 18, 291–301.
- Manning, A., 2003. The real thin theory: monopsony in modern labour markets. *Lab. Econ.* 10, 105–131.
- Manville, M., Shoup, D., 2003. People, parking and cities. *J. Urban Plann. Devel.* 131, 233–245.
- Margo, R.A., 1992. Explaining the postwar suburbanization of population in the United States: the role of income. *J. Urban Econ.* 31, 301–310.
- Markusen, J.R., 2002. *Multinational Firms and the Theory of International Trade*. MIT Press, Cambridge, MA.
- McDonald, J.F., 1989. Econometric studies of urban population density: a survey. *J. Urban Econ.* 26, 361–385.
- McDonald, J.F., Bowman, H.W., 1979. Land value functions: a reevaluation. *J. Urban Econ.* 6, 25–41.
- McGrath, D.T., 2005. More evidence on the spatial scale of cities. *J. Urban Econ.* 58, 1–10.
- McMillen, D.P., 1996. One hundred fifty years of land values in Chicago: a nonparametric approach. *J. Urban Econ.* 40, 100–124.
- McMillen, D.P., 2001. Nonparametric employment subcenter identification. *J. Urban Econ.* 50, 448–473.
- McMillen, D.P., 2006. Testing for monocentricity. In: Arnott, R.J., McMillen, D.P. (Eds.), *A Companion to Urban Economics*. Blackwell, Oxford, pp. 128–140.
- McMillen, D.P., 2010. Issues in spatial data analysis. *J. Reg. Sci.* 50, 119–141.
- McMillen, D.P., Singell Jr., L.D., 1992. Work location, residence location, and the intraurban wage gradient. *J. Urban Econ.* 32, 195–213.
- McMillen, D.P., Smith, S.C., 2003. The number of subcenters in large urban areas. *J. Urban Econ.* 53, 332–342.
- Merlo, A., Ortalo-Magné, F., Rust, J., 2013. *The Home Selling Problem: Theory and Evidence*. Processed, University of Pennsylvania.
- Mieszkowski, P., Mills, E.S., 1993. The causes of metropolitan suburbanization. *J. Econ. Persp.* 7, 135–147.
- Mieszkowski, P., Smith, B., 1991. Analyzing urban decentralization: the case of Houston. *Reg. Sci. Urban Econ.* 21, 183–199.
- Mills, E.S., 1967. An aggregative model of resource allocation in a metropolitan area. *Am. Econ. Rev. Pap. Proc.* 57, 197–210.
- Mills, E.S., 1969. The value of urban land. In: Perloff, H.S. (Ed.), *The Quality of the Urban Environment. Resources for the Future*, Baltimore, MD, pp. 231–253.
- Mills, E.S., 1972. *Studies in the Structure of the Urban Economy*. Johns Hopkins Press, Baltimore, MD.
- Mirrlees, J.A., 1972. The optimum town. *Swed. J. Econ.* 74, 114–135.
- Mohring, H., 1961. Land values and the measurement of highway benefits. *J. Pol. Econ.* 69, 236–249.
- Mohring, H., Harwitz, M., 1962. *Highway Benefits: An Analytical Framework*. Northwestern University Press, Evanston, IL.
- Montesano, A., 1972. A restatement of Beckmann's model on the distribution of urban rent and residential density. *J. Econ. Theory* 4, 329–354.
- Mortensen, D.T., Pissarides, C.A., 1994. Job creation and job destruction in the theory of unemployment. *Rev. Econ. Stud.* 61, 397–415.
- Muth, R.F., 1969. *Cities and Housing*. University of Chicago Press, Chicago, IL.
- Ogawa, H., Fujita, M., 1980. Equilibrium land use patterns in a nonmonocentric city. *J. Reg. Sci.* 20, 455–475.
- Ortalo-Magné, F., Prat, A., 2014. On the political economy of urban growth: homeownership versus affordability. *Am. Econ. J.* 6, 154–181.
- Ota, M., Fujita, M., 1993. Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Reg. Sci. Urban Econ.* 23, 695–729.
- Papageorgiou, Y.Y., Pines, D., 1990. The logical foundations of urban economics are consistent. *J. Econ. Theory* 50, 37–53.
- Paulsen, K., 2012. Yet even more evidence on the spatial size of cities: urban spatial expansion in the us, 1980–2000. *Reg. Sci. Urban Econ.* 42, 561–568.

- Pines, D., Sadka, E., 1985. Zoning, first-best, second-best, and third-best criteria for allocating land for roads. *J. Urban Econ.* 17, 167–185.
- Pissarides, C.A., 2000. *Equilibrium Unemployment Theory*. MIT Press, Cambridge, MA.
- Polinsky, A.M., Shavell, S., 1975. The air pollution and property value debate. *Rev. Econ. Stat.* 57, 100–104.
- Polinsky, A.M., Shavell, S., 1976. Amenities and property values in a model of an urban area. *J. Publ. Econ.* 5, 119–129.
- Postel-Vinay, F., Robin, J.M., 2002. Equilibrium wage dispersion with worker and employer heterogeneity. *Econometrica* 70, 2295–2350.
- Putnam, R.D., 2000. *Bowling Alone: The Collapse and Revival of American Community*. Simon & Schuster, New York, NY.
- Quigley, J.M., 1998. Urban diversity and economic growth. *J. Econ. Persp.* 12, 127–138.
- Raphael, S., 1998. The spatial mismatch hypothesis and black youth joblessness: evidence from the San Francisco Bay Area. *J. Urban Econ.* 43, 79–111.
- Redfearn, C.L., 2007. The topography of metropolitan employment: identifying centers of employment in a polycentric urban area. *J. Urban Econ.* 61, 519–541.
- Redfearn, C.L., 2009. Persistence in urban form: the long-run durability of employment centers in metropolitan areas. *Reg. Sci. Urban Econ.* 39, 224–232.
- Ricardo, D., 1817. *On the Principles of Political Economy and Taxation*. John Murray, London.
- Rogerson, R., Shimer, R., Wright, R., 2005. Search-theoretic models of the labor market: a survey. *J. Econ. Liter.* 43, 959–988.
- Rosenthal, S.S., 2014. Are private markets and filtering a viable source of low income housing? Estimates from a “repeat income” model. *Am. Econ. Rev.* 104, 687–706.
- Rossi-Hansberg, E., 2004. Optimal urban land use and zoning. *Rev. Econ. Dyn.* 7, 69–106.
- Rossi-Hansberg, E., Sarte, P.D., Raymond Owens, I., 2009. Firm fragmentation and urban patterns. *Int. Econ. Rev.* 50, 143–186.
- Saelens, B.E., Sallis, J.F., Black, J.B., Chen, D., 2003. Neighborhood-based differences in physical activity: an environment scale evaluation. *Am. J. Public Health* 93, 1552–1558.
- Samuelson, P.A., 1983. Thünen at two hundred. *J. Econ. Liter.* 21, 1468–1488.
- Sattinger, M., 1993. Assignment models of the distribution of earnings. *J. Econ. Liter.* 19, 831–880.
- Selod, H., Zenou, Y., 2001. Location and education in South African cities under and after Apartheid. *J. Urban Econ.* 49, 168–198.
- Siodla, J., 2014. *Razing San Francisco: The 1906 Disaster as a Natural Experiment in Urban Redevelopment*. Processed, Colby College.
- Small, K.A., Verhoef, E.T., 2007. *The Economics of Urban Transportation*. Routledge, New York, NY.
- Solow, R.M., 1972. Congestion, density and the use of land in transportation. *Swed. J. Econ.* 74, 161–173.
- Solow, R.M., 1973. On equilibrium models of urban location. In: Nobay, A.R., Parkin, M. (Eds.), *Essays in Modern Economics*. Longman, London, pp. 2–16.
- Solow, R.M., Vickrey, W.S., 1971. Land use in a long narrow city. *J. Econ. Theory* 3, 430–447.
- Strange, W.C., 1992. Overlapping neighborhoods and housing externalities. *J. Urban Econ.* 32, 17–39.
- Strange, W.C., 1995. Information, holdouts, and land assembly. *J. Urban Econ.* 38, 317–332.
- Stull, W.J., 1974. Land use and zoning in an urban economy. *Am. Econ. Rev.* 64, 337–347.
- Sweeney, J.L., 1974a. A commodity hierarchy model of the rental housing market. *J. Urban Econ.* 1, 288–323.
- Sweeney, J.L., 1974b. Quality, commodity hierarchies, and housing markets. *Econometrica* 42, 147–167.
- von Thünen, J.H., 1826. *Der Isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Perthes, Oxford, Hamburg, English Translation: *The Isolated State*: Pergamon Press, 1966.
- Tiebout, C.M., 1956. A pure theory of local expenditures. *J. Pol. Econ.* 64, 416–424.
- Turnbull, G.K., 1988. Residential development in an open city. *Reg. Sci. Urban Econ.* 18, 307–320.
- Turner, M.A., 2005. Landscape preferences and patterns of residential development. *J. Urban Econ.* 57, 19–54.
- Wasmer, E., Zenou, Y., 2002. Does city structure affect job search and welfare? *J. Urban Econ.* 51, 515–541.

- Wheaton, W.C., 1977. Income and urban residence: an analysis of consumer demand for location. *Am. Econ. Rev.* 67, 620–631.
- Wheaton, W.C., 1982. Urban spatial development with durable but replaceable capital. *J. Urban Econ.* 12, 1–21.
- Wheaton, W.C., 1990. Vacancy, search, and prices in a housingmarketmatching model. *J. Pol. Econ.* 98, 1270–1292.
- White, M.J., 1976. Firm suburbanization and urban subcenters. *J. Urban Econ.* 3, 323–343.
- Wildasin, D.E., 1986. Spatial variation of the marginal utility of income and unequal treatment of equals. *J. Urban Econ.* 19, 125–129.
- Yinger, J., 1979. Estimating the relationship between location and the price of housing. *J. Reg. Sci.* 19, 271–286.
- Zenou, Y., 2009. *Urban Labor Economics*. Cambridge University Press, Cambridge, MA.

CHAPTER 9

Neighborhood and Network Effects

Giorgio Topa^{*}, Yves Zenou[†]

^{*}Federal Reserve Bank of New York and IZA, NY, USA

[†]Stockholm University, IFN, and CEPR, Stockholm, Sweden

Contents

9.1. Introduction	562
9.2. Neighborhood Effects	566
9.2.1 Reduced-form empirical literature on neighborhood effects	566
9.2.1.1 <i>Experimental or quasi-experimental evidence</i>	566
9.2.1.2 <i>Nonexperimental evidence</i>	570
9.2.2 Neighborhood effects estimation using a structural approach	574
9.3. Network Effects	578
9.3.1 Network theory	578
9.3.1.1 <i>The local-aggregate model</i>	579
9.3.1.2 <i>The local-average model</i>	582
9.3.1.3 <i>Local aggregate or local average? theoretical considerations</i>	583
9.3.2 Empirical aspects of social networks: Structural approach	584
9.3.2.1 <i>Linear-in-means model: The reflection problem</i>	584
9.3.2.2 <i>Social networks: The local-average model</i>	586
9.3.2.3 <i>Social networks: The local-aggregate model</i>	588
9.3.2.4 <i>Testing the local-average model against the local-aggregate model</i>	589
9.3.2.5 <i>Endogenous network formation</i>	590
9.3.2.6 <i>Multiple equilibria</i>	591
9.3.3 Empirical results	593
9.3.3.1 <i>Local-average model</i>	593
9.3.3.2 <i>Local-aggregate model</i>	594
9.3.3.3 <i>Local-aggregate versus the local-average model</i>	595
9.4. Neighborhood and Network Effects	599
9.4.1 Theory: Spatial models with social networks	600
9.4.1.1 <i>Spatial models with social interactions</i>	600
9.4.1.2 <i>Spatial models with weak and strong ties</i>	605
9.4.1.3 <i>Spatial models with explicit social networks</i>	608
9.4.2 Discussion	611
9.4.3 Empirical results	614
9.5. Concluding Remarks	615
Acknowledgments	617
References	617

Abstract

In this chapter, we provide an overview of research on neighborhoods and social networks and their role in shaping behavior and economic outcomes. We include a discussion of empirical and theoretical analyses of the role of neighborhoods and social networks in crime, education, and labor-market outcomes. In particular, we discuss in detail identification problems in peer, neighborhood, and network effects and the policy implications of integrating the social and the geographical space, especially for ethnic minorities.

Keywords

Social networks, Neighborhoods, Group-based policies, Ethnic minorities, Labor economics

JEL Classification Codes

C23, D85, J15, J64, K42, R14, Z13

9.1. INTRODUCTION

Our environment, whether it includes our neighbors, our peers, or more generally, our social contacts, crucially affects many facets of our lives. For example, the decision of an agent of whether or not to buy a new product, study hard, commit a crime, smoke, or find a job is often influenced by the choices of his/her friends and acquaintances, neighbors, classmates, coworkers, professional contacts, etc. Economists—and before them, sociologists—have long recognized the importance of such nonmarket interactions in shaping behavior and outcomes in a large variety of contexts. A long, but only partial, list includes peer effects in the classroom and in the workplace, labor-market referrals, smoking, crime, and other social diseases, consumption externalities, herd behavior, and “contagion” in financial networks, bankruptcy and foreclosure decisions, risk sharing within households, communities, villages, sorting into residential neighborhoods, the adoption and diffusion of new technologies, the role of agglomeration economies in shaping cities and the location decisions of businesses, and the role of human capital externalities in economic growth.

At a very abstract level, we can think of these social interactions as taking place within a “social space,” defined by one’s reference group—be it one’s classmates, peers, neighbors, colleagues, other firms, and so on. There are various ways to model such a social space. Generally speaking, we need to define a set of individual agents (affecting each other), as well as the connections among them. These, in turn, help us define a notion of social or economic distance on the set of locations inhabited by agents in the abstract social space.

One large strand of literature has modeled the social space simply by characterizing the reference group of each agent. For peer effects in education, this is often the set of classmates or schoolmates. For the location decisions of households and firms, it is the set of residential neighbors or other firms in the same industrial district. Interactions are

assumed to be symmetric within each reference group, and the various reference groups often define a proper partition of the set of agents. One particular example of this approach consists of the literature on neighborhood effects, which attempts to study how the composition of one's residential neighborhood affects one's outcomes with regard to, for instance, educational achievement, the ability to find a job, or the propensity to engage in crime.

Another strand of the literature has focused on the structure of connections within the abstract social space. This is often modeled using the tools of social network theory. As we describe in more detail in [Section 9.3.1](#), a network is defined as a set of agents and a graph describing who is connected to whom. A growing empirical literature shows that the structure of the network, and the position of individual agents within it, plays an important role in shaping choices and outcomes.

The social space modeled by a network does not necessarily coincide with the physical space. Indeed, an individual may be closely connected to someone residing and working at the other end of the city or even in a different country. The social space—and the distance among agents—may be defined by ethnicity, race, age, nationality, tastes, and many attributes other than physical distance. At the same time, it seems reasonable to think that the costs of interaction increase with physical distance, so interactions may be easier and more frequent among agents who are physically close to each other. Therefore, in general, there will be some partial overlap between the social space modeled as a network and the physical space described by a residential neighborhood.

Finally, there are several mechanisms through which social interactions may affect behavior and outcomes. Social contacts may facilitate the flow of information about, for instance, job openings or the profitability of a new technology, thus influencing the choice set available to agents. Social contacts may also affect one's tastes for a certain good, influencing the likelihood that one will consume that good. One's network or reference group may provide risk-sharing devices and opportunities for cooperation. There may also be complementarities in production or consumption through which social interaction effects operate.

In this chapter, we review the literature on neighborhood effects and the literature on network effects. These two bodies of literature have developed largely separately: the neighborhood effects literature has mostly focused on how residential neighborhoods may shape opportunities, choices, and outcomes of individual agents living in them. This process has implications for urban policy, the evolution of neighborhoods and cities, and the dynamics of segregation and inequality—to mention just a few. The theoretical and empirical study of networks has largely focused on the social space of connections and its implications for outcomes, without including the physical space. We will review these two approaches separately first and then attempt to bring them together in a more unified setting.

It is worth noting here that the neighborhood effects literature has for the most part ignored the microstructure of connections underlying the social interactions occurring within the neighborhood. This is largely because of a data limitation problem: until recently, very few datasets were available that gave researchers information on both network connections and physical locations of agents. We will discuss recent advances in data collection efforts in what follows. Finally, neighborhood effects may arise not just because of social interactions within the neighborhood (or across adjacent neighborhoods) but also because of local shocks or institutions—such as a local business closing, or the presence of churches, clubs, and neighborhood associations. This is analogous to the education setting, where educational outcomes of students may be affected not only by their peers but also by the teacher or the school.

One important issue concerns the identification and estimation of neighborhood or network effects. Because agents are assumed to affect each other—through information exchanges, preferences, or actions—a telltale sign of the presence of such effects is the presence of co-movements in observed outcomes across agents. However, it is extremely difficult to separately identify these effects from other forces that also bring about co-movements. First, there is a simultaneity problem: I affect my social contact, and simultaneously she affects me. This is known in the literature as the “reflection problem,” and we will discuss it extensively in [Section 9.3.2.1](#). Second, agents may sort into neighborhoods or networks on the basis of similar tastes or attributes that are unobserved by the econometrician. Again, this poses identification challenges. Finally, agents residing in the same neighborhood or social network may be exposed to similar correlated shocks that are, again, unobserved by the econometrician: for instance, good or bad local institutions, environmental factors that affect an entire set of neighborhoods, or a plant closing and inducing a localized wave of unemployment. In what follows, we will discuss how each of the approaches developed below fares with regard to these identification and estimation challenges.

There exists a rich and long-standing neighborhood effects literature, developed both in the United States and in Europe. We first present the experimental approach, which mostly focuses on immigrants and refugees where the “natural” experiment comes from the fact that their location upon arrival in a new country is arguably “exogenous” because it is imposed by the local authorities of the host country. Other natural or randomized experiments include the relocation of families from public housing projects in poor neighborhoods to low-poverty neighborhoods, via housing vouchers. The Moving to Opportunity (MTO) program is perhaps the most well-known example.

We also present a nonexperimental approach to the analysis of neighborhood effects, where the identification strategy is clever and based on the smallest unit in the city—namely, the city block. By arguing that the assignment of agents to city blocks is quasi-random (i.e., driven by factors orthogonal to possible unobservable attributes),

researchers are able to separately identify neighborhood effects from other potential sources of co-movements.

Finally, we develop a structural approach where the theoretical models generate stationary distributions with well-defined properties over space. The parameters of these models can then be estimated by matching moments from the simulated spatial distribution generated by the model with their empirical counterparts from spatial data on neighborhoods or cities.

We then turn to the network literature. We first study settings in which the network is given. The main challenge in studying strategic interactions in social settings is the inherent complexity of networks. If we do not focus on specific structures in terms of the games, it is hard to draw any conclusions. We focus on strategic complementarities so that a player's incentives to take an action (or a "higher" action) increase with the number of his/her friends who take the (higher) action. We look, in particular, at quite tractable "linear-quadratic" settings where agents choose a continuous level of activity. This simple parametric specification permits an explicit solution for equilibrium behavior as a function of the network and thus leads to interesting comparative statics and other results that are useful in empirical work.

We then present the identification strategy based on the best-reply function of these models. This is mostly based on exclusion restrictions arising naturally from the partially overlapping nature of network connections: simply put, my friends' friends may not necessarily be my friends. We also show how identification may survive (and in some cases be strengthened) when one takes into account the endogenous formation of networks. A note of caution is brought by the introduction of nonlinear models of interaction, which may induce multiplicity of equilibria: we discuss some early attempts to estimate network models in the presence of such multiplicity. We conclude this section by reviewing different empirical results for crime, education, labor, health, etc.

In the last part of this chapter, we integrate the two previous bodies of literature by analyzing how the combined effect of neighborhoods and networks affects the outcomes of individuals, focusing mostly on the labor market. This literature is, unfortunately, in its infancy and we review the scarce evidence and theoretical models on this topic.

The rest of this chapter unfolds as follows. In the next section, we look at neighborhood effects, differentiating between the reduced-form empirical literature on neighborhood effects (Section 9.2.1) and the structural approach (Section 9.2.2). Section 9.3 focuses on network effects by first providing some theoretical background (Section 9.3.1), then analyzing the econometric issues related to the empirical study of networks (Section 9.3.2), and finally providing the main empirical results of this literature (Section 9.3.3). In Section 9.4, we study neighborhood and network effects together, looking first at the theoretical models (Section 9.4.1) and then discussing the theoretical results (Section 9.4.2) and the empirical results (Section 9.4.3). Finally, Section 9.5 concludes the chapter.

9.2. NEIGHBORHOOD EFFECTS

In this section, we first review the reduced-form empirical literature that aims at estimating neighborhood effects in a variety of settings. We examine both experimental and nonexperimental approaches. We then turn to more recent structural modeling and empirical work.¹

9.2.1 Reduced-form empirical literature on neighborhood effects

The reduced-form empirical work on neighborhood effects has a long tradition in both economics and sociology. Much of the early work focused on the effects of growing up in disadvantaged neighborhoods on educational attainment, employment, and other indicators of socioeconomic well-being. Public policy was an important component of this work, with a strong focus on poverty and inequality.² However, this work largely suffered from the [Manski \(1993\)](#) critique concerning the reflection problem. Most of the early work used simple regressions of individual outcomes on individual attributes, family and community attributes, and typically mean outcomes in the residential neighborhood. In the absence of an empirical strategy to separately identify the parameters of these models, most of this work suffered from a basic lack of identification.

Cognizant of these challenges, subsequent reduced-form work followed two broad strategies. The first is to exploit some natural variation arising from randomized or quasi-random experiments implemented in various cities to put into effect various policies. The second approach uses some innovative identification strategies to identify neighborhood effects using large datasets with detailed information on geography.

9.2.1.1 *Experimental or quasi-experimental evidence*

The first set of studies analyzes neighborhood effects by studying various randomized or natural experiments. The majority of these studies exploit housing relocation randomized experiments that allowed residents of low-income neighborhoods or in public housing projects to relocate to different neighborhoods. These experiments in principle allow the researcher to measure the effect of changing neighborhood characteristics on outcomes.

[Popkin et al. \(1993\)](#) study the impact of the Gautreaux program in Chicago, which helped relocate low-income families from public housing to private housing in the Chicago metropolitan area. While the selection of participants into the program was not random, the assignment to city versus suburban neighborhoods was quasi-random and was based on the availability of units. The authors of the study find that moving to a suburban residential location was associated with a significantly higher chance of

¹ For overviews of this literature spanning several decades, see [Jencks and Mayer \(1990\)](#), [Durlauf \(2004\)](#), and [Ioannides and Topa \(2010\)](#).

² [Jencks and Mayer \(1990\)](#) and [Brooks-Gunn et al. \(1997\)](#) provide nice surveys of this early literature. Prominent examples include the work of [Wilson \(1987\)](#), [Corcoran et al. \(1989\)](#), and [Brooks-Gunn et al. \(1992\)](#).

being employed than moving to a city location, even conditioning on observed personal characteristics. The employment gains are greater for those who never worked before.

Jacob (2004), on the other hand, exploits the quasi-random closing of high-rise public housing projects in Chicago during the 1990s. Families affected by the closings were offered Section 8 housing vouchers to move anywhere in the metropolitan area. Jacob compares school outcomes for students living in units affected by a closure with those for students in units in the same project that were not closed. Arguably the timing of building closures within a project is uncorrelated with unobserved characteristics of students. Contrary to the Gautreaux experiment, this article finds no evidence of any impact of the demolitions and subsequent relocations on student outcomes.

Oreopoulos (2003) focuses on another source of quasi-random variation in neighborhood quality—namely, the assignment of families to different housing projects in Toronto. By matching project addresses with an administrative panel of Canadians and their parents, this article can examine the impact of neighborhood quality on the long-run outcomes of adults who were assigned as children to different residential projects. Similar to Jacob (2004), Oreopoulos (2003) finds again no effect of neighborhood differences on a wide variety of outcomes, including unemployment, mean earnings, income, and welfare participation. Further, while neighborhood quality does not affect outcomes, family background explains about 30% of the total variation in income and wages.

A large set of studies focuses on the MTO program (Ludwig et al., 2001; Kling et al., 2005, 2007). This was a large, randomized experiment in which participants volunteered for the study, and was randomly assigned to one of three groups: a control group received no new assistance, a Section 8 group received a housing voucher without geographical restrictions, and a third group received a Section 8 voucher to move to a low-poverty neighborhood as well as mobility counseling. Relative to the control group, the two other groups indeed moved to neighborhoods with significantly lower poverty rates, with less crime, and in which residents reported feeling safer.

MTO studies generally find no significant evidence of treatment effects with regard to economic outcomes, such as earnings, welfare participation, or the amount of government assistance. However, these studies do find evidence of large and significant positive treatment effects on a variety of adult mental health measures. For outcomes of teenage youths, an interesting dichotomy appears: in general, treatment effects were positive with regard to mental health and risky behaviors for female youths, but were negative for male youths. These negative impacts for male youths were particularly large for physical health and risky behavior, suggesting that perhaps the neighborhood change induced a severe dislocation and social isolation, or rejection of the prevailing norms in the new neighborhood.

More recently, Ludwig et al. (2012) have studied the long-term effects of the MTO program, 10–15 years after the experiment. They look at intention-to-treat effects for a

variety of outcomes, grouped into economic self-sufficiency, physical health, mental health, and subjective well-being. Treatment effects are found not to be significant for economic outcomes, are positive but not statistically significant for physical health, are positive and marginally significant for mental health, and are significantly positive for subjective well-being.³

Our reading of this strand of the literature, that by and large employs careful program evaluation approaches, is that the estimated neighborhood effects tend to be small for educational and economic outcomes. Larger effects are found for mental health outcomes. The MTO-related literature represents perhaps the cleanest example of this approach.

However, it is important to note that there are important limitations in the extent to which the treatment effects identified through relocation experiments are informative about the nature of general forms of neighborhood effects *per se*. First, the individuals studied must be eligible for a relocation program in the first place; this typically implies that the resulting sample is somewhat “special” (i.e., so as to be a resident in public housing) and may not be as sensitive to neighborhood effects as other individuals. More generally, even if the eligible population is representative of the target population, the results of an experiment based on a small sample may not scale up to broader populations because of the strong possibility that general equilibrium effects may arise in that case.

Second, the experimental design involves relocation to new neighborhoods that are, by design, very different from baseline neighborhoods. This implies that the identified treatment effect measures the impact of relocating to a neighborhood where individuals initially have few social contacts and where the individuals studied may be very different from the average resident of the new neighborhood. In this way, the treatment effects identified with this design are necessarily a composite of several factors related to significant changes in neighborhoods that cannot be easily disentangled.

Another set of articles uses a different source of quasi-random variation in network composition and location—namely, the resettlement of refugees into various countries. Beaman (2012) studies refugees resettled into various US cities between 2001 and 2005 by the International Rescue Committee—a large resettlement agency. The location decision of the agency for refugees without family already in the United States is arguably exogenous. Beaman posits a dynamic model of labor-market networks inspired by Calvó-Armengol and Jackson (2004), where agents share information about jobs within their individual social networks. The model implies both a congestion effect due to competition for information among job seekers—which leads to negative correlation in

³ See, however, Bond and Lang (2014) for a discussion of “happiness scales.” Depending on the assumptions made regarding the underlying distribution of subjective well-being, the MTO treatment effects may be positive or null with regard to subjective well-being. Still, there is strong evidence that the MTO program reduced various other measures of well-being, such as symptoms of depression.

outcomes within networks—and a positive effect of network connections on employment outcomes, going from older to more recent cohorts. These effects are dynamic: an increase in the size of a given cohort will worsen the expected employment outcomes for subsequent cohorts that arrive immediately afterward, but will gradually improve the outcomes for later cohorts.

The empirical strategy exploits the variation in cohort size for different ethnicities in different cities at different points in time. The possibility of sorting or correlation between network size and unobserved city and ethnicity characteristics (possibly due to the agency's placement strategy) is addressed by controlling for individual characteristics that are observed by the agency, as well as city and nationality-cohort fixed effects. Beaman finds that a one standard deviation increase in the previous year's cohort for a newly arrived refugee lowers his employment probability by 4.9 percentage points. Conversely, an increase in longer-tenured network size improves employment outcomes by 4.3 percentage points. More senior social contacts also have a positive effect on expected wages. This study is notable for its emphasis on *dynamic* neighborhood effects. The model implications provide additional tools for identification.

Edin et al. (2003) and Åslund et al. (2011) exploit a similar source of quasi-random variation from a refugee resettlement program in Sweden during the late 1980s to study neighborhood effects on labor-market and education outcomes, respectively. Both authors argue convincingly that the initial assignment of refugee immigrants to neighborhoods within cities was uncorrelated with unobservable individual characteristics. In particular, "the individual could not choose his/her first place of residence due to the institutional setup, the practical limitations imposed by scarce housing, and the short time frame between the receipt of residence permit and placement." Further, there was no interaction between placement officers and immigrants, so any sorting could take place only on the basis of observable (to both the government officials and the econometrician) attributes.

The first study finds that a larger ethnic enclave in one's initially assigned location has a positive effect on earnings, especially for less-skilled immigrants: a one standard deviation increase in ethnic concentration raises earnings by 13% for less educated immigrants. These positive effects increase with the quality of the enclave as measured by earnings or self-employment rates. The second study focuses on school performance and finds that a one standard deviation increase in the share of highly educated adults (sharing the student's ethnicity) in the neighborhood of residence raises average grades in compulsory school by 0.8 percentile points.⁴

⁴ Åslund et al. (2010) also exploit this quasi-random assignment of immigrants to residential locations to revisit the "spatial mismatch" hypothesis. They find that local access to jobs does indeed have a statistically and economically significant impact on employment outcomes.

Damm (2009, 2014) and Damm and Dustmann (2014) also exploit a unique natural experiment between 1986 and 1998 when refugee immigrants to Denmark were assigned to neighborhoods quasi-randomly. The first articles focus on labor-market outcomes of ethnic minorities, while the last article looks at the effect of early exposure to neighborhood crime on subsequent criminal behavior of youths. In the latter, the authors find strong evidence that the share of young people convicted of crimes, in particular violent crimes, in the neighborhood increases convictions of individuals residing in the neighborhood later in life.⁵ Their findings suggest that social interaction is a key channel through which neighborhood crime is linked to individual criminal behavior. We will return to the issue of social interactions and crime in Section 9.3.3.3.

Finally, we wish to mention a separate strand of literature that also exploits natural experiments to evaluate the extent of residential neighborhood effects, in the context of housing and land prices. As an example, Rossi-Hansberg et al. (2010) examine how nonmarket interactions between residents of a given neighborhood (or across nearby neighborhoods) are reflected in land prices. They exploit a plausibly exogenous source of variation in the attractiveness of a given location provided by an urban revitalization program that was implemented in Richmond, Virginia, between 1999 and 2004. The program gave funding for housing investments in targeted neighborhoods, including demolition, rehabilitation, and new construction of housing. In addition, a “control” neighborhood was selected that was similar to the treated neighborhoods but did not receive any funding.

The study contains information on the location of homes that received funding, and the amount of the funding. Housing prices and characteristics before and after the program are also observed. This allows Rossi-Hansberg et al. (2010) to estimate land prices before and after the policy was implemented, using a hedonic approach. They can, therefore, estimate the spatial extent of neighborhood quality externalities on land prices. In addition, by comparing treated and control neighborhoods, they can compute the magnitude of these externalities. The study finds that increases in land values decline with the distance from the impact areas, as expected: housing externalities decline roughly by half every 1000 feet. Further, the increase in land values arising from externalities brought about by the revitalization ranges between \$2 and \$6 per dollar invested.

9.2.1.2 *Nonexperimental evidence*

As mentioned above, a more promising approach in our view has relied on very detailed spatial datasets and clever identification strategies to identify neighborhood effects in various settings. Essentially, this set of articles exploits either quasi-random assignment of individual agents to small geographical units (such as census blocks) or careful modeling

⁵ See also Jencks and Mayer (1990) and Gould et al. (2011) for the long-term effects of growing up in a poor and low-educated neighborhood.

of the mechanisms underlying social interaction effects that delivers clear testable implications that can be applied to the data.

Bayer et al. (2008) consider spatial clustering of individual work locations for a given residential location as evidence of local referral effects. In order to separately identify labor-market referrals from other spatially correlated effects, they estimate the excess propensity to work together (in a given city block) for pairs of workers who co-reside in the same city block (distinct from their work location), relative to the baseline propensity to work together for residents in nearby blocks (within a reference group of blocks). The key identifying assumption (which is tested on observable characteristics) is that there is no block-level correlation in unobserved attributes among block residents, after taking into account the broader reference group. An additional assumption underlying this research design is that a significant portion of interactions with neighbors are very local in nature—that is, they occur among individuals in the same block.⁶ We return to this question in Section 9.4.3. Bayer et al. (2008) find that residing in the same block raises the probability of sharing the work location by 33%, consistent with local referral effects. Inferred referral effects are stronger when they involve at least one individual who is more attached to the labor market, or individuals who are more likely to interact—for example, because they have children of similar ages. The observed variation in the excess propensity to work in the same block is then used to construct a measure of network quality available to each individual in a given neighborhood. A one standard deviation increase in this measure has a positive effect on various labor-market outcomes: labor force participation increases by about 3.4 percentage points for female workers, whereas hours worked increase by 1.8 h per week on average and earnings increase by about 3.4% for male workers.⁷

Hellerstein et al. (2011) build on the identification strategy of Bayer et al. (2008) using matched employer–employee data at the establishment level from the 2000 Decennial Employer–Employee Database. They use census tracts as the geographical unit of analysis, and compute the excess propensity to reside in the same tract for employees in a given *establishment*, relative to the likelihood of residing in the same tract for other

⁶ More generally, as discussed in Section 9.1, one important question concerns the extent of overlap between the social space spanned by individual social networks and the geographical space described by neighborhoods. Several sociological studies have examined this question, finding that a significant portion of social interactions occur at very close physical distance among agents. See, for instance, Wellman (1996), Otani (1999), and Lee and Campbell (1999).

⁷ Using an identification strategy similar to that of Bayer et al. (2008), Hawranek and Schanne (2014) look at how residential neighborhoods can serve as a pool of information for an informal labor market and investigate the effect of job referrals through one's residential location. They analyze the relationship between living and working together in the context of job referrals in the Rhine–Ruhr metropolitan area in Germany. They find effects very similar to those in Bayer et al. (2008). Indeed, Hawranek and Schanne (2014) find that sharing the same immediate neighborhood raises the propensity to work together by 0.14 percentage points.

employees who work in the same tract but not in the same establishment (which may be due to commuting patterns or the spatial distribution of jobs and workers). Hiring network effects at the neighborhood level can be inferred if the share of residential neighbors among one's coworkers is significantly higher than that predicted by random hiring. They find that indeed the hiring effect of residential networks is significant, and is especially strong for Hispanics and less-skilled workers, and for smaller establishments. They also find that residential labor-market network effects are stronger within than across races, suggesting racial stratification within residential social networks.

Hellerstein et al. (2014) extend this analysis using Longitudinal Employer-Household Dynamics (LEHD) data, which allow longitudinal observation of matched worker-employer pairs. This rich data source enables them to study additional features of labor-market networks, including wage and turnover effects.⁸ The main findings are that residence-based networks have a robust effect on worker-employee matches, lowering turnover. This effect is especially strong for neighbors within the same racial or ethnic group. For wages, while overall connectedness with residential neighbors tends to raise wages, within-group connectedness has the opposite effect, lowering wages. This is suggestive of overall residence-based networks being associated with more productive matches, while ethnic or racial residential network effects may capture nonwage amenities. In general, this work highlights the neighborhood-specific nature of social networks, at least in the context of labor-market networks.

Schmutte (2015) also uses matched employer-employee data from the LEHD. Adopting an identification strategy similar to that of Bayer et al. (2008), he studies whether residential labor-market networks lead to matches with higher-paying employers. In particular, he estimates a firm-specific wage premium (following Abowd et al., 1999) and finds that workers who live in neighborhoods with higher-quality networks (measured by the average employer-specific wage premium of network members) are more likely to move to jobs with higher wage premiums. This result holds for both employed individuals and unemployed individuals and is not driven by direct referrals from current employees at a given firm.

This study, together with the articles by Hellerstein et al. (2011) discussed above, brings important empirical insights into the nature of referral effects at the neighborhood level by combining the novel identification strategy of Bayer et al. (2008) with very rich data linking workers to firms at the establishment level. The longitudinal aspect of the LEHD is also important in enabling researchers to study dynamic implications such as turnover—in the case of Hellerstein et al. (2014)—as well as the quality of referral networks as in Schmutte (2015).

⁸ Dustmann et al. (2011) and Galenianos (2013) develop predictions for learning models of referrals with regard to wage trajectories and separations as a function of tenure. Datcher (1983) provides empirical evidence on turnover using Panel Study of Income Dynamics data. Brown et al. (2014) provide evidence consistent with learning models of referrals using a unique dataset on a single large US corporation.

In a different setting, [Helmert and Patnam \(2014\)](#) use spatial proximity within villages in Andhra Pradesh, India, to estimate neighborhood effects (spatial peer effects) in the production of cognitive skills for children between the ages of 8 and 12 years. Household locations are precisely mapped within villages, and the authors construct nearest-neighbor adjacency matrices, defined as $\mathbf{G}_r$ in [Section 9.3.1.1](#), to trace the village-level social network. The main idea is then to again use geographical proximity as a proxy for social distance within individual social networks—a theme that appears often in this literature.

The authors use a strategy developed by [Bramoullé et al. \(2009\)](#), among others, to address the reflection problem and to separately identify endogenous from contextual peer effects (see [Manski, 1993](#)). This strategy essentially involves exploiting the partially overlapping nature of individual networks to use friends of friends as valid instruments for one's direct social contacts.⁹ Helmert and Patnam also use various strategies to address the possibility of correlated unobservables or sorting into networks. They find that, on average, a one standard deviation increase in the growth in cognitive achievement of a child's peers increases cognitive achievement of the child by 0.4 standard deviations. Further, social networks help partially insure against idiosyncratic shocks that hit a household and tend to adversely affect the child's cognitive achievement.

[Patacchini and Zenou \(2012a\)](#) test how social networks affect the labor-market outcomes of ethnic minorities in England. They use a strategy similar to that of [Helmert and Patnam \(2014\)](#) by approximating *social proximity* between individuals by *geographical proximity*. Indeed, since ethnic communities tend to be more socially cohesive, a reasonable conjecture is that the density of people living in the same area is a good approximation for the number of direct friends one has (i.e., *strong ties*), especially if the areas are not too large and if people belong to the same ethnic group.¹⁰ In the same spirit, the density of individuals living in neighboring areas will be a measure of friends of friends (i.e., *weak ties*). Using this framework, Patacchini and Zenou look at the relationship between ethnic employment density and the probability of finding a job through social contacts and use *spatial data analysis techniques* to investigate the spatial scale of these effects. They find that the higher the percentage of a given ethnic group living nearby, the higher the probability of finding a job through social contacts. They also find that such an effect is, however, quite localized. It decays very rapidly with distance, losing significance beyond approximately 60 min travel time.¹¹

⁹ See [Section 9.3.2.2](#) for a precise description of this identification strategy.

¹⁰ A similar approximation of the social space (approximated by the physical space) is used in [Wahba and Zenou \(2005\)](#) for the case of Egypt.

¹¹ [Conley and Topa \(2002\)](#) use nonparametric methods to map out several dimensions along which social networks may exist in the context of urban unemployment, using mixtures of geographical, travel time, education, and ethnic distance to characterize social distance.

Conley and Udry (2010) use direct information on farmers' individual social networks in three villages in Ghana to estimate social learning in the adoption of new cultivation technologies. This article contains two important innovations that make it very noteworthy. First, it relies on actual observation of individual networks rather than using spatial proximity as a proxy for them. Second, it lays down an explicit learning model that yields specific implications for the shape of interactions, which enable the authors to identify social effects separately from other, spatially correlated, confounding factors. The sequential nature of plantings and harvests enables the authors to observe how a given farmer reacts to news about his social contacts' choices and outcomes. Consistent with the learning model, the authors find that farmers are more likely to change their fertilizer use when other farmers using similar amounts of fertilizer have lower than expected profits; increase (decrease) their fertilizer use after their social contacts achieve higher profits using more (less) fertilizer than they did; respond more to their neighbors' actions if they only recently started cultivating a particular crop; and respond more to the actions of veteran farmers.

Spatial neighborhood effects also play a role in recent literature on foreclosures, following the recent housing boom and bust cycle in the United States. Campbell et al. (2011) study the effect of sales of foreclosed properties (and more generally, forced sales) on the price of nearby houses in the same neighborhood. They use comprehensive house transactions data from Massachusetts over the 1987–2009 period, matched with information on deaths and bankruptcies of individuals. They find that forced sales in general, and those related to foreclosures in particular, are associated with significant price discounts. Further, local spillover effects from foreclosures are significant (foreclosures lower prices of nearby houses), but decline rapidly with distance. Harding et al. (2009) also find evidence of contagion effects in foreclosures. Several mechanisms can explain such spillovers, ranging from price discovery to the visual impact of run-down or vandalized properties, to a social interaction channel whereby individuals' valuations of their own homes are influenced by their neighbors' valuations (see Ioannides, 2003).

9.2.2 Neighborhood effects estimation using a structural approach

A family of articles uses structural models of social interactions to generate a rich stochastic structure that can be applied to data for estimation. Essentially, these models generate stationary distributions with well-defined properties over space (e.g., excess variance across locations, or positive spatial correlations). The parameters of these models can then be estimated by matching moments from the simulated spatial distribution generated by the model with their empirical counterparts from spatial data on neighborhoods or cities. The model parameters are locally identified (or, in some cases, set identification is attained).

Glaeser et al. (1996) explain the very high variance of crime rates across US cities through a model in which agents' propensity to engage in crime is influenced by neighbors' choices. In doing so, they provide estimates for the range of social interactions. The model is a version of the voter model, in which agents' choices regarding criminal activity are positively affected by their social contacts' choices. One important innovation in this article is to allow for "fixed agents," who are not affected by their neighbors' actions. The variance of crime outcomes across replications of the economy (i.e., cities) is inversely proportional to the fraction of fixed agents in an economy. The distance between pairs of fixed agents in the model yields a measure of the degree of interactions. By matching the empirical cross-city variance of various types of crime with that implied by the model, the authors estimate the extent of neighborhood effects for different types of crime.

Topa (2001) analyzes a structural model of transitions into and out of unemployment to estimate the impact of any local social interaction effects on employment outcomes. The model posits that individuals may receive useful information about job openings from their employed social contacts (the nearest neighbors) but not from their unemployed ones. Formally, the transition probability from employment to unemployment, P_{EU} , depends only on individual attributes and is given by

$$P_{EU} \equiv \Pr(y_{i,t+1} = 0 | y_{it} = 1; X_i) = \alpha(X_i),$$

where y_{it} is the employment status of agent i at time t (1 corresponds to employment and 0 to unemployment) and X_i is a vector of individual characteristics that may affect labor-market outcomes. The reverse probability of finding a job from unemployment, P_{UE} , depends not only on individual characteristics but also on information about job openings transmitted by agent i 's employed social contacts:

$$P_{UE} \equiv \Pr(y_{i,t+1} = 1 | y_{it} = 0; y_t, X_i) = \beta(X_i) + \phi_2(X_i)I_{it}(y_t),$$

where $I_{it}(y_t)$ is the information received about job openings, which depends on the average employment rate of the neighbors of agent i .

The model generates a first-order Markov process over the set of locations (defined at the census tract level), and the positive local feedback implies that the stationary distribution of unemployment in the simulated city exhibits positive spatial correlations. The model parameters are estimated via indirect inference, comparing the simulated spatial distribution of unemployment generated by the model with the empirical one, using census data for the city of Chicago in 1980 and 1990.

The identification strategy in this article relies on the assumption that neighboring census tracts can affect a given tract's employment outcomes only through their employment levels and not through their own attributes, and on the use of ethnic distance and local community boundaries (as identified by residents) to distinguish local social

interactions from other types of spatially correlated shocks. The key assumption is that social spillovers generated by information exchanges within networks are significantly weaker across tracts that are physically close but ethnically very different, or that belong to different local communities; on the other hand, other types of spatially correlated shocks may not be affected by such discontinuities across tracts. Indeed, the spatial correlation in crime outcomes across adjacent tracts does not depend on ethnic distance or on whether the two tracts belong to the same local community. Finally, detailed tract-level controls and fixed effects are also used in the estimation.

Conley and Topa (2007) extend the work of Topa (2001) in several directions, using data for the Los Angeles metropolitan area. First, the model of local interactions and employment transitions is defined at the level of individual agents rather than census tracts. This enables the authors to calibrate a subset of employment transition parameters from retrospective Current Population Survey (CPS) data. Further, the network structure is enriched by allowing for a small number of long “bridging” ties connecting artificial agents in the model that are physically distant from each other. This makes the network structure more realistic, since the sociological literature cited above shows that while many network connections are local in a geographical sense, a sizeable fraction of links occur between locations that are geographically far from each other. Finally, the value of information received about job openings is allowed to vary depending on whether the information is received from members of one’s own ethnic group or from members of other groups.

Formally, the probability of transition into unemployment is assumed to depend only on agents’ characteristics, race/ethnicity, and education:

$$\begin{aligned} \Pr(\gamma_{i,t+1} = 0 | \gamma_{i,t} = 1; A_i, H_i, W_i, X_i) \\ = \Lambda[(\alpha_{1A} + \alpha_{2A}X_i)A_i + (\alpha_{1H} + \alpha_{2H}X_i)H_i + (\alpha_{1W} + \alpha_{2W}X_i)W_i], \end{aligned}$$

where A , H , and W denote African-Americans, Hispanics, and whites, respectively, and $\Lambda(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. In contrast, the probability that an unemployed agent finds a job depends both on his/her own characteristics and on information flows concerning job opportunities that he/she receives from his/her currently employed social contacts at time t . The article takes the extreme modeling stand of allowing transitions out of unemployment to be affected by one’s network contacts, N_i , whereas transitions out of employment are affected by one’s personal characteristics alone. This is done in order to calibrate the parameters of the latter transition probabilities with CPS data.

Information received by agent i is assumed to be a function of the number of employed individuals in his/her set of neighbors. The authors distinguish between the number of employed individuals of an individual’s own race/ethnicity and those of

the other two groups using the notation $I_{i,t}^{\text{Own}}$ and $I_{i,t}^{\text{Other}}$. This allows them to investigate the possibility that information flow may depend on race/ethnicity. The definitions of $I_{i,t}^{\text{Own}}$ and $I_{i,t}^{\text{Other}}$ when agent i is African-American are

$$I_{i,t}^{\text{Own}} \equiv \sum_{k \in N_i} \gamma_{k,t} \times A_k \quad \text{and} \quad I_{i,t}^{\text{Other}} \equiv \sum_{k \in N_i} \gamma_{k,t} \times (1 - A_k).$$

The values of $I_{i,t}^{\text{Own}}$ and $I_{i,t}^{\text{Other}}$ are analogously defined for members of the remaining two racial/ethnic partitions. The probability of transition into employment for African-Americans is defined as

$$\begin{aligned} \Pr(\gamma_{i,t+1} = 1 | \gamma_{i,t} = 0; A_i = 1, X_i, I_{i,t}^{\text{Own}}, I_{i,t}^{\text{Other}}) \\ = \Lambda[\beta_{1A} + \beta_{2A}X_i + \phi_{2A}^{\text{Own}}I_{i,t}^{\text{Own}} + \phi_{2A}^{\text{Other}}I_{i,t}^{\text{Other}}]. \end{aligned}$$

The richer network structure poses an interesting estimation problem: the existence of long ties implies that cross-sectional data will potentially exhibit a strong dependence, with measures such as spatial correlations or mixing coefficients decaying only very slowly as the physical distance increases. This is in contrast to models with only nearest-neighbor interactions, which give rise to a weak cross-sectional dependence. Therefore, even large cross sections should be essentially viewed as a single observation from a vector time series process. [Conley and Topa \(2007\)](#) propose a minimum-distance estimator to obtain point estimates, and a test-statistic inversion method to obtain interval estimates using the minimum distance criterion function as the test statistic.

Thanks to the richer model structure, the parameter estimates can be used to evaluate how well unemployment spell distributions simulated from the model match the empirical ones from the CPS data. The authors find that the model generates too many long unemployment spells (with the estimated parameter values) relative to the data. They further present descriptive methods to illustrate model properties by simulating impulse response functions, in time and in space, to localized unemployment shocks that hit certain neighborhoods in the Los Angeles metropolitan area. They find that, at the stationary distribution, negative employment shocks take a long time to be fully absorbed (more than 2 years), but travel relatively little in space.

Finally, before turning to the literature on social network effects, we wish to mention the work of [Bayer et al. \(2007\)](#) that provides a framework for analyzing the extent and impact of sorting into neighborhoods on the basis of their socioeconomic composition and school quality. Their basic model is a rich discrete-choice model of household location decisions across residential neighborhoods, where household preferences are defined over housing and neighborhood characteristics. This model nests hedonic price regressions as well as traditional discrete-choice models.

The article addresses the endogeneity of school and neighborhood attributes by embedding a boundary discontinuity design into the model.¹² The idea is to use the geographical boundaries of school catchment areas to compare characteristics of households residing on opposite sides of a given boundary. Assuming that the underlying distribution of unobserved attributes affecting location choices is continuous, any observed discontinuity at the boundary in, say, household education or income enables the researcher to estimate the value of school quality.

The boundary discontinuity design is also used to identify and estimate the full distribution of household preferences over schools and neighbors. Household sorting across boundaries generates variation in neighborhood attributes that is related to an observable variable—namely, schools. Therefore, by controlling for differences in school quality on either side of the boundary, one can estimate the value to households of such neighborhood attributes. Thus, by embedding the boundary discontinuity design into a full sorting model, the article provides a strategy to estimate household preferences for housing and neighborhood attributes. This approach can be potentially very useful to jointly model sorting and social interaction effects, allowing the researcher to separately identify both channels.

9.3. NETWORK EFFECTS

We have seen the importance of neighborhood effects on different outcomes (crime, labor, etc.), using both natural experiments and a structural approach. We will now look at the network effects on different outcomes. Here the network will be modeled as a graph where nodes will be agents (workers, consumers, firms, etc.) and links will represent friendship relationships, R&D alliances, criminal interactions, etc.¹³

9.3.1 Network theory

We would like to develop some network theory that will be useful for the empirical estimation of network effects. There is a growing network literature in economics where researchers have been looking at both *network formation* and *games on networks*—that is, games in efforts for which the network is fixed. Here we will mainly describe the main results of games on networks—that is, when the network is taken as given, since there are

¹² This approach builds on the earlier work by Black (1999) and is a special case of the general regression discontinuity design developed by Hahn et al. (2001).

¹³ For overviews of the literature on the economics of networks, see, in particular, the surveys by Jackson (2003, 2004, 2005, 2011), Ioannides and Datcher-Loury (2004), De Martí and Zenou (2011), Zenou (2015a), Jackson and Zenou (2015), and Jackson et al. (2015), as well as the books by Vega-Redondo (2007), Goyal (2007), Jackson (2008), Benhabib et al. (2011), and Jackson and Zenou (2013).

no clear-cut results in the network formation literature. We will, however, return to network formation when we deal with the estimation of peer and network effects in economics.

Although there are many forms that games on networks can take, there are two prominent and broadly encompassing classes of games.¹⁴ The distinction between these types of games relates to whether a given player's relative payoff from taking an action versus not taking an action is *increasing* or *decreasing* in the set of neighbors who take the action. The first class of games on networks, of which coordination games are the canonical example, are games of *strategic complements*. In games of strategic complements, an increase in the actions of other players leads a given player's higher actions to have relatively higher payoffs compared with those of the player's lower actions. Examples of such games include the adoption of a technology, a search in the labor market, R&D efforts, human capital decisions, criminal efforts, smoking behaviors, etc. Games of *strategic substitutes* are such that the opposite is true: an increase in other players' actions leads to relatively lower payoffs from higher actions of a given player. Applications of strategic substitutes include local public good provision and information gathering.

We will here mainly describe games with strategic complements since their empirical applications are the most important in economics.¹⁵ There are two distinct models. In the first one, the *local-aggregate model*, it is the sum of active links that matters. In the second one, the *local-average model*, it is the average sum of active links that matters.

9.3.1.1 The local-aggregate model

Following Calvó-Armengol and Zenou (2004) and Ballester et al. (2006, 2010), we examine a simple model that can encompass any social network. For that, consider a game where $N_r = \{1, \dots, n_r\}$ is a finite set of agents in network $\mathbf{g}_r$ ($r = 1, \dots, \bar{r}$), where $\bar{r}$ is the total number of networks.¹⁶ We represent these social connections by a graph $\mathbf{g}_r$, where $g_{ij,r} = 1$ if agent i is connected to agent j , and $g_{ij,r} = 0$ otherwise. Links are taken to be reciprocal, so $g_{ij,r} = g_{ji,r}$.¹⁷ By convention, $g_{ii,r} = 0$. We denote by $\mathbf{G}_r$ the $n_r \times n_r$ adjacency matrix with entry $g_{ij,r}$, which keeps track of all direct connections. For example, if we consider criminal activities, then agents i and j share their knowledge of delinquent activities if and only if $g_{ij,r} = 1$. For the labor market, a link will indicate the exchange of job information between the individuals. Each agent i decides how much

¹⁴ For a complete overview on the literature on games on networks, see Jackson and Zenou (2014).

¹⁵ We refer to Allouch (2012), Bramoullé and Kranton (2007), Bramoullé et al. (2014), and Jackson and Zenou (2014) for an exposition of the games on networks with strategic substitutes.

¹⁶ Even though we consider only one network in the theoretical analysis, we keep the subscript r because it facilitates the transition to the econometric analysis.

¹⁷ This is only for the sake of the exposition. All the results go through with a directed and weighted network.

effort to exert in some activity, denoted $y_{i,r} \in \mathbb{R}_+$. This could be crime, education, labor search, R&D activities, etc. The utility of each agent i providing effort $y_{i,r}$ in network $\mathbf{g}_r$ is given by

$$u_{i,r}(\mathbf{y}_r, \mathbf{g}_r) = (a_{i,r} + \eta_r + \varepsilon_{i,r})y_{i,r} - \frac{1}{2}y_{i,r}^2 + \phi_1 \sum_{j=1}^{n_r} g_{ij,r} y_{i,r} y_{j,r}, \quad (9.1)$$

where $\phi_1 > 0$ and $\mathbf{y}_r$ is an n_r -dimensional vector of efforts. This utility has two parts. An individual part, $(a_{i,r} + \eta_r + \varepsilon_{i,r})y_{i,r} - \frac{1}{2}y_{i,r}^2$, where the marginal benefits of providing effort $y_{i,r}$ are given by $(a_{i,r} + \eta_r + \varepsilon_{i,r})y_{i,r}$ and increase with own effort $y_{i,r}$. $a_{i,r}$ denotes the *exogenous heterogeneity* of agent i that captures the *observable* characteristics of individual i (e.g., sex, race, age, parental education) and the observable *average* characteristics of individual i 's best friends—that is, the average level of parental education of i 's friends, etc. (*contextual effects*). To be more precise, $a_{i,r}$ can be written as

$$a_{i,r} = \sum_{m=1}^M \beta_m x_{i,r}^m + \frac{1}{g_{i,r}} \sum_{m=1}^M \sum_{j=1}^{n_r} g_{ij,r} x_{j,r}^m \gamma_m, \quad (9.2)$$

where $g_{i,r} = \sum_{j=1}^{n_r} g_{ij,r}$ is the number of direct links of individual i , x_i^m is a set of M variables accounting for observable differences in individual characteristics of individual i , and β_m and γ_m are parameters. In the utility function, η_r denotes the unobservable network characteristics—for example, the prosperous level of the neighborhood/network $\mathbf{g}_r$ —and $\varepsilon_{i,r}$ is an error term, which captures other uncertainty in the proceeds from the effort. Both η_r and $\varepsilon_{i,r}$ are observed by the agents (when choosing the effort level) but not by the econometrician.

The second part of the utility function, $\phi_1 \sum_{j=1}^{n_r} g_{ij,r} y_{i,r} y_{j,r}$, corresponds to the local-aggregate effect since each agent i is affected by the sum of the efforts of the agents for which he/she has a direct connection. The higher the number of active connections, the higher the marginal utility of providing his/her own effort. This is a game with strategic complementarities since

$$\frac{\partial^2 u_{i,r}(\mathbf{y}_r, \mathbf{g}_r)}{\partial y_{i,r} \partial y_{j,r}} = \phi_1 g_{ij,r} \geq 0.$$

At equilibrium, each agent maximizes his/her utility (9.1), and the best-reply function, for each $i = 1, \dots, n$, is given by

$$y_{i,r} = \phi_1 \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} + a_{i,r} + \eta_r + \varepsilon_{i,r}. \quad (9.3)$$

Denote by $\mu_1(\mathbf{g}_r)$ the largest eigenvalue of network $\mathbf{g}_r$ and $\alpha_{i,r} \equiv a_{i,r} + \eta_r + \varepsilon_{i,r}$ with the corresponding nonnegative n_r -dimensional vector $\boldsymbol{\alpha}_r$. It can be shown that if

$\phi_1 \mu_1(\mathbf{g}_r) < 1$, the peer effect game with payoffs (9.1) has a unique Nash equilibrium in pure strategies given by

$$\mathbf{y}_r^* \equiv \mathbf{y}_r^*(\mathbf{g}_r) = \mathbf{b}_{\alpha_r}(\mathbf{g}_r, \phi_1), \quad (9.4)$$

where $\mathbf{b}_{\alpha_r}(\mathbf{g}_r, \phi_1)$ is the weighted Katz–Bonacich centrality, a well-known measure defined by Katz (1953) and Bonacich (1987). Formally,

$$\mathbf{b}_{\alpha_r}(\mathbf{g}_r, \phi_1) = (\mathbf{I}_{n_r} - \phi_1 \mathbf{G}_r)^{-1} \alpha_r = \sum_{k=0}^{\infty} \phi_1^k \mathbf{G}_r^k \alpha_r, \quad (9.5)$$

where $\mathbf{I}_{n_r}$ is the $(n_r \times n_r)$ identity matrix, $\alpha_r = \mathbf{a}_r + \eta_r \mathbf{1}_{n_r} + \boldsymbol{\varepsilon}_{i,r}$, and $\mathbf{1}_{n_r}$ is an n_r -dimensional vector of ones. In words, the Katz–Bonacich centrality of agent i counts the total number of paths (not just the shortest paths) in $\mathbf{g}_r$ starting from i , weighted by a decay factor that decreases with the length of these paths. This is captured by the fact that the matrix $\mathbf{G}_r^k$ keeps track of the indirect connections in the network—that is, $g_{ij,r}^{[k]} \geq 0$ measures the number of paths of length $k \geq 1$ in $\mathbf{g}_r$ from i to j . This result shows that more central agents in the network will exert more effort. This is intuitively related to the equilibrium behavior, as the paths capture all possible feedbacks. In our case, the decay factor depends on how the effort of others enters into one's own effort's payoff. It is then straightforward to show that, for each individual i , the equilibrium utility is

$$u_{i,r}(\mathbf{y}_r^*, \mathbf{g}_r) = \frac{1}{2} [b_{\alpha_r}(\mathbf{g}_r, \phi_1)]^2,$$

so the equilibrium utility of each criminal is proportional to his/her Katz–Bonacich centrality. It is important to understand that there are magnifying or *social multiplying* effects due to network relationships, which are captured by the Katz–Bonacich centrality. To understand this last point, consider the case of a dyad for which $n_r = 2$ and, for simplicity, assume that $\alpha_{1,r} = \alpha_{2,r} = \alpha_r$. If there were no interactions—that is, $g_{12,r} = g_{21,r} = 0$ —then the unique Nash equilibrium would be $y_{1,r}^* = y_{2,r}^* = \alpha_r$. With social interactions (i.e., $g_{12,r} = g_{21,r} = 1$), if $\phi_1 < 1$, the unique Nash equilibrium is given by

$$y_{1,r}^* = y_{2,r}^* = \frac{\alpha_r}{1 - \phi_1}. \quad (9.6)$$

In the dyad, complementarities lead to an effort level above the equilibrium value for an isolated player. The factor $1/(1 - \phi_1) > 1$ is often referred to as a *social multiplier*. An important part of the empirical analysis of network effects would be to estimate ϕ_1 . If, for example, the estimated value of ϕ_1 is 0.5, then the social multiplier is equal to 2. Take the example of crime. This means that if a criminal would commit crimes alone, then he/she will commit α_r crimes, and this will be determined only by his/her observable characteristics. Now, if this criminal has only one criminal friend, compared with the case where he/she operates alone, he/she will increase his/her crime effort by 100%—that is,

he/she will commit $2\alpha_r$ crimes. This is not due to his/her characteristics but only to the fact that he/she interacts with another criminal.

9.3.1.2 The local-average model

Following [Patacchini and Zenou \(2012b\)](#), let us now develop the *local-average* model, where the *average effort level* of direct links affects utility. For that, let us denote the set of individual i 's direct connections as

$$N_{i,r}(g_r) = \{j \neq i, g_{ij,r} = 1\},$$

which cardinality is $g_{i,r}$. Let $g_{ij,r}^* = g_{ij,r}/g_{i,r}$ for $i \neq j$, and set $g_{ii,r}^* = 0$. By construction, $0 \leq g_{ij,r}^* \leq 1$. Note that $\mathbf{g}_r^*$ is a row-normalization of the initial network $\mathbf{g}_r$, as illustrated in the following example, where $\mathbf{G}_r$ and $\mathbf{G}_r^*$ are the adjacency matrices of $\mathbf{g}_r$ and $\mathbf{g}_r^*$, respectively.

Example 9.1. Consider the following network $\mathbf{g}_r$:

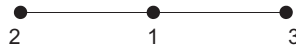


Figure 9.1 A star network.

Then,

$$\mathbf{G}_r = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{G}_r^* = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

As above, $y_{i,r}$ denotes the effort level of individual i in network r . Denote by $\bar{y}_{i,r}$ the average effort of individual i 's best friends. It is given by

$$\bar{y}_{i,r} = \frac{1}{g_{i,r}} \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} = \sum_{j=1}^{n_r} g_{ij,r}^* y_{j,r}. \quad (9.7)$$

Each individual i selects an effort $y_{i,r} \geq 0$ and obtains a payoff given by the following utility function:

$$u_{i,r}(\mathbf{y}_r, \mathbf{g}_r) = (a_{i,r}^* + \eta_r^* + \epsilon_{i,r}^*) y_{i,r} - \frac{1}{2} y_{i,r}^2 - \frac{\lambda_2}{2} (y_{i,r} - \bar{y}_{i,r})^2, \quad (9.8)$$

with $\lambda_2 > 0$. All the parameters have the same interpretation as in (9.1). Let us now interpret the peer-effect part of this utility function since it is the only aspect that differs from (9.1). Indeed, the last term, $\frac{\lambda_2}{2} (y_{i,r} - \bar{y}_{i,r})^2$, reflects the influence of the individual's friends' behavior on the individual's own action. It is such that each individual wants to minimize

the *social distance* between himself/herself and his/her reference group, where λ_2 is the parameter describing the *taste for conformity*. Here, the individual loses utility $\frac{\lambda_2}{2}(\gamma_{i,r} - \bar{\gamma}_{i,r})^2$ from failing to conform to others. This is the standard way economists have been modeling conformity (see, among others, Kandel et al., 1992; Bernheim, 1994; Akerlof, 1997; Fershtman and Weiss, 1998; Glaeser and Scheinkman, 2001).

Observe that beyond the idiosyncratic heterogeneity, $a_{i,r}^*$, there is a second type of heterogeneity, referred to as *peer heterogeneity*, which captures the differences between individuals due to network effects. Here it means that individuals have different types of friends and thus different reference groups $\bar{\gamma}_{i,r}$. As a result, the social norm each individual i faces is endogenous and depends on his/her location in the network as well as the structure of the network. Indeed, in a star-shaped network (as the one described in Figure 9.1) where each individual is at most distance 2 from each of the other individuals, the value of the social norm will be very different from a circle network, where the distance between individuals can be very large.

We now characterize the Nash equilibrium of the game where agents choose their effort level $\gamma_{i,r} \geq 0$ simultaneously. When $\phi_2 < 1$, the peer effect game with payoffs (9.8) has a unique interior Nash equilibrium in pure strategies for each $i = 1, \dots, n_r$ given by

$$\gamma_{i,r} = \phi_2 \sum_{j=1}^{n_r} g_{ij,r}^* \gamma_{j,r} + a_{i,r} + \eta_r + \varepsilon_{i,r}, \quad (9.9)$$

where $\phi_2 \equiv \lambda_2/(1 + \lambda_2)$, $a_{i,r} \equiv a_{i,r}^*/(1 + \lambda_2)$, $\eta_r \equiv \eta_r^*/(1 + \lambda_2)$, and $\varepsilon_{i,r} \equiv \varepsilon_{i,r}^*/(1 + \lambda_2)$. In matrix form, (9.9) can be written as

$$\mathbf{y}_r = (\mathbf{I}_{n_r} - \phi_2 \mathbf{G}_r^*)^{-1} \boldsymbol{\alpha}_r. \quad (9.10)$$

9.3.1.3 Local aggregate or local average? theoretical considerations

In the local-aggregate model, it is the *sum of the efforts of his/her peers* that affects the utility of individual i . So the more individual i has active (i.e., providing effort) friends, the higher is his/her utility. In contrast, in the *local-average* model, it is the deviation from the *average of the efforts of his/her peers* that affects the utility of individual i . So the closer individual i 's effort is to the average of his/her friends' efforts, the higher is his/her utility.

Consequently, the two models are quite different from an economic viewpoint, even though, from a purely technical point of view, they are not that different (compare the best-reply functions (9.3) and (9.9)). In particular, the adjacency matrix $\mathbf{G}_r$ of direct links of the network totally characterizes the peer effects in the *local-aggregate* model, whereas it is a transformation of this matrix $\mathbf{G}_r$ to a weighted stochastic matrix $\mathbf{G}_r^*$ that characterizes the peer effects in the *local-average* model. This means that, in equilibrium, in the former model, individuals are positively affected by the sum of their friends' effort

(non-row-normalized $\mathbf{G}_r$), while in the latter, they are positively affected by the average effort of their friends (row-normalized $\mathbf{G}_r$).

From an economic viewpoint, in the *local-aggregate* model, even if individuals were *ex ante* identical (in terms of $a_{i,r}$ and $\varepsilon_{i,r}$), different positions in the network would imply different effort levels, because it is the sum of the efforts that matter. This would not be true in the *local-average* model since, in that case, the position in the network would not matter since it is the deviation from the average effort of friends that affects the utility.

9.3.2 Empirical aspects of social networks: Structural approach

We now use the previous models from [Section 9.3.1](#) to estimate the empirical effects of networks. We will first start with the econometric issues and then state some empirical results, especially those relevant for policy issues.

9.3.2.1 Linear-in-means model: The reflection problem

In the standard *linear-in-means model*, each agent is affected by the average action of his/her reference group. This is the standard *peer-effect* model (see [Section 9.2](#)), where the reference group is the same for all individuals. For example, in crime, the criminal activity of individual i will depend on the average criminal activity of the neighborhood where he/she lives. As a result, the right-hand side of this equation will be same for all individuals living in the same neighborhood (typically a census tract in the United States). In education, this would mean that the grades of each student i will be determined by the average grades in the school or in the classroom to which that student belongs. Implicitly, when talking about neighborhood effects, we assume that each delinquent interacts in the same way with everybody in his/her neighborhood (if we think of a census tract then, on average, this would mean that individual i interacts with 4000 people). Similar assumptions have to be made for the classroom or school example. In contrast, in a network approach where the dyad is the unit of interest, one assumes that each individual interacts with only his/her direct friends. As we have seen in (9.4), the individual is also influenced by indirect links but he/she puts a lower weight on them. In (9.4), we showed that the weight is proportional to the distance in the network as captured by the Katz–Bonacich centrality of each individual. If an individual is five links away from individual i , then the weight is ϕ_1^5 , which is small given that ϕ_1 is less than 1.

Let us return to the linear-in-means model. From an econometric viewpoint, the simultaneity in behavior of interacting agents (i.e., the endogenous action of each agent is affected by the average endogenous action of the reference group) introduces a perfect collinearity between the expected mean outcome of the group and its mean characteristics. Therefore, it is difficult to differentiate between the effect of peers' choice of effort and peers' characteristics that do impact on their effort choice (the so-called *reflection problem*; [Manski, 1993](#)). Basically, the reflection problem arises because, in the standard approach, individuals interact in groups—that is, individuals are affected by all individuals

belonging to their group and by nobody outside the group. In other words, groups do not overlap. Let us explain formally the reflection problem in the linear-in-means model.

The reflection problem (Manski, 1993) arises when it is not possible to disentangle the endogenous effects from the contextual effects. The basic linear-in-means model can be written as

$$y_{i,r} = \phi_2 \mathbb{E}(y_r) + \gamma \mathbb{E}(x_r) + \beta x_{i,r} + \varepsilon_{i,r}, \quad (9.11)$$

where, as above, $y_{i,r}$ is the effort or outcome (e.g., education, crime, etc.) of individual i belonging to group r , $x_{i,r}$ is an observable characteristic of individual i 's (i.e., i 's characteristics such as his/her gender, age, education, etc.)¹⁸ in group r , $\mathbb{E}(y_r)$ denotes the average of the efforts/outcomes in the peer group r of individual i , $\mathbb{E}(x_r)$ denotes the average of the characteristics (or characteristics specific to group r) in the peer group r of individual i , and $\varepsilon_{i,r}$ is an error term. We want to identify $\phi_2 > 0$ (i.e., the *endogenous peer effect*) and separate it from $\gamma > 0$, the *exogenous contextual effect*. Observe that, contrary to (9.1) or (9.8), r refers to a group (i.e., neighborhood, school, class, etc.) and not to a network. Assume $\mathbb{E}(\varepsilon_{i,r} | y_r, x_r) = 0$. If we take the average over peer group r of Equation (9.11) and solve this equation, we obtain

$$\mathbb{E}(y_r) = \left(\frac{\gamma + \beta}{1 - \phi_2} \right) \mathbb{E}(x_r).$$

Plugging the value of $\mathbb{E}(y_r)$ into (9.11) yields

$$y_{i,r} = \left(\frac{\phi_2(\gamma + \beta) + \gamma(1 - \phi_2)}{(1 - \phi_2)} \right) \mathbb{E}(x_r) + \beta x_{i,r} + \varepsilon_{i,r}.$$

If one estimates this equation, there is an identification problem since ϕ_2 (endogenous peer effects) and γ (exogenous contextual effects) cannot be separately identified. There are three estimated coefficients and four structural parameters, and thus identification fails. This is the *reflection problem* (Manski, 1993). In terms of policy implications of peer effects, it is of paramount importance to *separately* identify peer or endogenous effects from contextual or exogenous effects (Manski, 1993, 2000; Moffitt, 2001) because endogenous effects generate a *social multiplier*, while contextual effects do not. Consider, for example, peer effects in crime. A special program targeting some individuals will have multiplier effects: the individual affected by the program will reduce his/her criminal activities and will influence the criminal activities of his/her peers, which, in turn, will affect the criminal activities of their peers, and so on. In contrast, any policy affecting

¹⁸ For the sake of the presentation, we consider only one characteristic of individual i and not the sum of characteristics $\sum_{j=1}^{n_r} x_{j,r}^m$ as in (9.2). The extension to more than one characteristics is straightforward.

contextual effects will have no social multiplier effects (e.g., improving the gender composition of students at school).

Let us show now that in the case of social networks the reflection problem nearly never arises because the reference group is the set of network contacts each individual has. Following [Bramoullé et al. \(2009\)](#), let us show how using a network approach, we can solve the reflection problem. We will also show how it can help solve the problem of endogenous network formation and, more generally, correlated effects.

9.3.2.2 Social networks: The local-average model

So far the reference group was the same for all individuals (the neighborhood, the class, etc.) since peer effects are an average intragroup externality that affects identically all the members of a given group. In particular, the group boundaries are arbitrary and at a fairly aggregate level. In contrast, social networks use the smallest unit of analysis for cross influences: the dyad (two-person group). In that case, the reference group of individual i is his/her direct links (e.g., friends). Furthermore, the reference group of individual j , who is a best friend of individual i , is not the same as for individual i because individual j may have some best friends that are not individual i 's best friends. As a result, Equation (9.11) can now be written as

$$y_{i,r_i} = \phi_2 \mathbb{E}(y_{r_i}) + \gamma \mathbb{E}(x_{r_i}) + \beta x_{i,r_i} + \varepsilon_{i,r_i}, \quad (9.12)$$

where r_i is now the reference group of individual i (see (9.7)), so $y_{i,r_i} \equiv y_{i,r}$ and $\mathbb{E}(y_{r_i}) \equiv \bar{y}_{i,r}$ where $\bar{y}_{i,r}$ is defined by (9.7). Similarly, if we consider more than one characteristics for individual i , then, using (9.2), we have $\gamma \mathbb{E}(x_{r_i}) + \beta x_{i,r_i} \equiv a_{i,r}$. As a result, adding the network fixed effect η_r , we find Equation (9.12) is exactly equivalent to (9.9), which corresponds to the unique Nash equilibrium of the *local-average model* where the utility function is given by (9.8).

Let us write (9.12) or (9.9) in matrix form (with network fixed effect). We have

$$\mathbf{Y}_r = \phi_2 \mathbf{G}_r^* \mathbf{Y}_r + \beta \mathbf{X}_r + \gamma \mathbf{G}_r^* \mathbf{X}_r + \eta_r \mathbf{1}_{n_r} + \boldsymbol{\varepsilon}_r, \quad (9.13)$$

where $\bar{r}$ is the total number of networks in the sample, n_r is the number of individuals in the r th network, $n = \sum_{r=1}^{\bar{r}} n_r$ is total number of sample observations, $\mathbf{Y}_r$ is an $n \times 1$ vector of observations on the dependent (decision) variable, $\mathbf{G}_r^*$ is the $n \times n$ row-normalized matrix of $\mathbf{G}_r$, $\mathbf{X}_r$ is an $n \times 1$ vector of observations on the exogenous variables, $\mathbf{1}_{n_r}$ is an n_r -dimensional vector of 1, and $\varepsilon_{i,r}$'s (whose corresponding vector is $\boldsymbol{\varepsilon}_r$) are independent and identically distributed innovations with zero mean and variance σ^2 for all i and r . Assume $E[\boldsymbol{\varepsilon} | \mathbf{G}_r, \mathbf{X}_r] = 0$. Then (9.13) is similar to a spatial autoregressive model ([Anselin, 1988](#)).

The network-specific parameter η_r is allowed to depend on $\mathbf{G}_r$, $\mathbf{G}_r^*$, and $\mathbf{X}_r$ as in a fixed effect panel data model. To avoid the incidental parameter problem when the number of groups $\bar{r}$ is large, we eliminate the term $\eta_r \mathbf{1}_{n_r}$ using the deviation from group mean

projector $\mathbf{J}_r = \mathbf{I}_{n_r} - \frac{1}{n_r} \mathbf{1}_{n_r} \mathbf{1}_{n_r}^T$. This transformation is analogous to the *within* transformation for a fixed effect panel data model. As $\mathbf{J}_r \mathbf{1}_{n_r} = 0$, the transformed network model is

$$\mathbf{J}_r \mathbf{Y}_r = \phi_2 \mathbf{J}_r \mathbf{G}_r^* \mathbf{Y}_r + \beta \mathbf{J}_r \mathbf{X}_r + \gamma \mathbf{J}_r \mathbf{G}_r^* \mathbf{X}_r + \mathbf{J}_r \varepsilon_r. \quad (9.14)$$

If $\phi_2 \beta + \gamma \neq 0$, [Bramoullé et al. \(2009\)](#) demonstrates that identification of the local-average model is possible since $[\mathbf{J}_r \mathbf{G}_r^{*2} \mathbf{X}_r, \mathbf{J}_r \mathbf{G}_r^{*3} \mathbf{X}_r, \dots]$ can be used as instrumental variables for the endogenous effect. Note, in a natural network, if individuals i and j are friends and individuals j and k are friends, it does not necessarily imply that individuals i and k are also friends. The *intransitivity in social connections* provides an exclusion restriction such that the characteristics of the friends' friends $\mathbf{G}_r^{*2} \mathbf{X}_r$ may not be perfectly correlated with the own characteristics $\mathbf{X}_r$ and the characteristics of the friends $\mathbf{G}_r^* \mathbf{X}_r$. Thus, one can use instrumental variables like $\mathbf{J}_r \mathbf{G}_r^{*2} \mathbf{X}_r$ to identify endogenous and contextual effects. On the basis of this important observation, [Bramoullé et al. \(2009\)](#) have shown that if the matrices $\mathbf{I}_{n_r}$, $\mathbf{G}_r^*$, and $\mathbf{G}_r^{*2}$ are linearly independent, social effects are identified. Thus, the natural exclusion restrictions induced by the network structure (existence of an intransitive triad) guarantee identification of the model.¹⁹

Although this setting allows us to solve the reflection problem, the estimation results might still be flawed because of the presence of unobservable factors affecting both individual and peer behavior. It is thus difficult to disentangle the *endogenous peer effects* from the *correlated effects*—that is, from effects arising from the fact that individuals in the same group tend to behave similarly because they face a common environment. If individuals are not randomly assigned into groups, this problem might originate from the possible sorting of agents. If the variables that drive this process of selection are not fully observable, potential correlations between (unobserved) group-specific factors and the target regressors are major sources of bias. In our case, two types of possibly correlated effects arise—that is, at the network level and at the peer group level.

The use of *network fixed effects* proves useful in this respect. Assume, indeed, that agents self-select into different networks in a first step, and that link formation takes place within networks in a second step. Then, as [Bramoullé et al. \(2009\)](#) observe, if linking decisions are uncorrelated with the observable variables, this two-step model of link formation generates network fixed effects. Assuming additively separable network heterogeneity, a within-group specification is able to control for these correlated effects. Indeed, by subtracting from the individual-level variables the network average, one can identify social effects and disentangle endogenous effects from correlated effects.

[Bramoullé et al. \(2009\)](#) also deal with this problem in the context of networks. They show that if the matrices $\mathbf{I}_{n_r}$, $\mathbf{G}_r$, $\mathbf{G}_r^2$, and $\mathbf{G}_r^3$ are linearly independent, then by subtracting from the variables the network component average (or the average over neighbors, i.e.,

¹⁹ [Cohen-Cole \(2006\)](#) and [Lee \(2007\)](#) present a similar argument—that is, the use of out-group effects—to achieve the identification of the endogenous group effect in the linear-in-means model.

direct friends), one can again identify social effects and disentangle endogenous effects from correlated effects. The condition is more demanding because some information has been used to deal with the fixed effects.²⁰

A number of articles using network data have used this strategy to deal with the identification and estimation of peer effects of (9.13) with correlated effects (e.g., Lee, 2007; Bramoullé et al., 2009; Calvó-Armengol et al., 2009; Lee et al., 2010; Lin, 2010; Liu and Lee, 2010; Liu et al., 2012; Patacchini and Zenou, 2012b; Boucher et al., 2014). As stated above, these articles exploit the architecture of network contacts to construct valid instrumental variables for the endogenous effect (i.e., the characteristics of indirect friends) and to use network fixed effects as a remedy for the selection bias that originates from the possible sorting of individuals with similar unobserved characteristics into a network. The underlying assumption is that such unobserved characteristics are common to the individuals within each network.

9.3.2.3 Social networks: The local-aggregate model

We have seen so far that the local-average model is well identified under some conditions on the adjacency matrix. Most researchers have used this model to estimate peer or network effects. However, in some cases, the local-aggregate model also seems to be a natural outcome of a game. In that case, do the identification conditions proposed by Bramoullé et al. (2009) still apply? Liu et al. (2012) show that they do not.

If we now consider the local-aggregate model presented in Section 9.3.1.1, then the matrix equivalent of the best-reply functions (9.3) in the theoretical model is

$$\mathbf{Y}_r = \phi_1 \mathbf{G}_r \mathbf{Y}_r + \beta \mathbf{X}_r + \gamma \mathbf{G}_r^* \mathbf{X}_r + \eta_r \mathbf{1}_{n_r} + \varepsilon_r, \quad (9.15)$$

where the only difference from the local-average model is that, for the endogenous effect, $\mathbf{G}_r$ is not row normalized.

Liu et al. (2012) demonstrate that the identification conditions for the *local-aggregate* model are weaker than those for the *local-average* model because one can use the sum of rows from the adjacency matrix as an additional instrument in the local-aggregate model, while this is not possible in the local-average model since it is always equal to 1. To be more precise, Liu et al. (2012) show that when $\mathbf{G}_r$ has *nonconstant row sums* for some network r , then if $\mathbf{I}_{n_r}$, $\mathbf{G}_r$, $\mathbf{G}_r^*$, and $\mathbf{G}_r \mathbf{G}_r^*$ are linearly independent and $|\beta| + |\gamma| + |\eta_r| \neq 0$, the model is identified.²¹

Figure 9.2 gives an example where identification is possible for the local-aggregate model but fails for the local-average model. Consider a dataset where each network is represented by the graph in Figure 9.2 (a star-shaped network). For the row-normalized adjacency matrix $\mathbf{G}_s^*$, it is easy to see that $\mathbf{G}_s^{*3} = \mathbf{G}_s^*$. Therefore, it follows from Bramoullé

²⁰ See Blume et al. (2011) for an overview of these econometric issues.

²¹ They also have some conditions for identification when $\mathbf{G}_r$ has *constant row sums*.

et al. (2009) that the local-average model (9.13) is not identified. On the other hand, as $\mathbf{G}_r$ in Figure 9.2 has *nonconstant row sums* and $\mathbf{I}_{n_r}$, $\mathbf{G}_r$, $\mathbf{G}_r^*$, and $\mathbf{G}_r \mathbf{G}_r^*$ are linearly independent, it follows that the local-aggregate model (9.15) can be identified for this network.

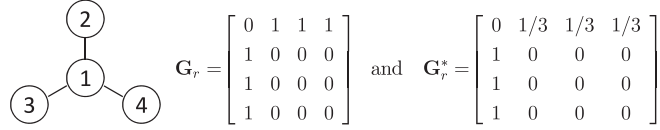


Figure 9.2 An example where the local-aggregate model can be identified but the local-average model cannot be identified.

9.3.2.4 Testing the local-average model against the local-aggregate model

Liu et al. (2014) propose a test to evaluate whether the local-average model is more relevant in some activities than the local-aggregate model, and vice versa. For that, they first develop a theoretical model by considering the following utility function:

$$u_{i,r}(\mathbf{y}_r, \mathbf{g}_r) = \underbrace{(\alpha_{i,r}^* + \lambda_1 \sum_{j=1}^{n_r} g_{ij,r} y_{j,r})}_{\text{benefit}} y_{i,r} - \underbrace{\frac{1}{2} [\gamma_{i,r}^2 + \lambda_2 (y_{i,r} - \sum_{j=1}^{n_r} g_{ij,r}^* y_{j,r})^2]}_{\text{cost}}. \quad (9.16)$$

This is the so-called *hybrid model* because it includes both local-aggregate and local-average aspects of preferences. The best-reply function of each individual i is given by

$$y_{i,r} = \phi_1 \sum_{j=1}^{n_r} g_{ij,r} y_{j,r} + \phi_2 \sum_{j=1}^{n_r} g_{ij,r}^* y_{j,r} + \alpha_{i,r}, \quad (9.17)$$

where $\alpha_{i,r} \equiv \alpha_{i,r}^* / (1 + \lambda_2)$, $\phi_1 \equiv \lambda_1 / (1 + \lambda_2)$, and $\phi_2 \equiv \lambda_2 / (1 + \lambda_2)$. It is easily verified that when $\lambda_1 = 0$, we are back to the local-average model (see (9.3)), while when $\lambda_2 = 0$, we are back to the local-aggregate model (see (9.9)).

Denote by $g_r^{\max}$ the highest degree in network r —that is, $g_r^{\max} = \max_i g_{i,r}$. If $\phi_1 \geq 0$, $\phi_2 \geq 0$, and $g_r^{\max} \phi_1 + \phi_2 < 1$, then the network game with payoffs (9.16) has a unique interior Nash equilibrium in pure strategies given by

$$\mathbf{Y}_r = (\mathbf{I}_{n_r} - \phi_1 \mathbf{G}_r - \phi_2 \mathbf{G}_r^*)^{-1} \boldsymbol{\alpha}_r. \quad (9.18)$$

In terms of econometrics, with network fixed effects, (9.17) can be written in matrix form as

$$\mathbf{Y}_r = \phi_1 \mathbf{G}_r \mathbf{Y}_r + \phi_2 \mathbf{G}_r^* \mathbf{Y}_r + \beta \mathbf{X}_r + \gamma \mathbf{G}_r^* \mathbf{X}_r + \eta_r \mathbf{1}_{n_r} + \varepsilon_r. \quad (9.19)$$

Liu et al. (2014) then test the local-aggregate model against the local-average model, and vice versa. For that, they extend Kelejian's (2008) J test for spatial econometric models to differentiate between the local-aggregate and the local-average endogenous peer effects in an econometric network model with network fixed effects. The idea of the J test is as

follows. If a given model contains the correct set of regressors, then including the fitted values of an alternative model (or of a fixed number of competing models) into the null model should provide no significant improvement.

9.3.2.5 Endogenous network formation

The instrumental variable strategy proposed by [Bramoullé et al. \(2009\)](#) and developed above, however, works if the network is exogenous (i.e., it works conditional on the exogeneity of the adjacency matrix $\mathbf{G}_r$), which is not usually the case unless one has a controlled field experiment so that the network was formed exogenously (e.g., see [Carrell et al., 2009, 2013](#)). Alternatively, one needs to be able to plausibly rule out unobserved factors or develop instruments that are clearly exogenous to the interaction structure, or else model network formation and try to account for factors that could have substantial influences on both behavior and network formation.²²

An approach to dealing with this comes from [Goldsmith-Pinkham and Imbens \(2013\)](#). Under homophily, linked individuals are likely to be similar not only in terms of *observed characteristics* but also in terms of *unobserved characteristics* that could influence their behavior. By failing to account for similarities in (unobserved) characteristics, one might mistakenly attribute similar behaviors to peer influence when they simply result from similar characteristics. In order to highlight the problem, let us write the model (9.13) as follows²³:

$$\mathbf{Y}_r = \phi_2 \mathbf{G}_r^* \mathbf{Y}_r + \beta \mathbf{X}_r + \gamma \mathbf{G}_r^* \mathbf{X}_r + \eta_r \mathbf{1}_{n_r} + \underbrace{\xi \mathbf{v}_r + \mathbf{e}_r}_{\mathbf{e}_r}, \quad (9.20)$$

where $\mathbf{v}_r = (v_{1,r}, \dots, v_{n_r,r})^T$ denotes a vector of *unobserved characteristics* at the individual level and $\mathbf{e}_r = (e_{1,r}, \dots, e_{n_r,r})^T$ is a vector of random disturbances. Let us consider a network formation model where the variables that explain the links between individuals i and j belonging to network r (i.e., $g_{ij,r}$) are the distances between them in terms of observed and unobserved characteristics—that is,

$$g_{ij,r} = \alpha + \sum_{m=1}^M \delta_m |x_{i,r}^m - x_{j,r}^m| + \theta |v_{i,r} - v_{j,r}| + \eta_r + u_{ij,r}. \quad (9.21)$$

Homophily behavior in the unobserved characteristics implies that $\theta < 0$ —that is, the closer two individuals are in terms of unobservable characteristics, the higher is the

²² Observe that this problem can be mitigated if one observes the network at different points in time. For example, [König et al. \(2014a\)](#) study R&D collaborations between firms for over 20 years and use time and firm fixed effects. In that case, if the unobservables that make firms create R&D collaborations do not change over time, this method should be satisfactory.

²³ For the argument, it does not matter if we apply the local-average or the local-aggregate model.

probability that they are friends. If ζ is different from zero, then the network $\mathbf{G}_r$ in model (9.20) is endogenous.

A testable implication of this problem would be to find a negative correlation between the predicted probability of forming a link (based on observable characteristics), as measured by $\widehat{g}_{ij,r}$, and the unobserved similarity in pairs, as measured by the difference in residuals from Equation (9.20), $|\hat{\varepsilon}_{i,r} - \hat{\varepsilon}_{j,r}|$.²⁴ Evidence against network endogeneity would be the finding of a zero correlation.²⁵

Another way of dealing with this problem is to simultaneously (or sequentially) estimate (9.21) and (9.20) as in Goldsmith-Pinkham and Imbens (2013). For example, König et al. (2014a) propose a three-stage least squares estimation, where in the first stage, a network formation model similar to (9.21) is estimated.²⁶ Then, using the *predicted value of the adjacency matrix*, the authors perform the other stages using a similar instrumental variable approach as in Bramoullé et al. (2009) and described above.²⁷

One of the challenges of the approach of Goldsmith-Pinkham and Imbens is that modeling network formation on a link-by-link basis is not very realistic because one must account for interdependencies (Chandrasekhar and Jackson, 2013; Jackson, 2013; Jackson et al., 2015). There is a powerful and natural formulation of network formation models that takes these interdependencies into account. They are known as *exponential random graph models*.²⁸ However, because the number of possible networks on a given number of nodes is an exponential function of the number of nodes, it is practically impossible to estimate the likelihood of a given network, and thus there is an important computational hurdle (see the discussion in Chandrasekhar and Jackson, 2013). Another possible approach is to model the network as an evolving process (see, e.g., Snijders, 2001; Christakis et al., 2010; Mele, 2013; König et al., 2014b) as such models allow for dependencies in that new links form on the basis of the network existing at the time.

9.3.2.6 Multiple equilibria

Whereas the previous sections focus mainly on linear models, we now consider nonlinear models of social interactions, which typically generate multiple equilibria, as they induce externalities.²⁹ Bisin et al. (2011a) use the Brock and Durlauf (2001) model of social interactions to study network effects in smoking, using National Longitudinal Survey of

²⁴ Under assortative matching (i.e., heterophily), the correlation should be positive.

²⁵ See Patacchini et al. (2014), who perform such a test.

²⁶ The idea to use the *predicted* adjacency matrix $\mathbf{G}_r^*$ to construct instruments can also be found in Kelejian and Piras (2014) and Comola and Prina (2014).

²⁷ In Section 9.4.3, we discuss the article by Del Bello et al. (2014), who also simultaneously estimate (9.21) and (9.20).

²⁸ See Jackson (2008) for background on these models.

²⁹ Glaeser and Scheinkman (2001) derive sufficient conditions on the strength of interactions to generate multiplicity.

Adolescent to Adult Health (Add Health) data on high schools. (For a description of the Add Health data, see Section 9.3.3.2). The model is an extension of the canonical random utility discrete choice model, where the utility of each choice is affected not only by individual attributes and a random term but also by a term that captures influences from network contacts. Thus, agents solve the following program:

$$\max_{y_i \in \{-1, 1\}} U(y_i, X_i, \pi_i, \varepsilon_i) = y_i(\beta X_i + \phi_2 \pi) + \varepsilon_i(y_i), \quad (9.22)$$

where π_i captures either average smoking among agent i 's direct social contacts (in the case of local interactions) or average smoking in the school as a whole (if we consider global interactions). The random term ε_i depends on the smoking choice y_i and follows an extreme value distribution:

$$\Pr(\varepsilon_i(-1) - \varepsilon_i(1) \leq z) = \frac{1}{1 + \exp(-z)}. \quad (9.23)$$

From the first-order conditions, the probability that agent i smokes is given by

$$\Pr(y_i = 1) = \frac{1}{1 + \exp(-2(\beta X_i + \phi_2 \pi_i))}. \quad (9.24)$$

Assuming that the number of agents in each school is large enough, then a law of large numbers argument applies and the following characterization of equilibrium is obtained for the case of global interactions:

$$\pi = \sum_{i \in I} \tanh(\beta X_i + \phi_2 \pi). \quad (9.25)$$

It is easy to show that very nonlinear effects may arise. Depending on which equilibrium a given school starts from, an increase in the utility cost of smoking (brought about, for instance, by a tobacco tax) may induce an increase or a decrease in equilibrium average smoking in the school. Similarly, an increase in the strength of social interactions or in the initial number of friends smoking in individual networks may cause—depending on the initial equilibrium—an increase or a decrease in eventual smoking. This is important from a policy perspective, since it emphasizes that a given policy may have counterintuitive effects because of the nonlinear feedbacks induced by network effects.

The model can be estimated using the techniques developed by [Moro \(2003\)](#).³⁰ As discussed in [Manski \(1993\)](#), the reflection problem is mitigated in nonlinear models; further, the possible presence of correlated unobservables can be addressed using a Heckman-style approach to correct for selection into networks. [Moro \(2003\)](#) developed a two-step approach to tackle the issue of estimating equilibrium models with multiple equilibria. In the first stage, summary statistics of the equilibrium for each school are

³⁰ See also [Aguirregabiria and Mira \(2007\)](#).

estimated, using nonparametric methods. In the second stage, the model parameters are estimated via maximum likelihood, conditioning the likelihood of the data on the first-stage estimates of the equilibrium. This allows the likelihood to be a well-behaved function, as opposed to a correspondence—as would be the case given the presence of multiple equilibria. This reduces the computational burden enormously.

Bisin et al. (2011a) find evidence of strong network effects in smoking, both school-wide and at the level of individual friendship networks. The parameter estimates are consistent with the widespread presence of multiple equilibria among the schools considered in the Add Health sample. As mentioned earlier, simulations of the model with the parameter estimates indicate that changes in attributes, the shape of networks, or various policies can have highly nonlinear and sometimes counterintuitive effects, with the possibility of large shifts in smoking prevalence because of the presence of multiple equilibria.

9.3.3 Empirical results

Let us describe the empirical results obtained on the basis of the theoretical models presented in Section 9.3.1 and discuss the policy implications.

9.3.3.1 Local-average model

This is the most tested model in the literature. Researchers have tested Equation (9.9) using the method developed in Section 9.3.2.2. There is usually no theoretical model for the microfoundation of Equation (9.9). Researchers have estimated this equation because it is similar to the one used in spatial econometrics (Anselin, 1988) and it is easier to test. The empirical results indicate that peer effects and network effects are important in education (Calvó-Armengol et al., 2009; De Giorgi et al., 2010; Lin, 2010; Bifulco et al., 2011; Boucher et al., 2014; Patacchini et al., 2014), crime (Patacchini and Zenou, 2012b), labor (Patacchini and Zenou, 2012a), consumption (De Giorgi et al., 2014), smoking (Fletcher, 2010; Bisin et al., 2011a), alcohol consumption (Fletcher, 2012), and risk sharing (Angelucci et al., 2014).³¹

Equation (9.9) has also been tested using another instrumental variable approach. The idea is to treat the composition of students in a given grade within a school as quasi-random and to isolate this quasi-random variation in the friendship network formation process. Using this approach, Fletcher and Ross (2012) find that students who have friends who smoke or drink are more likely to smoke or drink even when comparing observationally similar students who belonged to different cohorts in the same school and made exactly the same friendship choices on key student demographics. Fletcher et al. (2013) find that girls have higher grade point averages (GPAs) than very similar

³¹ There are also some tests of the local-average model (games played on networks) in laboratory settings (see Kosfeld, 2004; Jackson and Yariv, 2011; Charness et al., 2014 for additional background). There are also various field experiments that effectively involve games on networks (see, for example, Centola, 2010).

students in the same school when they belong to a cohort that implies more friends with a higher level of maternal education even after controlling for aggregate peer effects associated with maternal education. Finally, [Patacchini and Zenou \(2014\)](#) find strong peer effects in religion practice. They use the fraction of religious students of the same gender, religious affiliation, and ethnic group in the same grade and school as an instrument for the individual fraction of religious friends.

9.3.3.2 Local-aggregate model

There are very few tests of the local-aggregate model. Two notable exceptions are those provided by [Liu et al. \(2012\)](#) and [Lindquist and Zenou \(2014\)](#), who test peer and network effects in crime. Both estimate Equation (9.3) or its econometric equivalent (9.15) with instrumental variables and network fixed effects ([Section 9.3.2.3](#)). [Liu et al. \(2012\)](#) use Add Health data to estimate these network peer effects.³² The Add Health database has been designed to study the impact of the social environment (i.e., friends, family, neighborhood, and school) on adolescents' behavior in the United States by collecting data on students in grades 7–12 from a nationally representative sample of roughly 130 private and public schools in the 1994–1995 school year (wave I). Every pupil attending the sampled schools on the interview day was asked to compile a questionnaire (in-school data) containing questions on respondents' demographic and behavioral characteristics, education, family background, and friendships. This sample contains information on roughly 90,000 students. A subset of adolescents selected from the rosters of the sampled schools, about 20,000 individuals, were then asked to complete a longer questionnaire containing more sensitive individual and household information (in-home and parental data). Those subjects were interviewed again in 1995–1996 (wave II), in 2001–2002 (wave III), and in 2007–2008 (wave IV).

From a network perspective, the most interesting aspect of the Add Health data is the friendship information, which is based upon actual friend nominations. Indeed, pupils were asked to identify their best friends from a school roster (up to five males and five females). This information was collected in wave I and 1 year after, in wave II. As a result, one can reconstruct the whole geometric structure of the friendship networks. The Add Health dataset also contains information on 15 delinquency items, and the survey asked students how often they participated in each of these delinquent activities during the previous year.

Using the Add Health data, [Liu et al. \(2012\)](#) have estimated ϕ_1 for wave I for 1297 criminals distributed over 150 separate networks, with network size ranging between 4 and 77. They find an estimated value of ϕ_1 of 0.0457, which gives a social multiplier of 1.048 in the case of the dyad (see (9.6)). If we consider an average group of four best friends (linked to each other in a network), a standard deviation increase in the level

³² This dataset has also been used by [Calvó-Armengol et al. \(2009\)](#), [Fletcher \(2010\)](#), [Lin \(2010\)](#), [Bifulco et al. \(2011\)](#), [Fletcher \(2012\)](#), and [Patacchini et al. \(2014\)](#).

of delinquent activity of each of the peers translates into a roughly 17% increase of a standard deviation in the individual level of criminal activity.

Lindquist and Zenou (2014) also estimate ϕ_1 from Equation (9.15) with a very different dataset. They look at individuals in Sweden who are over 16 years old and who have been suspected (and convicted) of committing at least one crime. For that, they have access to the official police register of all people who are suspected of committing a crime in Sweden. In this register, the police keeps records of who is suspected of committing a crime with whom. In this context, a (criminal) link exists between two individuals if they are suspected of committing a crime together. Both the convictions data and the suspects data include crime type, crime date, and sanction received. One advantage of this dataset over the Add Health one is that links are not self-reported and are thus less subject to measurement errors. Another advantage is that information on links is available at each moment of time over a period of 20 years. As a result, Lindquist and Zenou (2014) can add individual lagged crime as one of the individual-level control variables.

They find an estimate of ϕ_1 of 0.167. For one individual, this means that having only one friend increases crime by 20%. If we consider the case of four individuals (their smallest network), then individual crime will increase by 100% compared with the case when the individual is committing crime alone.

9.3.3.3 Local-aggregate versus the local-average model

Instead of testing each model separately, one can test one model against the other using the method developed in Section 9.3.2.4. Using the Add Health data, Liu et al. (2014) find that for “studying effort” (i.e., how hard students study in schools), students tend to *conform* to the social norm of their friends (local-average model), while for sport activities, both the social multiplier (local-aggregate model) and the social norm effect (local-average model) matter. In contrast, for criminal activities, the local-aggregate model seems to be more appropriate (Liu et al., 2013). In terms of policy implications, an effective policy for the *local-average* model would be to change people’s perceptions of “normal” behavior (i.e., their social norm) so that a *group-based policy* should be implemented, while for the *local-aggregate* model, this would not be necessary, and an *individual-based policy* should be implemented instead.

9.3.3.3.1 Individual-based policies: Key players

Consider the case of crime, where we can show that the local-aggregate model is at work, at least for the Add Health data. In that case, a key-player policy (Ballester et al., 2006), whose aim is to remove the criminal whose removal reduces total crime in a network the most, would be the most effective policy since the effort of each criminal and thus the sum of one’s friends’ crime efforts will be reduced. In other words, the removal of the key player can have large effects on crime because of feedback effects or “social multipliers” (see, in particular, Glaeser et al., 1996; Verdier and Zenou, 2004; Kleiman, 2009). That is,

as the fraction of individuals participating in a criminal behavior increases, the impact on others is multiplied through social networks. Thus, criminal behaviors can be magnified, and interventions can become more effective.

Formally, consider the local-aggregate model presented in [Section 9.3.1.1](#) and denote by $Y_r^*(\mathbf{g}_r) = \sum_{i=1}^n \gamma_{i,r}^*$ the total equilibrium level of crime in network $\mathbf{g}_r$, where $\gamma_{i,r}^*$ is the Nash equilibrium effort given by (9.4). Also denote by $\mathbf{g}_r^{[-i]}$ the network $\mathbf{g}_r$ without individual i . Then, in order to determine the key player, the planner will solve the following problem:

$$\max \{ Y^*(\mathbf{g}_r) - Y^*(\mathbf{g}_r^{[-i]}) \mid i = 1, \dots, n \}.$$

When the original delinquency network $\mathbf{g}_r$ is fixed, this is equivalent to

$$\min \{ Y^*(\mathbf{g}_r^{[-i]}) \mid i = 1, \dots, n \}. \quad (9.26)$$

[Ballester et al. \(2006\)](#) and [Ballester and Zenou \(2014\)](#) have shown that if $\phi_1 \mu_1(\mathbf{g}_r) < 1$, then the key player that solves (9.26) is individual i^* if and only if he/she is a delinquent with the highest *intercentrality* in $\mathbf{g}_r$ —that is, $d_{i^*}(\mathbf{g}_r, \phi_1) \geq d_i(\mathbf{g}_r, \phi_1)$, for all $i = 1, \dots, n$, where³³

$$\begin{aligned} d_i(\mathbf{g}_r, \phi_1) &= \mathbf{1}_{n_r}^T \mathbf{M}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r - \mathbf{1}_{n_r}^T \mathbf{M}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r^{[i]} + \mathbf{1}_{n_r}^T \mathbf{M}^{[i]}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r^{[i]} \\ &= B(\mathbf{g}_r, \phi_1) - B(\mathbf{g}_r^{[i]}, \phi_1) + \frac{b_{\boldsymbol{\alpha}_r^{[i]}, i}(\mathbf{g}_r, \phi_1) \sum_{j=1}^n m_{ji}(\mathbf{g}_r, \phi_1)}{m_{ii}(\mathbf{g}_r, \phi_1)}. \end{aligned} \quad (9.27)$$

The intercentrality measure (9.27) highlights the fact that when a delinquent is removed from a network, two effects are at work. The first effect is the *contextual effect*, which indicates the change in the contextual effect $\boldsymbol{\alpha}_r$ (from $\boldsymbol{\alpha}_r$ to $\boldsymbol{\alpha}_r^{[i]}$) after the removal of the key player while the network $\mathbf{g}_r$ remains unchanged. The second effect is the *network*

³³ To understand (9.27), let $\mathbf{M}(\mathbf{g}_r, \phi_1) = (\mathbf{I}_{n_r} - \phi_1 \mathbf{G}_r)^{-1}$ and let its entries be $m_{ij}(\mathbf{g}_r, \phi_1)$, which count the number of walks in $\mathbf{g}_r$ starting from i and ending at j , where walks of length k are weighted by ϕ_1^k . Then, we know from (9.5) that the Katz–Bonacich vector of centralities is simply $\mathbf{b}_{\boldsymbol{\alpha}_r} = \mathbf{M}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r$. Thus, $b_{i,r}(\mathbf{g}_r, \phi_1)$ is the Katz–Bonacich centrality of i in network $\mathbf{g}_r$, $B(\mathbf{g}_r, \phi_1)$ is the *sum* of the Katz–Bonacich centralities in network $\mathbf{g}_r$ —that is, $B(\mathbf{g}_r, \phi_1) = \mathbf{1}_{n_r}^T \mathbf{M}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r$ (where $\mathbf{1}_{n_r}$ is an n -dimensional vector of ones and $\mathbf{1}_{n_r}^T$ is its transpose)—and $B(\mathbf{g}_r^{[-i]}, \phi_1) = \mathbf{1}_{n_r}^T \mathbf{M}^{[-i]}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r^{[-i]}$ is the *sum* of the Katz–Bonacich centralities in network $\mathbf{g}_r^{[-i]}$, where $\boldsymbol{\alpha}_r^{[-i]}$ is an $(n_r - 1) \times 1$ column vector in which $\alpha_{i,r}$ has been removed and $\mathbf{M}^{[-i]}(\mathbf{g}_r, \phi_1) = (\mathbf{I}_{n_r} - \phi_1 \mathbf{G}_r^{[-i]})^{-1}$ is an $(n - 1) \times (n - 1)$ matrix in which the i th row and i th column corresponding to i have been removed from $\mathbf{M}^{[-i]}(\mathbf{g}_r, \phi_1)$. Finally, let $\boldsymbol{\alpha}_r^{[i]}$ be an $(n \times 1)$ column vector where all entries but i are defined as $\alpha_r^{[i]} = m_{ji}^{[i]} / m_{ii}^{[i]}$ so that $B(\mathbf{g}_r^{[i]}, \phi_1) = \mathbf{1}_{n_r}^T \mathbf{M}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r^{[i]}$ and

$$\mathbf{1}_{n_r}^T \mathbf{M}^{[i]}(\mathbf{g}_r, \phi_1) \boldsymbol{\alpha}_r^{[i]} = b_{\boldsymbol{\alpha}_r^{[i]}, i}(\mathbf{g}_r, \phi_1) \sum_{j=1}^n m_{ji}(\mathbf{g}_r, \phi_1) / m_{ii}(\mathbf{g}_r, \phi_1).$$

effect, which captures the change in the network structure when the key player is removed. More generally, the intercentrality measure $d_i(\mathbf{g}, \phi_1)$ of delinquent i accounts both for one's exposure to the rest of the group and for one's contribution to every other exposure.

Liu et al. (2012) were the first to test the key-player policy using the Add Health data. As mentioned above, they find an estimate value of ϕ_1 of 0.0457. They then calculate the key player for each network using the intercentrality measure (9.27). They find that the key player is *not* necessarily the most active criminal in the network. They also find that it is *not* straightforward to determine which delinquent should be removed from a network by observing only his/her criminal activities or position in the network. Compared with other criminals, the key players are less likely to be a female, are less religious, belong to families whose parents are less educated, and have the perception of being more socially excluded. They also feel that their parents care less about them, are more likely to come from single-parent families, and have more trouble getting along with their teachers.

Lindquist and Zenou (2014) also test the key-player policy but with different data (the co-offending networks mentioned above). While Liu et al. (2012) observed the network at only one point in time, Lindquist and Zenou (2014) consider two periods of 3 years each (2000–2002 and 2003–2005). The period 1 dataset includes 15,230 co-offenders who are suspected of committing (on average) 5.91 crimes each and who are distributed over 1192 separate networks. The period 2 dataset includes 15,143 co-offenders who are suspected of committing (on average) 5.92 crimes each and who are distributed over 1185 networks. Their data also include 3881 individuals who are members of a network with four or more people in *both* periods. They show that 23% of all key players are not the most active criminals in their own networks, 23% do not have the highest eigenvector centrality, and 20% do not have the highest betweenness centrality.³⁴

As stated above, their estimate of peer effects ϕ_1 is 0.167. They show that the key-player model predicts that the (average) reduction in crime for the mean network (with size 80) is equal to 30%. Second, this reduction in crime is negatively related to network size. If one looks at a network that is twice as large as the mean network (i.e., with size 160), then the predicted percentage reduction in crime is 26%, while the predicted decrease for the smallest networks (with size 4) is 35%.

Given that the key-player policy can be controversial and can be costly to implement, we want to know by how much the key-player policy outperforms other reasonable policies. Because they have two periods of time (2000–2002 and 2003–2005), Lindquist and Zenou (2014) can test the prediction of crime reduction following the key-player policy against the true outcome observed in period 2 data. For that, they look at the

³⁴ Eigenvector centrality and betweenness centrality are well-known measures of centrality. See Wasserman and Faust (1994) and Jackson (2008) for a complete overview of the different existing centrality measures.

relative effect of removing the key player in those cases in which the key player is no longer a part of the active network. To do this, they create an indicator variable for each person indicating whether or not that person died during the relevant time period and if that person was put in prison. Their results indicate that, in the real world, the key-player policy outperforms the random-player policy by 9.58%. The key-player policy also outperforms the policy of removing the most active player by 3.16% and outperforms the policies of removing the player with the highest eigenvector and the highest betweenness centrality by 8.12% and 2.09%, respectively.³⁵

9.3.3.3.2 Group-based policies

As stated above, if the *local-average model* is at work, then a key-player policy would have a much smaller effect since it will not affect the social norm of each group of friends in the network. To be effective, one would have to change the social norm for each of the criminals, which is clearly a more difficult objective. In that case, one needs to target a group or gang of criminals to drastically reduce crime. It is indeed clearly much more complicated to implement a group policy than an individual policy since it is very difficult to change the social norm of a group. Consider education. Then, since the local-average model seems important (at least in the Add Health data), we should change the social norm in the school or the class and try to implement the idea that it is “cool” to work hard at school.³⁶ An example of a policy that has tried to change the social norm of students in terms of education is the *charter-school* policy. The charter schools are very good at screening teachers and at selecting the best ones. In particular, the “no excuses policy” (Angrist et al., 2010, 2012) is a highly standardized and widely replicated charter model that features a long school day, an extended school year, selective teacher hiring, and strict behavior norms, and emphasizes traditional reading and mathematical skills. The main objective is to change the social norms of disadvantaged children by being very strict on discipline. This is a typical policy that is in accordance with the local-average model

³⁵ Other articles have tested the key-player policies for other activities. For R&D networks, König et al. (2014a) calculate the key firms, which are the firms for which their removal will reduce total welfare the most. Banerjee et al. (2013) study a problem related to the key-player issue. Their data come from a survey on 75 rural villages in Karnataka, India, that they conducted to obtain information on network structure and various demographics. They look at the diffusion of a microfinance program in these villages and show that if the bank in charge of this program had targeted individuals in the village with the highest eigenvector centrality (a measure related to the Katz–Bonacich centrality), the diffusion of the microfinance program (i.e., take-up rates) would have been much higher. For an overview of key-player policies, see Zenou (2015c).

³⁶ This is related to the “acting white” literature where it is argued that African-American students in poor areas may be ambivalent about studying hard in school because this may be regarded as “acting white” and adopting mainstream identities (Fordham and Ogbu, 1986; Delpit, 1995; Ainsworth-Darnell and Downey, 1998; Austen-Smith and Fryer, 2005; Battu et al., 2007; Battu and Zenou, 2010; Fryer and Torelli, 2010; Bisin et al., 2011b; De Martí and Zenou, 2012).

since its aim is to change the social norm of students in terms of education. [Angrist et al. \(2012\)](#) focus on special needs students that may be underserved. Their results show average achievement gains of 0.36 standard deviations in mathematics and 0.12 standard deviations in reading for each year spent at a charter school called Knowledge is Power Program (KIPP) Lynn, with the largest gains coming from the limited English proficiency, special education, and low-achievement groups. They show that the average reading gains were driven almost entirely by special education and limited English proficiency students, whose reading scores rose by roughly 0.35 standard deviations for each year spent at KIPP Lynn.

Boarding schools could also be a way of changing the social norm in terms of education. For example, the SEED schools are boarding schools serving disadvantaged students located in Washington, DC, and Maryland. The SEED schools, which combine a “no excuses” charter model with a 5-day-a-week boarding program, are America’s only urban public boarding schools for the poor. The SEED schools serve students in grades 6–12. Like other “no excuses” charter schools—for example, KIPP or the Harlem Children’s Zone, SEED schools have an extended school day, provide extensive after-school tutoring for students who need support, rely heavily on data to alter the scope, pace, and sequence of instruction, and maintain a paternalistic culture with high expectations. [Curto and Fryer \(2014\)](#) provide the first causal estimate of the impact of attending SEED schools on academic achievement. Using admission lotteries, they show that attending a SEED school increases achievement by 0.211 standard deviations per year in reading and 0.229 standard deviations per year in mathematics.

9.4. NEIGHBORHOOD AND NETWORK EFFECTS

So far, we have described separately the literature on neighborhood and network effects. We have seen that there are some similarities, especially when researchers do not have data on the social space and approximate them by the geographical space (see, in particular, [Bayer et al., 2008](#); [Patacchini and Zenou, 2012a,b](#); [Helmers and Patnam, 2014](#)). However, these two spaces are different, and we need an explicit analysis of both of them in order to better understand their relationships and how they affect outcomes. For example, if we want to understand the adverse labor-market outcomes of ethnic minorities, we need to analyze each space and see how they reinforce each other. Unfortunately, this branch of the literature is still in its infancy and most research has been done from a theoretical perspective, with only a few empirical tests. Let us describe this research.³⁷

³⁷ In economics, [Ioannides \(2012\)](#) is a good starting point even though few analyses incorporate the two spaces. In sociology, there are some discussions of these issues. See, in particular, [Guest and Lee \(1983\)](#), [Wellman \(1996\)](#), [Otani \(1999\)](#), and [Mouw and Entwisle \(2006\)](#).

9.4.1 Theory: Spatial models with social networks

We will describe different models that integrate the urban and social space. We start with models with social interactions, then consider weak and strong ties, and end up with explicit networks using graph theory. As we enrich the social space, we model the urban space in a simpler way from a general urban model to a model with only two locations.

9.4.1.1 *Spatial models with social interactions*

In this section, the social network is not explicitly modeled, but is captured through social interactions. In contrast, the geographical space is explicitly modeled as in the standard urban economics literature (Fujita, 1989; Zenou, 2009; Fujita and Thisse, 2013). There is an early body of literature that deals with the endogenous location of firms and workers and the formation of cities by explaining why cities exist, why cities form where they do, and why economic activities agglomerate in a small number of places (Fujita and Thisse, 2013). The key articles in this literature are those of Ogawa and Fujita (1980) and Fujita and Ogawa (1982), who solve a more general model that includes both firms and households (see also Beckmann, 1976; Borukhov and Hochman, 1977; Papageorgiou and Smith, 1983). Their articles model the emergence of urban centers brought about by household and firm location decisions in the context of spatially differentiated labor and land market interactions. Consider, for example, the model of Fujita and Ogawa (1982). The key aspect of this model is to assume that productivity in a location is a function of the density of economic activity at various locations weighted by a decay function. In other words, the agglomeration force is the existence of informational spillovers among firms. An important characteristic of information is its public good nature: the use of a piece of information by a firm does not reduce its content for other firms. Hence, the diffusion of information within a set of firms generates *externality-like* benefits to each of them. Provided that the information owned by firms is different, *the benefits of communication generally increase as the number of firms involved rises*. Furthermore, since the quality of information involves distance-decay effects, *the benefits are greater if firms locate closer to each other*. Therefore, all other things being equal, each firm has an incentive to be close to others, thus fostering the agglomeration of firms. This is the social interaction aspect of these types of models (Beckmann, 1976 provides a similar model but for individuals rather than firms). Of course, there are also disagglomeration effects because the clustering of many firms in a single area increases the average commuting distance for their workers, which in turn increases the wage rate and land rent in the area surrounding the cluster. Consequently, the equilibrium distributions of firms and households are determined as the balance between these opposite forces.

In Fujita and Ogawa (1982), this type of specification yields a rich set of possible outcomes. Depending on the importance of the spatial decay function relative to commuting costs, many urban configurations are possible, from a purely monocentric city to

complete dispersion.³⁸ None of these articles, however, offer much detail regarding the information externality nor the spatial decay function.³⁹

Helsley and Strange (2007) propose an interesting spatial model of urban interactions where agents choose to visit a particular location to interact with others.⁴⁰ A critical component of the model is the decision taken by a city's firms or households to visit a particular location to interact with others. The greater the aggregate number of visits, the greater is the value derived from any given visit. Visits involve transportation costs, however, and this generates downward-sloping equilibrium housing rent, land rent, and population density functions. In equilibrium, all of these must be consistent with the interactions that take place in the center.

To be more precise, consider the location space as a long, narrow strip of land where there is one unit of land at each location. All interactions occur at a single location, the central business district (CBD). Locations are completely characterized by their distance from this CBD, given by the variable x . Consumers are identical and derive utility from residential (or commercial) space q , other goods z (the numeraire), and interaction according to the additively separable utility function:

$$u(y_i, S) = q_i + z_i + v(y_i, S),$$

where y_i is the number of visits to the center for agent i and S measures the quality of interactions there. Assume that $v(y_i, S)$ is increasing and strictly quasi-concave in both arguments, with $\partial^2 v(y_i, S) / \partial y_i \partial S > 0$. This last assumption means that the marginal value of a visit to the center increases with the quality of the interactions there. There are two costs associated with a visit to the center: a fixed cost T and transportation cost tx , $t > 0$. Since consumers are assumed to be all identical and have the same income w , we can skip the subscript i . The budget constraint for a consumer with income w at location x is

$$z = w - R(x)q - (T + tx)y, \quad (9.28)$$

where $R(x)$ is the rent per unit of space at distance x from the CBD. We assume that each consumer occupies one unit of space—that is, $q = 1$. Combining these two equations, the consumer chooses y that maximizes

$$u(y, S) = 1 + w - R(x) - (T + tx)y + v(y, S).$$

Solving this equation leads to a unique $y^* \equiv y(S, x)$ and it easily verified that the optimal number of visits y^* made to the center increases with the quality of interactions S and

³⁸ This type of model has been extended by Helsley (1990), Ota and Fujita (1993), Lucas (2001), Berliant et al. (2002), and Lucas and Rossi-Hansberg (2002).

³⁹ See Duranton and Puga (2004) for a critical overview of these issues.

⁴⁰ See also Brueckner et al. (2002) and Brueckner and Largey (2008).

decreases with distance x . The key new element here is to specify interaction quality, S . [Helsley and Strange \(2007\)](#) assume that the equilibrium level of interaction quality satisfies

$$S = \int_0^{x_f(S)} F(y(S, x)) n(S, x) dx, \quad (9.29)$$

where $x_f(\cdot)$ is the city fringe and $F(\cdot)$ is increasing and strictly concave, and $F(0) = 0$. Since each consumer occupies one unit of space, and there is one unit of land at each location, $n(\cdot)$ equals the population, population density (persons per unit land), and structural density (units of residential or commercial space per unit land). Here, each agent has the potential to benefit from interacting with any other agent. However, the value of interacting with any particular agent exhibits a diminishing marginal impact, captured by the concavity of $F(\cdot)$. In this model, the interdependence between agents arises from the endogeneity of interactions: agents choose jointly both how much to contribute to a location and how much to make use of that location. It is easily seen that the solution of S is a fixed point. The model is then easily closed by considering an open city with free migration and having a free-entry condition for builders.

[Mossay and Picard \(2011, 2013\)](#) propose a model in the same vein where the utility function is given by

$$u(q, z, S) = z + S(x) - \frac{\beta}{2q},$$

where β is the preference for residential space and where social interactions are given by

$$S(x) = A - \int n(x') T(x - x') dx',$$

where A denotes the total return from interacting with other agents and $\int n(x') T(x - x') dx'$ reflects the cost of reaching other agents from location x , where $n(x)$ is the population density with $\int n(x) dx = 1$. In this formulation of social interactions, the authors consider a linear cost function $T(x - x') = 2\tau|x - x'|$, where τ measures the intensity of traveling costs. In this model, each agent *interacts with all other agents* and A is assumed to be large enough so as to ensure that $S(x) \geq 0$, for any location x . Mossay and Picard have a similar budget constraint as in (9.28)—that is, $z = w - R(x)q$ so that consumers choose q and z that maximize $u(q, z, S)$. They then calculate a spatial equilibrium in a monocentric city so that no agent has an incentive to relocate. They show that there is a unique spatial equilibrium under the assumption of global social interactions where each agent interacts with all other agents residing in the same city.

In all these models, the interactions between the social and geographical spaces are explicitly modeled. However, apart from their residential location, the outcome of workers is *not* taken into account. [Picard and Zenou \(2014\)](#) extend the previous models

to introduce the labor-market outcomes of workers where it is assumed that social interactions are the main channel for finding employment. Indeed, consider two populations and assume that each individual of type i (i.e., belonging to population $i = 1, 2$) located at a distance x from the CBD can only *socially* interact with the members of his/her own population but must decide with how many of them he/she wants to interact, given that each social interaction implies a travel cost τ (per unit of distance) but leads to job information.

In this context, the expected utility of an individual of type i residing at location x is given by

$$u_i(x) = e_i(x)(w - t|x|) - T_i(x) - R(x), \quad (9.30)$$

where $e_i(x)$ is the individual's employment probability, $T_i(x)$ is the total travel cost at a distance x due to social interactions, and $R(x)$ is the land rent at a distance x from the CBD.⁴¹ In this expression, all individuals from the same group, employed and unemployed, socially interact with each other. The steady-state employment rate is

$$e_i(x) = \frac{\pi_i(x)}{\pi_i(x) + \delta}, \quad (9.31)$$

where δ is the exogenous destruction rate and $\pi_i(x)$ is the probability of finding a job at a distance x from the CBD for a worker from population i .

Let us be more precise about the meeting process between agents. Each individual of type i residing at x meets $n_i(x)$ individuals from his/her own population to socially interact with them. This means that each individual meets $n_i(x)$ times all his/her population mates in a deterministic way during the period considered in the model. Since social interactions occurs at the place of residence of the potential information holder, the cost of those social interactions is given by $T_i(x) = n_i(x)c_i(x)$, where

$$c_i(x) = \frac{1}{P_i} \int_{D_i} \tau|x - y| dy \quad (9.32)$$

measures the average cost of a single social interaction and P_i is the total population of individuals of type i . Observe that it is assumed that there is a uniform distribution of workers in the city, and this is why (given that each worker consumes one unit of land) the density of workers at each location is given by $1/P_i$. As a result, each worker i residing at x socially interacts with all members of his/her own population and each of these interactions implies a commuting cost of τ per unit of distance. Observe also that the location x of a worker i is crucial to determine $c_i(x)$. If, for example, a worker i lives close to the CBD, then his/her cost $c_i(x)$ will be relatively low since this worker will be at the same distance from the left and the right of x . But, if this worker is located at one end of the city, then $c_i(x)$ will be very high.

⁴¹ Unemployment benefits are normalized to zero.

Since each social interaction leads to job information, the individual's probability of finding a job for a worker of type i residing at x is given by

$$\pi_i(x) = \alpha n_i(x) \frac{E_i}{P_i}, \quad (9.33)$$

where α is a positive constant and E_i/P_i denotes the employment rate for workers of type i . This equation captures the fact that each individual i located at x meets $n_i(x)$ individuals from his/her own population, but only those who are employed provide some information about jobs. This highlights the random search process since the probability of employment of each person met by worker i is just E_i/P_i and is not specific to the person met. Quite naturally, the individual's probability of finding a job increases with the number of social interactions $n_i(x)$ and with the employment rate in his/her own population.

In this model, each individual chooses $n_i(x)$ that maximizes (9.30), which is obtained by plugging (9.33) into (9.31) and then into (9.30) and plugging (9.32) into $T_i(x) = n_i(x)c_i(x)$ and then into (9.30). When deciding the optimal level of social interactions, an individual i located at x trades off the benefits of an increase in $n_i(x)$, which raises his/her chance of obtaining a job with its costs, since more social interactions imply more traveling and thus higher $c_i(x)$.

Consider first a homogenous population. Then, in a monocentric city, one can easily close the model by solving for the land and labor equilibrium conditions and check that everything is consistent. In that case, it is easy to show that $c(x) = \frac{\tau}{p}(b^2 + x^2)$ on the city support $D = [-b, b]$, where b is the city border and $x = 0$ is the CBD. Picard and Zenou (2014) show that the employment probability $\pi(x)$ and the optimal number of social interactions $n(x)$ decrease with x , the distance from the city center.

If we now consider two populations that do not socially interact with each other, then it can be shown that there exists a spatially segregated equilibrium where population 1 resides around the city center, while population 2 is located at both ends of the city. In this equilibrium, the employment rate of population 1 is always higher than that of population 2 whatever their relative sizes, E_1/P_1 and E_2/P_2 . It can also be shown that each worker's employment probability $e_i(x)$ and the number of social interactions $n_i(x)$ decrease with x . Indeed, a residential location further away from the city center reduces the net gain from employment for both populations as well as each individual's average access to his/her social network. As a result, individuals have fewer incentives to find a job. This result is interesting because it highlights the feedback effect of space and segregation on labor-market outcomes. If we take two populations identical in all possible characteristics, then employment differences between these populations will result from the existence of spatial segregation and the resulting spatial organization of workers' social networks. Workers obtain job information through their social contacts that belong to the same type but organize in a different way through the urban area.

9.4.1.2 Spatial models with weak and strong ties

In the previous section, the modeling of social networks was implicit and was captured through social interactions. For example, in [Picard and Zenou \(2014\)](#), workers were interacting with all other workers of the same type in the city and each social interaction could lead to job information if one met someone who already had a job. We now enrich the social network aspect by differentiating between job information from strong ties (close and regular relationships such as family and friends) and from weak ties (random and irregular relationships). The notion of weak and strong ties was initially developed by [Granovetter \(1973, 1974, 1983\)](#),⁴² who stipulates and shows that weak ties are superior to strong ties for providing support in getting a job. Indeed, in a close network where everyone knows each other, information is shared and so potential sources of information are quickly shaken down, and so the network quickly becomes redundant in terms of access to new information. In contrast, Granovetter stresses the *strength of weak ties* involving a secondary ring of acquaintances who have contacts with networks outside the individual's network and therefore offer new sources of information on job opportunities.

[Montgomery \(1994\)](#), [Calvó-Armengol et al. \(2007\)](#), [Patacchini and Zenou \(2008\)](#), and [Zenou \(2013, 2015b\)](#) propose modeling the impact of weak and strong ties on workers' outcomes using a dyad model so that the social network is very simplified but keeps the interaction between the two types of ties. Formally, consider a population of individuals of size 1 and assume that individuals belong to mutually exclusive two-person groups, referred to as *dyads*. We say that two individuals belonging to the same dyad hold a *strong tie* to each other. We assume that dyad members do not change over time. A strong tie is created once and forever and can never be broken. Individuals can be in either of two different states: employed or unemployed. Dyads, which consist of paired individuals, can thus be in three different states,⁴³ which are the following: both members are employed—we denote the number of such dyads by d_2 ; one member is employed and the other is unemployed (d_1); and both members are unemployed (d_0). By denoting the employment rate and the unemployment rate at time t by $e(t)$ and $u(t)$, where $e(t), u(t) \in [0, 1]$, we have

$$\begin{cases} e(t) = 2d_2(t) + d_1(t), \\ u(t) = 2d_0(t) + d_1(t). \end{cases} \quad (9.34)$$

⁴² In his seminal articles, [Granovetter \(1973, 1974, 1983\)](#) defines *weak ties* in terms of lack of overlap in personal networks between any two agents—that is, weak ties refer to a network of acquaintances who are less likely to be socially involved with one another. Formally, two agents A and B have a weak tie if there is little or no overlap between their respective personal networks. Vice versa, the tie is *strong* if most of agent A's contacts also appear in agent B's network.

⁴³ The inner ordering of dyad members does not matter.

The population normalization condition can then be written as

$$e(t) + u(t) = 1, \quad (9.35)$$

or, alternatively,

$$d_2(t) + d_1(t) + d_0(t) = \frac{1}{2}. \quad (9.36)$$

Let us explain how social interactions are modeled. Time is continuous and individuals live forever. Matching can take place between dyad partners. At time t , each individual can meet a weak tie with probability $\omega(t)$ (thus $1 - \omega(t)$ is the probability of meeting the strong-tie partner at time t).⁴⁴ These probabilities are constant and exogenous, do not vary over time, and thus can be written as ω and $1 - \omega$. We refer to matchings inside the dyad partnership as *strong ties*, and to matchings outside the dyad partnership as *weak ties* or random encounters. Within each matched pair, information is exchanged in the following way. Each job offer is taken to be received only by employed individuals, who can then direct it to one of their contacts (through either strong or weak ties). This is a convenient modeling assumption, which stresses the importance of on-the-job information.⁴⁵ To be more precise, employed individuals hear of job vacancies at the exogenous rate λ , while they lose their job at the exogenous rate δ . All jobs and all workers are identical (*unskilled labor*), so that all employed individuals obtain the same wage. Therefore, employed individuals, who hear about a job, pass this information on to their current matched partner, who can be a strong or a weak tie. It can be readily checked that the net flow of dyads from each state between t and $t + dt$ is given by

$$\begin{cases} \dot{d}_2(t) = [1 - \omega + \omega e(t)]\lambda d_1(t) - 2\delta d_2(t), \\ \dot{d}_1(t) = 2\omega e(t)\lambda d_0(t) - \delta d_1(t) - [1 - \omega + \omega e(t)]\lambda d_1(t) + 2\delta d_2(t), \\ \dot{d}_0(t) = \delta d_1(t) - 2\omega e(t)\lambda d_0(t). \end{cases} \quad (9.37)$$

Take, for example, the first equation. Then, the variation of dyads composed of two employed individuals ($\dot{d}_2(t)$) is equal to the number of d_1 dyads in which the unemployed individual has found a job (through either his/her strong tie with probability $(1 - \omega)\lambda$ or his/her weak tie with probability $\omega e(t)\lambda$) minus the number of d_2 dyads in which one of the two employed individuals has lost his/her job. Observe that the urban spatial structure will be less rich here because, in all the models in [Section 9.4.1.1](#), the social interactions were localized and individuals had to commute to each other person in order to interact with him/her. In equilibrium, the choice of social interactions for each person had to be consistent with the global level of interactions in the city (see, e.g., [Equation 9.29](#)).

⁴⁴ If each individual has one unit of time to spend with his/her friends, then $\omega(t)$ can also be interpreted as the percentage of time spent with weak ties.

⁴⁵ [Zenou \(2015b\)](#) relaxes this assumption by studying a model where jobs can be found through social networks but also directly by unemployed individuals.

In the present model, social interactions or social networks are not localized. Workers meet their strong ties without commuting because either they live with them (e.g., if they are a couple) or they are close relatives or friends who can be reached without commuting (e.g., by telephone). Workers also meet their weak ties without having to pay extra commuting costs because they meet in common places (e.g., in the gym or at the tennis club or in a bar). As a result, if an individual is unemployed in a d_1 dyad, this means that, without commuting, he/she will meet his/her strong tie $1 - \omega$ percent of his/her time, and to obtain a job, it has to be that the strong tie has heard about a job, which occurs at rate λ . He/she will also meet his/her weak tie (without commuting) ω percent of his/her time, and to obtain a job this weak tie has to be employed and have heard about a job, which occurs with probability $e(t)\lambda$.

By solving the system of Equation (9.37) in the steady state, one can show that there exists an interior equilibrium where the employment rate is given by

$$e^* = \frac{\sqrt{\lambda[\lambda + 4\delta(1 - \omega)]} - 2\delta + 2\lambda\omega - \lambda}{2\lambda\omega}. \quad (9.38)$$

Moreover, it is easily verified that increasing ω , the time spent with weak ties, raises the steady-state employment rate e^* , confirming the initial idea of Granovetter that weak ties are superior to strong ties in providing information about jobs. Here, it is because workers stuck in a d_0 dyad can never find a job through their strong tie (who is unemployed) but only via their weak ties, while this is not true in a d_1 dyad.

Following Zenou (2013), we can then close the model by locating all workers in a monocentric city and assuming that they have an expected utility similar to (9.30)—that is,⁴⁶

$$u(x) = e^*(w - tx) - (1 - e^*)sx - T(x) - R(x), \quad (9.39)$$

where it is assumed that the employed individuals commute more to the CBD than the unemployed individuals ($0 < s < 1$ is the fraction of time the unemployed individuals commute to the CBD) and e^* is given by (9.38). The cost of social interaction $T(x)$ is defined as

$$T(x) = \omega(x) \int \tau|x - y|dy.$$

If social interactions ω are endogenized so that workers choose ω that maximizes (9.39) minus the social interaction costs, then workers face a trade-off between higher ω , which increases their chance of finding a job, and lower ω because of higher social interaction costs. It is straightforward to see that the optimal ω decreases with x , the distance to the CBD. This is because it is always more expensive to commute to the CBD when

⁴⁶ This utility function is similar to that of Picard and Zenou (2014). See (9.30).

employed than when unemployed (i.e., $t > st$), so the marginal gain of interacting with weak ties is higher for workers residing closer to jobs than for those located further away from the CBD.

This model can then be extended by introducing two populations, say black and white workers, where *strong ties* are always of the same race (family, best friends) and there are no spatial costs of interacting with them because they tend to live in the same neighborhood. In contrast, *weak ties* can be of either race, and meeting them implies a commute to the center of activities, here the CBD. Black and white workers are totally identical (in terms of characteristics, skills, etc.). If there is discrimination in the housing market (which is well documented; see, e.g., [Yinger, 1986, 1997](#)) against blacks so that they tend to reside further away from jobs than whites, then it can be shown that the former will experience a higher unemployment rate than the latter. Indeed, because black workers reside far away from the CBD, they will tend to interact less with weak ties, especially whites, and more with their strong ties. Weak ties are an important source of job information, and when black individuals do not obtain this information, they end up having a higher unemployment rate than whites. This is a vicious circle since blacks experience a higher unemployment rate and mostly rely on other blacks, who also experience a high unemployment rate, for example. Since jobs are mainly found through social networks via employed friends, black individuals are stuck in their location with no job. In particular, those residing far away from jobs will mainly rely on their strong ties. As a result, when they find themselves in a d_0 dyad, they have nearly no chance of leaving it since the only way out is to meet an employed weak tie. In the model, the lack of social contacts between blacks and whites⁴⁷ thus explains why the social network of black workers is not of good quality and why blacks experience high unemployment rates.⁴⁸

To summarize, in this framework, ethnic minorities experience higher unemployment rates because they are *separated both in the urban space and in the social space*.⁴⁹

9.4.1.3 Spatial models with explicit social networks

In this section, we describe an even richer structure of social networks by modeling them as in [Section 9.3.1](#). The seminal article of [Jackson and Wolinsky \(1996\)](#) was the first

⁴⁷ [Mouw and Entwisle \(2006\)](#) show empirically that about one-third of the level of racial friendship segregation in schools is attributable to residential segregation. Most of this effect is the result of residential segregation across schools rather than within them.

⁴⁸ American metropolitan areas are segregated by race, both by neighborhood and across jurisdiction lines. In 1980, after a century of suburbanization, 72% of metropolitan blacks lived in central cities, compared with 33% of metropolitan whites ([Boustant, 2010](#)).

⁴⁹ [Sato and Zenou \(2015\)](#) investigate the impact of urban structure on the choice of social interactions. They show that in denser areas, individuals choose to interact with more people and meet more weak ties than in sparsely populated areas.

article to model network formation in a game-theoretical framework. In their model, individuals benefit from direct links and also indirect links but with a decay. They pay, however, an exogenous cost for creating a link. [Johnson and Gilles \(2000\)](#) and [Jackson and Rogers \(2005\)](#) extend this model by assuming that the cost of creating a link is proportional to the *geographical distance* between two individuals so that agents living further away are less likely to form links because the costs are higher. These are interesting models that mainly show that geographical distance can hinder relationships and social interactions between agents. However, in these models, equilibrium networks are difficult to characterize and the focus is on network formation and not on individuals' outcomes.

Following [Helsley and Zenou \(2014\)](#), we develop a simple model where the impact of network structure and urban space on workers' outcomes is analyzed. On contrast to the previous models, there are only two locations, the center located at 0, where all interactions occur, and the periphery located at 1 (*geographical space*). Each agent is also located in a social network (*social space*), where, as in [Section 9.3.1](#), a network is captured by the $n \times n$ adjacency matrix $\mathbf{G}$ with entry g_{ij} , which keeps track of all direct connections so that $g_{ij} = 1$ if agent i is connected to agent j , and $g_{ij} = 0$ otherwise.⁵⁰

We study a two-stage game where the n agents first choose their geographical location and then, as in [Helsley and Strange \(2007\)](#),⁵¹ the number of visits to the center. Consider the local-aggregate model described in [Section 9.3.1.1](#) so that individuals in network $\mathbf{g}$ derive utility

$$U_i(y_i, \mathbf{y}_{-i}, \mathbf{g}) = w + \alpha_i y_i - \frac{1}{2} y_i^2 + \phi_1 \sum_{j=1}^n g_{ij} y_i y_j, \quad (9.40)$$

where $\phi_1 > 0$ and where w stands for income, y_i is the number of visits that agent i makes to the center, and $\mathbf{y}_{-i}$ is the corresponding vector of visits for the other $n - 1$ agents. Agents located in the periphery must travel to the center to interact with others. If we let t represent the marginal transport cost, then $\alpha_i = \alpha - t x_i$. Thus, for each agent i residing in the periphery (i.e., $x_i = 1$), $\alpha_i = \alpha - t$, while for agents living in the center (i.e., $x_i = 0$), $\alpha_i = \alpha$. We assume $\alpha > t$, so $\alpha_i > 0$, $\forall x_i \in \{0, 1\}$, and hence $\forall i = 1, 2, \dots, n$. We imagine that each visit results in one interaction, so the aggregate number of visits is a measure of aggregate interactivity. As in (9.1), utility (9.40) imposes additional structure on the interdependence between agents; under (9.40), the utility of agent i depends on his/her own visit choice and on the visit choices of the agents with whom he/she is directly connected in the network—that is, those for whom $g_{ij} = 1$.

⁵⁰ We skip subscript r since we consider only one network.

⁵¹ See [Section 9.4.1.1](#).

Each agent i chooses y_i to maximize (9.40) taking the structure of the network and the visit choices of other agents as given. With use of the results in Section 9.3.1.1, it is straightforward to see that if $\phi_1\mu(\mathbf{G}) < 1$, there is a unique Nash equilibrium in visits to the center given by

$$\mathbf{y}^* = (\mathbf{I}_n - \phi_1\mathbf{G})^{-1}\boldsymbol{\alpha} = \mathbf{M}\boldsymbol{\alpha} = \mathbf{b}_\alpha(\mathbf{g}, \phi_1), \quad (9.41)$$

where $\mathbf{b}_\alpha(\mathbf{g}, \phi_1)$ is the weighted Katz–Bonacich centrality defined in (9.5). The Nash equilibrium visit choice of agent i is thus

$$y_i^*(x_i, \mathbf{x}_{-i}, \mathbf{g}) = \sum_{j=1}^n m_{ij}\alpha_j = \sum_{j=1}^n \sum_{k=0}^{+\infty} \phi_1^k g_{ij}^{[k]} \alpha_j, \quad (9.42)$$

where $\mathbf{x}_{-i}$ is the vector of locations for the other $n - 1$ agents. The Nash equilibrium number of visits $y_i^*(x_i, \mathbf{x}_{-i}, \mathbf{g})$ depends on the position in the social network and the geographical location. An agent who is more central in the social network, as measured by his/her Katz–Bonacich centrality, will make more visits to the interaction center in equilibrium. Intuitively, agents who are better connected have more to gain from interacting with others, and so exert higher interaction effort for any vector of geographical locations.

Using the best-response function (see Section 9.3.1.1), we can write the equilibrium utility level of agent i as

$$U_i(y_i^*, \mathbf{y}_{-i}^*, \mathbf{g}) = w + \frac{1}{2} [y_i^*(x_i, \mathbf{x}_{-i}, \mathbf{g})]^2 = w + \frac{1}{2} [b_{\alpha_i}(\mathbf{g}, \phi_1)]^2, \quad (9.43)$$

where $y_i^*(0, \mathbf{x}_{-i}, \mathbf{g})$ and $y_i^*(1, \mathbf{x}_{-i}, \mathbf{g})$ are the equilibrium effort of individual i if he/she lives in the center and in the periphery, respectively.

This was the second stage. In the first stage, each agent i chose to live either in the center ($x_i = 0$) or in the periphery ($x_i = 1$) anticipating the utility (9.43) that agent will obtain at each location. There is an exogenous cost differential $c > 0$ associated with the central location. Assuming that the center has more economic activity generally, this cost differential might arise from a difference in location land rent from competition among other activities for center locations. Helsley and Zenou (2014) totally characterize the subgame-perfect Nash equilibria and show that this characterization depends on c , t , α , and the centralities of the agents determined by their m_{ii} and m_{ij} (i.e., their Katz–Bonacich centralities). In particular, more central agents always reside closer to the center than less central agents. If we define the type of an agent by his/her position in the network (in terms of Katz–Bonacich centrality), then it can be shown that the number of equilibria is equal to the number of types of agents plus one. For example, in a star network, there are two types of agents (the star and the peripheral agent) and thus, depending of the values of the parameters, there will be three equilibria: a central equilibrium, where

all agents live in the center; a peripheral equilibrium, where all agents live in the periphery; and a core-periphery equilibrium, where the stars live in the center and the peripheral agents reside in the periphery of the city.

An interesting result is that there is much more clustering in the center of the city in denser networks than in sparse networks. This is because there are many more interactions in a denser network and thus it is more beneficial for agents to live in the center and interact with other agents.

9.4.2 Discussion

In this theoretical presentation, we have seen how the urban space and the social space interact with each other and how they affect the labor-market outcomes of workers. We use this framework to explain the adverse labor-market outcomes of ethnic minorities, especially for black workers in the United States.

If we consider only *neighborhood effects* as in [Section 9.2](#), then there is an important literature in urban economics showing that the distance to jobs is harmful to workers, in particular, black workers. This is a particular form of neighborhood effects, in which the physical location of the neighborhood in relation to jobs, rather than the composition of the neighborhood, generates adverse effects. This is known as the “spatial mismatch hypothesis” ([Kain, 1968](#); [Ihlanfeldt and Sjoquist, 1998](#); [Gobillon et al., 2007](#); [Zenou, 2009](#)). In other words, it is because ethnic minorities reside in *neighborhoods* that are disconnected from jobs that they experience high unemployment rates. In the US context, where jobs have been decentralized and blacks have stayed in the central parts of cities, the main conclusion of the spatial mismatch hypothesis is that the distance to jobs is the main cause of their high unemployment rates.

If we consider only *network effects* as in [Section 9.3](#), then it is because ethnic minorities have “low”-quality social networks that they experience adverse labor-market outcomes.⁵² This is clearly shown by [Calvó-Armengol and Jackson \(2004\)](#),⁵³ where jobs can be found both directly and through other workers linked to each other in the social network. They show that a steady-state equilibrium with a clustering of workers with the same status is likely to emerge since, in the long run, employed individuals tend to be mostly friends with other employed individuals, and unemployed individuals tend to be mostly friends with other unemployed individuals. As a result, if because of some

⁵² There is strong evidence that indicates that labor-market networks are partly race based, operating more strongly within than across races ([Ioannides and Datcher-Loury, 2004](#); [Hellerstein et al., 2011](#)) and that the social network of black workers is of lower quality than that of whites ([Frijters et al., 2005](#); [Fernandez and Fernandez-Mateo, 2006](#); [Battu et al., 2011](#)).

⁵³ See also [Calvó-Armengol \(2004\)](#), [Calvó-Armengol and Zenou \(2005\)](#), [Calvó-Armengol and Jackson \(2007\)](#), and [Galenianos \(2014\)](#).

initial conditions, black individuals are unemployed, then in the steady state they will still be unemployed because both their strong and their weak ties will also be unemployed.

Here, we argue that both the neighborhood and the social network are important in explaining the high unemployment rates of blacks. Let us explain why this is so by considering the model of [Helsley and Zenou \(2014\)](#) ([Section 9.4.1.3](#)) and interpreting it in the following way. There are two locations, a center, where all jobs are located and all interactions take place, and a periphery. Here an interaction between two individuals means that they exchange job information with each other and thus each visit to the center implies a job-information exchange with someone else. As above, y_i is the number of visits that individual i makes to the center in order to obtain information about jobs, and each visit results in one interaction. As a result, the higher is the number of interactions, the higher is the quality of job information and the higher is the probability of being employed. There are two types of workers: black and white, and the only difference between them is their position in the network. We assume that whites have a more central position (in terms of Katz–Bonacich centrality) in the network than blacks. This captures the idea of the “old boy network” where whites grew up together, went through school together, socialized together during adolescence and early adulthood, and entered the labor force together ([Wial, 1991](#)).

In this interpretation of the model, it is straightforward to see that black workers will make fewer visits to the center and thus will interact less with other workers in the network, in particular, with very central agents such as white workers. Moreover, the black workers will also choose to locate further away from jobs than white workers because they interact less with central workers. At the extreme, we could have an equilibrium where all white workers live in the center of the city, while all black workers reside in the periphery. This would imply that whites will interact more with whites and less with black workers. Blacks will interact less and mostly with blacks and thus will have much less information about jobs. This will clearly have dramatic consequences in the labor market and will explain why blacks experience a lower employment rate than whites. In other words, *the lack of good job contacts would be here a structural consequence of the social isolation of inner-city neighborhoods*. Importantly, *the causality goes from the social space to the geographical space*, so it is the *social mismatch* (i.e., their “bad” location in the social network) of black workers that leads to their *spatial mismatch* (i.e., their “bad” location in the geographical space).

We saw in [Section 9.4.1.2](#) that the causality can go the other way. Indeed, in [Zenou \(2013\)](#), it is the spatial mismatch of black workers (due to housing discrimination) that leads to their social mismatch (i.e., less interaction with white weak ties) and thus their adverse labor-market outcomes.

For the policy implications of each model, it is crucial to know the sense of causality. If it is the geographical space that causes the social mismatch of black workers, then the policies should focus on workers’ geographical location, as in the spatial mismatch

literature. In that case, *neighborhood regeneration policies* would be the right tool to use. Such policies have been implemented in the United States and in Europe through the enterprise zone programs and the empowerment zone programs (e.g., Papke, 1994; Bondonio and Greenbaum, 2007; Ham et al., 2011; Busso et al., 2013). The enterprise zone policy consists in designating a specific urban (or rural) area, which is depressed, and targeting it for economic development through government-provided subsidies to labor and capital.

The aim of the empowerment zone program is to revitalize distressed urban communities, and it represents a nexus between social welfare policy and economic development efforts. By implementing these types of policies, one brings jobs to people and thus facilitates the flows of job information in depressed neighborhoods. Another way of reducing the spatial mismatch of black workers would be to implement a transportation policy that subsidizes workers' commuting costs (Pugh, 1998). In the United States, a number of states and counties have used welfare block grants and other federal funds to support urban transportation services for welfare recipients. For example, programs helping job takers (especially African-Americans) obtain a used car—a secured loan for purchase, a leasing scheme, a revolving credit arrangement—may offer real promise and help low-skilled workers obtain a job by commuting to the center where jobs are located.

If, in contrast, it is the social space that causes the spatial mismatch of black workers, then the policies should focus on workers' social isolation. Policies that promote social integration and thus increase the interracial interactions between black and white workers would also have positive effects on the labor-market outcomes of minority workers. Such policies, like the MTO program described in Section 9.2.1.1, have been implemented in the United States. Another way of reducing the unemployment rate of minorities in the context of our model is to observe that *institutional connections* can be engineered to create connections between job seekers and employers in ways that parallel social network processes. For example, scholars such as Granovetter (1979) and Wilson (1996) have called for poverty reduction programs to “create connections” between employers and poor and disadvantaged job seekers.⁵⁴

This is ultimately an empirical question of causality—whether people who are central in the network move to the city, or whether people who are less connected move to the city and then become more central. Such an empirical test is crucial, but one would need either a natural experiment with an exogenous shock or convincing instruments to break the sense of causality. In the labor-market interpretation, the key issue is whether black workers first choose to live in geographically isolated neighborhoods (or are forced to live

⁵⁴ This is related to the policy issues highlighted in Section 9.3.3.3, where we advocated a *group-based policy* for individuals who had preferences according to the *local-average* model and an *individual-based policy* for individuals who had preferences according to the *local-aggregate* model. Clearly, the MTO program, which gives vouchers to individual families, is an individual-based policy, while the enterprise zone program is a group-based policy.

there because of housing discrimination) and then become isolated in the social space because of the lack of contacts with white workers, or whether black workers prefer to interact mainly with other black individuals and as a consequence locate in areas where few whites live, which are isolated from jobs. In any case, we believe that the social and the geographical space are intimately related and policies should take into account both of them if they are to be successful.

9.4.3 Empirical results

Unfortunately, there are very few empirical studies that explicitly test the interactions between the urban space and the social space and their impact on the outcomes of individuals. We saw in [Section 9.2.1.2](#) that a significant portion of *social interactions with neighbors* are very local in nature—that is, occur among individuals in the same block.⁵⁵ [Bayer et al. \(2008\)](#) find that residing in the same block raises the probability of sharing the work location by 33%, which is consistent with a social network effect. Similarly, [Hellerstein et al. \(2011\)](#) and [Hellerstein et al. \(2014\)](#) also find that the hiring effect of *residential networks* is significant and is especially strong for Hispanics and less-skilled workers, and for smaller establishments. All this evidence highlights the neighborhood-specific nature of social networks, at least in the context of labor-market networks. [Ananat et al. \(2013\)](#) find that blacks get a higher return in wages from local agglomeration and human capital spillovers when more of the surrounding workers are black, suggesting that information flows occur along racial lines.

[Del Bello et al. \(2014\)](#) propose one of the few tests that aim to explicitly estimate the effect of the social and geographical space on two outcomes: education and crime. They use the Add Health data described above, which provides information on friendship networks for students in grades 7–12. This dataset also allows them to separate students in different census block groups and thus can determine whether two students who are friends (*social space*) also reside in the same neighborhood or not (*geographical space*). They consider two types of peers: *peers at school*, who are peers nominated at school but who do not live in the same neighborhood, and *peers in the neighborhood*, who are peers nominated at school and who also live in the same neighborhood. Using the local-aggregate model presented in [Section 9.3.2.3](#), they estimate Equation (9.15), which we rewrite for the sake of the exposition:

$$\mathbf{Y}_r = \phi_1 \mathbf{G}_r \mathbf{Y}_r + \beta \mathbf{X}_r + \gamma \mathbf{G}_r^* \mathbf{X}_r + \eta_r \mathbf{1}_{n_r} + \varepsilon_r.$$

⁵⁵ See also [Arzaghi and Henderson \(2008\)](#), [Rice et al. \(2006\)](#), and [Rosenthal and Strange \(2003, 2008\)](#), who show that interaction or agglomeration effects decay very quickly.

Del Bello et al. (2014) decompose the $\mathbf{G}_r$ matrix so that $\mathbf{G}_r = \mathbf{G}_{r,S} + \mathbf{G}_{r,N}$, where $\mathbf{G}_{r,S}$ keeps track only of peers at school in network r and $\mathbf{G}_{r,N}$ accounts for peers in the neighborhood in network r . Thus, the model estimated is

$$\mathbf{Y}_r = \phi_{1S} \mathbf{G}_{r,S} \mathbf{Y}_r + \phi_{1N} \mathbf{G}_{r,N} \mathbf{Y}_r + \beta \mathbf{X}_r + \gamma_S \mathbf{G}_{r,S}^* \mathbf{X}_r + \gamma_N \mathbf{G}_{r,N}^* \mathbf{X}_r + \eta_r \mathbf{1}_{n_r} + \varepsilon_r. \quad (9.44)$$

As in Section 9.3.2.3, Del Bello et al. (2014) estimate this equation using the characteristics of friends of friends as instruments for the endogenous peer effects and network fixed effects. However, as stated in Section 9.3.2.5, this empirical strategy works only if $\mathbf{G}_{r,S}$ and $\mathbf{G}_{r,N}$ are conditionally exogenous. If students sort themselves into neighborhoods and then into friendships according to some unobserved characteristics correlated with the error term, peer effects ϕ_{1S} and ϕ_{1N} in (9.44) are not identified. In order to address this issue, following the discussion in Section 9.3.2.5, one can simultaneously estimate Equation (9.44), the outcome equation, and Equation (9.21), the network formation equation.

Del Bello et al. (2014) find that the effect of peers (friends) on own education (measured by the average GPA of the student) are strong for both *peers at school* and *peers in the neighborhood*, although the effect of school friends is more than twice that of neighborhood peers. They obtain the opposite for the crime outcome, where only *peers in the neighborhood* appear to exhibit an endogenous multiplier effect on criminal activity. This suggests that friends at school (social space) are key for educational outcomes, while friends residing in the same neighborhood (social and geographical space) are the most important determinant of own criminal activities.

These results are important in light of our policy discussion in Section 9.4.2. According to these results, it seems that a key-player policy (see Section 9.3.3.3) as well as neighborhood policies (such as the neighborhood regeneration policies mentioned in Section 9.4.2) are crucial in reducing juvenile crime, while group-based policies at the school level such as the *charter-school* or *boarding-school* policies mentioned in Section 9.3.3.3 are the most efficient ones for improving education for young students.

9.5. CONCLUDING REMARKS

In this chapter, we have reviewed the literature on neighborhood effects, network effects, and neighborhood and network effects. We have seen that for the experimental evidence based on relocations or resettlements of individuals, the neighborhood effects are quite limited in the United States and Canada, while they are important in Europe, especially in Scandinavian countries. In the latter, we showed that ethnic enclaves can have positive effects on labor-market outcomes and education of immigrants, both in Sweden and in Denmark, especially for the less-skilled ones. Unfortunately, they seem to have a positive effect also on crime since growing up in a neighborhood with many criminals has a

long-term effect on crime for immigrants. Interestingly, when we look at nonexperimental evidence at the city block level in the United States, then there are strong neighborhood effects since workers who co-reside in the same city block are more likely to work together compared with residents in nearby blocks. In other words, a significant portion of interactions with neighbors are very local in nature—that is, they occur among individuals in the same block. This effect is especially strong for neighbors within the same racial or ethnic group. We also discussed the structural approach to the estimation of neighborhood effects: here the literature finds evidence of important neighborhood effects for crime and in the labor market.

We then turned to network effects and focused only on studies for which the network was explicitly studied and modeled as a graph. We mainly described (quasi) structural approaches where a model was first written and then tested. For that, we first developed a simple model where agents embedded in a network choose efforts in some activity (education, crime, labor, etc.) where the network is given,⁵⁶ the utility is linear-quadratic, and there are strategic complementarities in efforts. In one version of the model, the network effects of each individual i are captured by the sum of efforts of the agents who are directly connected to individual i (local-aggregate model) and, in the other, they are captured by the distance to the social norm from each agent i (local-average model). We calculated the Nash equilibrium of each of these models and showed the importance of the position in the network with regard to the outcomes of the agents. We then discussed the different empirical tests based on these models and their identification strategies. The results indicate that there are very strong network effects in different activities (education, crime, health, etc.) and that policies should take into account which model is more appropriate for the data. One interesting policy is the key-player policy, which aims to target an agent in a network in order to maximize total activity or welfare.

In the last part of this chapter, we studied the interaction between neighborhood and network effects. We first developed some models where the urban and the social space are integrated, and analyzed how the interaction between these two spaces affects the labor-market outcomes of workers, especially ethnic minorities. We then turned to the empirical tests and found that very few studies include both spaces in their analysis. This is clearly what should be done in the future since we are starting to have better data that can encompass both spaces. This will be very important for policies since it will help us understand the relative role of neighborhood versus peer and network effects on outcomes such as crime, education, and labor.

⁵⁶ There is an important literature on network formation that we do not survey here because these models are usually plagued by multiple equilibria, which are clearly difficult to test empirically. See [Jackson \(2008\)](#) for an overview.

ACKNOWLEDGMENTS

This chapter was prepared for the *Handbook of Regional and Urban Economics, Volume 5*, G. Duranton, V. Henderson, and W. Strange (Eds.). We thank Gilles Duranton, Jessie Handbury, Vernon Henderson, Steve Ross, and Will Strange for very helpful comments.

REFERENCES

- Abowd, J.M., Kramarz, F., Margolis, D.N., 1999. High wage workers and high wage firms. *Econometrica* 67, 251–333.
- Aguirregabiria, V., Mira, P., 2007. Sequential estimation of dynamic discrete games. *Econometrica* 75, 1–53.
- Ainsworth-Darnell, J.W., Downey, D.B., 1998. Assessing the oppositional culture explanation for racial/ethnic differences in school performance. *Am. Sociol. Rev.* 63, 536–553.
- Akerlof, G.A., 1997. Social distance and social decisions. *Econometrica* 65, 1005–1027.
- Allouch, N., 2012. On the private provision of public goods on networks. *Nota di Lavoro* 14.2012. Fondazione Eni Enrico Mattei.
- Ananat, E., Fu, S., Ross, S.L., 2013. Race-specific agglomeration economies: social distance and the black-white wage gap. NBER Working papers 18933.
- Angelucci, M., De Giorgi, G., Rasul, I., 2014. Resource pooling within family networks: insurance and investment. Unpublished manuscript, University of College London.
- Angrist, J.D., Dynarski, S.M., Kane, T.J., Pathak, P.A., Walters, C.R., 2010. Inputs and impacts in charter schools: KIPP Lynn. *Am. Econ. Rev. Pap. Proc.* 100, 239–243.
- Angrist, J.D., Dynarski, S.M., Kane, T.J., Pathak, P.A., Walters, C.R., 2012. Who benefits from KIPP? *J. Policy Anal. Manage.* 31, 837–860.
- Anselin, L., 1988. *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht.
- Arzaghi, M., Henderson, J.V., 2008. Networking off Madison Avenue. *Rev. Econ. Stud.* 75, 1011–1038.
- Åslund, O., Edin, P.-A., Fredriksson, P., Grönqvist, H., 2011. Peers, neighborhoods, and immigrant student achievement: evidence from a placement policy. *Am. Econ. J. Appl. Econ.* 3, 67–95.
- Åslund, O., Östh, J., Zenou, Y., 2010. How crucial is distance to jobs for ethnic minorities? Old question—improved answer. *J. Econ. Geogr.* 10, 389–422.
- Austen-Smith, D., Fryer Jr., R.D., 2005. An economic analysis of ‘acting white’. *Q. J. Econ.* 120, 551–583.
- Ballester, C., Zenou, Y., 2014. Key player policies when contextual effects matter. *J. Math. Sociol.* 38, 233–248.
- Ballester, C., Calvó-Armengol, A., Zenou, Y., 2006. Who’s who in networks. Wanted: the key player. *Econometrica* 74, 1403–1417.
- Ballester, C., Calvó-Armengol, A., Zenou, Y., 2010. Delinquent networks. *J. Eur. Econ. Assoc.* 8, 34–61.
- Banerjee, A., Chandrasekhar, A.G., Duflo, E., Jackson, M.O., 2013. The diffusion of microfinance. *Science* 341, 6144.
- Battu, H., Zenou, Y., 2010. Oppositional identities and employment for ethnic minorities. Evidence for England. *Econ. J.* 120, F52–F71.
- Battu, H., McDonald, M., Zenou, Y., 2007. Oppositional identities and the labor market. *J. Popul. Econ.* 20, 643–667.
- Battu, H., Seaman, P., Zenou, Y., 2011. Job contact networks and the ethnic minorities. *Labour Econ.* 18, 48–56.
- Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. *J. Polit. Econ.* 115, 588–638.
- Bayer, P., Ross, S.L., Topa, G., 2008. Place of work and place of residence: informal hiring networks and labor market outcomes. *J. Polit. Econ.* 116, 1150–1196.
- Beaman, L., 2012. Social networks and the dynamics of labor market outcomes: evidence from refugees resettled in the U.S. *Rev. Econ. Stud.* 79, 128–161.
- Beckmann, M.J., 1976. Spatial equilibrium and the dispersed city. In: Papageorgiou, Y.Y. (Ed.), *Mathematical Land Use Theory*. Lexington Books, Lexington, MA, pp. 117–125.

- Benhabib, J., Bisin, A., Jackson, M.O., 2011. *Handbook of Social Economics*, vols. 1 and 2. Elsevier Science, Amsterdam.
- Berliant, M., Peng, S.-K., Wang, P., 2002. Production externalities and urban configuration. *J. Econ. Theory* 104, 275–303.
- Bernheim, B.D., 1994. A theory of conformity. *J. Polit. Econ.* 102, 841–877.
- Bifulco, R., Fletcher, J.M., Ross, S.L., 2011. The effect of classmate characteristics on post-secondary outcomes: evidence from the Add Health. *Am. Econ. J. Econ. Policy* 3, 25–53.
- Bisin, A., Moro, A., Topa, G., 2011a. The empirical content of models with multiple equilibria in economies with social interactions. NBER Working paper No. 17196.
- Bisin, A., Patacchini, E., Verdier, T., Zenou, Y., 2011b. Formation and persistence of oppositional identities. *Eur. Econ. Rev.* 55, 1046–1071.
- Black, S.E., 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 114, 57–99.
- Blume, L.E., Brock, W.A., Durlauf, S.N., Ioannides, Y.M., 2011. Identification of social interactions. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), *Handbook of Social Economics*. Elsevier Science, Amsterdam.
- Bonacich, P., 1987. Power and centrality: a family of measures. *Am. J. Sociol.* 92, 1170–1182.
- Bond, T.N., Lang, K., 2014. The sad truth about happiness scales. NBER Working paper No. 19950.
- Bondonio, D., Greenbaum, R.T., 2007. Do local tax incentives affect economic growth? What mean impact miss in the analysis of enterprise zone policies. *Reg. Sci. Urban Econ.* 37, 121–136.
- Borukhov, E., Hochman, O., 1977. Optimum and market equilibrium in a model of a city without a pre-determined center. *Environ. Plan. A* 9, 849–856.
- Boucher, V., Bramoullé, Y., Djebbari, H., Fortin, B., 2014. Do peers affect student achievement? Evidence from Canada using group size variation. *J. Appl. Econom.* 29, 91–109.
- Boustant, L.P., 2010. Was postwar suburbanization white flight? Evidence from the black migration. *Q. J. Econ.* 125, 417–443.
- Bramoullé, Y., Kranton, R., 2007. Public goods in networks. *J. Econ. Theory* 135, 478–494.
- Bramoullé, Y., Djebbari, H., Fortin, B., 2009. Identification of peer effects through social networks. *J. Econom.* 150, 41–55.
- Bramoullé, Y., Kranton, R., D'Amours, M., 2014. Strategic interaction and networks. *Am. Econ. Rev.* 104, 898–930.
- Brock, W., Durlauf, S.E., 2001. Discrete choice models with social interactions. *Rev. Econ. Stud.* 68, 235–260.
- Brooks-Gunn, J., Duncan, G., Klebanov, P., Sealander, N., 1992. Do neighborhoods influence child and adolescent development? *Am. J. Sociol.* 99, 353–395.
- Brooks-Gunn, J., Duncan, G.J., Aber, J.L. (Eds.), 1997. *Neighborhood Poverty: Context and Consequences for Children*, vol. 1. Policy Implications in Studying Neighborhoods, vol. 2. Russell Sage Foundation, New York.
- Brown, M., Setren, E., Topa, G., 2014. Do informal referrals lead to better matches? Evidence from a firm's employee referral system. Unpublished manuscript, Federal Reserve Bank of New York.
- Brueckner, J.K., Largey, A.G., 2008. Social interactions and urban sprawl. *J. Urban Econ.* 64, 18–34.
- Brueckner, J.K., Thisse, J.-F., Zenou, Y., 2002. Local labor markets, job matching and urban location. *Int. Econ. Rev.* 43, 155–171.
- Busso, M., Gregory, J., Kline, P., 2013. Assessing the incidence and efficiency of a prominent place based policy. *Am. Econ. Rev.* 103, 897–947.
- Calvo-Armengol, A., 2004. Job contact networks. *J. Econ. Theory* 115, 191–206.
- Calvo-Armengol, A., Jackson, M.O., 2004. The effects of social networks on employment and inequality. *Am. Econ. Rev.* 94, 426–454.
- Calvo-Armengol, A., Jackson, M.O., 2007. Networks in labor markets: wage and employment dynamics and inequality. *J. Econ. Theory* 132, 27–46.
- Calvo-Armengol, A., Zenou, Y., 2004. Social networks and crime decisions. The role of social structure in facilitating delinquent behavior. *Int. Econ. Rev.* 45, 939–958.
- Calvo-Armengol, A., Zenou, Y., 2005. Job matching, social network and word-of-mouth communication. *J. Urban Econ.* 57, 500–522.

- Calvó-Armengol, A., Verdier, T., Zenou, Y., 2007. Strong and weak ties in employment and crime. *J. Public Econ.* 91, 203–233.
- Calvó-Armengol, A., Patacchini, E., Zenou, Y., 2009. Peer effects and social networks in education. *Rev. Econ. Stud.* 76, 1239–1267.
- Campbell, J.Y., Giglio, S., Pathak, P., 2011. Forced sales and house prices. *Am. Econ. Rev.* 101, 2108–2131.
- Carrell, S.E., Fullerton, R.L., West, J.E., 2009. Does your cohort matter? Estimating peer effects in college achievement. *J. Labor Econ.* 27, 439–464.
- Carrell, S.E., Sacerdote, B.I., West, J.E., 2013. From natural variation to optimal policy? The importance of endogenous peer group formation. *Econometrica* 81, 855–882.
- Centola, D., 2010. The spread of behavior in an online social network experiment. *Science* 329, 1194–1197.
- Chandrasekhar, A.G., Jackson, M.O., 2013. Tractable and consistent random graph models. Unpublished manuscript, Stanford University.
- Charness, G., Feri, F., Meléndez-Jiménez, M.A., Sutter, M., 2014. Experimental games on networks: underpinnings of behavior and equilibrium selection. *Econometrica* 82, 1615–1670.
- Christakis, N., Fowler, J., Imbens, G., Kalyanaraman, K., 2010. An empirical model for strategic network formation. NBER Working paper No. 16039.
- Cohen-Cole, E., 2006. Multiple groups identification in the linear-in-means model. *Econ. Lett.* 92, 157–162.
- Comola, M., Prina, S., 2014. Do interventions change the network? A dynamic peer effect model accounting for network changes. Unpublished manuscript, Paris School of Economics.
- Conley, T.G., Topa, G., 2002. Socio-economic distance and spatial patterns in unemployment. *J. Appl. Econom.* 17 (4), 303–327.
- Conley, T.G., Topa, G., 2007. Estimating dynamic local interactions models. *J. Econom.* 140, 282–303.
- Conley, T.G., Udry, C.R., 2010. Learning about a new technology: pineapple in Ghana. *Am. Econ. Rev.* 100, 35–69.
- Corcoran, M., Gordon, R., Laren, D., Solon, G., 1989. Effects of family and community background on men's economic status. NBER Working paper No. 2896.
- Curto, V.E., Fryer Jr., R.G., 2014. The potential of urban boarding schools for the poor: evidence from SEED. *J. Labor Econ.* 32, 65–93.
- Damm, A.P., 2009. Ethnic enclaves and immigrant labor market outcomes: quasi-experimental evidence. *J. Labor Econ.* 27, 281–314.
- Damm, A.P., 2014. Neighborhood quality and labor market outcomes : evidence from quasi-random neighborhood assignment of immigrants. *J. Urban Econ.* 79, 139–166.
- Damm, A.P., Dustmann, C., 2014. Does growing up in a high crime neighborhood affect youth criminal behavior? *Am. Econ. Rev.* 104, 1806–1832.
- Datcher, L., 1983. The impact of informal networks on quit behavior. *Rev. Econ. Stat.* 65, 491–495.
- De Giorgi, G., Pellizzari, M., Redaelli, S., 2010. Identification of social interactions through partially overlapping peer groups. *Am. Econ. J. Appl. Econ.* 2, 241–275.
- De Giorgi, G., Frederiksen, A., Pistaferri, L., 2014. Consumption network effects. Unpublished manuscript, Stanford University.
- Del Bello, C.L., Patacchini, E., Zenou, Y., 2014. Peer effects: social distance or geographical distance? Unpublished manuscript, Stockholm University.
- Delpit, L., 1995. *Other People's Children: Cultural Conflict in the Classroom*. The Free Press, New York.
- De Martí, J., Zenou, Y., 2011. Social networks. In: Jarvie, I., Zamora-Bonilla, J. (Eds.), *Handbook of Philosophy of Social Science*. SAGE Publications, London, pp. 339–361.
- De Martí, J., Zenou, Y., 2012. Friendship formation, oppositional identity, and segregation. CEPR Discussion Paper No. 7566.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Thisse, J.-F., Henderson, J.V. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier Science, Amsterdam, pp. 2063–2117.
- Durlauf, S., 2004. Neighborhood effects. In: Thisse, J.-F., Henderson, J.V. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier Science, Amsterdam, pp. 2173–2242.
- Dustmann, C., Glitz, A., Schoenberg, U., 2011. Referral-based job search networks. Unpublished manuscript, University College London.

- Edin, P.-A., Fredriksson, P., Aslund, O., 2003. Ethnic enclaves and the economic success of immigrants: evidence from a natural experiment. *Q. J. Econ.* 118, 329–357.
- Fernandez, R.M., Fernandez-Mateo, I., 2006. Networks, race, and hiring. *Am. Sociol. Rev.* 71, 42–71.
- Fershtman, C., Weiss, Y., 1998. Social rewards, externalities and stable preferences. *J. Public Econ.* 70, 53–73.
- Fletcher, J.M., 2010. Social interactions and smoking: evidence using multiple student cohorts, instrumental variables, and school fixed effects. *Health Econ.* 19, 466–484.
- Fletcher, J.M., 2012. Peer influences on adolescent alcohol consumption: evidence using an instrumental variables/fixed effect approach. *J. Popul. Econ.* 25, 1265–1286.
- Fletcher, J.M., Ross, S.L., 2012. Estimating the effects of friendship networks on health behaviors of adolescents. NBER Working paper No. 18253.
- Fletcher, J.M., Ross, S.L., Zhang, Y., 2013. The determinants and consequences of friendship formation. NBER Working paper No. 19215.
- Fordham, S., Ogbu, J.U., 1986. Black student' school success: coping with the burden of 'acting white'. *Urban Rev.* 18, 176–206.
- Frijters, P., Shields, M.A., Wheatley Price, S., 2005. Job search methods and their success: a comparison of immigrants and natives in the UK. *Econ. J.* 115, F359–F376.
- Fryer Jr., R.G., Torelli, P., 2010. An empirical analysis of 'acting white'. *J. Public Econ.* 94, 380–396.
- Fujita, M., 1989. *Urban Economic Theory: Land Use and City Size*. Cambridge University Press, Cambridge.
- Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Reg. Sci. Urban Econ.* 12, 161–196.
- Fujita, M., Thisse, J.-F., 2013. *Economics of Agglomeration: Cities, Industrial Location, and Regional Growth*, second ed. Cambridge University Press, Cambridge.
- Galenianos, M., 2013. Learning about match quality and the use of referrals. *Rev. Econ. Dyn.* 16, 668–690.
- Galenianos, M., 2014. Hiring through referrals. *J. Econ. Theory* 152, 304–323.
- Glaeser, E., Scheinkman, J.A., 2001. Measuring social interaction. In: Durlauf, S., Young, P. (Eds.), *Social Dynamics*. Brookings Institution Press and MIT Press, Cambridge, MA, pp. 83–102.
- Glaeser, E.L., Sacerdote, B., Scheinkman, J., 1996. Crime and social interactions. *Q. J. Econ.* 111, 508–548.
- Gobillon, L., Selod, H., Zenou, Y., 2007. The mechanisms of spatial mismatch. *Urban Stud.* 44, 2401–2427.
- Goldsmith-Pinkham, P., Imbens, G.W., 2013. Social networks and the identification of peer effects. *J. Bus. Econ. Stat.* 31, 253–264.
- Gould, E.D., Lavy, V., Paserman, D., 2011. Sixty years after the magic carpet ride: the long-run effect of the early childhood environment on social and economic outcomes. *Rev. Econ. Stud.* 78, 938–973.
- Goyal, S., 2007. *Connections: An Introduction to the Economics of Networks*. Princeton University Press, Princeton.
- Granovetter, M.S., 1973. The strength of short-lived ties. *Am. J. Sociol.* 78, 1360–1380.
- Granovetter, M.S., 1974. *Getting a Job: A Study of Contacts and Careers*. Harvard University Press, Cambridge, MA.
- Granovetter, M.S., 1979. Placement as brokerage: information problems in the labor market for rehabilitated workers. In: Vandergoot, D., Worrall, J.D. (Eds.), *Placement in Rehabilitation: A Career Development Perspective*. University Park Press, New York, pp. 83–101.
- Granovetter, M.S., 1983. The strength of short-lived ties: a network theory revisited. *Sociol. Theory* 1, 201–233.
- Guest, A.M., Lee, B.A., 1983. The social organization of local areas. *Urban Aff. Q.* 19, 217–240.
- Ham, J.C., Swenson, C., Imrohoroglu, A., Song, H., 2011. Government programs can improve local labor markets: evidence from State Enterprise Zones, Federal Empowerment Zones and Federal Enterprise Community. *J. Public Econ.* 95, 779–797.
- Hahn, J., Todd, P., van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69, 201–209.
- Harding, J.P., Rosenblatt, E., Yao, V.W., 2009. The contagion effect of foreclosed properties. *J. Urban Econ.* 66, 164–178.
- Hawranek, F., Schanne, N., 2014. Your very private job agency: job referrals based on residential location networks. IAB Discussion Paper No. 1/2014. The Research Institute of the German Federal Employment Agency.

- Hellerstein, J.K., McInerney, M.P., Neumark, D., 2011. Neighbors and co-workers: the importance of residential labor market networks. *J. Labor Econ.* 29, 659–695.
- Hellerstein, J.K., Kutzbach, M.J., Neumark, D., 2014. Do labor market networks have an important spatial dimension? *J. Urban Econ.* 79, 39–58.
- Helmers, C., Patnam, M., 2014. Does the rotten child spoil his companion? Spatial peer effects among children in rural India. *Quant. Econ.* 5, 67–121.
- Helsley, R.W., 1990. Knowledge production in the CBD. *J. Urban Econ.* 28, 391–403.
- Helsley, R.W., Strange, W.C., 2007. Urban interactions and spatial structure. *J. Econ. Geogr.* 7, 119–138.
- Helsley, R.W., Zenou, Y., 2014. Social networks and interactions in cities. *J. Econ. Theory* 150, 426–466.
- Ihlanfeldt, K.R., Sjoquist, D., 1998. The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. *Hous. Policy Debate* 9, 849–892.
- Ioannides, Y.M., 2003. Interactive property valuations. *J. Urban Econ.* 53, 145–170.
- Ioannides, Y.M., 2012. *From Neighborhoods to Nations: The Economics of Social Interactions*. Princeton University Press, Princeton.
- Ioannides, Y.M., Datcher-Loury, L., 2004. Job information networks, neighborhood effects and inequality. *J. Econ. Lit.* 42, 1056–1093.
- Ioannides, Y.M., Topa, G., 2010. Neighborhood effects: accomplishments and looking beyond them. *J. Reg. Sci.* 50, 343–362.
- Jackson, M.O., 2003. The stability and efficiency of economic and social networks. In: Koray, S., Sertel, M. (Eds.), *Advances in Economic Design*. Springer-Verlag, Heidelberg, pp. 319–362.
- Jackson, M.O., 2004. A survey of models of network formation: stability and efficiency. In: Demange, G., Wooders, M. (Eds.), *Group Formation in Economics. Networks, Clubs and Coalitions*. Cambridge University Press, Cambridge, UK, pp. 11–57.
- Jackson, M.O., 2005. The economics of social networks. In: Blundell, R., Newey, W., Persson, T. (Eds.), *Proceedings of the 9th World Congress of the Econometric Society*. Cambridge University Press, Cambridge, UK, pp. 1–56.
- Jackson, M.O., 2008. *Social and Economic Networks*. Princeton University Press, Princeton, NJ.
- Jackson, M.O., 2011. An overview of social networks and economic applications. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), *Handbook of Social Economics*, vol. 1A. Elsevier Science, Amsterdam, pp. 511–579.
- Jackson, M.O., 2013. Unraveling peers and peer effects: comments on Goldsmith-Pinkham and Imbens' social networks and the identification of peer effects. *J. Bus. Econ. Stat.* 31, 270–273.
- Jackson, M.O., Rogers, B.W., 2005. The economics of small worlds. *J. Eur. Econ. Assoc.* 3, 617–627.
- Jackson, M.O., Wolinsky, A., 1996. A strategic model of social and economic networks. *J. Econ. Theory* 71, 44–74.
- Jackson, M.O., Yariv, L., 2011. Diffusion, strategic interaction, and social structure. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), *Handbook of Social Economics*, vol. 1A. Elsevier Science, Amsterdam, pp. 645–678.
- Jackson, M.O., Zenou, Y., 2013. *Economic Analyses of Social Networks*. The International Library of Critical Writings in Economics. Edward Elgar Publishing, London.
- Jackson, M.O., Zenou, Y., 2015. Games on networks. In: Young, P., Zamir, S. (Eds.), *Handbook of Game Theory*, vol. 4. Elsevier Publisher, Amsterdam, pp. 91–157.
- Jackson, M.O., Rogers, B.W., Zenou, Y., 2015. The impact of social networks on economic behavior. SSRN Working paper 2467812.
- Jacob, B.A., 2004. Public housing, housing vouchers, and student achievement: evidence from public housing demolitions in Chicago. *Am. Econ. Rev.* 94, 233–258.
- Jencks, C., Mayer, S.E., 1990. The social consequences of growing up in a poor neighborhood. In: Lynn, L., McGeary, M. (Eds.), *Inner-City Poverty in the United States*. National Academy Press, Washington, DC, pp. 111–186.
- Johnson, C., Gilles, R.P., 2000. Spatial social networks. *Rev. Econ. Des.* 5, 273–299.
- Kain, J., 1968. Housing segregation, Negro employment, and metropolitan decentralization. *Q. J. Econ.* 82, 175–197.

- Kandel, E., Lazear, E.P., 1992. Peer pressure and partnerships. *J. Polit. Econ.* 100, 801–817.
- Katz, L., 1953. A new status index derived from sociometric analysis. *Psychometrika* 18, 39–43.
- Kelejian, H.H., 2008. A spatial J-test for model specification against a single or a set of nonnested alternatives. *Lett. Spat. Resour. Sci.* 1, 3–11.
- Kelejian, H.H., Piras, G., 2014. Estimation of spatial models with endogenous weighting matrices, and an application to a demand model for cigarettes. *Reg. Sci. Urban Econ.* 46, 140–149.
- Kleiman, M.A., 2009. *When Brute Force Fails. How to Have Less Crime and Less Punishment*. Princeton University Press, Princeton.
- Kling, J.R., Ludwig, J., Katz, L.F., 2005. Neighborhood effects on crime for female and male youth: evidence from a randomized housing voucher experience. *Q. J. Econ.* 120, 87–130.
- Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. *Econometrica* 75, 83–119.
- Kosfeld, M., 2004. Economic networks in the laboratory: a survey. *Rev. Netw. Econ.* 30, 20–42.
- König, M.D., Liu, X., Zenou, Y., 2014a. R&D networks: theory, empirics and policy implications. CEPR Discussion Paper No. 9872.
- König, M., Tessone, C., Zenou, Y., 2014b. Nestedness in networks: a theoretical model and some applications. *Theor. Econ.* 9, 695–752.
- Lee, L.F., 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *J. Econom.* 140, 333–374.
- Lee, B.A., Campbell, K.E., 1999. Neighbor networks of black and white Americans. In: Wellman, B. (Ed.), *Networks in the Global Village: Life in Contemporary Communities*. Westview Press, Boulder, CO, pp. 119–146.
- Lee, L.F., Liu, X., Lin, X., 2010. Specification and estimation of social interaction models with network structures. *Econom. J.* 13, 145–176.
- Lin, X., 2010. Identifying peer effects in student academic achievement by a spatial autoregressive model with group unobservables. *J. Labor Econ.* 28, 825–860.
- Lindquist, M.J., Zenou, Y., 2014. Key players in co-offending networks. CEPR Discussion Paper No. 9889.
- Liu, X., Lee, L.F., 2010. GMM estimation of social interaction models with centrality. *J. Econom.* 159, 99–115.
- Liu, X., Patacchini, E., Zenou, Y., Lee, L.-F., 2012. Criminal networks: who is the key player? CEPR Discussion Paper No. 8772.
- Liu, X., Patacchini, E., Zenou, Y., 2013. Peer effects: social multiplier or social norms? CEPR Discussion Paper No. 9366.
- Liu, X., Patacchini, E., Zenou, Y., 2014. Endogenous peer effects: local aggregate or local average? *J. Econ. Behav. Organ.* 103, 39–59.
- Lucas, R., 2001. Externalities and cities. *Rev. Econ. Dyn.* 4, 245–274.
- Lucas, R., Rossi-Hansberg, E., 2002. On the internal structure of cities. *Econometrica* 70, 1445–1476.
- Ludwig, J., Duncan, G.J., Hirschfield, P., 2001. Urban poverty and juvenile crime: evidence from a randomized housing-mobility experiment. *Q. J. Econ.* 116, 655–679.
- Ludwig, J., Duncan, G.J., Gennetian, L.A., Katz, L.F., Kessler, R.C., Kling, J.R., Sanbonmatsu, L., 2012. Neighborhood effects on the long-term well-being of low-income adults. *Science* 337, 1505.
- Manski, C.F., 1993. Identification of endogenous effects: the reflection problem. *Rev. Econ. Stud.* 60, 531–542.
- Manski, C.F., 2000. Economic analysis of social interactions. *J. Econ. Perspect.* 14, 115–136.
- Mele, A., 2013. A structural model of segregation in social networks. Unpublished manuscript, Johns Hopkins Carey Business School.
- Moffitt, R., 2001. Policy interventions low-level equilibria, and social interactions. In: Durlauf, S., Young, P. (Eds.), *Social Dynamics*. MIT Press, Cambridge, MA, pp. 45–82.
- Montgomery, J.D., 1994. Weak ties, employment, and inequality: an equilibrium analysis. *Am. J. Sociol.* 99, 1212–1236.
- Moro, A., 2003. The effect of statistical discrimination on black-white wage inequality: estimating a model with multiple equilibria. *Int. Econ. Rev.* 44, 457–500.
- Mossay, P., Picard, P.M., 2011. On spatial equilibria in a social interaction model. *J. Econ. Theory* 146, 2455–2477.

- Mossay, P., Picard, P.M., 2013. Spatial segregation and urban structure. CREA Discussion Paper Series 13-03. Center for Research in Economic Analysis, University of Luxembourg.
- Mouw, T., Entwisle, B., 2006. Residential segregation and interracial friendship in schools. *Am. J. Sociol.* 112, 394–441.
- Ogawa, H., Fujita, M., 1980. Equilibrium land use patterns in a non-monocentric city. *J. Reg. Sci.* 20, 455–475.
- Oreopoulous, P., 2003. The long-run consequences of living in a poor neighborhood. *Q. J. Econ.* 118, 1533–1575.
- Ota, M., Fujita, M., 1993. Communication technologies and spatial organization of multi-unit firms in metropolitan areas. *Reg. Sci. Urban Econ.* 23, 695–729.
- Otani, S., 1999. Personal community networks in contemporary Japan. In: Wellman, B. (Ed.), *Networks in the Global Village: Life in Contemporary Communities*. Westview Press, Boulder, CO, pp. 279–297.
- Papageorgiou, Y.Y., Smith, T.R., 1983. Agglomeration as local instability of spatially uniform steady states. *Econometrica* 51, 1109–1119.
- Papke, L., 1994. Tax policy and urban development: evidence from the Indiana enterprise zone program. *J. Public Econ.* 54, 37–49.
- Patacchini, E., Zenou, Y., 2008. The strength of weak ties in crime. *Eur. Econ. Rev.* 52, 209–236.
- Patacchini, E., Zenou, Y., 2012a. Ethnic networks and employment outcomes. *Reg. Sci. Urban Econ.* 42, 938–949.
- Patacchini, E., Zenou, Y., 2012b. Juvenile delinquency and conformism. *J. Law Econ. Org.* 28, 1–31.
- Patacchini, E., Zenou, Y., 2014. Social networks and parental behavior in the intergenerational transmission of religion. Unpublished manuscript, Stockholm University.
- Patacchini, E., Rainone, E., Zenou, Y., 2014. Heterogeneous peer effects in education. CEPR Discussion Paper No. 9804.
- Picard, P.M., Zenou, Y., 2014. Urban spatial structure, employment and social ties. CEPR Discussion Paper No. 10030.
- Popkin, S.J., Rosenbaum, J.E., Meaden, P.M., 1993. Labor market experiences of low-income black women in middle-class suburbs: evidence from a survey of Gautreaux program participants. *J. Policy Anal. Manage.* 12, 556–573.
- Pugh, M., 1998. Barriers to work: the spatial divide between jobs and welfare recipients in metropolitan areas. The Brookings Institution.
- Rice, P., Venables, A.J., Patacchini, E., 2006. Spatial determinants of productivity: analysis for the regions of Great Britain. *Reg. Sci. Urban Econ.* 36, 727–752.
- Rosenthal, S.S., Strange, W., 2003. Geography, industrial organization and agglomeration. *Rev. Econ. Stat.* 85, 178–188.
- Rosenthal, S.S., Strange, W., 2008. The attenuation of human capital spillovers. *J. Urban Econ.* 64, 373–389.
- Rossi-Hansberg, E., Sarte, P.-D., Owens III, R., 2010. Housing externalities. *J. Polit. Econ.* 118, 485–535.
- Sato, Y., Zenou, Y., 2015. How urbanization affect employment and social interactions. *European Economic Review*, forthcoming.
- Schmutte, I.M., 2015. Job referral networks and the determination of earnings in local labor markets. *J. Labor Econ.* 33, 1–32.
- Snijders, T., 2001. The statistical evaluation of social network dynamics. *Sociol. Method.* 31, 361–395.
- Topa, G., 2001. Social interactions, local spillovers and unemployment. *Rev. Econ. Stud.* 68, 261–295.
- Vega-Redondo, F., 2007. *Complex Social Networks*. Cambridge University Press, Cambridge.
- Verdier, T., Zenou, Y., 2004. Racial beliefs, location and the causes of crime. *Int. Econ. Rev.* 45, 727–756.
- Wahba, J., Zenou, Y., 2005. Density, social networks and job search methods: theory and applications to Egypt. *J. Dev. Econ.* 78, 443–473.
- Wasserman, S., Faust, K., 1994. *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge.
- Wellman, B., 1996. Are personal communities local? A Dumptarian reconsideration. *Soc. Netw.* 18, 347–354.

- Wial, H., 1991. Getting a good job: mobility in a segmented labor market. *Ind. Relat.* 30, 396–416.
- Wilson, W.J., 1987. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press, Chicago.
- Wilson, W.J., 1996. *When Work Disappears: The World of the New Urban Poor*. Knopf, New York.
- Yinger, J., 1986. Measuring racial discrimination with fair housing audits. *Am. Econ. Rev.* 76, 881–893.
- Yinger, J., 1997. Cash in your face: the cost of racial and ethnic discrimination in housing. *J. Urban Econ.* 42, 339–365.
- Zenou, Y., 2009. *Urban Labor Economics*. Cambridge University Press, Cambridge.
- Zenou, Y., 2013. Spatial versus social mismatch. *J. Urban Econ.* 74, 113–132.
- Zenou, Y., 2015a. Networks in economics, In: Wright, J.D. (Ed.), *International Encyclopedia of Social and Behavioral Sciences*. second ed. Elsevier Publisher, Amsterdam.
- Zenou, Y., 2015b. A dynamic model of weak and strong ties in the labor market. *J. Labor Econ.* forthcoming.
- Zenou, Y., 2015c. Key players. In: Bramoullé, Y., Rogers, B.W., Galeotti, A. (Eds.), *Oxford Handbook on the Economics of Networks*. Oxford University Press, Oxford, forthcoming.

CHAPTER 10

Immigration and the Economy of Cities and Regions

Ethan Lewis*, Giovanni Peri[†]

*Dartmouth College, Hanover, NH, USA, and NBER, Cambridge, MA, USA

[†]University of California-Davis, CA, USA, and NBER, Cambridge, MA, USA

Contents

10.1. Introduction	626
10.2. Immigrants' Distribution and Native Exposure	632
10.3. Theoretical Framework: The Skill Cells Approach at the National and Local Level	637
10.3.1 Basic framework: Production and labor demand	637
10.3.2 Education- and age-based skill cells in a CES production function: The national approach	638
10.3.2.1 <i>Most commonly used nests</i>	642
10.3.2.2 <i>Partial and total wage effects of immigrants in the CES model</i>	643
10.3.3 The area approach and the labor supply response	644
10.3.4 Occupations and tasks	648
10.3.5 The margin of technological choice	653
10.3.6 Scale externalities, heterogeneity, and search	655
10.4. Empirical Approaches to Identify Causal Effects on Local Economies	657
10.5. Estimates of Native Responses and Effects on Outcomes	661
10.5.1 Geographic mobility of natives	662
10.5.2 Mobility across skill categories	665
10.5.3 Immigrant impacts	667
10.5.4 Summarizing the productivity impacts	675
10.6. Recent Evolutions: Employer–Employee Panel Data and Historical Data	675
10.7. Conclusions	680
References	681

Abstract

In this chapter, we analyze immigration and its effect on urban and regional economies focusing on productivity and labor markets. While immigration policies are typically national, the effects of international migrants are often more easily identified on local economies. The reason is that their settlements are significantly concentrated across cities and regions, relative to natives. Immigrants are different from natives in several economically relevant skills. Their impact on the local economy depends on these skills. We emphasize that to evaluate correctly such impact, we also need to understand and measure the local adjustments produced by the immigrant flow. Workers and firms take advantage of the opportunities brought by immigrants and respond to them trying to maximize their welfare. We present a common conceptual frame to organize our analysis of the local effects of immigration, and we describe several applications. We then discuss the empirical literature that has tried to isolate and identify a

causal impact of immigrants on the local economies and to estimate the different margins of response and the resulting outcomes for natives of different skill types. We finally survey promising recent avenues for advancing this research.

Keywords

Immigration, Labor markets, Skill complementarities, Innovation, Endogenous technical change, Labor supply, Immigrant enclaves, Firms, Productivity

JEL Classification Codes

Labor: J2, J3, J61; International: F16, F22; Production technique and innovation: O31, O33; Regional: R11, R12

10.1. INTRODUCTION

International migrants to the United States and to other rich countries have grown in number and as share of the population during the last four decades. As of 2010, about 10% of the population in the average OECD country (the club of most economically advanced nations) was foreign born. In the United States, this percentage was 12.9, only slightly above that average. The increase over recent decades of such share was also significant as immigrants comprised only 4.7% of the US population in 1970. While this aggregate number is not negligible, what makes immigration particularly interesting to urban and regional economists is its remarkable concentration in some regions and cities. The United States is a good example of this. Immigrants are more geographically concentrated than natives no matter what geographic unit we choose. We will illustrate this fact with more detailed statistics in the next section. For now, let us just mention that California, the top immigration state, hosts 25% of all US foreign born but only 9% of its natives. New York, the top immigration metropolitan area, hosts 14.5% of all US foreign born but only 5.5% of natives (authors' calculations using the 2010 American Community Survey (ACS; [Ruggles et al., 2010](#))).

As a consequence, native individuals have a very different degree of exposure (in any aspect of their life) to immigrants depending on where they live. Among California residents, in 2011, for every two US-born, there was one foreign born. Hence, it was very likely that the effects of those foreign-born individuals, through their economic and labor market transactions, were felt, in some form, by natives. At the other hand of the spectrum, among West Virginia's residents, for every 99 natives, there was 1 immigrant. This makes it much less likely that those few immigrants produced any noticeable economic or labor market impact on most native West Virginians. Even more extremely, Miami and Los Angeles counted more than 40% of foreign-born residents¹ (almost one

¹ The percentage of foreign-born residents was 62% in Miami and 43% in Los Angeles (authors' calculations using [Ruggles et al., 2010](#)).

foreign born for each native) in 2011, while other metropolitan areas (such as Johnstown, PA, and Billings, MT) had less than 1% of foreign-born residents.²

The very uneven distribution of immigrants across regions, relative to the native population, makes for a very good “prima facie” setup to study the differential impact of immigration on the local economies. Different geographic areas and the native workers and firms within them have been exposed to very different inflows of immigrants over the last decades. Hence, by appropriately tracking their economic performance (wage and employment of native workers and productivity of firms) subsequent to the inflow of immigrants, we may be able to identify the effects of immigration on these economies.

Certainly, one has to be very careful in drawing causal inference from statistical association. The location of immigrants is not random but is itself the result of decisions that depended on local economic conditions. A booming economy attracts more workers and more firms. If immigrants respond more vigorously to economic incentives than natives (and there is some evidence of this; see, for instance, [Cadena and Kovak, 2013](#)), an increase in their share in the population may be a consequence (and not a cause) of regional economic success. Caution is also required in identifying the total economic effect of immigrants by analyzing regions, as those are interconnected: the effects of an inflow of immigrants in one region can spill to others through labor mobility, capital mobility, or trade. Nevertheless, exploiting the massive differences in migrant settlements across regions and cities and correlating those differences with local economic outcomes have been the foundation of the largest part of the empirical studies that have focuses on the local effects of immigrants.³

Let us emphasize right away that the features of geographic concentration (and skill concentration, as described below) of immigrants relative to natives are typical not only of the United States but also of most industrialized countries. European cities (such as London, Paris, and Barcelona) have an immigrant density comparable to the top US cities. Our chapter, in fact, will analyze features of immigrants and local economies that can be considered as very general across industrial countries. While we will begin by reviewing several studies that focus on the United States, where this literature originated because of data availability, we will also discuss and analyze many studies and results for other countries, especially in Europe, where immigration flows have been particularly large during the years since 2000, feeding a very contentious policy debate, and where very good administrative data have become available in the recent years, making empirical analysis much more detailed and interesting.

² These percentages are calculated from ACS microdata ([Ruggles et al., 2010](#)) including all working individuals aged 18–65 not residing in group quarters.

³ In the 1990s and part of the 2000s, studies using the geographic variation of immigrants to estimate their economic effects were identified as using the “area approach.” The more competent of those studies, however, always accounted also for the skill distribution of immigrants across area units not simply their density.

Considering immigrants as one group and focusing only on their uneven geographic distribution cannot by itself provide good insight into their economic impact on natives. Immigrants, in fact, are also distributed differently than natives across other dimensions that we broadly define as “skills.” Considering the specific “skill” distribution of immigrants is crucial. For one, it provides us with another dimension of variation to analyze the effect of immigrants. More importantly, it forces us to develop a theoretical approach to analyzing productivity and economic effects of migrants in a context of productive specialization and complementarities across skills. First, we need to identify the more appropriate cells that best correspond to homogeneous skills (or “factors of production”). Then, we need to specify how they are combined in production exhibiting certain patterns of complementarity and substitutability with each other. The immigrant and native distribution across these skill cells and the ability of natives to move across them in response to immigration (as well as to move across geographic units) will be very important factors in determining productivity, wage, and employment effects of immigrants.

There are three dimensions of the native–immigrant difference in skill characteristics that have been used in the literature, leading to somewhat different strategies to identify and analyze the immigrants’ effect. First, immigrants differ in their educational composition vis-à-vis natives. They are relatively more represented among very high (PhD degrees) and very low (less than high school diploma) levels of schooling. Second, they differ in their age distribution, as they are overrepresented among young individuals in the labor force (18- to 35-year old). Third, they are employed in some occupations much more than in others with a clear and specific pattern: They are overrepresented in manual–physical-intensive jobs among the less educated and they are overrepresented in science–technology–engineering–math (STEM)–intensive jobs among the highly educated. In contrast, they are relatively rare in white-collar, communication-intensive, bureaucratic types of jobs. This is possibly because their language skills provide them with a comparative disadvantage in those jobs as their physical/manual skills (on one hand) or mathematical/analytic skills (on the other) are more internationally transferrable.

This concentration of immigrants in some skill groups produces three interesting theoretical consequences that we need to consider when analyzing the impact on the native economy. First, the effect on natives will depend on a native’s characteristics: individuals with skills and in occupations similar to those where immigrants concentrate will experience their competition more strongly. Individuals in other jobs will experience a beneficial effect or no effect at all, depending on the productive interactions (complementarity) between skills. Second, this uneven concentration will introduce differential incentives for natives to “move” out of their cells. While they may move across local economies toward or away from the areas where immigrants concentrate (depending on their competition or complementarity), they can also move away from the skill cells in which immigrants are concentrated to those skill cells benefiting from immigrants. While workers cannot change their age, they can change their education, occupation,

and job specialization, and they typically do over their working career. When exposed to immigration, natives will have economic incentives to specialize, upgrade, and direct their career in order to maximize returns and minimize losses from immigration. Finally, firms are also important players. When faced with a changing concentration of potential workers across skill cells, they may adopt differential technologies or techniques or they may change product combination so as to use more intensely and more efficiently those skills that have become more abundant.

Traditionally, the economic analysis has distinguished between short-run and long-run effects of immigration. However, the so-called short-run effects are mostly a theoretical device to decompose a complex effect. When economists analyze the “short-run effects” of immigrants, they try to isolate the consequences of immigration when all other variables (including the stock of capital, the skill supply of natives, and the technology and productive structure) are fixed. This should be called “partial” effect. It is a way to understand and isolate a specific effect, not a way to forecast what happens, even in the short run. The adjustments in skill supply of natives, the adaptation of technologies and the related capital investments, and the change in product composition described above have typically been associated to the long-run response to immigration. However, bar some exceptional cases, immigration has been a slow and consistent force in the last decades for most countries. It has rarely (if ever) been a temporary 1-year burst followed by slow adjustment. Typically, the yearly inflow of immigrants in countries with fast-growing foreign population has been between 0.3% and 0.6% of the resident population. These inflows have produced significant changes over time, but the horizon to observe these consequences is decades, not years. Hence, the speed of these inflows and their progression and relative predictability imply that the correct perspective is a “long-run” one. Within this time horizon, the described adjustment margins (changes of native skill supply, of capital, of technology, and of output composition) have also played important roles and need to be analyzed as part of the effect of immigration. Let us also add that a focus on the “long-run” consequences of migration implies that the most relevant measure of immigration flows in a country is the change in the stock of foreign born, hence net migration. This implies that short-run temporary flows of migration and return are not central in our chapter. While there is an interesting literature devoted to the selection of returnees and to how this affects the features of remaining migrants (e.g., [Abramitzky et al., 2014](#)), we are simply focusing on the characteristic of nonreturning migrants in the long run and their effects on the receiving economies.

The “long-run” nature of the migration phenomenon and the skill characteristics of migrants, at the top and bottom of the receiving-country human capital distribution, imply that at the national level, immigration could have an important role in economic growth and economic inequality. In particular, due to the increase in economic inequality in the United States during the last three to four decades, immigration has been sometimes scrutinized as a potential determinant of it, through its labor market competition

effects on less educated natives. Card (2009) and Blau and Kahn (2012) did not find a significant role of immigration in the increase in US inequality during the recent decades. The relatively balanced inflow of immigrants between college-educated and noncollege-educated and the response of local markets and native workers (that we will analyze below) imply a small effect of immigration on native wage (and income) inequality. Also, while some immigrants themselves are at the bottom of the income distribution, their number as share of population is relatively small.⁴ Several of the studies at the national level that we will review in Section 10.3.2⁵ have quantified the contribution of immigration to inequality in the United States, and none of them has found more than a very small role. Dustmann et al. (2013) considered more directly the effect of immigration on the UK wage distribution and found a mild positive effect on inequality, mainly through an increase in high wages due to complementarity with immigrants.

On the other hand, very limited research exists, at the national level, on immigration and growth. Ortega and Peri (2014) are among the few to tackle the issue of estimating the impact of immigration on average GDP per person and aggregate productivity using cross-country analysis. They use geographic features predicting immigration and control for an array of institutional, cultural, and historical determinants, to isolate the effect of immigration. They find a strong positive effect of the immigrant share on productivity across countries, and they document that this derives in part from more innovation and from other benefits of diversity. Alesina et al. (2013) adopted a similar approach to analyze the effect of “country of birth” diversity on GDP per person and productivity and also found a positive and significant effect. Also promising are those studies analyzing the impact of highly skilled immigrants (scientists and engineers) on average wage and productivity growth in US cities (such as Peri et al., 2014 described in Sections 10.3 and 10.5). The aggregate studies mentioned are interesting and quantitatively useful. However, our approach in this chapter will look in greater detail at mechanisms and models that help us understand the working of immigration on economic activity, productivity, and labor markets. Focusing on local economies and shedding light on those mechanisms have clearly important implications on the role of immigration on aggregate inequality and growth.

After presenting in Section 10.2 some statistics about the distribution of immigrants across geography and skill space, Section 10.3 introduces a rather general “production-function approach” to the economic effects of immigration. We will focus on wage and employment effects of immigrants, and we will also discuss productivity effects that need to be considered as we analyze specialization and choice of technique. The analyzed approaches model the skill interactions across different types of workers in a city

⁴ Peri (2013) analyzed directly the impact of immigrants on native poverty rates in the United States during the years 1990–2010, through the labor market competition channel. He found extremely small effects.

⁵ See in particular Borjas (2003), Borjas and Katz (2007), and Ottaviano and Peri (2012).

(or region) using a production-function approach. We devote special attention to the nested-CES approach that organizes native and immigrants workers in education, age, and nativity cells and then into production tasks. Variations of this model have been widely used in the recent literature. Using this framework, we derive effects on productivity and wages, and they also produce predictions on changes in specialization, skill supply, and choice of production technology, consequent to immigration-induced changes to the distribution of skills.

Then, [Section 10.4](#) analyzes the empirical strategies used to isolate immigration as an exogenous change in skill supply at the local level. We will consider the potential challenges to identification and the proposed remedies. In particular, the methods based on preexisting settlements and current aggregate inflow by nationality and those focusing on some sudden and large migration shocks or policy changes are considered. Then, in [Section 10.5](#), we review estimates of the effects produced by immigration on local economies in terms of wages and productivity outcomes, and we will pay attention to native responses to immigration and to a general equilibrium effect. The inflow of immigrants, in fact, appears to trigger a mobility response of natives (as immigration changes the relative rewards for them). It turns out that the most significant responses are not represented by net outflows or inflows of natives in geographic areas (what [Card and DiNardo, 2000](#) called the “skating-rink” hypothesis) but by increased mobility across skill cells (specialization, occupation upgrading, and education improvement). This is important because mobility in the skill space affects native wages in a different way than mobility across regions. In particular, if natives move from skills (occupation, tasks, and jobs) that are more substitutable to skills that are more complementary to immigrants’ in response to their inflow, this response would increase the native gains from immigration, and those gains can be captured within an economic area. If they instead move out of the area simply to avoid competition, they may not gain from immigrants and the area analysis may miss some of the total effect. Another aspect that we emphasize is that firms may be induced to adopt technology and capital in order to take advantage of immigrant skills. This is even more important in the long run because it may change the productivity of specific skills. We will analyze studies that combine the direct effects and the induced responses (of native workers and firms) to determine the observed productivity and wage outcomes.

Analyzing recent contributions, we think that the differentiated skill cell approach, using variation across regions and cities, is emerging as dominant in the study of immigration to the United States and to other developed countries. Recently, individual-level and firm-level data from different developed countries (mainly in Europe) have also been tackled to analyze these effects. The most interesting European data are from administrative sources and make available to the researcher panels of individuals over time and panel of establishments over time. The ability to identify firm outcomes and the possibility of following individual workers make those data sets able to reveal in more detail

the microlevel mechanism of adjustment of local economies to immigration. While several empirical and identification issues still exist when using these data, we think they add very interesting tools to our understanding of the role of immigrants, and in particular, they allow a closer inquiry of the mechanisms at work within labor markets. In [Section 10.6](#), we will analyze the possibilities opened in terms of methods and analysis by the availability of these individual panel data sets, and we also review some recent studies using historical microdata to analyze the productive response to immigrants in historical large migration episodes. Finally, in [Section 10.7](#), we briefly summarize and conclude the chapter.

10.2. IMMIGRANTS' DISTRIBUTION AND NATIVE EXPOSURE

Immigration affects the geographic and the skill distribution of productive resources (workers) in a country. Defining the relevant cells to analyze the economic and productive consequences of immigrants is important. First, however, we describe how the distribution of immigrants in the United States differs on dimensions of geography and skills relative to the distribution of natives. These differences are what create economic opportunities and incentives to implement changes and adjustment by native agents. We use data from the American Community Survey 2011, and we select only individuals who are currently working.⁶ A few simple statistics help us to see that the largest variation in native exposure to immigrants is in the geographic dimension, using metropolitan areas as units of analysis. Not only does immigrants' share of employment vary hugely across units, but also immigrants exhibit a much stronger absolute concentration in the top locations than natives do. Then, we analyze the distribution of immigrants and natives across occupations, using the census occupational classification, and finally, we describe the distribution of immigrants across education and age groups. All empirical studies we are aware of use one or more of these dimensions as unit of analysis to identify the productive and labor market effects of immigrants.

The upper part of [Table 10.1](#) shows some simple statistics on the overall concentration of immigrants, relative to natives, across different dimensions. The lower part of [Table 10.1](#) shows instead statistics representing the variation in exposure of natives to immigrants across cells in that dimension. Column 1 of the table considers 284 metropolitan areas as cells, column 2 considers 50 states, column 3 considers 333 occupations, and column 4 uses 7 schooling groups.⁷ Finally, column 5 considers 70 education-by-age groups (7 education groups each divided into 10 age bins, for workers between 18 and 65).

⁶ Specifically, we consider individuals 18–65 years of age, not living in group quarters who have worked at least a week.

⁷ The groups are no diploma, high school diploma, some college, associate degree, college degree, master, and PhD.

Table 10.1 Measures of concentration and exposure native–immigrants for different cell structures

Cell	Metropolitan areas	States	Occupations	Education cells	Education–age cells
Number	284	50	333	7	70
Measures of concentration across cells: relative values immigrant/native					
Foreign/native Herfindahl	3.42	2.76	0.87	0.93	0.99
Foreign/native percentage of population in top unit	2.48	2.74	1.09	1.03	0.98
Foreign/native percentage of population in top 5% of units	1.86	2.07	0.92	N.A.	1.05
Foreign/native percentage of population in top 10% of units	1.60	1.85	0.88	0.94	1.06
Measure of variation in native exposure to immigrants across cells					
(Immigrant per resident) Top 10%/bottom 10%	11.1	6.6	4.3	2.6	5.2
(Immigrant per resident) Top 5%/bottom 5%	17.2	10.3	7.3	N.A.	6.8
(Immigrant per resident) Ratio of largest to smallest	65.1	21.8	63.0	4.6	10.0

Note: All statistics use all US resident individuals, not living in group quarters and working at least 1 week, aged 18–65. The data are from ACS 2011.

The production models that we analyze below consider a stronger direct competition effect of immigrants on natives when they are in the same cell. Hence, the variation in exposure across cells is a crucial dimension to identify the direct competition effect. On the other hand, it is also very important to consider different degrees of interaction between cells and also different ability of natives to move across cells. The more recent empirical studies have been careful in accounting for these cross effects and responses. The interaction between skill cells is typically analyzed within the context of complementarities/substitutability of skills in production, while the interaction across geographic cells is usually considered in the context of the native migration response to immigrants.

The distribution across metropolitan statistical areas (MSAs) shows the strongest difference in concentration of immigrants relative to natives. The Herfindal index of concentration across MSAs, which is calculated as the sum of squared share of total population in each unit, captures the degree of concentration of a population in cells. The Herfindal index of urban population (between 0 and 1) would be close to 1 if most of the urban population (immigrant or native) in the United States were concentrated in the largest metropolitan area. It would be essentially 0 if the urban population was instead equally distributed across metropolitan areas of the same size. The table reports the immigrants–natives ratio of such Herfindal index and implies a 3.5 times larger value for immigrants than for natives denoting a significantly larger concentration of their urban population in the largest metropolitan areas. Similarly (second row), the percentage of immigrant employment in the top metro area is 2.5 times as large as the percentage of natives employed in the top metro area. The percentage of immigrants in the top 5% of metro areas is 1.8 times the percentage of natives in top 5% of metro areas, and immigrants in the top 10% MSAs are 1.6 times the percentage of natives in the top 10% metro areas. The stronger concentration of immigrants across metropolitan areas relative to natives is also shown in [Figure 10.1](#). In that figure, we see that the percentage of total immigrant employment in the top 15 metropolitan areas is significantly larger than that of

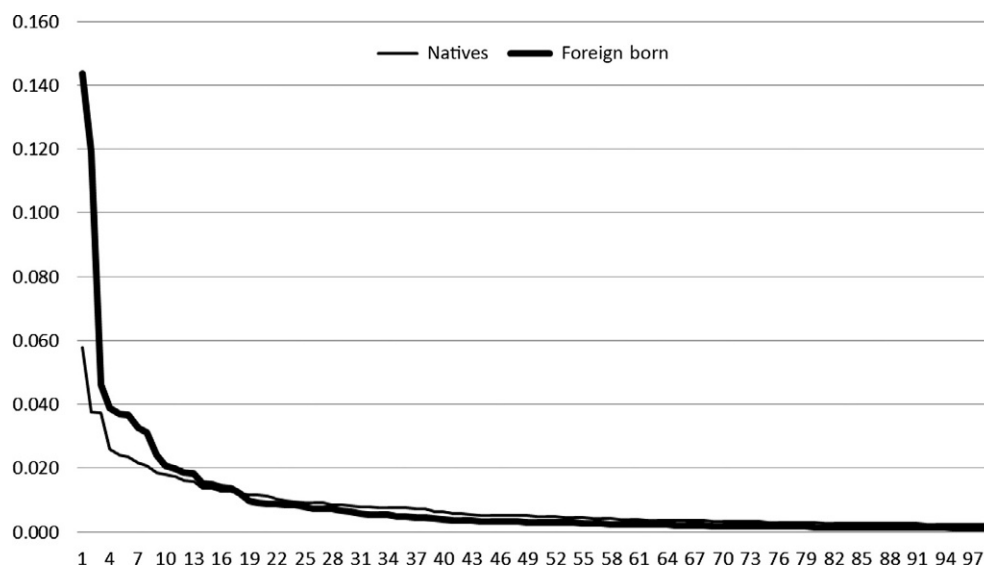


Figure 10.1 Sorted share of urban employment in each of top 100 metro areas—native and foreign born. *Note: The share of employment in each metro area is calculated based on 2011 ACS data, excluding people in group quarters and including only people 18–65 who worked at least 1 week. We consider as urban population is the total population of the top 284 metro areas in the United States.*

natives. When ranking metropolitan areas based on their percentage of total urban employment, the mass of immigrants is strongly shifted toward the very top areas, relative to the mass of natives. A similar pattern of stronger geographic concentration is also revealed in column 2 when we consider states. California, the top immigrant state, had 25% of all US working immigrants but only 9% of all US native workers. Similarly, Figure 10.2 shows much larger concentration of immigrants in the top five states relative to natives.

The lower part of Table 10.1 shows even more interesting statistics. We show the range in the exposure of natives to immigrants as the ratio between the share of foreign born in the most exposed and that of in the least exposed cells. The bottom row is the ratio of the highest exposed cell to the lowest exposed cell; the next row up is the ratio of the cell at the 95 percentile (top 5%) and the one at the fifth. The row above that shows the ratio of the 90–10 percentile. Remarkably, the share of foreign born in the city with the highest concentration (Miami) was 65 times the share in Johnstown (PA), the city with the lowest relative presence of immigrants. Even the 90–10 percentile ratio was a very large, 11. This means that in metropolitan areas with high concentration of immigrants, their density relative to natives was more than 10 times larger than in metropolitan areas with low concentration. Across states, the variation was also remarkable with a top–bottom ratio of almost 22. While these differences are certainly not random, comparing wage, productivity, employment, and other economic outcomes across cells that experience such a drastically different presence of immigrants, if done carefully, could reveal important implications of their presence.

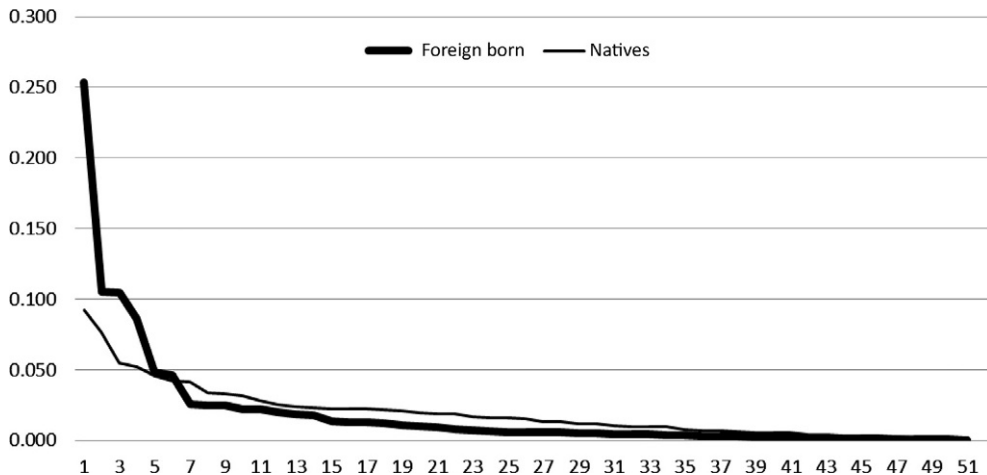


Figure 10.2 Sorted share of total employment in each state—native and foreign born. Note: The share of employment in each state is calculated based on 2011 ACS data, excluding people in group quarters and including only people 18–65 who worked at least 1 week. We consider 50 US states plus DC (hence 51 units).

Let us then analyze similar statistics calculated across occupations (column 3) and education and age (columns 4 and 5). In terms of occupations, we notice that immigrants are not more concentrated across them, in absolute terms, than natives are. The indices of relative concentration in the upper part of the table are, in fact, close to 1. However, their distribution across occupations is very different from that of natives, and it generates very different degrees of exposure of natives to immigrants depending on the occupation they are in. The top–bottom ratio is 63 and the 95–5 percentile ratio is 7.3. A native working as “sorter of agricultural products” (the top occupation as share of immigrants) is exposed to a share of immigrants 63 times larger than one working as “funeral director (sic!)” (the occupation with the lowest share of immigrants). Occupations, therefore, as metropolitan areas, vary enormously in the presence of immigrants. Moreover, both dimensions exhibit a significant intercell mobility of native workers over their lifetime, especially when young. Hence, the differential immigration “pressure” across cells may produce a significant response of natives in flowing across cells. This does not imply that they cannot be used as units of analysis of the effects of immigrants, but one certainly needs to account for flows of natives between them as potential response to immigration.

The last two columns show the relative concentration and distribution of immigrants across education and education–age cells. In both dimensions, immigrants and natives have similarly concentrated distributions (ratios are close to 1). Moreover, the educational grouping, as it only includes seven cells that are much larger than those of other grouping, does not exhibit the extreme differences in native–immigrant exposure as the other groups. Even when we consider 70 education–age groups, the range of exposure to immigrants is significantly smaller than for the geographic dimension. Column 5 in [Table 10.1](#) shows a range of exposure of 10 in the top–bottom comparison and of 5 in the 90–10 percentile; both values are well below the corresponding ratio in the geographic units (metropolitan areas and states). An interesting feature of education–age as skill groups is that the intercell mobility of natives in response to immigrants may be significantly smaller than for the geographic–occupation cells. As we will see, the “given” native supply (nationally) within each cell, even as immigrant pressure may vary across them, has contributed to the success of this cell structure in analyzing the effect of immigration.

Let us, finally, emphasize that there is a key economic difference between the geographic and the skill cell units. In the first case, one can treat cells as separate units (in production and as labor markets) and worry later about potential interactions across them because of native mobility or trade of goods and capital. This has been the approach of regional and labor economists, assuming at first independent units (cities and states) and then checking whether the linkages (through internal migration or trade) would affect the findings. In the skill approach, instead, cells are considered as factors interacting within one same production process, and hence, one cannot analyze each cell in isolation. Economists have clearly understood the need to model right away linkages and interactions among them as a first-order concern. The approaches we prefer combine skill cells

as factors of production and geographic cells as different production units. It is time to introduce a framework for organizing workers in skill cells and a simple structure to analyze cross cell interactions and potential cross cell mobility.

10.3. THEORETICAL FRAMEWORK: THE SKILL CELLS APPROACH AT THE NATIONAL AND LOCAL LEVEL

10.3.1 Basic framework: Production and labor demand

The commonly used framework to think about the impact of immigrants within the skill cell approach considers an area (typically a region, state, or a city) as producing a homogeneous tradable final good by combining different production skills and physical capital through a production function. This final good (output) is the numeraire, and we can think of the production function of a region as the reduced form of a multigood economy in which different nontradable intermediate goods (and services) each provided by a skill type are combined in the typical final consumption basket (the final output). The simplification is that all local economies produce and consume the same final good, y . They may, however, have different supplies of each intermediate factor (skill) and different techniques in production and hence different marginal productivities and returns to skills.

An alternative framework is one in which individual localities produce a number of different varieties and they partially specialize in the production and trade of varieties. This would generate a Heckscher–Ohlin type of model with a further margin of adjustment to changes in skills due to immigration, represented by changes in the variety composition of production. An increase of a type of skill due to immigrants could be fully absorbed by a change in production composition toward goods intensive in the use of that skill (the so-called Rybczynski effect). However, [Lewis \(2003\)](#) and [Card and Lewis \(2007\)](#), among others, showed that the adjustment in the variety composition of output is not an important margin of adjustment to immigration. This implies that the constant output composition model (the one-good model, used here) does not miss an important margin of adjustment and is a reasonable working model.

For area (region and city) r , the production function of output can be represented as follows:

$$y_r = F(\mathbf{A}_{K,r}\mathbf{K}_r, \mathbf{L}(\mathbf{A}_{1,r}\mathbf{L}_{1,r}, \mathbf{A}_{2,r}\mathbf{L}_{2,r}, \dots, \mathbf{A}_{n,r}\mathbf{L}_{n,r})) \text{ for } r = 1, 2, \dots, R \quad (10.1)$$

where $\mathbf{L}_{n,r}$ is the amount of factor (skill/task) n used in the production of area r . Similarly, $\mathbf{A}_{n,r}$ is the productivity of factor n in area r . In general, we allow for factor-specific productivity (determined by the chosen technology) to vary across localities (hence the subscript r). Notice that we included the physical capital K (and its productivity $\mathbf{A}_K$) as a factor separable from an aggregate labor factor ($\mathbf{L}$) that, in turn, combines all the skill groups $\mathbf{L}_1, \dots, \mathbf{L}_n$ and their productivity $\mathbf{A}_1, \dots, \mathbf{A}_n$. This implies that physical capital is combined with the labor aggregate and has the same degree of substitutability/complementarity with

all skill cells. An alternative to this assumption is entertained by [Lewis \(2013a\)](#) who explored the consequences of considering different degrees of complementarity between physical capital and different skill groups. In particular, in the more relevant case of complementarity between capital and college-educated workers, [Lewis \(2013a\)](#) showed that the capital response to immigration of college-educated will attenuate its wage impact.⁸ Capital–skill complementarity is an interesting and important avenue to pursue. However, the current literature on regional impact of immigrants mostly relies on the assumption of separability between capital and aggregate labor.⁹

The capital separability assumption, combined with the assumption of long-run mobility of capital and constant long-run returns for capital, implies that we can solve physical capital out of the function to obtain a reduced form:

$$\mathbf{y}_r = \mathbf{f}(\mathbf{A}, \boldsymbol{\theta}_{1,r} \mathbf{L}_{1,r}, \boldsymbol{\theta}_{2,r} \mathbf{L}_{2,r}, \dots, \boldsymbol{\theta}_{n,r} \mathbf{L}_{n,r}) \text{ for } r = 1, 2, \dots, R \quad (10.2)$$

In (10.2), the parameter $\mathbf{A}$ is a combination of parameters including the return and productivity of physical capital and total factor productivity, while the terms $\boldsymbol{\theta}_n$ capture relative productivity of factor (skill) n standardized so that $\sum_n \boldsymbol{\theta}_n = 1$. In the long run, competition among workers and firms ensures that each factor is paid its marginal product. Hence, the compensation to each skill in region r $w_{n,r}$ is as follows:

$$w_{n,r} = \frac{\partial \mathbf{F}}{\partial \mathbf{L}_{n,r}} = \mathbf{f}_n(\mathbf{A}, \boldsymbol{\theta}_{1,r} \mathbf{L}_{1,r}, \boldsymbol{\theta}_{2,r} \mathbf{L}_{2,r}, \dots, \boldsymbol{\theta}_{n,r} \mathbf{L}_{n,r}) \quad (10.3)$$

If the reduced form production function is constant return to scale in the labor aggregate, then the sum of compensation to skill equal total output in region r .

10.3.2 Education- and age-based skill cells in a CES production function: The national approach

While early studies (such as [Grossman, 1982](#)) experimented with different functional forms for the production function in (10.2), such as the flexible translog specifications, the more recent research on the local (and national) impact of immigrants has focused

⁸ This can be shown with the derivative identity, $\frac{d \ln(w_S/w_U)}{d \ln(L_S/L_U)} = \frac{\partial \ln(w_S/w_U)}{\partial \ln(L_S/L_U)} + \frac{\partial \ln(w_S/w_U)}{\partial \ln K} \frac{\partial \ln K}{\partial \ln(L_S/L_U)}$, which says that the total relative wage response to a change in the supply of skilled labor (S) relative to unskilled labor (U) is equal to its partial direct effect— $\frac{\partial \ln(w_S/w_U)}{\partial \ln(L_S/L_U)}$, the (negative of the) inverse elasticity of substitution—plus indirect effects working through the adjustment of capital. Under capital–skill complementarity, both $\frac{\partial \ln(w_S/w_U)}{\partial \ln K}$ and $\frac{\partial \ln K}{\partial \ln(L_S/L_U)}$ are positive, so the adjustment of capital attenuates wage impacts $\left(\frac{d \ln(w_S/w_U)}{d \ln(L_S/L_U)} > \frac{\partial \ln(w_S/w_U)}{\partial \ln(L_S/L_U)} \right)$. When capital is instead assumed separable from labor inputs in production $\frac{\partial \ln(w_S/w_U)}{\partial \ln K} = 0$, so $\frac{d \ln(w_S/w_U)}{d \ln(L_S/L_U)} = \frac{\partial \ln(w_S/w_U)}{\partial \ln(L_S/L_U)}$ (which makes it convenient to make this assumption).

⁹ To partial defense of this approach, many of the insights from capital–skill complementarity are recovered in the literature through the introduction of endogenous choice of techniques (hence technology–skill complementarity) that we will review in [Section 10.3.6](#).

on CES and specifically on nested-CES functions. The reason is that the nested CES provides a simple expression of the (log) marginal productivity of each skill as a function of the supply of the same skill, of simple aggregators of other skill supply, and of a small number of parameters. Hence, observing skill supply and compensation (wages) and accounting for the factor aggregators (also easily constructed), one can use (10.3) to estimate empirically the few parameters regulating the response of wages to changes in skill supply.

It is useful to describe in some detail, following Ottaviano and Peri (2012), how the nested-CES approach can be used to estimate important elasticity parameters and to calculate effects of the change in immigrant supply on wages. This approach has been used by several recent empirical papers. The most relevant characteristics used to organize cells in the nested-CES framework have been education levels, age groups (or experience groups), and nativity groups (foreign–native). These have provided the grid to organize workers into cells. Adopting a CES structure, one could represent production function (10.2) with a small number of parameters. And one would be able to estimate those parameters using the whole country as relevant area, simply exploiting the variation of immigrant supply over time and across skill cells.

The cell structure we describe here, originally proposed by Borjas (2003) and Card and Lemieux (2001), has then been followed and enriched by Ottaviano and Peri (2012) and Manacorda et al. (2012) and then followed by other studies after those. All those studies have considered the whole country, rather than local areas, as units of analysis, and hence, we omit the area subscript (r) in this section. One appealing feature of this approach is that considering relatively fixed characteristics (such as age, education, and nativity) and a national market makes the operational assumption that skill supply by natives did not respond to immigration more plausible. In this national approach, the typical assumption is that the supply of skills by natives is totally inelastic (given).

We describe the flexible nested-CES structure that embeds various alternative models studied in the literature, using general notation and allowing for recursive expressions of general results. Consider four characteristics numbered $n=0, \dots, 3$. Characteristic 0 is common to all workers and defines them as such. Characteristic 1 is education and can be used to partition workers into groups $i(1)=1, \dots, M_1$ that differ according to educational attainment (e.g., high school dropouts, high school graduates, and college graduates). Then, each of these education groups can itself be partitioned into groups $i(2)=1, \dots, M_2$ that differ according to characteristic 2, which is age (say, age intervals in the range 18- to 65-year old). Finally, each of those can be partitioned into two groups “natives” and “foreign born” according to characteristic 3, which is “nativity.”¹⁰

¹⁰ Studies that are focused on the diversity of immigrants consider “country of birth” as relevant characteristics and include several countries (or groups) as categories for this partition. See, for instance, Ottaviano and Peri (2005, 2006).

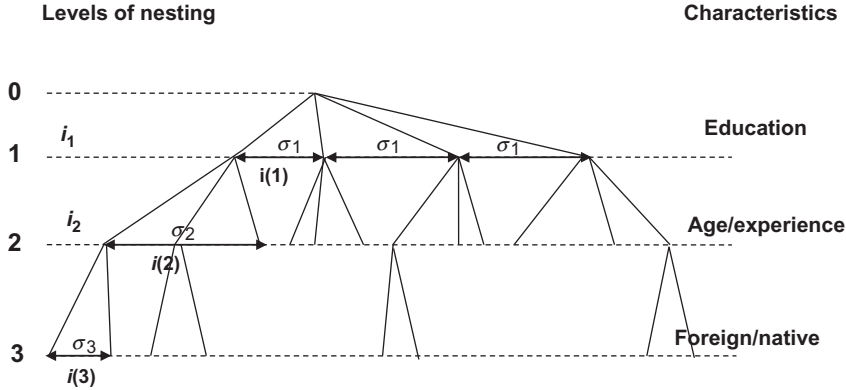


Figure 10.3 General scheme of the CES nests and relative notation education.

This sequential partitioning and its relative notation are illustrated in Figure 10.3. The figure shows how groups are nested into each other with n indexing the nesting level.

The nested-CES structure allows us to define production function (10.2) in the following recursive form. Let us call $i(n)$ a group (cell) of workers defined by common characteristics up to n and define as $L_{i(n)}$ the corresponding factor supply. The CES aggregator at the level n is then defined as

$$L_{i(n)} = \left[\sum_{i(n+1) \in i(n)} \theta_{i(n+1)} (L_{i(n+1)})^{\frac{\sigma_{n+1}-1}{\sigma_{n+1}}} \right]^{\frac{\sigma_n+1}{\sigma_n+1-1}}, \quad n = 0, 1, 2, 3 \quad (10.4)$$

where $\theta_{i(n+1)}$ is the relative productivity level of type $i(n)$ standardized so that $\sum_{i(n) \in i(n-1)} [\theta_{i(n)}] = 1$. Any common multiplying productivity factor is absorbed in the TFP parameter $\mathbf{A}$ shown in expression (10.2). Both the parameter $\mathbf{A}$ and $\theta_{i(n+1)}$ depend on exogenous technological factors only. The parameter $\sigma_n > 0$ is the elasticity of substitution between types $i(n)$. Hence, σ_1 is the elasticity of substitution across education group, σ_2 is the elasticity across age groups within education category, and σ_3 is the elasticity between natives and immigrants in the same education–age group. Given the ordering of characteristics and sequential partitioning that leads to less and less heterogeneous groups $i(n)$ as n increases, a reasonable assumption is that $\sigma_3 > \sigma_2 > \sigma_1 > 1$. As type $i(0)$ includes all workers, we can embed the nested structure defined by (10.4) into (10.2) by writing that equation as $\mathbf{y}_i = \mathbf{f}(\mathbf{A}, \mathbf{L}_0)$, where $\mathbf{L}_0$ is the top-level aggregator in the nesting.

Using this structure and notation, we can express the wage of a worker of type $i(3)$, where $i(3)$ indicates a cell for specific values of education, age, and nativity, as the value of her marginal productivity:

$$\begin{aligned} \ln(w_{i(3)}) = & \ln(A) + \frac{1}{\sigma_1} \ln(L_0) + \ln\theta_{i(1)} - \left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2}\right) \ln L_{i(1)} + \ln\theta_{i(2)} \\ & - \left(\frac{1}{\sigma_2} - \frac{1}{\sigma_3}\right) \ln L_{i(2)} + \ln\theta_{i(3)} - \frac{1}{\sigma_1} \ln L_{i(3)} \end{aligned} \quad (10.5)$$

First, focusing on the last level of nesting and considering native (nat(3)) and foreign born (for(3)) sharing the same characteristics of the first two nests, education and experience ($i(2)$ and $i(1)$), Equation (10.5) implies

$$\ln\left(\frac{w_{\text{nat}(3)}}{w_{\text{for}(3)}}\right) = \ln\left(\frac{\theta_{\text{nat}(3)}}{\theta_{\text{for}(3)}}\right) - \frac{1}{\sigma_3} \ln\left(\frac{L_{\text{nat}(3)}}{L_{\text{for}(3)}}\right) \quad (10.6)$$

Therefore, $\frac{1}{\sigma_3}$ the inverse elasticity of substitution between natives and immigrants in the same education–age cell can be estimated from observations on wages and employment levels of natives and immigrants over time, using fixed effects to control for $\ln\left(\frac{\theta_{\text{nat}(3)}}{\theta_{\text{for}(3)}}\right)$. Second, for higher nesting level $m=1, 2$, we can define $w_{i(m)}$ as the average wage of a specific group of workers $i(m)$ sharing characteristics up to m . Then, substituting m instead of 3 as the highest nest level in expression (10.5) gives the profit maximizing relation between $w_{i(m)}$ and $L_{i(m)}$. In this case, using observations over time, the estimation of $\frac{1}{\sigma_m}$ can be achieved by regressing the logarithmic wage of group $i(m)$ on the logarithmic CES aggregate $L_{i(m)}$ with the inclusion of fixed time effects to capture the variation of the aggregate terms $\ln(A)$ and $\ln(L_0)$, when estimating $\frac{1}{\sigma_1}$. In the case of $m=2$, when estimating the elasticity of substitution across age groups, we should also include education by year effects in order to absorb the terms $\ln\theta_{i(1)} - \left(\frac{1}{\sigma_1} - \frac{1}{\sigma_2}\right) \ln L_{i(1)}$ that do not change with characteristic 2 (age).

Once we have estimated the elasticity of substitution between different types of workers at each level of the nest, the wage equation (10.5) can also be used to compute the percentage change in the wage of workers of a certain type j (defined by a specific combination of education–age–nativity) caused by a percentage change in the labor supply of workers of another type i (defined by another combination of characteristics). To show this in a compact way, let us denote by s_i^m the type i 's share of labor income among workers exhibiting the same characteristics up to m as that type. Then, we can write the percentage impact of a change in labor supplied by workers of type i on the wage of a worker of type j who share the same characteristics up to m as follows:

$$\frac{\Delta w_j^0 / w_j^0}{\Delta L_i / \Delta L_i} = \frac{s_i^0}{\sigma_1} > 0 \text{ for } m=0 \text{ and } \frac{\Delta w_j^m / w_j^m}{\Delta L_i / \Delta L_i} = - \sum_{n=0}^{m-1} \left(\frac{s_i^{n+1} - s_i^n}{\sigma_{n+1}} \right) < 0 \text{ for } m=1, 2, 3 \quad (10.7)$$

Three remarks are in order. First, an increase in the labor supply of a certain type i causes an increase in the wage of another type j only if the two types differ in terms of

characteristic 1 (education in our case) as shown in the first expression of (10.7). In that case, the factors are complements. Second, if the two types share at least characteristic 1, then a rise in the labor supply of i always depresses the wage of j (second expression) as $s_i^{n+1} > s_i^n$ if groups are ordered in increasing level of substitutability. This effect is stronger the larger the number of differentiating characteristics j has in common with i , because this implies more terms in the summation in (10.7). Third, and specific to the effect of immigrants, while the partial effect of immigrants in the same education–experience group as natives is negative, this is only a partial effect. The impact of immigrants in other education–age groups on native wages may be positive and the total effect may therefore be positive. The production function described above allows us to use the easy formulas in (10.7) to calculate the wage impact of immigrants in each education–age group on the wages of natives in each education–age group, once we have the elasticity and the wage shares.

10.3.2.1 Most commonly used nests

Within the general structure described above, the literature based on a nested-CES function has converged toward one (or a few) most commonly used partitions at each level of the nesting. Beginning with the lower level (nativity), most of the literature since Ottaviano and Peri (2012) has allowed two imperfectly substitutable groups of workers: natives and foreign born. There are several reasons for this simple partition. First, even when considering workers with equivalent education and experience, natives and immigrants differ in detailed abilities, motivations, and tastes that may set them apart. Second, in manual and intellectual work, they have culture-specific skills (e.g., cooking, crafting, artistic abilities, and sport talent) and limits (e.g., limited knowledge of the language or culture of the host country), which create comparative advantages in some tasks. Third, due to comparative advantage, migration networks, or historical accidents, immigrants tend to choose different occupations with respect to natives, even for given education and experience levels. Finally, there is no need to impose perfect substitutability between natives and immigrants *ex ante* as, within the structure proposed, this elasticity can be estimated. While one could envision a larger number of nativity groups, based on areas of origin, the most common studies only separate natives and foreign born.

In terms of the second level of the nest (characteristic 2 is age or experience), the literature has been rather openhanded. Some studies allow 4 or 8 age groups partitioning experience between 0 and 40 years of work (Card and Lemieux, 2001; Borjas, 2003). Others only include two groups (young and old). As it turns out (see Ottaviano and Peri, 2012), this partition and the relative elasticity of substitution are not very relevant in determining wage effects between immigrants and natives and between skilled and unskilled (usually associated with educational differences). In some cases (Peri and Sparber, 2009; Docquier et al., 2011), this level of the nest is omitted altogether, because it does not affect much the consequences of immigrants on native wage distribution.

Finally, and importantly, characteristic 1 determines the grouping according to education. The partition more frequently used in the labor literature is a division into two broad educational characteristics, “high school equivalents,” which include individuals up to a high school diploma, and “college equivalents,” which include individuals with some tertiary education and a those with college degree. Several papers, most notably Goldin and Katz (2008) and Katz and Murphy (1992) (but others as well¹¹), have emphasized that college-educated and high school-educated are hard to substitute and their relative supply, combined with technological progress and an elasticity of substitution around 1.5–2, explains well their relative wage movements in the United States post 1960. The further distinction between high school graduates and high school dropouts does not seem useful to understand relative wages in the United States (see Card, 2009; Ottaviano and Peri, 2012) because those two groups seem close substitutes to each other in production, at least after 1950 (Goldin and Katz, 2008). Hence, we will consider the college–high school partition of education and the foreign–native partition of nativity as the most common features of this approach, with a less clear preference for 2, 4, 8 or even omitting the level altogether, for age–experience groups.

10.3.2.2 Partial and total wage effects of immigrants in the CES model

The nested-CES model described above allows us to distinguish partial and total wage effects of immigrants. The former is the wage impact on native workers due to a change in the supply of immigrants with the same education–age characteristics, while keeping constant the labor supplies of all other workers. This effect has been the main or only coefficient of interest in many “reduced form” approaches that regress native wages on the employment of immigrants in the same skill groups.¹² However, this effect is only an “artificial” partial effect as it misses the entire set of cross effects. The total wage effect, instead, accounts also for the indirect impact of immigration among all groups of workers and is what one would be interested in when analyzing the impact of changes in immigration flows (or immigration policies).

The direct partial wage effect can be estimated by panel regressions of $\ln(w_{j(N)}^{N-1})$ the logarithmic wage of natives, sharing characteristics up to $N-1$ (namely, education and age) on the supply of immigrants, $\ln L_{i(N)}$, in the same age–education group. Careful econometric specifications (such as Borjas, 2003) control for year-specific effects (to absorb the variation of L_o , the labor aggregate over time) and characteristic-by-year-specific effects (to absorb the variation of $L_{i(n)}$ for $n = 1, 2$) where characteristics are

¹¹ Examples are Autor et al. (1998), Krusell et al. (2000), Card and Lemieux (2001), Acemoglu (2002), and Caselli and Coleman (2006), among others.

¹² For instance, in Borjas (2003, sections II–VI) or in Borjas (2006) and in the studies inspired by these seminal papers, the direct partial wage effect of immigration is the main estimated wage effect. Even the recent meta-study by Longhi et al. (2005) considers this partial effect as the relevant estimate across studies.

education and age groups, when running these regression. Using the notation defined above, the resulting partial elasticity can be written as

$$\varepsilon_i^{\text{PART}} = -\left(\frac{1}{\sigma_2} - \frac{1}{\sigma_3}\right) s_i^{N-1} \quad (10.8)$$

The term s_i^{N-1} represents the wage share of immigrants among workers within the same education–age cell as native group i . Note that the direct partial wage effect (10.8) coincides only with the last among the several terms composing the summation in (10.7), which includes both direct and indirect wage effects. This happens because, by construction, the elasticity $\varepsilon_i^{\text{PART}}$ captures only the wage effect of a change in immigrant labor supply operating through the term $\left(\frac{1}{\sigma_2} - \frac{1}{\sigma_3}\right) \ln L_{i(2)}$ in (10.5).

Hence, two important observations are in order. First, $\varepsilon_i^{\text{PART}}$ is negative whenever the elasticity of substitution between age groups, σ_2 , is smaller than the substitutability between native and foreign born in the same education–experience group, σ_3 . If those elasticities of substitution are close to each other, the partial effect can be 0 or close to 0 (a point emphasized in Peri, 2011 and discussed in Section 10.5). Second, the value and the sign of $\varepsilon_i^{\text{PART}}$ give incomplete information about the overall effect of immigrant supply changes on the wages of domestic workers of type j . Indeed, (10.8) includes only the last term of (10.7). In order to evaluate the total wage effect of immigrants on natives of type j , one has to combine the impacts generated by (10.7) across all the $i(3)$ cells that include foreign-born workers for which $L_{i(3)}$ changes due to immigration. This implies that the total wage effect of immigrants cannot be directly estimated from a regression: one can, however, estimate the elasticities $\sigma_1, \sigma_2, \sigma_3$ and combine them with the income shares in (10.7) and aggregate across all groups for which $L_{i(N)}$ changes due to immigrants.

This detailed analysis of the CES model exemplifies well the importance of recognizing the crucial role of indirect general effects (cross complementarities in this case) in order to capture the total impact of immigrants. We will consider in the next section another important indirect effect of immigrants, namely, their effect on native skill supply.

10.3.3 The area approach and the labor supply response

The nested-CES approach, described in the previous section, can be used at the national level or at the local area level. However, the tenability of the assumption of a fixed skill supply of natives, vis-à-vis changes in the supply of immigrants, is what has moved several researchers to criticize the area approach and prefer a national market one (e.g., Borjas, 1994, 2003; Borjas et al., 1997). They have argued that, while at the national level, the assumption of a rigid labor supply by native workers is tenable, at the local level, mobility of people between cities and regions would cause labor market opportunity differentials to be arbitrated away. Hence, any potential effect of immigrant skills on demand for

native skills would be matched by changes in their supply (through internal migration), leaving no effect (or much attenuated effects) on local wages and making cross area wage comparison uninformative.

This criticism is valid. It is not, however, a good reason to abandon the city and regional data that still contain rich variation of immigrant flows and of their labor market effects. First, of all the impact of an exogenous change in foreign born, skills on wages will be uninformative only if native people are perfectly mobile in the long run, and they fully undo the change in skill supply generated by immigrants. Several empirical papers show that this seems hardly to be the case as there is not strong evidence of native internal migratory response to immigrants (Card and DiNardo, 2000; Card, 2001, 2005; Peri and Sparber, 2011a,b). Also, the skill distribution of immigrants seems to affect permanently the skill distribution of a metro area, and it is not undone by differential migration of natives. For instance, as shown in Card (2009) and as reproduced in Figure 10.4 for 283 US metropolitan areas as of 2011, the percentage of high school dropouts in an MSA's labor force is strongly positively correlated with the share of immigrants. This illustrates that cities receiving a large share of immigrants are likely to be permanently affected in their relative skill composition vis-à-vis cities not receiving them. Second, even if perfect (or large) mobility takes place, we simply need to account for it. In particular, we should consider the labor supply of natives at the area level as a potential margin of adjustment to immigration and analyze the impact of immigrants on it. Within the skill cell model sketched above, if we are analyzing two regions, r and s , for each specific skill n , which

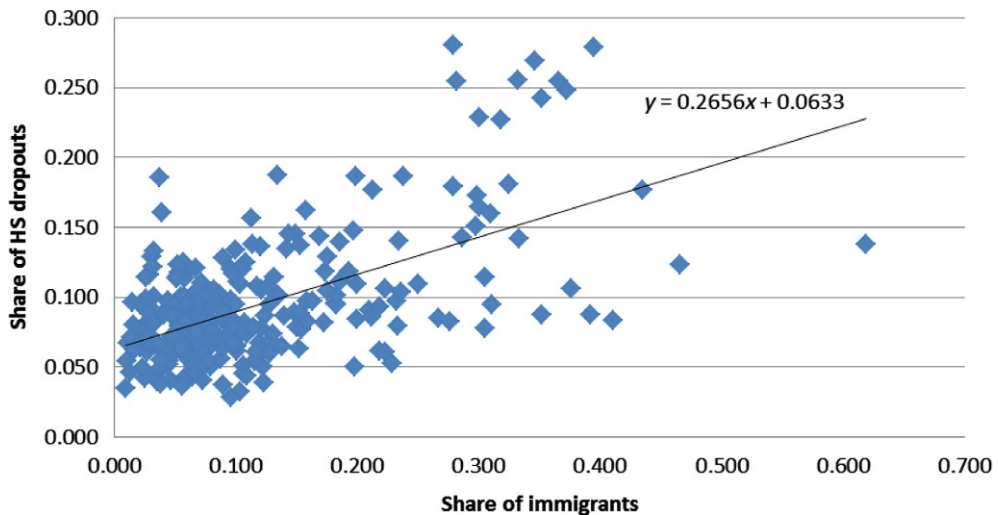


Figure 10.4 Share of immigrants and of dropouts 283 MSA (2011). *Note: Authors' calculations based on 2011 ACS data, excluding people in group quarters and including only people 18–65 who worked at least 1 week.*

we can think as an education–age group, we may model the change in native relative supply of that skill between regions r and s (driven by net migration between the two) as

$$\Delta \ln \left(\frac{L_m}{L_{sn}} \right) = \gamma_n \Delta \ln \left(\frac{w_m}{w_{sn}} \right) \text{ for } n = 1, 2, \dots, N \quad (10.9)$$

This represents a log–linear version of the relative native labor supply of skill n , between areas r and s . The parameter γ_n is the elasticity of relative labor supply capturing the response of interarea migration by native workers of skill n , to wage differentials between the two regions. Namely, an increase in the wage differential will move people toward the high-paying region and change the relative supply of skills in favor of that region. The extreme cases would be represented by $\gamma_n = 0$, when prohibitive moving cost would imply no response to wage differential (vertical labor supply) and by $\gamma_n = \infty$ that would imply wage equalization across areas.

If we observe an *exogenous inflow of immigrants* in a large number of localities (and/or periods of time) and we also observe the native employment (population) and wage change associated to it, for each skill group, we could, in principle, identify the parameters of labor demand and labor supply by solving the sets of Equations (10.2) and (10.9). Using the nested-CES production-function approach, we restrict the number of cross elasticity parameters to estimate, and we can derive log–linear demand functions as (10.3) for each skill, across regions. Then, introducing skill-specific supply would simply add one extra parameter (the elasticity) for each extra equation (log–linear labor supply as in (10.9)). Information on relative changes in native employment (population) and native wages will allow estimation of demand and supply parameters.

An important variation of the supply function described by (10.9) accounts for the fact that the migration response of natives between regions r and s may depend on real rather than nominal relative wages. In particular, the change in local housing price, which several studies have shown to be one of the consequences of immigrants (e.g., Ottaviano and Peri, 2006; Saiz, 2007), can be a separate channel in adjusting relative real wages and hence may affect the supply response of natives. In most cases, however, the housing price (or rent) effect is a common area-level effect and does not vary across skills. It can be due to changes in local amenities (as in Ottaviano and Peri, 2006) or to an upward sloping housing supply (Saiz, 2007), but as long as it affects skill group similarly, it will be absorbed by a common area effect across skills. This approach—identifying skill cell effects in the presence of supply response of natives and allowing for common housing price effects—is followed by Peri (2011), and we will describe his findings in Section 10.5. Peri et al. (2014), on the other hand, use a skill cell approach allowing for skill-supply and (one of the very few cases we know of) skill-specific housing prices to evaluate the demand/supply and productivity effect of STEM immigrants using an exogenous change across US cities due to the introduction of the H1B visa.

Card (2009) used also a skill cell analysis across US cities to show that the employment of natives does not respond much to immigration-driven changes in skill supply and that the estimated labor demand parameters (from a nested CES) are broadly consistent with those estimated at the national level (e.g., in Ottaviano and Peri, 2012). Peri (2013) used a simulation of the same nested-CES education–age skill model across US cities, with nationally estimated parameters, to show that the immigrant inflow to US metropolitan areas during the period 2000–2009 had very small effect on wages of less educated natives (and usually positive effect from complementarity with college-educated immigrants). The area analysis of the impact of immigrants, therefore, has benefited much from a more careful treatment of immigrant skill composition and skill complementarities in production. Accounting for the heterogeneity between immigrants and natives is crucial in finding these complementarity effects.

Let us also notice that some recent papers such as Smith (2012), Hunt (2012), and Jackson (2013) find that immigration may affect the probability that natives stay in school and hence their distribution across age–education cells. We will describe the details in Section 10.5.2. Here, it suffices to say that the impact of immigration, by changing relative returns to each skill cell, pushes natives to respond. They tend to move toward cells that are complements (rather than substitutes) of immigrant skills. The skill cell upgrading, which we will analyze more in detail below, implies that we should not assume fixed native supply of skills even when analyzing national markets.

A simple way to represent the supply response across skill groups of natives is to rewrite (or reinterpret) Equation (10.9) as describing the response of supply across skill groups, rather than regions, to wage differentials. In particular, omitting the regional subscripts and considering two skill cells, n_1 and n_2 , we can write

$$\Delta \ln \left(\frac{L_{n1}}{L_{n2}} \right) = \gamma \Delta \ln \left(\frac{w_{n1}}{w_{n2}} \right) \quad (10.10)$$

This relationship can be seen as a basic consequence of a Roy (1951) selection model, in which individuals, given their abilities, choose the skill group that maximizes their returns and respond to changes in relative compensation of those groups. We develop this point further in the next section, when talking about production tasks, but the main working of it is clear already within the skill cell model.

The native population may change its relative supply of skills in response to immigration. Moreover, not accounting for this margin of adjustment, we would underestimate the positive wage effect of immigrants on natives. Think, for instance, of a simple model with two skill levels, high and low, as the one we will introduce below. If an inflow of immigrants in the low-skill cell pushes natives toward the high-skill cell (because its relative compensation increases), then, at the end of the period, fewer natives are exposed to competition and more natives benefit from complementarity from immigrants (assuming the two skills are complementary). If we do not account for this and we

evaluate wage gains and losses of natives using the *initial distribution* of natives, we overestimate the competition and underestimate the complementarity effect of immigrants on natives.

10.3.4 Occupations and tasks

The CES model with three levels of nesting (education, age, and nativity) described above is an excellent framework to discuss wage effect of migrants across regions. However, it has two limitations that we address in this section. First, the “age–experience” tier adds complication and number of cells to the model especially in a cross area analysis, but it is not very relevant to understand the impact of immigrants on native wage and their distribution, as shown in [Ottaviano and Peri \(2012\)](#). This is because the largest differences in native–foreign distribution of skills are among education groups and regions rather than across age groups and also because age groups are often close substitutes for each other. Moreover, as experience in the country of origin may have a different labor market value than experience in the destination, a fine partition on this dimension may generate grouping together of rather different natives and immigrant workers. Hence, while allowing individuals of different ages to have different productivity levels, we combine age groups and eliminate the second level of the nest. Second, and more importantly, the described CES nesting assumes a difference between natives’ and immigrants’ skills, by partitioning them in different groups within the third nest. This way, we can estimate their substitutability, but we do not have a theory of why and how immigrants and natives are different in production for given observable skills. In this section, therefore, we substitute the nativity nest with difference in productivity in performing different tasks, linked to language ability, as the basis of the productive difference between foreign born and natives.

We maintain the top tier of the CES, as above, namely, a partition between two education groups, college and high school equivalents. This partition, we argued in [Section 10.5.2](#), captures fundamental skill differences in production. Within each of those two groups, we include native and foreign born of all ages so that the production function, reintroducing the region index and keeping notation consistent with before, can be written as

$$\gamma_r = A_r \left[(\theta_{CO,r} L_{CO,r})^{\frac{\sigma-1}{\sigma}} + (\theta_{HS,r} L_{HS,r})^{\frac{\sigma-1}{\sigma}} \right]^{\frac{\sigma}{\sigma-1}} \text{ for } r = 1, \dots, R \quad (10.11)$$

where the factors $L_{CO,r}$ and $L_{HS,r}$ represent employment of college equivalents and high school equivalents, respectively, and the parameters $\theta_{CO,r}$ and $\theta_{HS,r}$ represent their relative productivity and add up to one. The term A_r captures total factor productivity, and σ (>1) represents the elasticity of substitution between them. Within each of these groups, age and detailed education (as well as other attributes) may affect relative

productivity. Hence, one can use relative wages to convert workers into “high school” or “college” equivalents. Namely, $L_{HS} = \sum \theta_{HSj} L_{HSj}$ and $L_{CO} = \sum \theta_{COj} S_{COj}$, where the coefficients θ_{nj} are the relative productivity (wages) of workers in subgroup j relative to a “pure” college or high school equivalent. Notice that a two-cell model as (10.11) implies that for the United States as a whole, immigrants did not change much the cell distribution of skills, as it turns out that the immigrant composition between college-educated and noncollege-educated is similar to the native’s one.¹³ This approach argues that if immigrants affect the wage distribution within each of these two broad skill groups, this is because of different characteristics, not captured by education and age only.

A prominent difference between native and foreign born among high school equivalent is that the second group is highly concentrated in occupations characterized by the use of manual and physical abilities much more intensively than communication and interactive skills. In agriculture, construction, and personal and household services (which are sectors attracting large percentages of immigrants workers), immigrants are builders, sorters, maidens, cooks, and waiters. Rarely are they coordinators, supervisors, or salespersons. In part, this is because of their schooling (no high school degree). However, also controlling for observable characteristics, immigrants with a high school degree or less are disproportionately concentrated in manual jobs. This may be due in part to their worse language skills (Lewis, 2013b), in part to their larger tolerance for (lower disutility from) manual labor (D’Amuri and Peri, 2014).¹⁴ In any case, this type of specialization is typical of less educated immigrants in most rich countries. D’Amuri and Peri (2014) show that this holds also for Europe and it is reasonable to think that immigrants have a comparative advantage and hence specialize in manual jobs.

¹³ As of 2011, 31% of US-born workers had a college degree or more, while for foreign born, 29%. The figures are from our calculations on ACS data.

¹⁴ There are several pieces of evidence supporting that language skills drive immigrant comparative advantage. First, in the United States, Lewis (2013b) showed that the elasticity of substitution between natives and subgroups of immigrants with stronger English skills, such as those who arrived at younger ages, is larger than between natives and those with worse English skills (such as those who arrived at older ages). Second, you can find some cross-country evidence in support of this view. You cannot reject that immigrants and natives are perfect substitutes in Puerto Rico (Lewis, 2013b) or Costa Rica (Castillo et al., 2009), where both speak Spanish. Not all of the cross-country evidence is supportive. Amuedo-Dorantes and de La Rica (2011) found, if anything, larger occupational specialization among immigrants in Spain—where a large share of immigrants speak Spanish—than what Peri and Sparber (2009) found in the United States. Another outlier is Manacorda et al. (2012), who found a small elasticity of substitution between immigrants and natives in the United Kingdom. UK immigrants may have particular issues with skills transferability, however, not accounted for in the Manacorda study: while on paper, “high-skill” UK immigrants appear to compete mainly with natives at the lower end of the skill distribution (Dustmann et al., 2013). Future research might do more of these cross-country comparisons of imperfect substitutability, ideally with harmonized methods, although obtaining the wage data to do so would be challenging.

Peri and Sparber (2009) showed this fact and modeled it in a simple way that we describe here. When considering high school-educated equivalents in the two-skill partition model described above, they suggest a further partition based on manual (M) versus communication (C) skills, and they also suggest to nest those two skills within L_{HS} as a lower-level CES aggregate, as follows:

$$L_{HS} = \left[(\beta M_r)^{\frac{\sigma_{MC}-1}{\sigma_{MC}}} + ((1-\beta) C_r)^{\frac{\sigma_{MC}-1}{\sigma_{MC}}} \right]^{\frac{\sigma_{MC}}{\sigma_{MC}-1}} \quad (10.12)$$

Hence, M_r and C_r are the aggregate amount of manual and communication skills supplied in area r , β captures the relative demand for manual skills, and σ_{MC} represents the elasticity of substitution between manual and communication skills. The relative supply of skills of each individual is derived by O*NET data (from the US Bureau of Labor Statistics) that describe the type of skills used in each occupation and allow a classification between manual (strength, coordination, and manipulation) and communication (spoken and written interactions) tasks. The occupation distribution of natives and immigrants, therefore, determines their supply of manual and communication skills. In particular, the much larger employment of immigrants in manual-intensive occupations reveals that they have a comparative advantage for manual jobs, as they are less proficient than natives in their language skills. We formalize this concept in a simple way.

We assume that natives (N) have efficiency levels μ_N and ξ_N in performing manual and communication tasks, while foreign-born efficiency levels are μ_F and ξ_F . The comparative advantage of foreign born in manual tasks is expressed as $(\mu_F/\xi_F) > (\mu_N/\xi_N)$. Each individual chooses to divide one unit of labor supply (time) into l_j and $(1-l_j)$ units, performing manual and communication tasks, respectively, at a return equal to w_M and w_C per unit of manual and communication service performed. With decreasing returns in performing each type of task, we can write the labor income of individual $j(=F, N)$ as

$$w_j = (l_j)^\delta \mu_j w_M + (1-l_j)^\delta \xi_j w_C \quad (10.13)$$

In (10.13), the amount of effective units of task-service provided is a function of time spent in the task and efficiency in the task as follows: $C_j = (1-l_j)^\delta \xi_j$ and $M_j = (l_j)^\delta \mu_j$. Assuming $\delta < 1$, the income-maximizing choice of workers implies that each worker type supplies relative tasks C/M as a positive function of her relative ability (ξ_F/μ_F) and of task's relative compensation. In particular, the logarithm of relative supply of communication/manual tasks by workers of nativity $j(=N, F)$ in region r is given by

$$\ln \left(\frac{C_{jr}}{M_{jr}} \right) = \frac{1}{1-\delta} \ln \left(\frac{\xi_j}{\mu_j} \right) + \frac{\delta}{1-\delta} \ln \left(\frac{w_{Cr}}{w_{Mr}} \right) \quad (10.14)$$

Expression (10.14) shows that high school natives would have a higher communication/manual relative supply than foreign born ($C_N/M_N > C_F/M_F$) because they have comparative advantages in it ($\xi_N/\mu_N > \xi_F/\mu_F$). It is also easy to show that (10.14) implies that an inflow of immigrants in region r decreases the overall relative supply C/M and, hence, it increases the relative compensation w_{Cr}/w_{Mr} . This improves the wages of occupations using communication skills, and it decreases wages for those occupations using manual skills. Natives who are more concentrated in communication-intensive jobs will mainly benefit from the complementarity effect. Moreover, natives will move their choice of occupations further toward communication-intensive ones as their relative returns (w_{Cr}/w_{Mr}) increase and they have comparative advantages in those.

Let us emphasize that the shift in relative supply by natives obtained taking the difference before and after immigration of Equation (10.14) would be a supply response very similar to the one described in (10.10), except that native reallocation takes place across supplied tasks (manual–communication) rather than across skill groups. Equation (10.14) fully reflects the selection of skill supply of natives, in response to relative compensation, typical of the Roy (1951) model. Hence, high school-educated natives benefit in relative terms, from an inflow of high school-educated immigrants through two channels. First, as they are already more specialized in communication skills, their compensation increases. Second, in response to immigration, they supply more communication skills and those skills are complementary to (and usually better paid than) the manual ones supplied by immigrants. The stronger competition effect is instead experienced by existing immigrants, more specialized in manual-intensive occupations.

In general, the fact that high school equivalent immigrant workers filled manual-intensive jobs that are often at the bottom of the career ladder for natives implies that in locations with large inflows of immigrants, native workers move more rapidly toward communication-intensive and more complex type of jobs. Peri and Sparber (2009) showed this mechanism at work across US states in the period 1960–2000 by estimating a regression as (10.14), using the change in foreign born as exogenous shifter of relative compensation $\frac{w_C}{w_M}$, and found a significant response of native relative task supply. D’Amuri and Peri (2014) show a similar push toward more “complex” jobs, when exposed to immigrant competition for European workers. Fogel and Peri (2013) identified the same effect in response to non-EU immigrants for Danish workers. Immigrants provide the incentives and the complementary manual factors for natives to specialize in better remunerated communication-intensive job. Those individuals, therefore, move more rapidly out of manual- and physical-intensive occupations. Accounting for this change in relative skill supply of natives is crucial to measure the overall wage effects of immigrants on natives. The upward mobility generated as part of this mechanism shields native wages from competition.

A similar mechanism implying immigrant specialization in the occupation (task) spectrum and subsequent change in the relative skill supply by natives may also take place

among college-educated workers. Within that group, immigrants are particularly concentrated in occupations that are STEM-related. The international selection of highly skilled immigrants and the high transferability of mathematical–analytic skills imply that foreign-born immigrants are particularly productive in those skills. Natives, instead, are relatively more specialized in supervisory, managerial, interactive type of occupations, and they further move toward those as more immigrants arrive. [Peri and Sparber \(2011b\)](#) considered such a mechanism of specialization response of natives to college-educated immigrants. [Borjas and Doran \(forthcoming\)](#) showed a similar margin of adjustment in the much smaller field of “mathematician specialization.” In response to the large inflow of Russian mathematicians after the collapse of the Soviet Union, US mathematicians moved toward the fields of mathematics that were more complementary and less crowded by Russian mathematicians.

High-skill STEM immigrants may also have a particularly important role in innovation and technological growth, in part, because of their greater connections to the global economy ([Saxenian, 2002b](#)).¹⁵ That foreign skilled workers are particularly concentrated in STEM jobs and contribute substantially to patented innovation was shown by [Hunt and Gauthier-Loiselle \(2010\)](#) and [Kerr and Lincoln \(2010\)](#). [Peri et al. \(2014\)](#) look directly at the productive effects of STEM immigrants and find that they generated a positive contribution, localized at the metropolitan area level, to the productivity of college-educated natives. First, they show that in metropolitan areas with large inflow of foreign STEM workers (determined by the change in aggregate H1B visa entry, the main channel of entry of highly educated immigrants), the wage and employment rate of native college-educated workers were substantially higher than in metropolitan areas with small inflows. The use of instrumental variables based on aggregate change in visa and preexisting localization of foreign STEM workers shows that the correlation can be causal. Then, they show that such an increase is only compatible with a significant increase in A_H , the specific productivity of the college-educated equivalents. They also show that A_L , the productivity of high school equivalents, increased as well but not as much as A_H . They emphasize, therefore, that the productivity improvements introduced by STEM (foreign) workers appear to be “skill-biased” in that they increase the productivity of college-educated equivalents more than that of high school equivalents.¹⁶

The possibility that immigrants may affect productivity through their contributions to technology and science (or through other channels) opens new and very important

¹⁵ [Saxenian \(1994\)](#) had also written about the fact that, in addition to a deep concentration of high-skill workers (including immigrants), institutions and culture that support idea sharing may also be a necessary input into innovation and successful entrepreneurship.

¹⁶ [Table 10.2](#) reports the magnitudes of the effects that they estimate per percentage-point increase of STEM worker employment share.

potential channels through which immigration affects production. Immigration may not only change the supply of foreign skills, inducing a response in the supply of native skills, but also change their productivity A_j by affecting technology, techniques, or efficiency. If this is the case, the overall surplus of immigration can be much larger than calculated before. In particular, increasing the share of STEM or college-educated workers may have positive local externalities on the region. Papers by [Ciccone and Peri \(2006\)](#) and [Moretti \(2004a,b\)](#) have emphasized the importance of productive externalities in US cities from increasing the share of college-educated workers. Those effects go beyond the complementarity effects analyzed above. In particular, local learning and the diffusion of better ideas and of better technologies can be affected by the concentration of college-educated workers, many of whom are foreign born. At least, one paper ([Iranzo and Peri, 2009](#)) has directly connected the higher share of college-educated workers in US cities, to a higher share of some immigrants (Indian and Europeans) showing their positive impact on productivity. Another recent paper ([Docquier et al., 2011](#)) also emphasizes the potential importance of human capital externalities, driven by immigration in OECD countries, in positively affecting wages. That paper considers OECD countries as units, and it emphasizes the fact that immigration to those countries was prevalently college-biased during the 1990s and 2000s. Adopting a two-tier nested-CES model similar to the one described in [Section 10.5.2](#), but allowing for externalities due to increased share of college-educated workers, that study simulates positive wage effect of immigrants in most countries.

Let us mention, in closing of this section, another way used in a recent paper by [Dustmann et al. \(2013\)](#) of organizing skill cells in a CES model. That study considers workers as belonging to the same skill group if their productivity (wage) is similar. Hence, skills groups are interdecile intervals of the wage distribution. While this requires strong assumptions (such as a unidimensional representation of skills), it proposes to analyze the more direct competition effects of immigrants on natives at a similar level of wage. The paper finds evidence of stronger competition of immigrants with natives of similar wages. It also finds a strong positive effect of immigrants in raising the average native wage, possibly indicating an aggregate productivity (or strong aggregate complementarity) effect. We will focus on potential productivity effects in the next section.

10.3.5 The margin of technological choice

The simple production model of high school and college equivalents, illustrated in Equation (10.11), had widespread diffusion in the literature. One reason for its success is that, when combined with skill-based technological progress, it explains parsimoniously and reasonably well the evolution of relative college-high school wages (e.g., [Katz and Murphy, 1992](#)) in the United States during the last four decades. Even more interestingly, [Acemoglu \(1998, 2002\)](#) had argued that the type of technology adopted in a market

depends on the relative supply of skills. When one type of skill (say college graduates) becomes more abundant, technologies that increase its productivity (skill-complementary or skill-biased) become more profitable and hence more frequently adopted. The rise in college-educated workers in the United States during the last 40 years can explain, therefore, the adoption of skill-biased technologies that in the long run have increased the productivity of college-educated workers and even increased their wages relative to high school graduates. Against this long-run increase in college education and adoption of skill-biased technology, the fluctuations of relative college–high school supply have then determined the shorter-run change in relative wages.

Notice a very important implication of directed technological change. For a given change in relative supply of skills, the adoption of directed technology, increasing the productivity of the factor whose supply increases more, will attenuate the effect on wages, relative to the case with unchanged technology. This is seen very simply by deriving the relative compensation of skills from (10.11):

$$\frac{w_{CO}}{w_{HS}} = \left(\frac{\sigma - 1}{\sigma} \right) \frac{A_{CO}}{A_{HS}} - \left(\frac{1}{\sigma} \right) \frac{L_{CO}}{L_{HS}} \quad (10.15)$$

An increase in the relative supply $\frac{L_{CO}}{L_{HS}}$ would reduce the college–high school wage ratio, everything else constant, as long as the two factors are imperfect substitutes. However, if the relative productivity $\frac{A_{CO}}{A_{HS}}$ is also positively affected by the relative skill supply, this negative wage effect can be attenuated or even reversed.

Lewis (2011) tested the idea of directed technological adoption at the local (metropolitan area) level when the shift in relative skills is caused by immigration. In particular, in MSAs with a large inflow of less educated immigrants, Lewis (2011) found that firms have fewer economic incentives to adopt techniques that substitute for manual labor (such as automation) and complement human capital relative to metropolitan areas with small inflows of immigrants. Reducing automation and maintaining, instead, techniques that use more efficiently manual and less skilled workers will produce an increase in the relative efficiency A_{HS}/A_{CO} and hence could attenuate or eliminate the relative wage effects of an increase in L_{HS}/L_{CO} . Lewis (2011) was able to identify the adoption of specific mechanization and automation procedures by firms in different metropolitan areas and relate that to the inflow of less educated immigrants. As predicted by the directed technological adoption framework, the study finds that mechanization and automation are faster in metropolitan areas with low immigration and hence lower supply of high school equivalents.

While Lewis (2011) provided evidence on directed technological adoption using microdata, Peri (2012) estimated in a panel of US states the effect of immigration on $\frac{A_{CO}}{A_{HS}}$ within an aggregate production function model, which, per (10.15), means imposing a value of σ (the college–noncollege elasticity of substitution) established by the literature. The study finds strong evidence of a negative correlation between the inflow of

immigrants and the change in $\frac{A_{co}}{A_{HS}}$, which is consistent with the theory of directed technological change. More details of the empirical approach of this study are in [Section 10.5.2](#).

Overall, the recent literature has emphasized several margins of response to immigration taking place within area economies. A change in relative skills caused by immigrants induces a change in relative skill supply of natives (also characterized as specialization, occupational upgrading, and possibly education upgrading of natives). It may also induce a change in technologies/techniques adopted, which results in a change in relative productivity. Both responses reduce the “competition effect” of immigrants on similarly skilled natives. And they both increase the surplus received by natives as a consequence of immigration. Hence, they can help to explain a smaller negative wage effect of immigrants on wages of comparable natives than predicted by the “partial” effect (everything else equal). We mentioned that the other margins, such as native migration across areas and change in variety composition of output, have been investigated but do not seem to play a major role. While no study accounts explicitly for each margin of adjustment, we will overview the empirical findings on each of them, and we will assess what their combined effect implies on wage and productivity of natives.

10.3.6 Scale externalities, heterogeneity, and search

In the models considered so far, the increase in scale of the local economy due to inflow of immigrants plays no role in productivity and hence wages (because of constant return to scale). The simple increase in density of economic activity due to immigrants (representing simply an increase in the number of worker) and to the fact that they tend to concentrate in cities may have beneficial productivity effects (e.g., [Ciccone and Hall, 1996](#); [Greenstone et al., 2010](#)). However, we review here some potential local externalities, from density, that are more specific to immigration. Some recent papers, based on the trade and growth literature, have considered an even finer differentiation of skills between immigrant and natives and among immigrants themselves. Allowing for each country of birth to represent a different skill group (producing a differentiated nontradable intermediate) and combining them in a CES for final production these studies have derived that an index of diversity (fractionalization) of immigrants in the area is positively related with local aggregate productivity (and hence average wages and possibly average rents). In this framework, more immigrants from more countries have a direct positive productivity effect. [Ottaviano and Peri \(2006\)](#) estimated such a model for US metro areas. [Alesina et al. \(2013\)](#) estimated the effect of such country of birth diversity index across countries in the world. [Trax et al. \(2013\)](#) explored this relation at the firm/plant level. Within the production-function context and treating each nationality as a different skill, a more differentiated and larger population of immigrants has positive productivity effects at the local level as it increases the variety of intermediates. There is significant

evidence of this positive effect in the mentioned papers, and place of birth is a potentially important dimension of differentiation.¹⁷ However, this research is still at its early stage, and we need a better understanding of the channels through which and the level (firm, area, and sector) at which this “place of birth diversity” effect operates. Certainly, an approach that considers in greater detail the skill differences between immigrants and natives and looks at the microlevel (firm and plant) to identify these effects can shed light on important margins.

Most of the research on the effect of immigrants has focused on their skill differences with natives and has used a competitive approach to labor market equating marginal productivity to wages. This is reasonable in the long run. An interesting new line of research considers instead frictions in labor markets and employer–employee matching that create match-specific surplus. In particular, following the huge success of search models pioneered by [Mortensen and Pissarides \(1994\)](#) to analyze important labor market features, some very recent studies have incorporated immigrants in search and matching models. In this search context, differences between natives and immigrants on the labor supply side (especially in their bargaining power and their outside options) may generate surplus that native workers appropriate. In particular, [Chassamboulli and Palivos \(2014\)](#) show that when immigrants have a worse outside option and/or a lower bargaining power than natives, they will be paid less in equilibrium, even if they have identical productivity as natives. This, in turn, implies that in a market with a larger percentage of immigrants, firms will earn larger average surplus per vacancy filled. Hence, they will create and post more vacancies. If they cannot discriminate natives out of those vacancies, but they can pay different wages, some of these jobs will be filled by natives, and this will increase their employment and also increase their wages (by making their bargaining position stronger).

While the model is somewhat specific, it emphasizes a very general idea. If firms save on their costs by hiring immigrants (who are paid less due to their lower bargaining power), they will appropriate a larger surplus and they will be willing to create more jobs and expand because of that. The new jobs created as a consequence of this profit-seeking expansion will also benefit natives. Hence, differences on the supply side may also result in a “complementarity” of a different kind between natives and immigrants and imply beneficial effects due to stronger job creation when more immigrants are in the labor market. In a similar vein, [Chassamboulli and Peri \(2014\)](#) analyze the effect of different policies aimed at reducing undocumented immigrants, in a search and matching model of the labor market and apply it to the US–Mexico case. This flexible frame allows them to characterize documented and undocumented immigrants as having different outside options and hence to study their job creation effect on the US economy. It also allows

¹⁷ Large part of the gains from trade is predicated on similar type of gains from varieties. See, for instance, [Broda and Weinstein \(2006\)](#).

them to internalize the decision to migrate (from the United States to Mexico) as a search decision and hence analyzes the effect of policies on the incentives to migrate.

Having defined the main frameworks and a unifying approach used in the literature to analyze the economic consequences of immigrants in local economies, we now focus on the empirical implementation and on the issues related to identifying an exogenous change in immigrants and their skill supply.

10.4. EMPIRICAL APPROACHES TO IDENTIFY CAUSAL EFFECTS ON LOCAL ECONOMIES

As summarized in the previous section, the skill cell approach implies that immigration may affect the absolute and relative productivity of skills. However, differential productivity growth (possibly skill-specific) may also attract and select immigrants. The economic conditions in receiving countries are a major motivation for migration (e.g., [Clark et al., 2007](#); [Mayda, 2010](#)), and, at least in the United States, evidence suggests that immigrants can be very responsive to different conditions across labor markets ([Borjas, 2001](#); [Cadena and Kovak, 2013](#)). This potential endogeneity problem has been addressed with two related strategies: (1) flows, often of refugees, arising from shocks in sending countries (and often flowing to a small number of destinations) and (2) “shift-share”-type instruments that largely take advantage of the autocorrelation in the regional distribution of immigrant flows by origin. Both approaches were pioneered by [Card \(1990, 2001\)](#). A useful recent development is the linking of the “shift-share” instrument to variation driven by actual immigration policies.

The first, “natural experiment” approach was pioneered in [Card’s \(1990\)](#) study of the Mariel Boatlift, the influx of about 125,000 Cuban refugees in 1980, on Miami, the largest location in which they settled. There have been several studies since that time using this approach, including Jennifer Hunt’s study of repatriates from Algeria to France (1992) and [Carrington and Lima’s \(1996\)](#) study of African repatriates to Portugal. Other studies have examined the impact of the refugee flows from the breakup of Yugoslavia ([Angrist and Kugler, 2003](#)), among other refugee flows (e.g., [Foged and Peri, 2013](#)), and flows that are the result of natural disasters ([Kugler and Yuskel, 2008](#)). These studies largely involve immigrants that flow into the lower end of the labor market. Finding similar such events that generate very high-skill immigrant flows is more difficult, but not impossible. Several studies have investigated the impact of the Former Soviet Union (FSU) immigrants to Israel including [Friedberg \(2001\)](#), [Lach \(2007\)](#), and [Paserman \(2013\)](#).¹⁸ The impact of the dismissal of Jewish scientists from Nazi Germany has also received attention ([Waldinger, 2012](#); [Moser et al., 2013](#)).

¹⁸ [Borjas and Doran \(2012\)](#) studied the impact of the FSU flows on the academic labor market for mathematicians in the United States.

This approach has at least two challenges. The first challenge is that while a refugee crisis may generate exogenous emigration, the location and occupations taken by refugees may very well be endogenous to economic opportunities in the receiving country. Hence, defining a credible comparison group may be challenging. This is not a trivial problem and not all of the comparisons above necessarily get to credible causal inference.¹⁹ A second challenge for some of these studies, raised in [Donald and Lang \(2007\)](#), is obtaining proper and meaningful inference with a small number of “treatment” cells.²⁰ Both challenges may have become a bit easier with the spread of [Abadie et al.’s \(2010\)](#) synthetic control technique, which uses data-intensive techniques to construct a matched comparison group and also allows inference by placing the estimates in a distribution of similar structured “placebo” regressions. However, to date, this approach has seen little use in the immigration literature.²¹

An additional concern with this approach is about external validity: studying a group of immigrants that are narrow in type, often to a narrow set of destinations, raises concerns that the results may not generalize to more common types of immigrant flows. Most high-immigration countries have received significant (but not catastrophic) flows for a decade or more, and we are more interested in the effects of those.

A second approach, beginning with [Card \(2001\)](#), parallels [Bartik style \(1991\)](#) of widely used instrument for demand shocks. It essentially predicts flows of immigration in region r based on the lagged locations of similar immigrants.²² The basic structure of the key “predicted immigration flows” component of the instrument is typically of the form

$$\hat{F}_r = \sum_c (\text{Lag\_sh}_c^r \times F_c)$$

where $\text{Lag\_sh}_c^r = M_c^r / M_c$ is the share of the stock of immigrants, M_c , from source country “ c ” living in destination area “ r ” (usually a within-country region or metropolitan area) at

¹⁹ For example, many of the studies that take advantage of the breakup of the FSU use the variation in location of FSU immigrants within Israel across firms ([Paserman, 2013](#)) or cities ([Lach, 2007](#)). This is likely to be endogenous, and so, it is not clear that these analyses are more credible than OLS. To address this, [Friedberg \(2001\)](#), who used variation across occupations, used occupation prior to arrival as an instrument.

²⁰ This point is driven home by [Angrist and Krueger’s \(1999\)](#) examination of what they call the 1994 “Mariel Boatlift that did not happen” as the Clinton administration blockaded what appeared to be shaping up to be flotilla from Cuba similar in size to the 1980 boatlift. This analysis found a marginal significant positive effect of the nonevent on unemployment rate of blacks (using standard inference techniques that Donald and Lang argue give incorrectly sized tests). There is also the question of whether meaningful inference is even possible with such a small number of cells: Donald and Lang’s reestimation of the confidence intervals in [Card \(1990\)](#) suggests that they are uninformative. Even Donald and Lang’s approach is made under restrictive assumptions; [Conley and Taber \(2011\)](#) provided a more general procedure for inference in panel data with a small number of treatment cells, which tends to produce even wider confidence intervals.

²¹ One example is [Bohn et al. \(2014\)](#).

²² The idea of using lagged immigrant shares as instrument for current immigration is older, going back to [Altonji and Card \(1991\)](#), but this “shift-share” formulation began with [Card \(2001\)](#).

some point prior to the period of analysis (often the initial period of the analysis), and F_c is the aggregate flows from country c during the periods of study. Note that the formula above can be specialized to skill cells. In that case, F_c would not simply be the total count of immigrants from country c , but instead, the flow of immigrants in a specific skill cell. Often, the constructed variable $\hat{F}_r$ is called the “imputed” inflow of immigrants in area r (and skill cell i , if it is skill-specific). $\hat{F}_r$ is also almost always normalized by some measure of the size of the local economy (or of the cell), such as initial employment so that the final instrument, for areas r and skill cell i , is $Z_{ri} = \hat{F}_{ri} / \text{Emp}_{ri}$ (where Emp_{ri} is initial employment in that area cell).²³

The basic argument for this instrument’s validity begins with the idea that the aggregate component of the instrument is not driven by demand conditions in the destination regions, but rather by conditions in the sending country and possibly by aggregate conditions in the destination country; hence, this is sometimes referred to as the “supply/push” instrument (e.g., Card, 2001). These aggregate flows are apportioned to destination regions by Lag_sh_c^r , the historical destinations of the same immigrant groups. Mechanically, the strength of the instrument derives from the tendency of new immigrants to choose destinations with existing concentrations of the same ethnicity (Bartel, 1989). The argument for validity rests on the idea that this is driven, rather than by autocorrelation in demand for the labor of that group in a particular location, by labor supply factors like family ties or a preference to settle in a culturally or linguistically familiar environment.²⁴

²³ A common approach is to normalize the predicted immigration variable in a way that mirrors the endogenous variable of interest. For example, the endogenous variable in both Card (2009) and Lewis (2011) is the ratio of two different education levels of workers, and so, the predicted immigration expression is calculated separately for high and low education levels (with the same Lag_sh_c^r but separate F_c s for high- and low-skill aggregate flows by country), and the final instrument is structured as the ratio of the two in the same way it is in the endogenous variable. Peri et al.’s (2014) endogenous variable is high-skill workers/current employment; the numerator is constructed with predicted immigration, and the denominator is constructed by inflating base year employment by national employment growth. Smith’s (2012) independent variable is the growth in the number of immigrants, so he normalizes by a number that imposes that all markets start off at the same fraction immigrant and then grow at the national rate. He also considered alternative versions of the instrument construction, including one that considers the mix of ethnic groups within markets rather than their distribution across markets. The manner in which the predicted immigration variable is normalized may affect the strength of the instrument or the plausibility of the exclusion restriction.

²⁴ One way to think of this is that an ethnic cluster is a group-specific amenity, an assumption supported with wage and house price data in Gonzalez (1998) for Mexicans in the United States. Additional evidence supporting validity is that a version of this instrument used to predict Mexican flows to particular metro areas in the United States appears uncorrelated with forecasts of employment growth (Card and Lewis, 2007). Lafortune and Tessada (2013) compared the ability of the size of ethnic enclaves and the local occupation mix to predict where immigrants will settle in historical data and find evidence that ethnic networks dominate in most cases.

Note that this approach is not completely independent of the natural experiment approach, which also tends to rely on the historical settlement locations of the immigrant groups under study (when the analysis is regional in nature) such as Cubans to Miami. Indeed, one concern with this instrument is that when destinations and sources are tightly linked—such as Cubans to Miami or Algerians to France—the aggregate flows might be partly driven by regional demand conditions at the destination. Demand conditions in LA and Chicago, for instance, might significantly affect Mexican arrivals to the United States.

One way researchers have attempted to refine this instrument is to try to uncover more exogenous sources of variation in the aggregate component of the instrument. [Pugatch and Yang \(2011\)](#) used rainfall shocks in Mexico as a component of an instrument for Mexican flows to the United States. In their study of the impact of immigration in Malaysia, [Ozden and Wagner \(2013\)](#) used the age structure of the population in the Philippines and Indonesia, two major sending countries, in the instrument, with the idea that migration tends to be concentrated in relatively young age groups.

A very promising recent development is the use of policy variation for the aggregate component of the instrument. In addition to being potentially helpful for addressing the endogeneity problem, this approach makes the estimates more policy-relevant. A number of studies have directly or indirectly used variation in the sometimes restrictive cap on the number of H1-B visas (a high-skill visa issued in the United States) including [Peri et al. \(2014\)](#) and [Kerr and Lincoln \(2010\)](#).²⁵ [Kato and Sparber \(2013\)](#) examined the aggregate impact of the large drop in the H1-B visa cap after 2003 on the quality of foreign students coming to the United States, using countries that have other visa options as a comparison group.

Much less has been done to make the “shares” (Lag_sh_i^t) part of the instrument more credibly exogenous beyond using longer time lags.²⁶ But, again, policy-driven variation has begun to be used: some studies have used the so-called dispersal policies, used in some European countries to give initial, often random, placement of immigrants through public housing. [Glitz \(2012\)](#) used this to study the labor market impact of immigration in Germany. Another example is [Damm \(2009\)](#), who used dispersal policies to study the impact of living an “ethnic enclave” in Denmark at the individual level. Although it is not used directly in their analysis, much of the variation in [Foged and Peri \(2013\)](#)—who studied the effect of non-EU and refugee immigrant share on native wages

²⁵ [Kerr and Lincoln \(2010\)](#) did not use the “country” (c) element of the instrument described above—they aggregate together all immigrant groups—and estimated a reduced form regression. The H1-B visa cap is not country-specific, but aggregate.

²⁶ Part of [Pugatch and Yang’s \(2011\)](#) variation derives from the historical ties of three major border crossing areas in Mexico to destination markets in the United States via historic railroad routes. Another aspect of the instrument that has not received much attention is the level of detail with which the immigrant groups are constructed, which for practical (small cell size) reasons are often not individual countries but groups of similar countries.

and occupation using variation across Danish municipalities—likely comes from the effect of these dispersal policies.

In the absence of definitive *a priori* grounds for lagged origin shares being random, another approach has been to subject the instrument to various “falsification” tests. One that is often available is to ask if trends in outcomes or treatment are correlated with the instrument prior to the analyzed period. The challenge is that often many of the same patterns of immigration tend to hold prior to the analyzed period, albeit at a lower magnitude. After all, the instrument in fact *exploits* trends in the regional patterns of immigration, so finding zero correlation in a just prior period is not necessarily realistic, but certainly, the correlation should be much weaker for the instrument to be credible. It also helps to be studying a period that begins with a sharp break from prior patterns of immigration, such as Foged and Peri (2013), who examined the impact of immigration from the genesis of large refugee flows to Denmark, in the 1990s.²⁷ Although harder to come by, when available, examining contemporaneous outcomes that arguably should be unaffected by the treatment is helpful.²⁸ Although doubts about this approach will likely never totally go away, in the absence of a new approach that captures enough of the variation in immigration to precisely estimate its economy-wide impacts, it seems like that the shift-share instrument will remain a major analysis tool in the near term.

10.5. ESTIMATES OF NATIVE RESPONSES AND EFFECTS ON OUTCOMES

The considerable variation in immigrant density across cells defined by geography or skill categories, as was described in Table 10.1, is potentially useful in estimating immigration’s impact on native outcomes. However, as we illustrated in Section 10.3, a challenge in both identifying and understanding the impact of immigration using this variation is that natives may move across these cells in response. This section considers more in detail the empirical estimates of how natives respond to immigration by moving across both geographic locations and skill cells. With that understood, it then turns to organizing the estimates of the impact of immigration on wages, firm productivity, technology choice, and externalities.

²⁷ The dispersal policy that motivates their approach was in place before their period of study, but Foged and Peri (2013) showed that there was little non-EU immigration to Denmark prior large refugee events in the 1990s (Yugoslavia and Somalia). In their study of native school district choice response to Mexican school-aged arrivals, Cascio and Lewis (2012) examined changes from 1970, which is near the beginning of the large influx of Mexican migrants to California, although Mexican immigration had been rising more slowly for decades before that. Their pretrend test examined changes in the 1960s.

²⁸ For example, Lewis (2011) examined whether plans for technology adoption (asked in the baseline period) were correlated with later immigration-driven skill mix shocks and found that they were not.

10.5.1 Geographic mobility of natives

The earliest papers that attempted to estimate the impact of immigration, such as Grossman (1982), used variation across geographic space. However, it was not long before the mobility response of natives was raised as a potential confounder of these estimates. Borjas' often-cited review article (Borjas, 1994) argued that estimates of the impact of immigration across space had "no structural interpretation" (p. 1699). In fact, as Equation (10.9) described, a careful approach that allows for native supply mobility response can uncover structural demand parameters, by estimating together wage and employment regressions. Since the time of that review, a considerable literature on the native mobility response to immigration has developed.

Using the "Bartik style" of instrument described in Section 10.4, Card (2001) and Card and DiNardo (2000) found no evidence of any native mobility response across US metropolitan areas to immigrant inflows within broad occupation classes.²⁹ Borjas (2006) countered with evidence of a native mobility response to immigration, which, sensibly, is increasing in going from large (census divisions) to small (metropolitan areas) geographic units: he found that for every 10 immigrant inflows to a metropolitan area, 6 natives move out.³⁰ This study differs from the other two in examining responses within relatively narrow (education $\times$ experience) skill cells, in using the observed inflow of immigrants as explanatory variable in an OLS, rather than an IV, approach and in how the mobility response relationship is specified.

Indeed, significant concerns about the specification used in Borjas (2006) are raised in Peri and Sparber (2011a). They argued that because the stock of natives appears both as the dependent variable (native population) and in the denominator of the right-hand-side explanatory variable (immigrants as share of population including natives) in Borjas' specification, estimates may be biased toward finding a negative relationship. Using simulations designed to match aggregate population moments, they confirm that Borjas' approach would tend to find strong evidence of native mobility response even in its absence. Specifically, the authors assume that the change in native population in region (r) $\times$ skill (j) cell, ΔN_{rj} , is linked to foreign arrivals in the cell, ΔF_{rj} , via the structural equation:

$$\Delta N_{rj} = \alpha + \beta \Delta F_{rj} + \varepsilon_{rj}$$

They took random normal draws of ΔF_{rj} and ε_{rj} (and chose α) to match the aggregate mean and standard deviation of the observed data on ΔF_{rj} and ΔN_{rj} across state $\times$ education $\times$ experience cells in the US decennial census since 1960, under varying

²⁹ Card and Lewis (2007) found a similar lack of native mobility response specifically in response to Mexican inflows.

³⁰ An older demographic literature also claims to find evidence of "native flight" (e.g., Filer, 1992; Frey, 1995). However, this literature generally does not specify the relationships in a way that is consistent with the recent understanding of how immigration impacts the labor market (described in Section 10.3). In particular, it fails to differentiate by skill category and takes the observed choices of immigrant location as random.

assumptions about β , including $\beta = -1$ (full displacement or “crowding out”), $\beta = 0$ (no displacement) and values up to $\beta = 1$ (“crowding in”). To restate their findings, regardless of the value of β chosen, Borjas’ specification (log of native employment on the share of immigrants in employment) finds “evidence” of displacement.³¹ Through additional simulations, they also find that the bias in Borjas’ specification worsens in the relative variance of ΔN_{jc} compared with ΔF_{jc} , and with the number of cells in the regression, which may help account for how Borjas’ estimates vary with geographic scale.

So, what is the specification producing the most reliable estimate of the magnitude of native displacement? Peri and Sparber (2011a) found that the specification used in Card (2007) performs best. In that study, the author regresses $\frac{\Delta L_{tj}}{L_{tj0}}$ on $\frac{\Delta F_{tj}}{L_{tj0}}$, where ΔL_{tj} is the change in and L_{tj0} is the initial size of the labor force (immigrants + natives). Notably, this is similar to specifications used in Card (2001) and Card and DiNardo (2000), and all three studies found little sign of native displacement. In this specification, a coefficient of 1 indicates no displacement: immigration has a one-for-one impact on total skill supply. The authors also point out that, equivalently, one could regress $\frac{\Delta N_{tj}}{L_{tj0}}$ on $\frac{\Delta F_{tj}}{L_{tj0}}$, in which case the coefficient would directly indicate the level of displacement, with a coefficient of 0 implying no displacement.

Peri (2011) further advanced the literature by characterizing the mobility response in terms of the underlying parameters of an aggregate production function. Specifically, it mirrors the CES structure developed and supported in Ottaviano and Peri (2012) and discussed in Section 10.3.2. In such a setup, using Equation (10.5) and taking total differentials with respect to changes in employment of natives and immigrants, one can show that wage growth of native skill group j in local economy r will be

$$\frac{\Delta w_{tj}}{w_{tj0}} = \phi_r + \phi_{r, \text{edu}} + \left(\frac{1}{\sigma_2} - \frac{1}{\sigma_3} \right) \frac{\Delta F_{tj}}{L_{tj0}} + \left(\frac{1}{\sigma_2} - \frac{1}{\sigma_3} + \frac{1}{\bar{x}\sigma_3} \right) \frac{\Delta N_{tj}}{L_{tj0}} - \Delta \ln \theta_{tj} \quad (10.16)$$

where $\bar{x}$ is the native wage bill share; σ_2 and σ_3 are, consistent with the notation of Section 10.3.2, the cross experience cell and immigrant–native elasticities of substitution, respectively; and the rest captures education-specific aggregators and unobserved productivity terms.

The key assumption motivating Peri’s (2011) approach is perfect mobility: Natives move to equilibrate wages across labor markets, so $\frac{\Delta w_{tj}}{w_{tj0}}$ is the same in all markets. So Peri (2011) took differences of (10.16) between market r and the rest of the country and denoted with a tilde “ $\sim$ ” above a variable such a difference. Then, exploiting full mobility and wage equalization, he set the differenced equation equal to zero and solved for $\frac{\widetilde{\Delta N_{tj}}}{L_{tj0}}$ as a function of $\frac{\widetilde{\Delta F_{tj}}}{L_{tj0}}$. Recall that this is the displacement specification recommended by Peri and Sparber (2011a,b) simulations, but now, the coefficient of that

³¹ Borjas also includes an alternative specification meant to reduce this bias, but the authors show that this specification is also strongly biased toward finding evidence of native displacement: again, in all of their simulations, it always finds displacement.

regression is characterized structurally as proportional to $(\sigma_3 - \sigma_2)$, the difference between the elasticity of substitution between native and immigrants and the cross experience elasticity. Comparing California to the rest of the country and using national net arrivals of Central Americans as an instrument for $\frac{\Delta F_{ij}}{L_{j0}}$, Peri found little evidence of native mobility response within these skill cells since 1960 (in either OLS or IV), despite the massive and constant inflows of immigrants to California over this period. This can thus be rationalized in the model by $\sigma_2 \approx \sigma_3$. What about his perfect mobility assumption? This is not rejected: Peri (2011) also looks directly at wages and finds no response, which per (10.16) again could be explained by $\sigma_2 \approx \sigma_3$.³² In plainer terms, the substitution/complementarity structure is such that immigration ends up having little impact on natives' wages within skill cells, providing them no incentive to move out of California in response to immigration. Thus, it simultaneously rationalizes the "zero mobility" finding and explains why immigration appears to have little impact on the wage structure.

A reasonable summary of this literature, then, is that the native mobility response to immigration across geographic space is quite small, an outcome that may be incentive-compatible as the equilibrium effect of immigrants on native wages appears to be small. Nevertheless, the best approach in geographically based studies of immigration's impact is, per (10.9), to allow for the possibility of native mobility response in order to give estimates the proper structural interpretation.

Incidentally, one specification that helps simplify this problem is (when appropriate) to use as the right-hand-side variable the total supply in the relevant skill cells, L_{ij} or a transform of it, and then instrument total changes in skill supply with immigration (or immigration instrument). Put differently, the right-hand-side variable is immigrants + natives, rather than the old approach of treating the size of the immigrant stock itself as the "treatment" variable. In this way, the first stage (regressing labor supply on immigrants by skill across regions) implicitly adjusts for any skill-specific native mobility response and is also a direct indicator of the level of displacement. This approach is, in fact, standard in area studies since Card (2001).³³

³² A reasonable question to ask is whether this is consistent with the direct estimates of these parameters in the national data, such as those in Ottaviano and Peri (2012). Peri (2011) asked this as well. Estimates of σ_2 tend to be around 10, while estimates of σ_3 tend to be larger, around 20. Peri argued, however, that existing direct estimates of $1/\sigma_3$ may be biased downward in national regressions by endogeneity. One caveat on this interpretation, which Peri pointed out, is that the regional wage impacts may be small for some other reason, such as some remaining endogeneity (that is, despite having instruments) or adjustments of nonlabor inputs (see Lewis, 2013a).

³³ A nice improvement going forward would be the use of an estimate of the within-skill category imperfect substitutability between immigrants and natives, σ_3 , to generate a more general skill group supply measure that allows for such imperfect substitutability, that is, to use

$$\left([(1 - \theta_{j, \text{IMM}})N_{ij}]^{\frac{\sigma_3 - 1}{\sigma_3}} + [\theta_{j, \text{IMM}}F_{ij}]^{\frac{\sigma_3 - 1}{\sigma_3}} \right)^{\frac{\sigma_3}{\sigma_3 - 1}} \text{ in place of } L_{ij} = N_{ij} + F_{ij}.$$

10.5.2 Mobility across skill categories

An alternative approach to using variation across geographic space to identify the impact of immigration is to use variation across skill cells over time. As described in [Section 10.3.2](#), this approach was pioneered by [Borjas \(2003\)](#), who described it as a way around the problems of endogeneity of immigrant flows to particular areas and native mobility response. However, depending on one's definition of skill categories, this approach may not be totally immune to native mobility. Indeed, in some cases, such as across occupation, the native mobility response is interesting *per se* in revealing elements of the structure of the labor market. In addition, as was emphasized earlier, estimates that fail to take account of native mobility may lead to an understatement of native gains from immigration.

Native mobility response to immigration across occupations has probably received the most empirical attention. [Peri and Sparber \(2009\)](#) were the first to show that native comparative advantage played a role in this response. Specifically, they test the hypothesis that, among less skilled workers, natives have a comparative advantage in “communication” task-intensive jobs (e.g., sales) and immigrants have a comparative advantage in manual task-intensive jobs (e.g., construction). If so, their model (see Equations 10.12–10.14) has three predictions that they evaluate: (1) Immigration induces natives to shift to more communication-intensive jobs; (2) immigration induces the overall shift to more manual-intensive jobs; and therefore, (3) it raises the relative price of communication tasks. As was mentioned above, they measure task intensity with occupation-level data from O*NET merged to census occupations. To measure the “price” of tasks, they run state time-specific regressions of adjusted occupation-level wages on the average manual and communication intensity of each occupation.³⁴ They run their analysis across US states over time using 1960–2000 decennial censuses, using both “imputed Mexican immigration” (see [Section 10.4](#)) and time-varying functions of the distance to the Mexican border as instruments. They find support for all three predictions, and from the third, they also obtain estimates of the elasticity of substitution between tasks, σ_{MC} in Equation (10.12), which ranges between 0.6 and 1.4. Through simulations, they show that this level of substitutability is of the right order to account for the magnitude of the directly estimated elasticity of substitution between immigrants and natives, σ_3 (per [Ottaviano and Peri, 2012](#)).

³⁴ As is common, they combine several similarly themed O*NET measures into a “percentile” index giving the share of employment with a value at that level or lower. It is worth highlighting at least one unavoidable measurement issue: the occupation-level averages in O*NET are not time-varying, but measured at a single point in time (2000). Although occupational attributes have been measured in other, older surveys including the Dictionary of Occupational Titles, the overlap in the available measures across surveys is basically zero.

A similar pattern emerges among the highly educated: in that skill range, immigration is associated with native shifts toward communication-intensive occupations and away from analytic- or quantitative-intensive occupations, which immigrants tend to specialize in [Peri and Sparber \(2011b\)](#). Both are consistent with comparative advantage driven by natives' superior English language skills ([Lewis, 2013b](#)), although other unobserved skill or taste differences might also help rationalize this pattern.

Combining the results of the two [Peri and Sparber \(2009, 2011b\)](#) studies, [Foged and Peri \(2013\)](#) defined job “complexity” as $\ln((\text{communication} + \text{analytical})/\text{manual})$, using similar O*NET occupation communication, analytic, and manual measures. Taking advantage of the large influx of non-EU immigrants to Denmark since the mid-1990s due to major refugee events (e.g., breakup of Yugoslavia)—and prior to that non-EU immigration to Denmark was trivial—and using the “ethnic enclave” style of instrument, they used detailed firm worker-level data to show that non-EU immigration tends to push both high-skill and low-skill native Danes into more complex jobs, partly through occupational transitions. Since the non-EU immigrants are employed in largely low-skilled occupations, the fact that these gains and occupational transitions occur for both low-skill and high-skill groups (albeit, larger for low-skill groups) is again consistent with gains from specialization. Another nice feature of this study is that the “enclaves” in Denmark were largely the creation of random assignment of the previous small stock of non-EU immigrants due to a dispersal policy that distributed them around the country. Consistent with this, the authors are able to show that there are no significant “pretrends” (1991–1994) in outcomes.

An interesting recent finding is that immigration is not just associated with native specialization across occupations, but more broadly native occupational “upgrading,” shifts, that is, to occupations requiring higher skills and granting higher pay. [Cattaneo et al. \(2013\)](#) used rare individual-level panel data for natives in several European countries and found that immigration is associated with the natives of that country moving to higher-skill occupations in order to avoid competition. [Foged and Peri \(2013\)](#) found the same in Denmark. The additional value that comes from the panel-level feature of these studies will be described in [Section 10.6](#).

As was already discussed above, even within narrow occupations, one can sometimes find evidence of specialization. [Borjas and Doran \(forthcoming\)](#) analyze the interesting case of Soviet mathematicians in US departments, following the collapse of the Soviet Union. Soviet mathematicians were particularly concentrated in certain specialties, and these authors show that non-Soviet mathematicians tended to move out of these specialties after the influx.

Other studies, like [Borjas \(2003\)](#) and [Ottaviano and Peri \(2012\)](#), define skill cells in terms of education and (potential) work experience. While it is not possible to move across potential experience cells, there is new evidence that (young) natives may move up education cells in response to low-skill immigration. Using variation across US

“commuting zones,” which are smaller than metropolitan areas and cover the entire the United States, [Smith \(2012\)](#) showed that youth employment rates are particularly sensitive to immigration, a result that can partly be accounted for by a school enrollment response. Using cross-state variation and carefully accounting for the effect of immigrant children on the school system, [Hunt \(2012\)](#) found that a supply of immigrant adults that pushes up the relative supply of high school dropouts tends to increase the high school completion rates of the native born. Similarly, using cross-state variation and controlling separately for the impact of immigrants as students, [Jackson \(2013\)](#) found that increases in the relative proportion of adult immigrants without a college degree are associated with increases in college enrollment rates of native born.³⁵ This latter finding also reinforces the consensus that dividing the workforce into two broad categories based on college education is a reasonable first approximation of the labor market (see [Section 10.3.4](#) and [Card, 2009](#)).

10.5.3 Immigrant impacts

A very simple two-factor competitive model of the labor market, such as was presented in Equation (10.11), predicts that immigration can raise the average wages of native-born workers (e.g., [Borjas, 1999](#)), although there will still be “winners” and “losers” as usual in economics. Note that in this simple, constant return to scale model, the impact of immigration on overall average wages—that is, both immigrants and natives together—is by construction approximately 0. As was discussed above, richer models have been developed that allow for other potential sources of gains from immigration, such as direct impacts on productivity or production technology.

During the past decade, economists have produced many different estimates of the impact of immigration on wages and productivity. The richness of the mechanisms that they have considered has increased as researchers have considered more detailed measures of immigrants’ skills and additional outcomes besides wages, including investment and technology choices, and they have considered various sources of complementarities and “spillovers.” In [Table 10.2](#), we summarize the estimates in some (but certainly not all) recent studies that try to get to the overall impact of immigrants on productivity of native-born workers. While several studies in the past have focused on the “partial” impact of immigration, that is, on natives within narrow skill cells,³⁶—such as in Equation (10.8)—here, we are interested in the general equilibrium impacts of immigration on native productivity and wages, accounting for adjustments, which have been the focus of several studies in the recent decade.

A useful benchmark, reported in the first line of [Table 10.2](#), comes from [Ottaviano and Peri \(2012\)](#), who estimated the elasticity of substitution in a nested-CES production

³⁵ All of these studies use the “ethnic enclave”/“Bartik style” of instrument.

³⁶ These studies were summarized in [Longhi et al. \(2005\)](#) and [Kerr and Kerr \(2011\)](#).

Table 10.2 Total productivity effects of immigrants, estimates from the literature

Study	Outcome	Source of variation (cells)	Treatment; instrument	OLS	IV
Ottaviano and Peri (2012)	Native-born average $\ln(\text{wage})$	Education $\times$ experience	Actual immigration 1990–2006 ($\approx 10\%$ of pop); N/A simulation	0–0.007	
Ottaviano and Peri (2006)	Native-born average $\ln(\text{wage})$	US metropolitan areas	Diversity of country of birth index (shift-share diversity index)	1.27 (0.27)	0.98 (0.50)
Ottaviano and Peri (2006)	Native-born average $\ln(\text{wage})$	US metropolitan areas	Share foreign born; OLS only	0.57 (0.11)	
Peri (2012)	TFP (residual GSP/worker)	US states	$\Delta \text{imms/pop}$; ethnic enclave shift share	0.80 (0.39)	1.37 (0.27)
Peri (2012)	$\ln(\text{GSP/worker})$	US states	$\Delta \text{imms/pop}$; ethnic enclave shift share	0.62 (0.43)	0.88 (0.25)
Lewis (2011)	Output/worker, manufacturing	US metropolitan areas	HS dropouts/HS graduate; ethnic enclave shift share	–0.14 (0.10)	–0.03 (0.24)
Peri et al. (2014)	Native-born <i>college</i> $\ln(\text{wage})$	Metropolitan areas	STEM share of employment; ethnic enclave $\times$ change in H1-Bs	4.10 (1.86)	8.03 (3.02)
Peri et al. (2014)	Native-born <i>noncollege</i> $\ln(\text{wage})$	Metropolitan areas	STEM share of employment; ethnic enclave $\times$ change in H1-Bs	1.16 (1.24)	3.78 (1.75)
Foged and Peri (2013)	Native-born <i>college</i> $\ln(\text{hourly wage})$	Metropolitan areas, but within worker and firm	Non-EU imms/pop; ethnic enclave	0.254 (0.121)	0.864 (0.271)
Foged and Peri (2013)	Native-born <i>noncollege</i> $\ln(\text{hourly wage})$	Metropolitan areas, but within worker and firm	Non-EU imms/pop; ethnic enclave	0.236 (0.114)	0.460 (0.234)

Trax et al. (2013)	TFP (value-added residual)	Plant (conditional on metro area)	Immigrant diversity index; lagged values of inputs	Manufacture: 0.046(0.027) Service: 0.090(0.042)	Manufacture: 0.310(0.142) Service: 0.033(0.280)
Trax et al. (2013)	TFP (value-added residual)	Region (conditional on plant)	Immigrant diversity index; lagged values of inputs	Manufacture: 0.193(0.101) Service: 0.613(0.168)	Manufacture: 1.617(0.705) Service: 1.187(0.829)
Paserman (2013)	Output/worker, manufacturing	Plant	Share immigrant; OLS only	−0.073 (0.030)	
Paserman (2013)	Output/worker, manufacturing	Three-digit industry	Share immigrant; shift share	−0.028 (0.040)	0.216 (0.554)

function (Equation 10.4) and then used those to simulate the wage impacts of recent immigration on native wages. Depending on the choice of parameters, they found that immigration between 1990 and 2006 (amounting to roughly 10% of the initial workforce) should have raised the wages of the average native-born worker somewhere between 0% and 0.7%. While this model is rich in terms of skill categories—it includes four education groups $\times$ eight experience groups and allows for immigrant-native imperfect substitutability within those cells—it may not account for all of the beneficial impacts of immigrant skill diversity, let alone impacts through other mechanisms such as spillovers and native responses.

Ottaviano and Peri (2006) is an early paper suggesting that immigration is associated with higher average wages through a skill diversity mechanism. They found both a general positive association between immigration and wages and, conditional on this, an association with a measure the country diversity of immigrants, a Herfindahl index. Part of this gain may be in the form of cultural amenities, as the index is also associated with higher housing rents (also found in Saiz, 2007). Table 10.2 reports the effect of an increase in the diversity of country of birth index estimated in that article.

Ottaviano and Peri (2006) and Peri (2012) also produced useful “reduced form” benchmarks for the impact of immigration on wages and productivity. The first study estimates that the elasticity of native average wages with respect to immigrant share is 0.57. Since Ottaviano and Peri (2006) used data from 1970 to 1990, we have reestimated a version of what they did using more recent (2000–2010) data. Scatterplots and regression lines are shown in Figure 10.5. We find a native-born wage elasticity with respect to immigrant population share of 0.64 (with a standard error of 0.30) using variation in the percentage change of yearly earnings regressed on changes in share of foreign born across 219 metropolitan areas, weighted by initial population.³⁷

Peri (2012) examined productivity. Using gross state product (GSP) data merged to imputed state estimates of capital stock and skill mix, he log-linearized a generalized version of (10.11) (in particular, adding capital as a factor of production) and then regressed each component of the aggregate production function on immigration-driven population change.³⁸ After imposing an elasticity of substitution between college and noncollege workers of $\sigma = 1.75$ (compatible with consensus estimates, including Katz and Murphy, 1992; other studies include Hamermesh, 1993; Borjas, 2003; Ottaviano and Peri, 2012), he was able to estimate the impact of immigration on TFP. His basic OLS estimate implies that TFP rises 8% for each 10 percentage-point increase in

³⁷ If one splits employment between college-educated and noncollege-educated, one finds that the positive effect is driven by the college-educated group (elasticity of native college-educated wage to college-educated immigrants share of 0.93, standard error 0.58), while immigration among noncollege-educated has no significant effect on their wages (elasticity of -0.14 with standard error of 0.19).

³⁸ That is, his “treatment” variable is $\Delta F/\text{Pop}$, where ΔF is the change in the number of foreign-born residents and Pop is population.

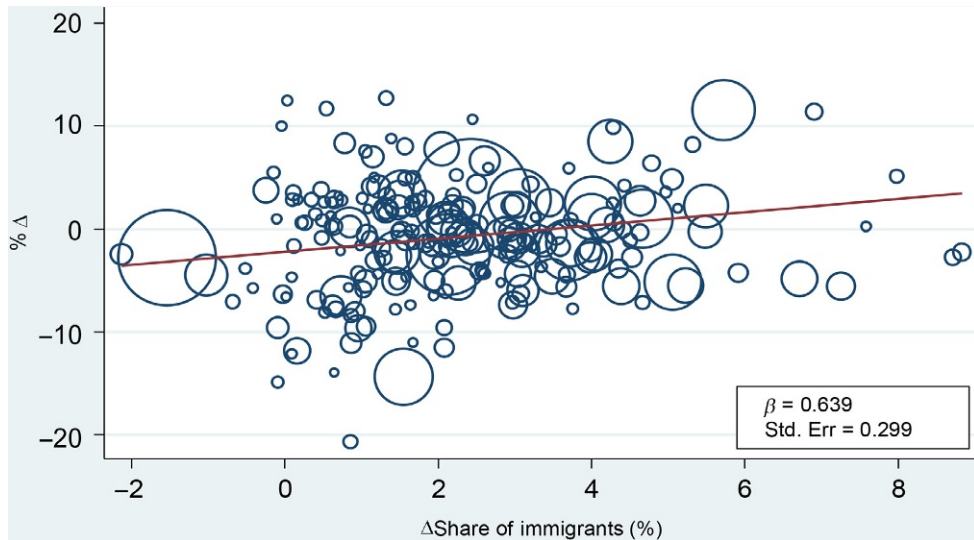


Figure 10.5 Native wages versus immigrant share. US cities 2000–2010. *Note: Each circle represents one of the 219 MSAs (cities). Data are calculated from the 2000 census and 2008–2010 3-year combined American Community Survey. Beta coefficient is from an OLS regression of the percentage change in wages against the change in the immigrant share, and standard errors are corrected for heteroskedasticity. All results are weighted by initial city population in 2000, and thus, the size of circles reflects the city size in 2000.*

immigration-induced population growth. This is confirmed in the 2SLS regression using distance from Mexico interacted with national immigration flows as instruments: his IV estimates are close to 14%. In short, states with a large increase in high school equivalents driven by Mexican immigrants have experienced a very strong relative increase in the productivity/efficiency of that factor in production. Note the critical role of imposing a value of σ to identify TFP impacts, which cannot be separately identified from a direct effect of labor supply otherwise. Reassuringly, although the estimates show some sensitivity to the choice of elasticity, all of the estimates are positive. In addition, the impact on the cruder GSP/worker is roughly the same magnitude (Table 10.2).

It is noteworthy that even the OLS estimates of the reduced form relationship between immigration and average wages in Peri (2012) and in Ottaviano and Peri (2006) are an order of magnitude larger than the simulated impact found in Ottaviano and Peri (2012). While the potential of omitted variable bias is a concern in spite of the 2SLS estimation, the larger measured effects suggest the potential presence of other mechanisms, besides the mechanical changes in observed skill mix, at work in affecting the wages of native-born workers. One mechanism is the working of efficient specialization and upgrades documented in Peri and Sparber (2009, 2011b), Foged and Peri (2013), D'Amuri and Peri (2014), and Cattaneo et al. (2013). Indeed, Peri (2012)

provided preliminary support for such a mechanism: controlling for the “communication intensity” of the average native’s job—the specialization measure used in [Peri and Sparber \(2009\)](#)—the productivity impact of immigration becomes smaller and statistically insignificant.³⁹ Another mechanism is a price effect: using a similar cross area approach and ethnic enclave instrument, [Cortes \(2008\)](#) showed that immigration lowers the price of nontraded goods.⁴⁰

As was discussed in [Section 10.3.5](#), firms may also respond to immigration by changing their production technology or, more narrowly, their capital intensity (e.g., [Lewis, 2011; Peri, 2012](#)). [Peri \(2012\)](#) showed that immigration (associated with a decrease in the college share of workers) is associated with an increase in the efficiency of high school equivalents in production.⁴¹ Going back to Equation (10.15), this says that immigration lowers both L_{CO}/L_{HS} and A_{CO}/A_{HS} ; as (10.15) shows, the latter attenuates any relative wage impacts of immigration due to the shift in L_{CO}/L_{HS} . Equation (10.15) shows the identification problem: impacts of relative labor supply on A_{CO}/A_{HS} are confounded with the direct of labor supply on wages (second term of (10.15)). As before, [Peri](#) broke this endogeneity by imposing a value of σ . So, what [Peri](#) found, in plainer terms, is that the cross-state impacts on the relative wage of immigration-induced changes in labor mix are smaller than what external estimates of the elasticity of substitution between college and noncollege workers would predict. Qualitatively similar findings are found in [Lewis \(2003\)](#) and [Card and Lewis \(2007\)](#), both of which use variation across metropolitan areas and “ethnic enclave”-style instruments (the latter focuses on the impacts of Mexicans only) and find little relative wage impact of immigration. One strength of the latter two studies is that they account for shifts in industry mix that would occur in a multisector model. These are found to be small, and so, skill mix changes are nearly as large “within industry” as overall. A weakness is that, unlike [Peri \(2012\)](#), they fail to focus on the college/noncollege relative supply, which newer research suggests is the main relative supply measure of interest.

In light of the identification challenge revealed by (10.15), are there some ways to get at the productivity terms or at least their determinants more directly? As was mentioned above, [Lewis \(2011\)](#) looked at the use of automation and capital intensity in the manufacturing sector and found that low-skill immigration reduces it. These changes in production technology and investment might partially account for the impact of

³⁹ Such evidence should only be taken as preliminary, as it attempts to partial out the impact of an endogenous regressor.

⁴⁰ See also [Lach \(2007\)](#).

⁴¹ The finding that changes in skill mix are associated with changes in the skill intensity of production is similar to a cross-country patterns found in [Caselli and Coleman \(2006\)](#) (which did not study immigration).

TFP that [Peri \(2012\)](#) found. However, Lewis did not find any significant association between low-skill immigration and output per worker. The difference in estimated productivity response between [Peri \(2012\)](#) and [Lewis \(2011\)](#) could be a result either of Lewis' narrower focus on the manufacturing sector or of his focus on the high school completion skill margin.⁴² A nice direction for future work would be to repeat a study like [Lewis \(2011\)](#) with a focus on the college completion margin.

Immigration may also affect productivity and wages through innovation and possibly through entrepreneurship. A couple of recent studies have focused on immigrants' disproportionate role in patenting and innovation. [Hunt and Gauthier-Loiselle \(2010\)](#) showed that among college graduates, immigrants have much higher patenting rates, which appears to be due to the fact that foreign college graduates have more education and they specialize in larger proportions in scientific and technological fields. Similarly, [Brunello et al. \(2007\)](#) showed that in a regression across US states over time, increases in foreign science PhD density are associated with a greater increase in patent counts than domestic science PhD density. Part of this finding may be due to a spillover from foreign scientists to the innovative productivity of domestic ones: [Hunt and Gauthier-Loiselle \(2010\)](#) found that in a similar panel regression, an increase in foreign college share in a state is associated with an increase in the patenting rate in a state that exceeds what one would expect "mechanically" from the higher patenting rate of immigrants in cross-sectional data. Hunt and Gauthier-Loiselle can only speculate that this is due to "spillovers," however, because the patent count data are not broken out by nativity in their panel data.⁴³

To partly address this problem, [Kerr and Lincoln \(2010\)](#) linked the names of patent holders to an ethnic names database, which allows them to divide patent counts, not by nativity, but into "Indian," "Chinese," and "Anglo-Saxon" patents. They studied specifically the role of the US high-skill "H1-B" program, and they took advantage of the fact that most H1-B visa holders are Indian and Chinese, making their ethnic groups a reasonable proxy for nativity.

They used variation across US labor markets. Though they cannot measure the quantity of H1-B holders at a local level, they essentially estimate a "reduced form" version of the ethnic enclave instrument: their right-hand-side variable is the interaction of the

⁴² [Paserman \(2013\)](#) also found little sign of a positive association between immigration and productivity using variation at the firm and industry levels in Israel's manufacturing sector. The immigrants in Paserman's study were largely "high-skill" immigrants from the FSU, although many did not end up in high-skill positions.

⁴³ The cross-sectional data are the National Survey of College Graduates, 2003 wave, while the panel data are tabulations from the US Patent and Trademark Office from 1940 to 2000. Thus, the difference in the timing of the two surveys is one confounder of the apparent difference in the association between immigrants and patenting in the two surveys.

stock of H1-B visas issued nationally with a measure of local “dependence” on H1-B type workers, which is a kind of imputed stock. They found that areas with more H1-B dependence have moderately higher rates of Anglo-Saxon patenting.⁴⁴

While higher patenting rates are associated with higher productivity at the country and sector level (Eaton and Kortum, 1996; Furman et al., 2002), patent counts are only imperfectly related to productive innovations (e.g., Griliches, 1990; Jaffe and Trajtenberg, 2002). So, a complementary approach is to examine the direct relationship between local high-skill share and wages, as Peri et al. (2014) do. Using variation across 219 US metropolitan areas, these authors estimate the relationship between changes in the STEM share of employment—that is, the share in high-tech occupations that they take to represent STEM—and wages. Their identification derives from a version of the ethnic enclave instrument: it is the sum of national changes in the country composition of H1-B immigrants interacted with the lagged size of the foreign STEM workforce in the area. Both their OLS and IV estimates suggest that STEM share is associated with higher wages for college graduates and, to a lower degree, for high school graduates, too.

These studies corroborate influential anecdotal evidence arguing a role for highly skilled engineers and entrepreneurs, especially immigrants (Saxenian, 2002a,b), in innovation and growth (Saxenian, 1994). Indeed, in addition to producing more patents, immigrants have significantly higher rates of entrepreneurship (Hunt, 2011), and immigration is associated with the creation of more small firms (Olney, 2013). Whether this matters for productivity growth is an open question; however, one tantalizing fact is that immigrant-owned businesses seem to be about 12% more productive, on average, than native-owned businesses (Garcia-Perez, 2008).

One final channel by which immigration may affect average productivity—which is related to the trade diversity model described in Section 10.3.5—is by increasing product diversity. di Giovanni et al. (2013) simulated the impact of the increase in-product diversity that comes from an increase in the scale of the economy associated with immigration. They found that it has a substantial positive impact on welfare in many immigrant-receiving developed countries. Empirically, immigration is associated with greater product diversity (Mazzolari and Neumark, 2012), though an increase in place of birth diversity (Ottaviano and Peri, 2005, 2006), rather than a pure scale effect, may account for this fact. Mazzolari and Neumark (2012) found that the strongest association is between immigration and an increase in restaurant diversity; other forms of retail diversity actually decline.

⁴⁴ A larger literature uses other (nongeographic) approaches to measuring the impact of immigration on innovation, including the dismissal of Jewish scientists from Nazi Germany (Waldinger, 2012; Moser et al., 2013). This broader literature is reviewed in Kerr (2013). This review also covers the impact of immigrant entrepreneurship.

10.5.4 Summarizing the productivity impacts

Though the number of studies of the direct association between immigration and productivity or wages summarized in Table 10.2 is not large, some interesting patterns emerge that may be helpful in charting a path for future research. The first pattern is that the association seems to be larger for high-skill (e.g., Peri et al., 2014) than low-skill immigrants (Lewis, 2011; Foged and Peri, 2013). The second is that there seems to be an additional impact of “origin diversity” separate from immigration’s impact generally (Ottaviano and Peri, 2006; Alesina et al., 2013; Trax et al., 2013). This combined with the fact that the “mechanical” impacts of observed skill mix on the wage structure seem to be small (Ottaviano and Peri, 2012), suggesting that unmeasured skills of immigrants contribute to their productivity impacts, perhaps in part through the sort of specialization directly documented in Peri and Sparber (2009, 2011b). The association with the wages of more educated natives also seems to be stronger (Foged and Peri, 2013; Peri et al., 2014), consistent with a directed technical change story (among others). Immigration may also affect productivity through prices (Lach, 2007; Cortes, 2008), product diversity (Mazzolari and Neumark, 2012; di Giovanni et al., 2013), innovation (Hunt and Gauthier-Loiselle, 2010; Kerr and Lincoln, 2010), or entrepreneurship (Hunt, 2011).

Recently available and richer data—for example, covering firm-level outcomes—may be helpful in uncovering the mechanisms, which rationalize these patterns.⁴⁵ We now turn to a discussion of such data and how they are being used.

10.6. RECENT EVOLUTIONS: EMPLOYER–EMPLOYEE PANEL DATA AND HISTORICAL DATA

Two valuable recent developments in the research on the impact of immigration are the use of individual-level panel data and, to be discussed later, the application of similar tools to historical data. The panel data confer several advantages. First, by following individuals over time, panel data allows controls for unobserved sources of heterogeneity. As may have been indirectly evident from the discussion about “displacement effects” in Sections 10.5.1–10.5.2, most of the studies above that employ a cell-based approach do not literally track the same individuals’ wages in a cell over time, but rather a representative sample of whoever is in that cell at each point in time.⁴⁶ While the potential for immigrants to affect the aggregate quantity of natives in each cell is the first-order concern with this approach (the subject of Sections 10.5.1–10.5.2), even after addressing this

⁴⁵ For example, it is notable that the impacts in Table 10.2 seem to be external to the firm (Paserman, 2013; Trax et al., 2013), though, as will be discussed below, firm-level impacts may so far not be as well identified.

⁴⁶ The use of repeated, representative cross-sectional sample not necessarily including the same individuals is also sometimes referred to as a “pseudo panel.” Note that this description applies to both area studies and “national” approaches to the study of the labor market impact of immigration.

problem, estimates could be biased by more subtle changes in the composition of workers due to selection or attrition from a cell; panel data address such concerns by tracking the same individuals over time. Indeed—and second—such compositional changes may be of direct interest in understanding the labor market dynamics, revealing the dynamics of individual workers' adjustment to immigration.⁴⁷ For example, one can estimate the benefits of the type of immigrant–native specialization documented in [Peri and Sparber \(2009, 2011b\)](#) directly, by tracking how individual native workers' productivity changes as they shift across occupations in response to immigration. Third, to the extent that workers are tracked across firms, we can look at outcomes at the level of the firm for the first time. This is a fortuitous feature of the fact that modern panel data often come from a sample of administrative records (social security earnings records), which, for many European countries, is the only reliable source of wage data over a long time frame.⁴⁸

The study by [Cattaneo et al. \(2013\)](#) was mentioned above. They use individual-level panel data to study both the impact of immigration on the wages of native-born workers and the dynamics of adjustment—specifically, the movement of incumbent natives across occupations over time—to immigration. Usefully, their data cover a number of European countries but are from survey data, not from administrative records that identify the firm of employment. They found significant occupation “upgrading” in response to immigration, namely, increased mobility of incumbent natives to jobs associated with higher skills and higher pay. They also found small wage effects on natives. [Foged and Peri \(2013\)](#) had access to a full panel of matched firm–worker data for Denmark from 1991 to 2008.⁴⁹ This allows them to study the dynamics of adjustment to immigration and to decompose any impacts into those that occur within a worker–firm match and those due to movements across firms. The focus of that study is on the impact of the non-EU immigration wave that swept over Denmark starting in the mid-1990s, largely due to refugees flows. These immigrants were on average less educated than the native-born

⁴⁷ Other studies exist on the dynamics of the response to immigration but observe the adjustment at the aggregate, rather than individual level, including [Barcellos \(2010\)](#), [Cohen-Goldner and Paserman \(2011\)](#), and [Murray and Wozniak \(2012\)](#).

⁴⁸ Until recently, European labor force surveys have tended to lack wage data, unlike in the United States, which was a challenge for research on the labor market impact of immigration. Interestingly, the willingness of many European countries to share social security earnings records with researchers now means that European data are of higher quality than those of the United States for many cutting-edge immigration-related research questions. While similar data are now available in the United States in the form of the Longitudinal Employer–Household Dynamics (LEHD) database, access is more difficult, use is more cumbersome, coverage is not as complete, and records are not as detailed when compared with the data available in many European countries. See [Kerr et al. \(2013\)](#) for a description of these data.

⁴⁹ [Malchow-Møller et al. \(2011, 2012\)](#) used the same Danish data to analyze the impact of immigrants on firm-specific wages and productivity. However, they do not analyze other outcomes, they do not follow individuals, and their identification strategy is not as convincing.

population, a trait they have in common with the flows analyzed in [Peri \(2012\)](#) and [Ottaviano and Peri \(2006\)](#). Interestingly, the coefficients in all of these studies are a similar order of magnitude despite the richer controls in [Foged and Peri \(2013\)](#)—see [Table 10.2](#). Given the large differences between them, this may simply be a coincidence, but it is nevertheless striking. It would be useful to apply these methods to similar data in other countries.

[Trax et al. \(2013\)](#) was an early attempt to use German social security data to separately estimate the impact of immigrant diversity (measured in the same manner as in [Ottaviano and Peri, 2006](#)) at the plant and region level, using lagged independent variables as instruments. They found effects at both levels, though the effects at the metropolitan area level are much larger in magnitude. At the metro level, they found that a 0.1 unit increase in diversity is associated with 16% higher wages in manufacturing and 18% in services, which is about 50% larger than what [Ottaviano and Peri \(2006\)](#) found. One caveat is that Trax et al.'s instrument set seems dubious, though their metropolitan area results remain similar when adding a more standard ethnic enclave-type instrument. [Paserman \(2013\)](#) also used variation at the firm level interacted with the large inflow of highly educated Russian immigrants to Israel to study the relationship between immigration and productivity at Israel manufacturing firms. He found little to no association in OLS estimates.

[Dustmann and Glitz \(forthcoming\)](#) use German plant-level data from the same source to ask how firms adjust to local immigration-driven changes in skill mix in the manufacturing sector. They find—consistent with the adjustments in production technology that [Peri \(2012\)](#) found at an aggregate level—a surprising level of responsiveness of unit efficiency to regional skill mix changes. In principle, [Peri's \(2012\)](#) result on unit efficiency might have been driven by the composition of firms and industries, rather than a change in production technology *per se*. However, Dustmann and Glitz find that, at least for Germany between 1985 and 1995, 70% or more of the skill mix changes at the region level are passed through to the plant level, despite there being no change in relative wages associated with immigration.⁵⁰ This is consistent with fully offsetting changes in unit efficiency (Equation 10.6) due to the adoption of techniques appropriate to the type of skills available. They use standard “ethnic enclave”-type instruments for identification.

Data linking employee characteristics with establishments are harder to come by in the United States. One recent study uses the best available to study the impact of high-skill immigration at the firm (not establishment) level. [Kerr et al. \(2013\)](#) used a subsample of 319 large firms in the so-called LEHD database, which is created from

⁵⁰ This figure is for firms that exist over the whole 10-year period. They find that net entry of firms also helps to push the skill mix toward the new level in the region. The authors also find a modest role for changes in product mix in absorbing skill mix changes, though larger than what previous studies had found (including [Lewis, 2003](#)).

unemployment insurance records from US states, but which the authors supplement with data from other firm databases (including Compustat).⁵¹ Using a panel regression, the authors find that in response to an influx of young (under age 40), high-wage (above \$50,000 in 2008 dollars) immigrants, firms hire more high-wage natives but especially young high-wage native workers. When they estimate using instrumental variables using, among other things, a version of the “ethnic enclave” instrument applied to the firm level, the effect for older native-born workers is often insignificant.⁵² Using a simple model of production adapted from [Desai et al. \(2009\)](#), they showed that under some assumptions, this finding is sufficient to show that young immigrants complement young natives.⁵³ This is not, however, sufficient to demonstrate that these immigrants actually substitute for older workers, but the authors produce some additional inconclusive evidence that these immigrants may actually displace older native-born STEM workers from the firm.⁵⁴ If so, the authors argue that this would be inconsistent with the way in which age categories are nested in the CES production structure used in [Borjas \(2003\)](#).

Another interesting recent development is the application of cutting-edge methods to historical data, which allows an investigation of the impact of immigration in the context of much different set of production choices facing firms and potentially differences in market structure. History can thus potentially help reveal how the impact of immigration may depend on context (such as the cultural factors and institutions that [Saxenian, 1994](#) discussed). An additional advantage of historical data is the relative ease of access to business and individual-level data, compared with modern data, in light of the lack of confidentiality concerns.⁵⁵ A weakness is often that there is less detail than modern data on things like compensation (wages) and even occupation, and so, researchers are left with the usual challenge of historical research of deriving credible proxies for the desired measures.

[Kim \(2007\)](#) used plant level data taken from 1860 to 1880 (at that time, decennial) US Censuses of Manufactures and ran regressions similar to what are described in [Table 10.2](#)— $\ln(\text{output per worker})$ and $\ln(\text{average wages})$ on immigrant share—using

⁵¹ In order to have their analysis go back to 1995, the authors are limited to firms with a significant presence in 18 US states.

⁵² Another instrument interacts the size of the H1-B “cap” with a measures of the firm’s H1-B “dependency,” for example, the number of “labor conditions applications” (LCA) a firm filed in 2001 per high-wage worker. LCAs are a precursor to hiring a worker on an H1-B visa, among other visa programs. [Ghosh and Mayda \(2013\)](#) used LCA data linked to Compustat to study the impact of H1-Bs at the firm level.

⁵³ Complementarities between factors are more clearly identified from the cross elasticities of factor prices or output shares (e.g., [Lewis, 2013a](#)), but these authors only have data on employment.

⁵⁴ If this result holds up, the interesting question would be what happens to these workers, which in principle could be answered with the sort of data that [Kerr et al. \(2013\)](#) were using.

⁵⁵ The individual-level records for many historic population and industry censuses, for example, are publicly available.

variation in immigrant share across US counties.⁵⁶ The regressions control for industry and state effects but, importantly, not for county fixed effects. In other words, unlike most of the studies in Table 10.2, Kim's results are cross-sectional. He found consistent significant relationships in OLS, with coefficients ranging from 0.5 to 1.5, which are not unlike the estimates in Table 10.2. IV estimates, which use the 1850 share foreign born as an instrument, are similar in magnitude. Kim also showed that immigrant share was associated with larger plants, a proxy for "factory production." This was the important innovation of the era, and it may be partly responsible for these productivity benefits.⁵⁷ A serious caveat is the cross-sectional nature of these results. In particular, the author shows that places with more immigrants also were also closer to New York and had better water transportation. These factors would have enabled access to larger markets, which is thought to be a prime driver of the adoption of factory production and productivity gains in this era.⁵⁸

Other recent history papers use variation across US counties but condition on county effects, rather than relying on cross-sectional variation; two studies also use an "enclave" style of instrument. Gonzalez-Velosa et al. (2013) examined the impact on the agricultural sector between 1900 and 1940, using data tabulated from the Census of Agriculture. They found little evidence that immigration affected agricultural productivity but found that immigration may have been mostly accommodated by shifts away from less labor-intensive crops (e.g., wheat) in places where the land could accommodate multiple crops.⁵⁹ In places where it could not, immigration is associated with markers of a shift toward more labor-intensive production techniques (from tractors to mules). The latter is consistent with evidence from the natural experiment of the US shutting down immigration in the 1920s (Lew and Cater, 2010).⁶⁰

Lafortune et al. (2014) examine the impact of changes in skill ratios (share literate), induced by immigration, on the manufacturing sector between 1860 and 1940, using data tabulated from the Census of Manufactures to the county or city $\times$ industry level. They

⁵⁶ Note that the wage data are averaged at the plant level, so the more conceptually appropriate specification average $\ln(\text{wage})$ is simply unavailable.

⁵⁷ Factory production is believed to be unskilled-labor-intensive compared to the predecessor technology of "artisan" production. Thus, if immigrants raised the relative abundance of unskilled labor, they might have induced adoption of factory production. This specification may oversimplify a bit the role of immigrants, however, many of whom were high skill.

⁵⁸ For example, see simulations in Donaldson and Hornbeck (2013) suggesting that market access via waterways access significantly raised local land values. See also Chandler (1977).

⁵⁹ The industry mix result contrasts with modern findings that industry mix adjustments play a trivial role in the absorption of immigration-driven skill mix changes (e.g., Lewis, 2003; Card and Lewis, 2007; Gonzalez and Ortega, 2011; Dustmann and Glitz, forthcoming).

⁶⁰ Lew and Cater (2010) examined agricultural counties on opposite either side of the Canada-US border during the 1920s, when the United States shut down inflows of foreign workers. This is associated with a sharp uptick in labor-saving tractor use on the US side of the border relative to the Canadian side, on what should be very similar agricultural land.

find a positive association between immigration-induced increases in the skill ratio (that is, using variation from an “enclave” instrument) and wages and productivity, a result that may be compositional only.⁶¹ More interestingly, unskilled immigration is associated with the adoption of more, not less, capital-intensive production techniques between 1860 and 1880. This contrasts sharply with twentieth-century capital–skill complementarity (e.g., Griliches, 1969; Goldin and Katz, 1998; Lewis, 2011) but is consistent with the view that advances in nineteenth-century manufacturing were “deskilling” (e.g., Atack et al., 2004). Thus, the finding is—similar to what Kim (2007) found—unskilled immigration may have induced faster adoption of new production methods in manufacturing in the nineteenth century, though the productivity gains from this may have been much more modest than what Kim found.

10.7. CONCLUSIONS

Across a wide range of settings and research approaches, immigration is associated with higher wages for most native-born workers and with higher productivity, especially when analyzing immigration across geographic areas. This simple fact, which has been subject to a large number of tests but on average has survived, indicates that immigrants represent more than a simple change in total labor supply at the local level. First of all, their skills and occupations are crucial to understand their impact. Second, the responses of native workers, firms, sectors, and potentially local consumers are also important margins to understand the equilibrium effects of migration on local economies. A more careful consideration of these aspects and the development of model-based ways of testing local responses to immigration have been crucial parts of the recent developments in this area of research. While framing the analysis of the effects of immigrants in a model that allows for different skills is very important and allows for a better understanding of complementarities and adjustment margins, it also important to allow for the possibility of productivity effects of immigrants.

The productivity effects of immigration have been found in a variety of European countries and in the United States and measured at the local level (cities or metropolitan areas) up to the country level. The positive impact of immigration on “skill diversity,” broadly construed, appears to be a key driving force behind this productivity impact. The adoption of new and efficient technologies in response to this richer set of skills may also be another important force underlying the adjustment. The exact mechanism underlying this impact, however, is still not well pinned down, but evidence suggests that

⁶¹ Their IV estimates are in the range 0.2–0.3, and their skill mix measure is $\ln(\text{literate workers}/\text{illiterate workers})$. Thus, if literate workers were on the order 20–30% more productive than illiterate workers, these results could be fully accounted for with compositional change. Unfortunately, no data exist to measure the productivity or wages by literacy status at the individual level in this era.

immigration induces natives to specialization in more complex jobs, which complement immigrants' skills and that it induces higher levels of innovation, both of which may contribute to the observed productivity impacts. More research taking advantage of plant- or firm-level data would likely be helpful in achieving a deeper understanding of how the impact of immigration materializes. At the same time, data that follow individuals and firms allow us to identify how these productivity effects diffuse across firms and labor markets and how they interact with firm and workers characteristics to determine winners and losers from these changes.

REFERENCES

- Abadie, A., Diamond, A., Hainmueller, J., 2010. Synthetic control methods for comparative case studies: estimating the effect of California's Tobacco Control Program. *J. Am. Stat. Assoc.* 105 (490), 493–505.
- Abramitzky, R., Boustan, L.P., Eriksson, K., 2014. A nation of immigrants: assimilation and economic outcomes in the age of mass migration. *J. Polit. Econ.* 122 (3), 467–506.
- Acemoglu, D., 1998. Why do new technologies complement skills? Directed technical change and wage inequality. *Q. J. Econ.* 113 (4), 1055–1089.
- Acemoglu, D., 2002. Technical change, inequality and the labor market. *J. Econ. Lit.* 40 (1), 7–72.
- Alesina, A., Harnoss, J., Rapoport, H., 2013. Birthplace diversity and economic prosperity. NBER Working paper 18699.
- Altonji, J.G., Card, D., 1991. The effects of immigration on the labor market outcomes of less-skilled natives. In: Abowd, J.M., Freeman, R.B. (Eds.), *Immigration, Trade, and the Labor Market*. University of Chicago Press, Chicago, pp. 201–234.
- Amuedo-Dorantes, C., de la Rica, S., 2011. Complements or substitutes? Task specialization by gender and nativity in Spain. *Labour Econ.* 18 (5), 697–707.
- Angrist, J.D., Krueger, A.B., 1999. Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3. Elsevier, Amsterdam, pp. 1277–1366.
- Angrist, J.D., Kugler, A.D., 2003. Protective or counter-productive? Labour market institutions and the effect of immigration on EU natives. *Econ. J.* 113, F302–F331.
- Atack, J., Bateman, F., Margo, R., 2004. Skill intensity and rising wage dispersion in nineteenth-century American manufacturing. *J. Econ. Hist.* 64 (1), 172–192.
- Autor, D.H., Katz, L.F., Krueger, A., 1998. Computing inequality: have computers changed the labor market? *Q. J. Econ.* 113 (4), 1169–1214.
- Barcellos, S., 2010. The dynamics of immigration and wages. RAND Working paper #WR-755.
- Bartel, A., 1989. Where do the new US immigrants live? *J. Labor Econ.* 7 (4), 371–391.
- Bartik, T., 1991. Who Benefits from State and Local Economic Development Policies? W.E. Upjohn Institute for Employment Research, Kalamazoo, Michigan.
- Blau, F.D., Kahn, L.M., 2012. Immigration and the distribution of incomes. IZA Discussion Papers 6921.
- Bohn, S., Lofstrom, M., Raphael, S., 2014. Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population? *Rev. Econ. Stat.* 96 (2), 258–269.
- Borjas, G.J., 1994. The economics of immigration. *J. Econ. Lit.* 32 (4), 1667–1717.
- Borjas, G.J., 1999. The economic analysis of immigration. In: Ashenfelter, O., David, C. (Eds.), *Handbook of Labor Economics*, vol. 3A. Elsevier, Amsterdam, pp. 1697–1760.
- Borjas, G.J., 2001. Does immigration grease the wheels of the labor market? *Brook. Pap. Econ. Act.* 1, 69–119.
- Borjas, G.J., 2003. The labor demand curve is downward sloping: reexamining the impact of immigration on the labor market. *Q. J. Econ.* 118 (4), 1335–1374.
- Borjas, G.J., 2006. Native internal migration and the labor market impact of immigration. *J. Hum. Resour.* 41 (2), 221–258.

- Borjas, G.J., Doran, K.B., 2012. The collapse of the Soviet Union and the productivity of American mathematicians. *Q. J. Econ.* 127 (3), 1143–1203.
- Borjas, G.J., Doran, K.B., forthcoming. Cognitive mobility: native responses to supply shocks in the space of ideas. *J. Labor Econ.*
- Borjas, G.J., Katz, L.F., 2007. The evolution of the Mexican-Born workforce in the United States. In: Borjas, G. (Ed.), *Mexican Immigration to the United States*. NBER, Cambridge, MA.
- Borjas, G.J., Freeman, R.B., Katz, L., 1997. How much do immigration and trade affect labor market outcomes? *Brook. Pap. Econ. Act.* 1, 1–67.
- Broda, C., Weinstein, D.E., 2006. Globalization and the gains from variety. *Q. J. Econ.* 121 (2), 541–585.
- Brunello, G., Fredriksson, P., Lamo, A., Messina, J., Peri, G., 2007. Higher education, innovation and growth. In: Brunello, G., Garibaldi, P., Wasmer, E. (Eds.), *Education and Training in Europe*. Oxford University Press, New York, pp. 56–70.
- Cadena, B.C., Kovak, B.K., 2013. Immigrants equilibrate local labor markets: evidence from the great recession. NBER Working paper 19272.
- Card, D., 1990. The impact of the Mariel Boatlift on the Miami labor market. *Ind. Labor Relat. Rev.* 43 (2), 245–257.
- Card, D., 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *J. Labor Econ.* 19 (1), 22–64.
- Card, D., 2005. Is the new immigration really so bad? *Econ. J.* 115 (507), F300–F323.
- Card, D., 2007. How Immigration Affects U.S. Cities. CREAM Discussion Paper no. 11/07.
- Card, D., 2009. Immigration and inequality. *Am. Econ. Rev.* 99 (2), 1–21.
- Card, D., DiNardo, J., 2000. Do immigrant inflows lead to native outflows? *Am. Econ. Rev.* 90 (2), 360–367.
- Card, D., Lemieux, T., 2001. Can falling supply explain the rising return to college for younger men? A cohort-based analysis. *Q. J. Econ.* 116 (2), 705–746.
- Card, D., Lewis, E., 2007. The diffusion of Mexican immigrants during the 1990s: explanations and impacts. In: Borjas, G.J. (Ed.), *Mexican Immigration to the United States*. University of Chicago Press, Chicago, pp. 193–228.
- Carrington, W.J., de Lima, P., 1996. The impact of 1970s repatriates from Africa on the Portuguese labor market. *Ind. Labor Relat. Rev.* 49 (2), 330–347.
- Cascio, E., Lewis, E., 2012. Cracks in the melting pot: immigration, school choice, and segregation. *Am. Econ. J. Econ. Policy* 4 (3), 91–117.
- Caselli, F., Coleman II, W.J., 2006. The world technology frontier. *Am. Econ. Rev.* 96 (3), 499–522.
- Castillo, F., Gilles, J.K., Raphael, S., 2009. Comparing the Domestic Labor Market Impacts of a South-North and South-South Migration: The Cases of Costa Rica and the United States. Mimeo, UC, Berkeley.
- Cattaneo, C., Fiorio, C.V., Peri, G., 2013. What happens to the careers of European workers when immigrants ‘take their jobs’? IZA Discussion Papers 7282.
- Chandler Jr., A.D., 1977. *The Visible Hand: The Managerial Revolution in American Business*. Harvard University Press, Cambridge.
- Chassamboulli, A., Palivos, T., 2014. A search-equilibrium approach to the effects of immigration on labor market outcomes. *Int. Econ. Rev.* 55 (1), 111–129.
- Chassamboulli, A., Peri, G., 2014. The labor market effects of reducing undocumented immigrants: NBER Working paper 19932.
- Ciccone, A., Hall, R.E., 1996. Productivity and the density of economic activity. *Am. Econ. Rev.* 86 (1), 54–70.
- Ciccone, A., Peri, G., 2006. Identifying human-capital externalities: theory with applications. *Rev. Econ. Stud.* 73 (2), 381–412.
- Clark, X., Hatton, T.J., Williamson, J.G., 2007. Explaining US immigration, 1971–1998. *Rev. Econ. Stat.* 89 (2), 359–373.
- Cohen-Goldner, S., Paserman, M.D., 2011. The dynamic impact of immigration on natives’ labor market outcomes: evidence from Israel. *Eur. Econ. Rev.* 55 (8), 1027–1045.
- Conley, T.G., Taber, C.R., 2011. Inference with ‘difference in differences’ with a small number of policy changes. *Rev. Econ. Stat.* 93 (1), 113–125.
- Cortes, P., 2008. The effect of low-skilled immigration on US prices: evidence from CPI data. *J. Polit. Econ.* 116 (3), 381–422.

- Damm, A.P., 2009. Ethnic enclaves and immigrant labor market outcomes: quasi-experimental evidence. *J. Labor Econ.* 27 (2), 281–314.
- D'Amuri, F., Peri, G. 2014. Immigration, jobs and employment protection: evidence from Europe before and during the great recession. *J. Eur. Econ. Assoc.* 12 (2), 432–464.
- Desai, M., Friz Foley, C., Hines, J., 2009. Domestic effects of the foreign activities of US multinationals. *Am. Econ. J. Econ. Policy* 1 (1), 181–203.
- di Giovanni, J., Levchenko, A., Ortega, F., 2013. A Global View of Cross-Border Migration. Mimeo, UPF, Barcelona, Spain.
- Docquier, F., Ozden, C., Peri, G., 2011. The labor market effects of immigration and emigration in OECD countries. IZA Discussion Papers 6258.
- Donald, S.G., Lang, K., 2007. Inference with difference-in-differences and other panel data. *Rev. Econ. Stat.* 89 (2), 221–233.
- Donaldson, D., Hornbeck, R., 2013. Railroads and economic growth: a 'market access' approach. NBER Working paper 19213.
- Dustmann, C., Frattini, T., Preston, I.P., 2013. The effect of immigration along the distribution of wages. *Rev. Econ. Stud.* 80 (1), 145–173.
- Dustmann, C., Glitz, A., forthcoming. How do industries and firms respond to changes in local labor supply? *J. Labor Econ.*
- Eaton, J., Kortum, S., 1996. Trade in ideas: patenting and productivity in the OECD. *J. Int. Econ.* 40 (3–4), 251–278.
- Filer, R., 1992. The effect of immigrant arrivals on migratory patterns of native workers. In: Borjas, J., Freeman, R.B. (Eds.), *Immigration and the Workforce: Economic Consequences for the United States and Source Areas*. National Bureau of Economic Research, Cambridge, MA, pp. 245–270.
- Foged, M., Peri, G., 2013. Immigrants and native workers: new analysis using longitudinal employer-employee data. NBER Working paper 19315.
- Frey, W.H., 1995. Immigration and internal migration 'flight': a California case study. *Popul. Environ.* 16 (4), 353–375.
- Friedberg, R.M., 2001. The impact of mass migration on the Israeli labor market. *Q. J. Econ.* 116 (4), 1373–1408.
- Furman, J.L., Porter, M.E., Stern, S., 2002. The determinants of national innovative capacity. *Res. Policy* 31 (6), 899–933.
- Garcia-Perez, M., 2008. Does It Matter Who I Work For and Who I Work With? The Impact of Owners and Coworkers on Hiring and Wages. Mimeo, University of Maryland, College Park, MD.
- Ghosh, A., Mayda, A.M., 2013. The Impact of Skilled Migration on Firm-Level Productivity: An Investigation of Publicly Traded U.S. Firms. Mimeo, Georgetown University, Washington, DC.
- Glitz, A., 2012. The labor market impact of immigration: a quasi-experiment exploiting immigrant location rules in Germany. *J. Labor Econ.* 30 (1), 175–213.
- Goldin, C., Katz, L.F., 1998. The origins of technology-skill complementarity. *Q. J. Econ.* 113 (3), 693–732.
- Goldin, C., Katz, L.F., 2008. *The Race Between Education and Technology*. Harvard University Press, Cambridge, MA.
- Gonzalez, A., 1998. Mexican enclaves and the price of culture. *J. Urban Econ.* 43 (2), 273–291.
- Gonzalez, L., Ortega, F., 2011. How do very open economies absorb large immigration flows? Evidence from Spanish regions. *Labour Econ.* 18, 57–70.
- Gonzalez-Velosa, C., Lafortune, J., Tessada, J., 2013. More hands, more power? The impact of immigration on farming and technology choices in US agriculture in the early twentieth century. Mimeo, Pontificia Universidad Católica de Chile, Santiago, Chile.
- Greenstone, M., Hornbeck, R., Moretti, E., 2010. Identifying agglomeration spillovers: evidence from winners and losers of large plant openings. *J. Polit. Econ.* 118 (3), 536–598.
- Griliches, Z., 1969. Capital-skill complementarity. *Rev. Econ. Stat.* 51 (4), 465–468.
- Griliches, Z., 1990. Patent statistics as economic indicators: a survey. *J. Econ. Lit.* 28 (4), 1661–1707.
- Grossman, J.B., 1982. The substitutability of natives and immigrants in production. *Rev. Econ. Stat.* 64 (4), 596–603.
- Hamermesh, D., 1993. *Labor Demand*. Princeton University Press, Princeton.
- Hunt, J., 1992. The impact of the 1962 repatriates from Algeria on the French labor market. *Ind. Labor Relat. Rev.* 45 (3), 556–572.

- Hunt, J., 2011. Which immigrants are most innovative and entrepreneurial? Distinctions by entry visa. *J. Labor Econ.* 29 (3), 417–457.
- Hunt, J., 2012. The impact of immigration on the educational attainment of natives. NBER Working paper 18047.
- Hunt, J., Gauthier-Loiselle, M., 2010. How much does immigration boost innovation? *Am. Econ. J. Macroecon.* 2 (2), 31–56.
- Iranzo, S., Peri, G., 2009. Schooling externalities, technology, and productivity: theory and evidence from U.S. States. *Rev. Econ. Stat.* 91 (2), 420–431.
- Jackson, O., 2013. Does Immigration Crowd Natives Into or Out of Higher Education? Mimeo, Northeastern University, Boston, MA.
- Jaffe, A., Trajtenberg, M., 2002. *Patents, Citations and Innovations: A Window on the Knowledge Economy*. MIT Press, Cambridge, MA.
- Kato, T., Sparber, C., 2013. Quotas and quality: the effect of H-1B visa restrictions on the pool of prospective undergraduate students from Abroad. *Rev. Econ. Stat.* 95 (1), 109–126.
- Katz, L.F., Murphy, K.M., 1992. Changes in relative wages, 1963–1987: supply and demand factors. *Q. J. Econ.* 107 (1), 35–78.
- Kerr, W., 2013. U.S. high-skilled immigration, innovation, and entrepreneurship: empirical approaches and evidence. National Bureau of Economic Research Working paper 19377.
- Kerr, S.P., Kerr, W.R., 2011. Economic impacts of immigration: a survey. *Finn. Econ. Pap.* 24 (1), 1–32.
- Kerr, W., Lincoln, W.F., 2010. The supply side of innovation: H-1B visa reforms and U.S. ethnic invention. *J. Labor Econ.* 28 (3), 473–508.
- Kerr, S.P., Kerr, W.R., Lincoln, W.F., 2013. Skilled immigration and the employment structures of U.S. firms. NBER Working paper 19658.
- Kim, S., 2007. Immigration, industrial revolution and urban growth in the United States, 1820–1920: factor endowments, technology and geography. NBER Working paper 12900.
- Krusell, P., Ohanian, L., Rios-Rull, J.-V., Violante, G., 2000. Capital-skill complementarity and inequality: a macroeconomic analysis. *Econometrica* 68 (5), 1029–1053.
- Kugler, A.D., Yuskel, M., 2008. Effects of low-skilled immigration on U.S. natives: evidence from Hurricane Mitch. NBER Working paper 14293.
- Lach, S., 2007. Immigration and prices. *J. Polit. Econ.* 115 (4), 548–587.
- Lafortune, J., Tessada, J., 2013. Smooth(er) Landing? The Role of Networks in the Location and Occupational Choice of Immigrants. Mimeo, Pontificia Universidad Católica de Chile, Santiago, Chile.
- Lafortune, J., Lewis, E., Tessada, J., 2014. People and Machines: A Look at the Evolving Relationship Between Capital and Skill in Manufacturing Using Immigration Shocks. Mimeo, Pontificia Universidad Católica de Chile, Santiago, Chile.
- Lew, B., Cater, B., 2010. Farm Mechanization on an Otherwise ‘Featureless’ Plain: Tractors on the Great Plains and Immigration Policy in the 1920s. Mimeo, Trent University, Peterborough, ON.
- Lewis, E., 2003. Local, open economies within the U.S.: how do industries respond to immigration? Federal Reserve Bank of Philadelphia Working papers 04–1.
- Lewis, E., 2011. Immigration, skill mix, and capital-skill complementarity. *Q. J. Econ.* 126 (2), 1029–1069.
- Lewis, E., 2013a. Immigration and production technology. *Annu. Rev. Econ.* 5, 165–191.
- Lewis, E., 2013b. Immigrant-native substitutability and the role of language. In: Card, D., Raphael, S. (Eds.), *Immigration, Poverty and Socio-Economic Inequality*. Russell Sage Foundation, New York.
- Longhi, S., Nijkamp, P., Poot, J., 2005. A meta-analytic assessment of the effects of immigration on wages. *J. Econ. Surv.* 19 (3), 451–477.
- Malchow-Moller, N., Munch, J.R., Skaksen, J.R., 2011. Do foreign experts increase the productivity of domestic firms? IZA Discussion Paper 6001.
- Malchow-Moller, N., Munch, J.R., Skaksen, J.R., 2012. Do immigrants affect firm-specific wages? *Scand. J. Econ.* 114 (4), 1267–1295.
- Manacorda, M., Manning, A., Wadsworth, J., 2012. The impact of immigration on the structure of wages: theory and evidence from Britain. *J. Eur. Econ. Assoc.* 10 (1), 120–151.
- Mayda, A.M., 2010. International migration: a panel data analysis of the determinants of bilateral flows. *J. Popul. Econ.* 23 (4), 1249–1274.
- Mazzolari, F., Neumark, D., 2012. Immigration and product diversity. *J. Popul. Econ.* 25 (3), 1107–1137.

- Moretti, E., 2004a. Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *J. Econ.* 121 (1–2), 175–212.
- Moretti, E., 2004b. Workers' education, spillovers, and productivity: evidence from plant-level production functions. *Am. Econ. Rev.* 94 (3), 656–690.
- Mortensen, D., Pissarides, C., 1994. Job creation and job destruction in the theory of unemployment. *Rev. Econ. Stud.* 61 (3), 397–415.
- Moser, P., Voena, A., Waldinger, F., 2013. German Jewish Émigrés and U.S. Invention. Mimeo, Stanford University, Stanford, CA.
- Murray, T.J., Wozniak, A., 2012. Timing is everything: short-run population impacts of immigration in U.S. cities. *J. Urban Econ.* 72 (1), 60–78.
- Olney, W., 2013. Immigration and firm expansion. *J. Reg. Sci.* 53 (1), 142–157.
- Ortega, F., Peri, G., 2014. Openness and income: the roles of trade and migration. *J. Int. Econ.* 92 (2), 231–251.
- Ottaviano, G.I.P., Peri, G., 2005. Cities and cultures. *J. Urban Econ.* 58 (2), 304–337.
- Ottaviano, G.I.P., Peri, G., 2006. The economic value of cultural diversity: evidence from US cities. *J. Econ. Geogr.* 6 (1), 9–44.
- Ottaviano, I.P., Peri, G., 2012. Rethinking the effects of immigration on wages. *J. Eur. Econ. Assoc.* 10 (1), 152–197.
- Ozden, C., Wagner, M., 2013. Immigrants vs Natives: Displacement and Job Creation. Mimeo, Boston College, Boston, MA.
- Paserman, M.D., 2013. Do high-skill immigrants raise productivity? Evidence from Israeli manufacturing firms, 1990–1999. *IZA J. Migration.* 2 (6).
- Peri, G., 2011. Rethinking the area approach: Immigrants and the labor market in California. *J. Int. Econ.* 84 (1), 1–14.
- Peri, G., 2012. The effect of immigration on productivity: evidence from U.S. states. *Rev. Econ. Stat.* 94 (1), 348–358.
- Peri, G., 2013. Immigration, native poverty and the labor market. In: Card, D., Raphael, S. (Eds.), *Immigration, Poverty and Socio-Economic Inequality*. Russell Sage Foundation, New York.
- Peri, G., Sparber, C., 2009. Task specialization, immigration, and wages. *Am. Econ. J. Appl. Econ.* 1 (3), 135–169.
- Peri, G., Sparber, C., 2011a. Assessing inherent model bias: an application to native displacement in response to immigration. *J. Urban Econ.* 69 (1), 82–91.
- Peri, G., Sparber, C., 2011b. Highly-educated immigrants and native occupational choice. *Ind. Relat.* 50 (3), 385–411.
- Peri, G., Shih, K.Y., Sparber, C., 2014. Foreign STEM workers and native wages and employment in U.S. cities: NBER Working papers 20093.
- Pugatch, T., Yang, D., 2011. The Impact of Mexican Immigration on U.S. Labor Markets: Evidence from Migrant Flows Driven by Rainfall Shocks. Mimeo, University of Michigan, Ann Arbor, MI.
- Roy, A.D., 1951. Some thoughts on the distribution of earnings. *Oxf. Econ. Pap.* 3, 135–146.
- Ruggles, S., Trent Alexander, J., Genadek, K., Goeken, R., Schroeder, M.B., Sobek, M., 2010. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. University of Minnesota, Minneapolis.
- Saiz, A., 2007. Immigration and housing rents in American cities. *J. Urban Econ.* 61 (2), 345–371.
- Saxenian, A.L., 1994. *Regional Advantage: Culture and Competition in Silicon Valley and Route 128*. Harvard University Press, Cambridge.
- Saxenian, A.L., 2002a. Silicon Valley's new immigrant high growth entrepreneurs. *Econ. Dev. Q.* 16, 20–31.
- Saxenian, A.L., 2002b. Local and Global Networks of Immigrant Professionals in Silicon Valley. Public Policy Institute of California, San Francisco, CA.
- Smith, C.L., 2012. The impact of low-skilled immigration on the youth labor market. *J. Labor Econ.* 30 (1), 55–89.
- Trax, M., Brunow, S., Suedekum, J., 2013. Cultural Diversity and Plant-Level Productivity. Mimeo, Mercatur School of Management, University of Duisburg-Essen, Essen, Germany.
- Waldinger, F., 2012. Peer effects in science—evidence from the dismissal of scientists in Nazi Germany. *Rev. Econ. Stud.* 79 (2), 838–861.

This page intentionally left blank

INDEX

Note: Page numbers followed by *f* indicate figures and *t* indicate tables.

A

Add Health database, crime commitment, 595

Agglomeration. *See also* Innovation)

agglomeration economies, 388–390

city size and industrial composition, 386–388

clusters, 366, 392–394

employer–employee dataset, 397

firms (*see* Firms)

geographic concentration, 366–367

industrial clusters, 366

innovative activity, 367

knowledge spillovers (*see* Knowledge spillovers)

Marshallian equivalence, 386

Marshallian externalities, 367

Silicon Valley initiatives, 397

spatial concentration (*see* Geographic concentration)

Agglomeration and land use, 503–504

bid-rent function, 505, 509–510

communication spillovers, 505, 507–509, 508*f*

monocentric equilibrium, 513, 513*f*

secondary employment centers, 504

spillovers, 513–514

Agglomeration economies

characteristics, 249

determinants, 270–271

developing countries, 339

economic geography, 296–297

elasticity, 178, 189, 218, 297

employment growth, 251

endogeneity issues, 250, 284–286

entrepreneurship, 251–252

housing, 339

human capital externalities, 278–282

industrial employment, 262–263

industrial specialization and diversity, 274–278

mechanisms, 174 (*see also* Agglomeration mechanisms)

price indices, 339

productivity growth, 262–263

productivity, local determinants

(*see* Productivity, local determinants)

public policies, 248–249

skills, 340

spatial lag models, 297–298

spatial scale, 294–295

typology, 249

Agglomeration mechanisms

case studies, 336–338

coagglomeration, 331–336

economic forces, 361–366

matching, 369–371

MSA population, mean household income and median rent, 179*f*

sharing, 368–369

sustaining mechanisms, 390–391

Agglomeration theory

cities, 172–173

heterogeneity, 173

heterogeneous agents, 173–174, 239–240

housing stock, 174

inequality (*see* Inequality)

innovation clusters, 351

microeconomic mechanisms, 174, 187

selection (*see* Selection, agglomeration theory)

sorting (*see* Sorting, agglomeration theory)

urban productivity premium, 172–173

Zipf's law, 173–174

Air quality, 413–417

environmental trends, 413–421

random utility models, 416

spatial heterogeneity, 414–415

Alonso–Muth–Mills model, 198

“Ashenfelter dip”, difference in differences (DD), 32

B

Biases

omitted variable biases

contextual effects, 159

Biases (*Continued*)

- endogeneity, 160, 657
- social interactions, 158

Bid-rent approach, 475–477

Binary treatment environment

- causal inference, 9–10
- marginal treatment effect (MTE), 10
- urban and regional applications, 9

Bottom-up approach, 447, 448, 449

Building stock's energy consumption

- bottom-up approach, 447–449
- carbon pricing, 456–457
- commercial real estate sector, 450–456
- electricity consumption, 446–447
- energy and urban economics, 445–446
- greenhouse gas production, 445–446
- weak price incentives, 449–450

C

Causal inference

- binary treatment, 9–10
- continuous treatments, 15
- difference in differences (DD), 30–37
- endogeneity bias, 5
- equilibrium effects, 4–5
- fixed effects, 24–30
- instrumental variables (IV), 6, 43–53
- matching methods, 37–43
- regression discontinuity (RD), 53–62
- selection on observables, 23–43
- spatial aggregation, 20–23
- treatment effect, 4–6, 11–14

CES. *See* Constant elasticity of substitution (CES)

City size distribution

- agglomeration economies, 193–195
- Zipf's law, 184, 193–195, 220

Clean Air Act, 414–416

Clusters

- entrepreneurial, 387, 394–396
- geographic, 361
- innovation, 351
- innovation projects, 368
- spatial movements, 366

Cobb–Douglas function, 252–253, 283

Commercial real estate sector

- big data, 455
- dwellings, 455
- green buildings, 451–454

marketing, 454–455

regressions, 450–451

Composition of cities

- functional composition, 206–210
- industry composition, 202–206
- skill composition, 210–211

Constant elasticity of substitution (CES)

- education, 641
- effects, 641–642, 643–644
- entrepreneurs, 231
- flexible translog specifications, 638–639
- labor supply, 641–642, 644–648
- multi-industry model, 225–226
- between natives and immigrants, 641
- nested-CES approach, 638–640, 640f
- preferences, 230–231
- quadratic preferences model, 230–231
- selection, 229–230
- structure, 230–231, 640–641

D

Density

- agglomeration effects, 271–274, 294, 298–303, 304, 312, 337–338
- location, 265, 336–337
- spatial extension, 271–274, 306–307

Difference in differences (DD) methods

- assumption, 30
- control group, 32–33, 34
- empowerment zone (EZ), 33
- estimator, 31
- first-differencing data, 8
- identification strategies, 35–36
- implementation, 30–31
- pretreatment trends, 34
- synthetic controls, 36–37
- treatment on the treated (TT), 31
- triple difference, 33

Dual approach, monocentric model

- Alonso–Muth condition, 477–478
- closer city, 480–481
- comparative statics, 481
- construction industry, 479–480
- house price, 478, 479
- open city, 480–481
- population density, 480
- residential equilibrium, 477–478, 480–481
- transport costs, 481

- Durable housing
 - comparative statics, 497
 - demand shock, 503
 - developers' behavior, 501
 - dynamic monocentric model, 499
 - expectations, 497–498
 - implications, 503
 - lags, 500
 - monocentric model, 497–499
 - myopic foresight, 498–499
 - redevelopment, 501–502
 - supply and demand, 502–503, 502*f*
- E**
- Econometrician
 - household and, 75
 - household location choice model, 429
 - information set of, 88
 - spatial, 137
- Economic activity, space
 - distance decay, 99
 - entry models, 106–110
 - internal structure, cities, 100–103
 - policy analysis, 103–106
 - productivity distribution, 99
 - specialization of regions, 96–100
 - transportation cost, 97
- Elasticity, log-linear relationship, 296–297
- Empirical approaches
 - endogeneity bias, 657
 - exploits* trends, 661
 - natural experiment, 657, 660
 - supply/push instrument, 659
- Empirical investigation
 - data-generating process, 7
 - structural, 7
 - treatment effects, 8–9
 - treatment variables, 6–7
- Employer–employee panel data, 675–680
- Employment decentralization
 - commuting, 546–547
 - firm location, 544, 549–550
 - metropolitan areas, 544–545
 - spatial mismatch, 547–549
 - subcenter detection method, 545
 - transport networks, 549
 - zoning, 550
- Empowerment zone (EZ), 21, 33, 34, 41
- Endogenous variables
 - fixed effects, 286–287
 - generalized method of moments (GMM)
 - approach, 289–290
 - historical and geological variables, 287–289
 - instrumentation, 286
 - natural experiments, 290–292
- Entrepreneurship
 - culture, 391
 - local firm size distributions, 387
- Environmental amenities
 - air pollution, 435–436
 - climate, 427–428, 437
 - climate change, 438–440
 - firm location, 431–432
 - green cities, 440–442
 - heterogeneous workers, 427
 - land use regulation/zoning, 436–437
 - local public goods, 430–431
 - location choice model, 428–430
 - Rosen/Roback model, 427–428
- Environmental trends
 - air quality, 413–417
 - center city land, 421
 - economic activity, 413
 - economic incidence, 421
 - National Priorities List (NPL), 419
 - open space preservation, 417–419
 - price effects, 421
- Equilibrium
 - amenities and heterogeneity, 87
 - city size distribution, 192
 - computation, 86
 - dynamics, 87
 - intercommunity equilibrium, 84–85
- Explicitly model, local public good
 - valuation, 445
- F**
- Firms
 - agglomeration clusters, 396
 - characteristics, 292–293
 - creation and entrepreneurship, 327–328
 - location choice, 324–327
 - strategies, 322–324
 - venture capital (VC) firms, 390–391
 - workers, 252–253
- Fiscal competition. *See* Public good provision

Fixed effects methods
 data-generating process, 28
 fixed effects, 25–26
 hedonic models, 27
 heterogeneity, 24–25
 heterogeneous treatment effects, 26–27
 local treatment effects, 27–28
 microgeographic fixed effects, 29–30
 randomization, 27
 Friendship networks, Add Health data, 594

G

Geographical information systems (GIS),
 122–123, 144
 Geographic concentration
 agglomeration, 366–367
 knowledge spillovers, 371
 research and development (R&D) activity, 383
 Geographic mobility
 aggregate production function, 663
 Bartik instrument, 662
 Borjas' approach, 662
 native displacement, 663
 native mobility, 664
 perfect mobility, 663
 Green buildings
 ENERGY STAR program, 452
 hedonic regressions, 453
 investment decisions, 453–454
 Leadership in Energy and Environmental Design
 (LEED) certification, 454
 solar panels, 453
 Group-based policies, 598–599

H

Heckscher–Ohlin–Vanek trade models, 392–393
 Heterogeneity
 agglomeration economies, 173–174
 assignment models, 491
 assignment problems, 489
 demand, housing, 487–488
 housing market, 495
 labor search models, 496–497
 land and property markets, 494, 495
 levels of income, 491
 parcels (*see* Parcel heterogeneity)
 poor and rich residents, 487–488
 supply and demand, housing, 490

Heterogeneous agents
 agglomeration theory (*see* Agglomeration theory)
 inequality, 234
 sorting, 173–174, 178–181, 216*f*
 Heterogeneous individuals, city size distribution,
 219–220
 Household and firm composition impact, 445
 Household location choice model, 428–430
 Household sorting, income, 89–91

I

Immigrant impacts, native responses and effects
 automation, 672–673
 communication, 671–672
 complementarities, 667
 entrepreneurship, 673
 ethnic enclave instruments, 672, 674
 firms, 672
 native wages, 670, 671*f*
 productivity effects, 667–670, 668*t*
 spillovers, 673
 substitution, 670–671
 trade diversity, 674
 two-factor model, 667
 Immigration
 American Community Survey 2011, 632
 census occupational classification, 632
 concentration and exposure, 632, 633*t*
 education, 636
 foreign-born, 626–627, 635
 individual-level and firm-level data, 631–632
 inflow, 631
 local effects, 627
 mechanism, 631–632
 metropolitan statistical areas (MSAs), 634–635
 national level, 630
 pressure, 636
 production function, 630–631, 633
 productivity impacts, 675
 short-run effects, 629
 “skating-rink” hypothesis, 631
 skills, 628 (*see also* Skill and production)
 wage and productivity growth, 630
 Individual-based policies
 contextual effect, 596–597
 crimes, 597
 network effect, 596–597
 social multipliers, 595–596

- Industrial composition, city size distribution, 386–388
- Industrial employment
 - dynamic specifications, 321–322
 - productivity externalities, 315–319
 - specialization and diversity, 319–321
- Inequality
 - composition (selection and sorting), 234
 - earnings distribution, 234
 - heterogeneous agents, 234
 - size (agglomeration economies), 234
 - urban (*see* Urban inequality)
- Innovation. *See also* Agglomeration)
 - advantages, 388–390
 - description, 350
 - endogenous growth theory, 350
 - entrepreneurial clusters, 394–396
 - globalization, 351
 - invention, 352–353
 - literature-based indicators, 357–358
 - patents and citations, 355–357
 - R&D expenditures/venture capital investment, 354–355
 - research, 351
 - technological, 354
 - types of, 353–354
- Instrumental variables (IV) estimators
 - “Bartik” instruments, 48
 - control function, 45–46
 - correlation, 44–45
 - exclusion restriction, 43
 - exogenous variation, 48
 - generalized method of moments (GMM), 46, 48–49
 - historical plans and networks, 51
 - instrument exogeneity, 46
 - labor markets, 50
 - local ATE (LATE), 47
 - local labor supply, 50–51
 - metropolitan areas, 52
 - Moving to Opportunity (MTO), 53
 - political power, 52
 - random variation, 43, 46
 - and regional economics literature, 47–48
 - self-selection, 43–44
 - treatment effect, 47
 - urban and regional literature, 52–53
 - WRI, 49–50
- Intention to treat (ITT), 13
- Interconnections, spatial models
 - aggregating, 131–132
 - contiguity matrices, 130–131, 132
 - transitivity property, 130–131
 - weights, 131
- Internal structure of cities
 - commuting costs, 100–101
 - description, 100
 - equilibrium, 102
 - estimation and identification, 101–103
 - productivity, 100–101
 - wages, 102
- Intracity model, household location choice, 428
- Intrametropolitan area location choice
 - energy, 423–424
 - greenhouse gas emissions, 423, 424
 - sorting, 423
 - suburbs, 423, 424–427
- Invention
 - description, 350
 - disadvantages, 355–356
 - innovation, 350, 352–353
 - patent publication, 355
- K**
 - “Kitchen sink” feel, local public good valuation, 442–443
- Knowledge spillovers
 - agglomeration literature, 384–385
 - economic geography models, 372
 - endogenous growth, 372–376
 - geographic concentration, 371
 - human capital spillovers, 384
 - patents, 380–385
 - prominent scientists, 385
 - Silicon Valley, 371–372
 - skills, 384
 - wages, 376–380
- L**
 - Labor mobility, agglomeration mechanisms, 329–331
 - Labor supply
 - immigrants, 645–646
 - initial distribution, 647–648
 - national/local area level, 644–645
 - native population, 647–648

- Labor supply (*Continued*)
 - selection, 647
 - skills, 646, 647
 - supply function, 646
 - Land use regulation
 - commercial activity, 515–516
 - commuting costs, 518
 - exclusionary zoning, 521
 - externalities, 516–517
 - fiscal federalism, 520–521
 - intensity, development, 517, 518
 - land development, 515
 - market equilibrium, 516
 - road congestion, 519–520
 - second best, 519
 - Tiebout model, 521
 - zoning and restrictions, 515, 520
 - Learning effects
 - agglomeration economies, 249
 - localization effects, 265
 - technological spillovers, 262–263
 - Linear models
 - exogenous contextual effect, 585–586
 - peer-effect model, 584
 - reflection problem, 584–586
 - social multiplier, 585–586
 - Linear spatial models
 - description, 124–125
 - endogenous effects, 128
 - geographical space, 126
 - neighboring individuals, 125
 - social interactions, 128
 - spatial econometrics, 127–128
 - vector of characteristics, 125–126
 - Localization economies, 176–178, 202, 205–207, 208, 210
 - Local public good valuation
 - air quality, 443–444
 - improvements, 444–445
 - local environmental quality, 445
 - metropolitan area, 442–443
 - sorting, 443
 - Locational equilibrium models
 - Ceteris paribus*, 412
 - crime, 409–410
 - environmental trends, 413–421
 - household preferences, 409–410
 - pollution, 412
 - price and income effects, 412–413
 - quality of life, 421–423
 - school quality, 409–410
 - single-crossing property, 411
 - sorting, 410
 - spatial heterogeneity, 411
 - Location choices
 - endogeneity, 286
 - firms (*see* Firms)
 - individual, 250–251
 - intertemporal, 268–270
 - workers, 250, 340
 - Log-normal distribution, city size distribution, 195–197, 196f
 - Long-term distribution, city size distribution, 375
- ## M
- Marginal treatment effect (MTE), 10
 - Marshallian approach, 474–475
 - Marshallian equivalence, 386
 - Marshallian spillovers, Marshallian broad families, 328
 - Matching methods
 - agglomeration mechanisms, 329–331
 - data-generating process, 37
 - difference in differences (DD), 37, 41
 - empowerment zone (EZ), 41
 - Oaxaca–Blinder estimator, 41
 - ordinary least squares (OLS), 38
 - propensity score, 37, 40, 41
 - selection, 38
 - treatment and control group, 38–39, 39f, 41–42
 - MAUP. *See* Modifiable area unit problem (MAUP)
 - Metropolitan statistical areas (MSAs)
 - amenity, 175–176, 177f, 196f
 - incomes and rents, 179f
 - inequality, 184, 185f
 - selection, 182f
 - size distribution, 186f
 - sorting, 180f
 - Migration costs
 - at-risk cities, 439
 - household location choice model, 429–430
 - Mobility, skill categories
 - commuting zones, 666–667
 - education, 666–667
 - ethnic enclave, 666
 - predictions, 665

Modifiable area unit problem (MAUP), 360–361

Monocentric model

- assumptions, 473–474
- bid-rent approach, 475–477
- central business district (CBD), 473
- consumer problem, 473
- dual approach, 477–482
- durable housing, 497–503
- Marshallian approach, 474–475
- quantitative predictions, 482–483
- travel costs and accessibility, 484–486

Monte Carlo simulations, city size distribution, 195

Moving to Opportunity (MTO) experiment, 14

MSAs. *See* Metropolitan statistical areas (MSAs)

N

Neighborhood effects. *See also* Network effects)

- labor market, 147
- quasi-experimental/experimental evidence, 566–570
- structural approach, 574–578

Network effects

- empowerment zone (EZ) program, 613
- formation, 578–579
- labor market, 613–614
- nonmarket interactions, 562
- peers, 614–615
- policy implications, 612–613
- reflection problem, 564
- schools, 615
- social contacts, 563
- social space, 562–563
- spatial mismatch hypothesis, 611
- steady-state equilibrium, 611–612
- workers, 612

Network formation, 590–591

Network peer effects, Add Health data, 594

Nonexperimental evidence

- contextual peer effects, 573
- labor-market outcomes, 571
- learning model, 574
- local spillover effects, 574
- social proximity, 573, 574
- spatial clustering, 571
- spatial data analysis techniques, 573
- spatial datasets and clever identification strategies, 570–571

O

Occupations and tasks, 648–653

Olley–Pakes estimation, 302–303

OLS estimation. *See* Ordinary least squares (OLS) estimation

Open space preservation

- economic development, 417
- external effects, 418
- hedonic methods, 418–419
- monocentric city model, 417–418

Ordinary least squares (OLS) estimation, 137–138, 255–256, 260, 284, 313–314

P

Parcel heterogeneity

- exogenous supply, 493–494
- property, 492–493
- shape, 492

Pareto distribution, city size distribution, 193–195

Policy analysis

- decentralization *vs.* centralization, 95–96
- parameter, 104–105
- productivity, 104
- regional specialization, 103

Price, urban land

- accessibility, 527–528
- commuting, 529
- cross-city heterogeneity, 524
- internal city structure, 528–529
- land use models, 530
- parcel and tract level, 525
- population density gradients, 522, 523, 525–526
- production and consumption, 529
- subcenters, 524–525
- switching regression methods, 526–527

Productivity, local determinants

- developing economies, 311–314
- economies of density, 298–303
- estimation strategies, 298
- firms' location (*see* Firms)
- heterogeneous effects, 303–306
- human capital externalities, 310–311
- industrial employment (*see* Industrial employment)
- market potential, 308–309
- natural experiments, 307–308
- spatial decay, 306–307
- specialization and diversity, 309–310

Public good provision
 budgets, 82–84, 83f
 description, 80
 household sorting, 81–82, 82f, 89–91
 housing supply function, 92–93
 identification, 88–93
 observed and unobserved, 91
 predictions, 88
 preferences, 80–81
 voting, 92
 Public policy, dispersal policies, 660–661
 Public regulatory programs
 cost–benefit analyses, 93–94
 equilibrium, 94
 households, 94
 sorting model, 94

Q

Quality, life
 downtown environmental progress, 422
 gentrification, 422
 Quantile treatment effects (QTEs), 11
 Quasi-experimental/experimental evidence
 hedonic approach, 570
 housing and land prices, 570
 housing investments, 570
 housing projects, 567
 job seekers, 568–569
 labor market, 570
 Moving to Opportunity (MTO) program,
 567–568
 neighborhoods, 568
 network composition and location, 568–569
 refugee resettlement program, 569
 relocation program, 568
 school performance, 569
 sorting, 569

R

Randomization
 data-generating process, 19
 economic models, 15–16
 internal and external effects, 18
 Moving to Opportunity (MTO), 17–18
 population and data-generating process, 16
 randomized controlled trials (RCT), 17
 stable unit treatment value assumption
 (SUTVA), 18–19

students, 19
 treatment intensity, 19
 treatment variables, 16
 Randomized controlled trials (RCT) experiments
 cluster, 162
 identification, 156–157
 instrument, 164
 intervention, 163–164
 linear interactions, 163
 spatial clusters, 161
 RD. *See* Regression discontinuity (RD)
 Regional economics, causal inference (*see* Causal
 inference)
 Regional specialization, 96, 102, 103
 Regression discontinuity (RD)
 average treatment effect (ATE), 54
 boundary discontinuity research design, 59, 60
 design, 53
 implementation, 56–59
 intention to treat (ITT), 60
 local ATE (LATE), 55–56
 randomization, 60–61
 regression kink (RK) design, 62
 selection threshold, 54
 spending, 59–60
 treatment on the treated (TT), 60
 treatment rule, 54
 Regression kink (RK) design, 62
 Research and development (R&D) activity
 academic, 388–389
 advantages, 355
 geographic concentration, 383
 innovation process, 354
 Jaffe, Trajtenberg and Henderson (JTH)
 examination, 380
 location, 359f
 United States Patent and Trademark Office
 (USPTO) patent data, 384
 venture capital (VC) firms, 355
 Residential choice
 Berry–Levinsohn–Pakes instrumentation
 strategy, 78
 decision-theoretical framework, 75
 demand estimation, 77
 discrete choice models, 79
 econometric model, 75
 housing investments, 78–79
 location choice model, 77

- logit model, 75–76
- probabilities, 75
- random coefficient model, 76–77
- Residential land development
 - central city population, 540
 - commuting costs, 538–540
 - decentralization, 541
 - density gradients, 537–538
 - fragmentation, 543
 - leapfrog development, 541–542
 - municipal services, 543
 - physical landscape, 542–543
 - scattered development, 538, 543
 - “sprawl”, 537, 538–539
- Residential sorting
 - cost of time, 533
 - filtering mechanism, 535
 - heterogeneous residents, 530
 - idiosyncratic preferences, 536
 - income, 531, 532–533
 - public transport, 534
 - rich and poor households, 531
 - rich and poor municipalities, 532
- S**
- Selection, agglomeration theory
 - agglomeration economies, 173–174
 - constant elasticity of substitution (CES)
 - illustration, 229–230
 - effects, 233
 - empirical evidence, 233
 - firms, 233
 - full-fledged model, 227
 - individual characteristics and education, 233
 - location choice *vs.* occupation, 232–233
 - monocentric city setting, 230–231
 - quadratic preferences model, 230–231
 - quasi-linear setting, 231
 - simple model, 227–229
 - and sorting, 231–232
 - talent composition, 232–233
 - types, 226–227
- Skill and production
 - CES (*see* Constant elasticity of substitution (CES))
 - “competition and partial effect”, 655
 - “complementarity”, 656–657
 - firms, 656–657
 - high school and college equivalents, 653–654
 - labor markets and employer–employee
 - matching, 656
 - mechanization and automation
 - procedures, 654
 - microdata, 654–655
 - production and labor demand, 637–638
 - scale, local economy, 655–656
- Skills, talented agents, 212–213
- Small Business Administration (SBA), 357, 359
- Social networks
 - best-response function, 610
 - black and white workers, 608
 - central business district (CBD), 601–602
 - core-periphery equilibrium, 610–611
 - dyad partnership, 606
 - endogeneity, interactions, 601–602
 - endogenous network formation, 590–591
 - geographical distance, 608–609
 - homogenous population, 604
 - interior equilibrium, 607
 - Katz–Bonacich centrality, 610
 - labor and land market interactions, 600
 - linear-in-means model, 584–586
 - local-aggregate model, 588–589
 - local-average model, 586–588
 - meeting process, 603
 - multiple equilibria, 591–593
 - public good nature, 600
 - random search process, 604
 - social interaction costs, 607–608
 - spatial equilibrium, 601–602
 - steady-state employment rate, 603
 - strength, weak ties, 605
 - strong tie, 605–606
- Sorting, agglomeration theory
 - agglomeration economies, 173–174
 - distribution of talent matters, 222–226
 - firm characteristics, 292
 - heterogeneous agents, 211–212
 - implications, city sizes, 219–220
 - individual unobserved heterogeneity, 298–299
 - limitations and extensions, 220–222
 - and selection models, 211–212
 - simple model, 212–213
 - skills and endogeneity, 306–307
 - spatial equilibrium with continuum of cities, 217–219

- Sorting, agglomeration theory (*Continued*)
 - spatial equilibrium with discrete set of cities, 213–217
 - unobservable worker characteristics, 305
 - workers and firms, 338
- Sorting and spatial unobservables
 - component, 150–151
 - estimation techniques, 150
 - group membership, 149
 - individual-level and city-level fixed effects, 150
 - social network data, 150–151
 - urban economics literature, 150
- Spatial aggregation
 - binary treatment, 20
 - data-generating process, 20–21
 - empowerment zone (EZ), 21
 - equilibrium effects, 22
 - metropolitan area, 22–23
 - stable unit treatment value assumption (SUTVA), 20
 - treatment effect, 21–22
 - urban and regional empirical applications, 23
- Spatial concentration
 - agglomeration, 358–361
 - agglomeration mechanisms, 331–336
- Spatial econometrics literature, 117–118
- Spatially autocorrelated unobservables
 - agglomeration literature, 145
 - “boundary-discontinuity”, 147–148
 - city-level productivity, 146–147
 - endogenous effects, 148–149
 - fixed effects estimator, 146
 - group membership, 145
 - incomplete interaction matrix., 149
 - neighborhood effects, 147
 - standard nonexperimental approaches, 146
 - transformation matrix, 146–147
 - urban economics literature, 146
 - weight matrices, 146
- Spatial methods
 - correlated effects, 137
 - cumulative density functions, 121–122
 - description, 116
 - discretization, 122
 - economics literature, 121, 143
 - geographical information systems (GIS), 122–123, 144
 - hypothesis, 120
 - and identification, 151–152
 - interaction matrix, 141, 145
 - interaction scheme, 118–119
 - interpretation, 132–136
 - measurement error, 143–144
 - neighborhood effects, 141
 - network structure, 143–144
 - nonrandomness, 116–117, 124
 - non-zero-diagonal matrices, 139
 - observation, 116
 - ordinary least squares (OLS), 137–138
 - parameters, 138
 - randomness *vs.* nonrandomness, 120, 120f
 - R&D spending, 118
 - reflection problem, 118
 - simple network, 142, 142f
 - social interactions, 117–118
 - solutions, reflection problem, 140–145
 - spatial autocorrelation, 136
 - structural assumptions, 140–141
 - structural parameters, 136
 - weak instruments problem, 144–145
 - zero-diagonal grouping matrix, 139–140
- Spatial models
 - linear spatial models, 124–128
 - regression, 124
- Spatial *vs.* social interactions
 - econometrics literature, 135
 - parameters, 134
 - sales prices, neighboring houses, 134–135
- Specialization, agglomeration mechanisms, 329–331
- Spearman rank correlation, city size distribution, 195–197
- Stable unit treatment value assumption (SUTVA), 18–19
- Static agglomeration effects
 - advantages, 257, 258
 - dispersion forces, 254
 - and dynamic effects, 263–265
 - equilibrium relationships, 254
 - estimators of parameters, 262
 - heterogeneous effects, 261–262
 - heterogeneous workers, 256–257
 - identification of local effects, 257–258
 - individual characteristics, 256–257
 - individual fixed effect, 262
 - industries, 260
 - local labor skills, 255

- local traffic congestion/housing/land prices, 254
- local unobserved effects, 255–256
- location–industry–time random effects, 260
- locations, movers and nonmovers, 258–259
- market externalities, 261
- market-mediated agglomeration effects, 255
- productivity effect, 252–253
- pure externalities, 261
- skilled workers value, 255
- static gains, 250
- unobserved skills, 256–257
- urbanization economies, 261
- Structural approach
 - boundary discontinuity design, 578
 - employment and unemployment, 575
 - endogeneity, school and neighborhood attributes, 578
 - “fixed agents”, 575
 - identification strategy, 575–576
 - Los Angeles metropolitan area, 576
 - Markov process, 575
 - rich discrete-choice model, 577
 - richer network structure, 577
 - social spillovers, 575–576
 - traditional discrete-choice models, 577
- Structural estimation
 - description, 70
 - economic activity, 74
 - economic model, 71
 - equilibrium, 72
 - fiscal competition, 74
 - full-solution estimation approach, 72
 - model fits, 73
 - model selection and development, 71
 - parametric approach, 73
 - “partial solution” approaches, 72–73
 - prospective analysis, 73–74
 - public good provision, 79–96
 - residential choice, 74–79
 - retrospective analysis, 73–74
 - in urban and regional economics, 74
- “Superstar” green cities
 - coal-fired power plants, 441
 - coastal high-human-capital liberal cities, 441
 - industrial activity and population, 440
 - social multiplier/environmental multiplier, 441
 - state environmental policy, 441
- T**
 - Trade-offs, air quality, 413
 - Total-factor productivity (TFP)
 - agglomeration effects, 302–303
 - elasticities, 303
 - firm, 293, 302–303, 305, 307–308
 - food industry in Chile, 312, 314
 - Olley–Pakes estimation, 302–303
 - static/dynamic specifications, 289–290
 - vs.* wages, 282–284
 - Travel costs and accessibility
 - budget constraint, 484–485
 - first-order importance, 485–486
 - railway stations, 486
 - spatial equilibrium condition, 484–485
 - two-dimensional cities, 486
 - Treatment effects
 - assumption, 155–156
 - average treatment effect (ATE), 11–12
 - binary, 11
 - cluster randomized trials, 153
 - components, 155
 - control group, 153
 - group membership, 154
 - intention to treat (ITT), 13
 - local ATE (LATE), 12–13
 - marginal treatment effect (MTE), 11, 14
 - Moving to Opportunity (MTO), 14
 - policy evaluation, 153–154
 - quantile treatment effects (QTE), 11
 - quasi-random variation, 154–155
 - randomization, reflection problem, 152
 - randomized controlled trials (RCT), 152
 - treatment on the treated (TT), 13–14
 - within-group correlation, 154
- U**
 - United States data, city size distribution, 184
 - United States Patent and Trademark Office (USPTO) data, 355, 357, 360*f*, 384
 - Urban economics toolkit, 351
 - Urban inequality
 - and agglomeration, 236–237
 - causes and effects, 234
 - and selection, 237–239
 - and sorting, 235–236

Urbanization economies, 176–178, 202–203
 Urban land use
 accessibility, 471
 American households, 468
 built-up land, 468–470
 description, 468
 heterogeneity, 471, 472
 integrated summary, 470–471
 monocentric model (*see* Monocentric model)
 neighborhood interactions, 472
 Paris, 468–470, 469*f*
 residential land development
 (*see* Residential land development)

residential sorting, 530–537
 trade-off, 471–472
 urban travel, 472

W

Wages
 compensation model, 254
 computation, 340
 inequalities, 254
 knowledge spillovers, 376–380
 and total-factor productivity (TFP), 250–251,
 282–284
 Weights matrix, 124, 127–128, 131