Integrated Transport & Land Use Modeling for Sustainable Cities

Michel Bierlaire, André de Palma,
Ricardo Hurtubia, Paul Waddell (Eds.)



EPFL Press
Distributed by Routledge

Integrated Transport & Land Use Modeling for Sustainable Cities

Integrated Transport & Land Use Modeling for Sustainable Cities

Michel Bierlaire, André de Palma,
Ricardo Hurtubia, Paul Waddell (Eds.)

EPFL Press
Distributed by Routledge

Routledge
Taylor & Francis Group
www.routledge.com/builtenvironment

Taylor & Francis Group Ltd
2 Park Square, Milton Park
Abingdon, Oxford, OX14 4RN, UK

Routledge is an imprint of the Taylor & Francis Group,
An informa business.

Simultaneously published in the USA and Canada by Routledge,
711 Third Avenue, New York, 10017

www.routledge.com

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

This book is published under the editorial direction of Professor Vincent Kaufmann (EPFL).

The authors and publisher express their thanks to the
Ecole polytechnique fédérale de Lausanne for its generous support
towards the publication of this book.

EPFL Press

is an imprint owned by the Presses polytechniques et universitaires romandes,
a Swiss academic publishing company whose main purpose is to publish the teaching
and research works of the Ecole polytechnique fédérale de Lausanne (EPFL).

Presses polytechniques et univeristaires romandes
EPFL – Rolex Learning Center
Post office box 119
CH-1015 Lausanne, Switzerland
E-mail: ppur@epfl.ch

www.epflpress.org

© 2015, First edition, EPFL Press
ISBN 978-2-940222-72-8 (EPFL Press)
ISBN 978-0-415-72910-9 (Routledge)
Printed in Italy

All rights reserved, including those of translation into other languages. No part of this
book may be reproduced in any form – by photoprint, microfilm, or any other means –
nor transmitted or translated into a machine language without written permission from
the publisher.

Contents

Contents v

1 SustainCity: Overview and introduction 1
Paul Waddell, André de Palma, Michel Bierlaire, and Ricardo Hurtubia
1.1 Introduction . . . 1
1.2 Challenges for integrated planning and modeling . . . 4
1.3 Organization of the book . . . 12

I Introduction 15

2 Agent-based land use transport interaction modeling: state of the art 17
Christof Zöllig Renner, Thomas W. Nicolai, and Kai Nagel
2.1 Introduction . . . 17
2.2 A framework of urban systems . . . 18
2.3 Modeling the urban system . . . 20
2.4 Overview of LUTI models . . . 22
2.5 A comparative overview of ILUTE, ILUMASS, and UrbanSim . 23
2.6 ILUTE . . . 30
2.7 ILUMASS . . . 33
2.8 UrbanSim . . . 36
2.9 Conclusions . . . 39

3 Microsimulation for land use modeling: implementation challenges 41
Ricardo Hurtubia and Michel Bierlaire
3.1 Introduction . . . 41
3.2 UrbanSim overview . . . 43
3.3 Implementing UrbanSim . . . 45
3.4 Brussels case study . . . 51
3.5 Conclusions . . . 53

4 Spatial challenges in the estimations of LUTI models: some lessons from the SustainCity project **55**
Isabelle Thomas, Camille Cotteels, Jonathan Jones, Alain Pholo Bala and Dominique Peeters
4.1 Introduction . . . 55
4.2 How to represent the studied urban environment? . . . 56
4.3 What kind of bias can be expected in the econometrical estimations? . . . 63
4.4 Choosing is biasing? Biasing is misinterpreting? . . . 69
4.5 Conclusion . . . 72

II Modeling and methodological contributions **75**

5 Methodologies for synthesizing populations **77**
Bilal Farooq, Kirill Müller, Michel Bierlaire, and Kay W. Axhausen
5.1 Introduction . . . 77
5.2 Literature review . . . 78
5.3 Methodology . . . 81
5.4 Data . . . 86
5.5 Application and results . . . 91
5.6 Discussion . . . 93

6 Simulation based generation of a synthetic population for Brussels **95**
Bilal Farooq, Ricardo Hurtubia, and Michel Bierlaire
6.1 Introduction . . . 95
6.2 Data . . . 96
6.3 Population generation . . . 103
6.4 Results . . . 107
6.5 Concluding remarks . . . 110

7 Agent-based microsimulation of population dynamics **113**
Lorenzo Turci, Sophie Pennec, Laurent Toulemon, Arnaud Bringé, Rodolfo Baggio, and Elisabeth Morand
7.1 Introduction . . . 113
7.2 State of the art . . . 114
7.3 Demographic model for the SustainCity project . . . 121
7.4 Parameters estimation and calibration for the Paris case study 125
7.5 Results . . . 131
7.6 Conclusion . . . 134

8 Modeling real estate investment decisions in households 137
André de Palma, Matthieu de Lapparent, and Nathalie Picard
8.1 Introduction . . . 137
8.2 A microeconomic model of residential choices . . . 140
8.3 Application . . . 148
8.4 Conclusion . . . 159

9 Intra-household decision models of residential and job location 161
Nathalie Picard, André de Palma, and Ignacio A. Inoa
9.1 Introduction . . . 161
9.2 Literature review . . . 163
9.3 Modeling intra-household decisions in two specific examples . . . 166
9.4 Conclusion . . . 179

10 A real estate development model with heterogeneous agents 181
Christof Zöllig Renner, and Kay W. Axhausen
10.1 Introduction . . . 181
10.2 Literature review . . . 182
10.3 A concept for a real estate development model with heterogeneous agents . . . 186
10.4 Interviews with developers in Zürich . . . 187
10.5 Real estate development models considering heterogeneity of developers . . . 191
10.6 Conclusions . . . 198

11 Modeling the life-cycle of firms and its effect on relocation choice 201
BR Bodenmann and KW Axhausen
11.1 Introduction . . . 201
11.2 Conceptual framework and data . . . 202
11.3 Events in the life-cycle of a company . . . 209
11.4 Conclusion . . . 217

12 Econometric methods for land use microsimulation 219
Constantinos Antoniou, and Nathalie Picard
12.1 Introduction . . . 219
12.2 Suitable techniques . . . 221
12.3 Models to be estimated . . . 231
12.4 Conclusions and recommendations . . . 234

13 Spatial issues in econometric analyses: a hedonic estimation of rents in Brussels **237**
Alain Pholo Bala, Dominique Peeters, and Isabelle Thomas
13.1 Introduction . . . 237
13.2 Delineation of the study area and basic spatial unit . . . 239
13.3 Data description . . . 242
13.4 Estimation strategy . . . 244
13.5 Results . . . 246
13.6 Conclusion . . . 256
13.A List of dwelling structural characteristics . . . 259
13.B Further details on the econometric strategies . . . 260
13.C Two-stage estimation with fixed effects: first stage . . . 263

14 Incorporating equilibrium aspects in microsimulation models **265**
André de Palma, Stef Proost, and Saskia van der Loo
14.1 Model description . . . 267
14.2 Equilibrium without government investments in transport . . . 273
14.3 Optimum with transport investment . . . 283
14.4 The developers, the transport agency, their expectations, and the resulting equilibrium . . . 288
14.5 Concluding comments . . . 290
14.A Chapter appendix . . . 291

15 Indicators of sustainable development for microsimulation models **293**
Stef Proost, Saskia Van der Loo, Constantinos Antoniou, and Dimitrios Efthymiou
15.1 Introduction . . . 293
15.2 What is meant by sustainability and what is a sustainable city? 294
15.3 The economic approach to sustainability . . . 296
15.4 Operationalization at the city level . . . 298
15.5 Evaluating sustainable urban policies with a microsimulation model . . . 300
15.6 Overview of the indicator module . . . 302
15.7 Conclusion . . . 305

IIIIntegration of Transport and Land Use Models **307**

16 Agile modeling: adapting UrbanSim to the European context using the Open Platform for Urban Simulation **309**
Paul Waddell, Liming Wang, Hana Ševčíková, and Alan Borning
16.1 Introduction . . . 309

16.2 OPUS architecture 312
16.3 Extending OPUS 323
16.4 Adaptations for SustainCity applications 326
16.5 Status and further development 329

17 Integration of agent-based transport and land use models 333
Thomas W. Nicolai, Kai Nagel
17.1 Introduction 333
17.2 MATSim 334
17.3 Special MATSim features used for the case studies 339
17.4 Integration approach to use MATSim as a travel model plugin to UrbanSim 341
17.5 Access and accessibility indicators 347
17.6 Some accessibility results 351
17.7 Computational issues 351
17.8 Discussion and outlook for the SustainCity project 353

18 Integration of dynamic transport models and agent-based land use models 355
André de Palma, Mohammad Saifuzzaman, and Kiarash Motamedi
18.1 Introduction 355
18.2 UrbanSim and METROPOLIS 357
18.3 Integrated model: modeling approach 360
18.4 Structure of the interface 361
18.5 Conclusions 368

IV Case studies 371

19 Integrated land use and transport microsimulation for Brussels 373
Inês Cabrita, Sylvie Gayda, Ricardo Hurtubia, Dimitrios Efthymiou, Isabelle Thomas, Dominique Peeters, Jonathan Jones, Camille Cotteels, Kai Nagel, Thomas Nicolai, and Daniel Röder
19.1 Introduction 373
19.2 The Brussels context 374
19.3 General structure 375
19.4 Data collection 378
19.5 Sub-model estimation results 386
19.6 The MATSim model for Brussels 394
19.7 Validation of the business-as-usual scenario 395
19.8 Definition of the scenarios 398

19.9 Simulation results 403
19.10 Conclusions 409

20 Application of UrbanSim in Paris (Ile-de-France) 413
André de Palma, Nathalie Picard, and Kiarash Motamedi
20.1 Scope and challenges 413
20.2 Introduction 414
20.3 Study area 416
20.4 Data 419
20.5 Data analysis: spatial issues 425
20.6 Model structure 431
20.7 Transportation model 440
20.8 Integrated model of METROPOLIS and UrbanSim 441
20.9 Policy scenarios and results: The Grand Paris project 442
20.10 Conclusions 458

21 Land use and transport microsimulation in the canton of Zürich using UrbanSim 461
Patrick M. Schirmer, Christof Zöllig Renner, Kirill Müller, and Kay W. Axhausen
21.1 Study area and setup 461
21.2 Data processing and data model 465
21.3 Modeling 470
21.4 Scenarios: simulation and results 492
21.5 Summary 506
21.6 Discussion 508

V Conclusion 511

22 Future challenges in transport and land use modeling 513
André de Palma, Michel Bierlaire, Ricardo Hurtubia, and Paul Waddell
22.1 Introduction 513
22.2 Setting the stage 514
22.3 Collecting and processing data 518
22.4 Agent behavior 521
22.5 Integration 524
22.6 Policy Insights 527
22.7 Closing 528

Bibliography 531

Index 577

Contributors

Constantinos Antoniou, Laboratory of Transportation Engineering, School of Rural and Surveying Engineering, National Technical University of Athens, antoniou@central.ntua.gr

Kay W. Axhausen, Institute for Transport Planning and Systems (IVT), ETH Zurich, axhausen@ivt.baug.ethz.ch

Rodolfo Baggio, Dondena Centre for Research on Social Dynamics, Bocconi University, rodolfo.baggio@unibocconi.it

Michel Bierlaire, Transport and Mobility Laboratory, School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne, michel.bierlaire@epfl.ch

Balz R. Bodenmann, regioConcept AG and Institute for Transport Planning and Systems (IVT), ETH Zurich, bbodenmann@regioconcept.ch

Alan Borning, Department of Computer Science & Engineering, University of Washington, borning@cs.washington.edu

Arnaud Bringe, Institut National d'Études Démographiques, bringe@ined.fr

Ines Cabrita, STRATEC, i.cabrita@stratec.be

Camille Cotteels, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, camille.cotteels@uclouvain.be

Matthieu de Lapparent, Université Paris-Est, IFSTTAR & École Normale Supérieure de Cachan, matthieu.de-lapparent@ifsttar.fr

André de Palma, Centre d'économie de la Sorbonne, École Normale Supérieure de Cachan and CECO, École Polytechnique, K. U. Leuven andre.depalma@ens-cachan.fr

Dimitrios Efthymiou, Laboratory of Transportation Engineering, School of Rural and Surveying Engineering, National Technical University of Athens, dimefthim@gmail.com

Bilal Farooq, Department of Civil, Geological and Mining Engineering, École Polytechnique de Montréal, bilal.farooq@polymtl.ca

Sylvie Gayda, STRATEC, s.gayda@stratec.be

Ricardo Hurtubia, Institute of Urban Studies and Department of Transport Engineering and Logistics, Pontificia Universidad Catolica de Chile, rhg@ing.puc.cl

Ignacio Inoa, THEMA, Université de Cergy-Pontoise, ignacio.inoa@u-cergy.fr

Jonathan Jones, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, jonathan.jones@uclouvain.be

Kirill Müller, Institute for Transport Planning and Systems (IVT), ETH Zurich, kirill.mueller@ivt.baug.ethz.ch

Elisabeth Morand, Institut National d'Études Démographiques, elisabeth.morand@ined.fr

Kiarash Motamedi, THEMA, Université de Cergy-Pontoise and École Normale Supérieure de Cachan, kiarash.motamedi@u-cergy.fr

Kai Nagel, Transport Systems Planning and Transport Telematics (VSP), Institute for Land and Sea Transport Systems, Berlin Institute of Technology (TU Berlin), nagel@vsp.tu-berlin.de

Thomas W. Nicolai, Transport Systems Planning and Transport Telematics (VSP), Institute for Land and Sea Transport Systems, Berlin Institute of Technology (TU Berlin), nicolai@vsp.tu-berlin.de

Dominique Peeters, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, dominique.peeters@uclouvain.be

Sophie Pennec, Institut National d'Études Démographiques, pennec@ined.fr

Alain Pholo Bala, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain and Faculty of Economic and Financial Sciences, University of Johannesburg, apholo@uj.ac.za

Nathalie Picard, THEMA, Université de Cergy-Pontoise and CECO, École Polytechnique, nathalie.picard@polytechnique.edu

Stef Proost, Center for Economic Studies, K. U. Leuven, Stef.Proost@kuleuven.be

Daniel Röder, Transport Systems Planning and Transport Telematics (VSP), Institute for Land and Sea Transport Systems, Berlin Institute of Technology (TU Berlin) and Senozon Deutschland GmbH, roeder@senozon.de

Mohammad Saifuzzaman, University of Brisbane and Centre d'économie de la Sorbonne, École Normale Supérieure de Cachan, saifuzzaman@qut.edu.au

Patrick Schirmer, Institute for Transport Planning and Systems (IVT), ETH Zurich, patrick.schirmer@ivt.baug.ethz.ch

Hana Ševčíková, University of Washington, hanas@uw.edu

Isabelle Thomas, Center for Operations Research and Econometrics (CORE), Université Catholique de Louvain, isabelle.thomas@uclouvain.be

Laurent Toulemon, Institut National d'Études Démographiques, toulemon@ined.fr

Lorenzo Turci, Dondena Centre for Research on Social Dynamics, Bocconi University,lorenzo.turci@unibocconi.it

Saskia van der Loo, Center for Economic Studies, K. U. Leuven, saskiavdloo@gmail.com

Paul Waddell, Department of City and Regional Planning, University of California, Berkeley, waddell@berkeley.edu

Liming Wang, Toulan School of Urban Studies & Planning, Portland State University, limwang@pdx.edu

Christof Zöllig Renner, Institute for Transport Planning and Systems (IVT), ETH Zurich, christof.zoellig@ivt.baug.ethz.ch

Acknowledgments

The research presented in this book is the work of the SustainCity project (2010 - 2013), funded by the Socio-economic Sciences and Humanities chapter of the Seventh Framework Programme for Research and Technological Development of the European Commission (FP7-SSH, reference number 244557). The editors and authors thank the European Commission for the funding and support, in particular its research officer Domenico Rossetti-Di-Valdalbero and also the attendees of the SustainCity workshops and conferences for their active critiques and inputs.

Chapter 1

SustainCity: Overview and introduction

Paul Waddell, André de Palma, Michel Bierlaire, and Ricardo Hurtubia

1.1 Introduction

As the rapid pace of urbanization has tilted the balance of the Earth's population to urban places, the challenges of planetary sustainability are increasingly likely to be addressed through efforts to make urban regions more sustainable. The research project that is the focus of this book emerged in response to an increasingly urgent need to develop analysis and decision-support tools that can assist local and regional governments in coordinating the planning of infrastructure investments and urban development to reduce their carbon footprint substantially, while assuring that their prosperity can grow efficiently and equitably.

To address these challenges, an international research team submitted a proposal to the European Commission FP7 program in 2009, entitled *SustainCity*. This book reflects many of the research contributions made over the course of the project, and represents the state of the art in integrated modeling to address sustainability challenges in European metropolitan regions.

The dynamic processes of urban development, transport, and environmental impact result from the cumulative results of decisions of individual citizens, households, firms, developers, and governments. Infrastructure plans and investments are coordinated by one set of institutions, often working at a metropolitan scale, while land use controls such as zoning that restrict development outcomes are generally managed by local governments. Thus there are significant mis-alignments among institutions that influence the conditions for urban development, and complex behavioral processes that emerge from largely independent market choices made by private actors. The major challenges faced in developing integrated models that can be useful as decision support tools to

design policies to improve urban sustainability include the need to adequately represent the heterogeneity of these agents and their choices, to account for the effects of the externalities generated by their decisions, to represent how their individual choices accumulate to shape metropolitan outcomes, and to realistically represent the temporal dynamics of those impacts without relying on strong simplifying assumptions of market equilibrium.

Most prior work in urban modeling is highly aggregate, and emphasized theoretical formulations dating from the early origins of the field of urban economics, that emphasized analytically tractable mathematical models of the monocentric city form that predominated until the 20th century. These modeling approaches, however, leave much to be desired for understanding and modeling complex, polycentric cities that have become the norm since the mid 20th century. Simulation approaches that emphasize modeling individual choices using a microsimulation approach have emerged in recent years in response to these emerging urban processes and spatial dynamics, and these approaches are now becoming more widely adopted in formal planning processes. The SustainCity project leverages these recent developments in agent-based modeling by using the UrbanSim agent-based land use model (Waddell 2002), the MATSim agent-based travel model (Cetin et al. 2002), and the mesoscopic dynamic traffic assignment model, METROPOLIS (de Palma et al. 1997, Kilani & Lindsay 2005), to form a suite of models that have been integrated and applied to simulate sustainability policies in three case study cities in Europe: Brussels, Paris, and Zürich.

The objectives of the SustainCity project are ambitious and important. The research collaborators who coordinated their efforts for this project span much of Europe and also draw partners from the U.S. The researchers represented among the collaborators draw heavily from civil engineering, but also include demographers, economists, geographers and urban planners. Such interdisciplinary collaboration is an important aspect of a project of such an ambitious scope, and involves many challenges of coordinating across institutions, geographies and disciplines. By involving models and researchers working in both the European and North American contexts, we also had opportunities to reflect on the differences in urban development patterns, institutional process, and cultural norms that might influence the design of models tailored to the European context that is the focus of this project.

To set the stage for this book, it is useful to review the original scope and objectives that framed the SustainCity project, which were initially described in the abstract of the research proposal submitted in June, 2009 to the European Commission FP7 program:

Increasing concerns about sustainable development and the growth of urban areas have brought forth in recent years a renewed enthusiasm and need for the use of quantitative models in the field of transportation and spatial planning. This project proposes to improve urban simulation mod-

els and their interaction with transport models. Unified operational models that favour a microscopic approach, such as UrbanSim and ILUTE (Integrated Land Use, Transportation, and Environment Modelling System), have recently gained a lot of interest both in the land use and transport communities. Nevertheless, in their current forms these models still require further development to support a comprehensive analysis of the main environmental and socioeconomic questions of the sustainability of urban growth and the relevant public policies.

The goal of this project is to address the modelling and computational issues of integrating modern mobility simulations with the latest micro-simulation land use models. The project intends to advance the state-of-the-art in the field of the micro-simulation of prospective integrated models of Land-Use and Transport (LUTI). On the modelling side, the main challenges are to integrate a demographic evolution module, to add an environmental module, to improve the overall consistency and, last but not least, to deal with the multi-scale aspects of the problem: several time horizons and spatial resolutions are involved.

Theoretical models (such as the monocentric textbook model) have proven useful in stating normative results, but remain of limited usefulness, when it comes to empirical applications that include any degree of disaggregation. However, they are useful benchmarks. Simulation tools in this case are more suitable for dealing with realistic policy analysis. Integrated urban modelling is still a complex task and developing a complete and reliable tool remains out of reach of many professional organizations. The Open Platform for Urban Simulation (OPUS), is an Open Source modelling platform for a suite of behavioural-based models addressing (housing location decisions, employment decisions, and urban form). UrbanSim is an integrated model system implemented within the OPUS Platform.

UrbanSim has proved to be a powerful and efficient tool in the context of American cities (Waddell, Ulfarsson, Franklin & Lobb (2007), Waddell & Borning (2004)). However, various recent projects conducted in Europe point to the need to adapt the available modeling platform, to develop new tools consistent with the historical, political and social characteristics of European cities and with specific lifestyles and preferences. High-level academic research in transportation, planning, economics and in econometric modelling and in other fields such as econometric, new economic geography, computer science and operations research happen to be particularly relevant for analyzing urban development and building LUTI models. These developments will aim at:

- *Improving the integrated urban model specification, estimation and validation (and therefore the accuracy of predictions),*
- *Adapting UrbanSim to European contexts,*

- *Adding to UrbanSim the simulation of travel and daily behavior,*
- *Adding to UrbanSim a sustainability and environmental module,*
- *Adding a demographic microsimulation module to the integrated platform,*
- *Introducing some elements of collective decision process within households (taking into account potential conflicts of interests among different members of a household).*

These developmentscales will be analyzed, tested and validated for three case studies in Europe. We also intend to promote this new prototype modeling platform UrbanSimE for Europe among the professional and academic communities in order to ease its application by European local and regional communities.

After several years of research effort, these objectives have largely been met. This book reflects much of the product of this research effort. Some things that were not anticipated became more important over the course of the project, such as the challenges in compiling regional databases for the three case studies, and making the data usable for modeling purposes. Other things that initially were assumed were later found to be incorrect, such as the assumption that there would be a need to create a new UrbanSimE software platform to adapt UrbanSim to the European context. And some objectives were not realized as fully as initially anticipated, such as the creation of an environmental module for UrbanSim, or the integration of a demographic microsimulation module into UrbanSim, while more effort was put into advanced econometric guidance and theoretical experimentation. As with any large-scale project of this type, the project adapted over the course of its development, to overcome obstacles and to take advantage of opportunities and research interests of its participants.

Some early success has also been achieved outside of this project in the final, overarching objective, which was to bring the use of integrated urban models into use for actual planning and decision-making in the European context. A clear example of this is the use of UrbanSim to help the Société du Grand Paris evaluate the plans for construction of the Grand Paris Express metro system on urban development and economic agglomeration within the Ile de France region, an applied project which builds on research done before and within the SustainCity project (de Palma, Picard & Waddell 2005, 2007).

1.2 Challenges for integrated planning and modeling

The challenges in integrating land use and transportation planning and modeling are substantial, and have been examined previously in the North American context (Waddell 2011). It is worth reflecting on those challenges in the European context of this project, and assessing the degree to which they apply

or need adaptation to reflect the substantial differences in context, institutions, norms and history. This section draws on the assessment in Waddell (2011), and examines the degree to which these issues also remain relevant in the European context.

1.2.1 Challenges for integrated planning

There are many reasons that integrated land use and transportation planning has not proceeded as fast or as far as some would have expected, not the least of which are that the domain is intensely controversial, involves many institutional and non-institutional stakeholders with divergent values and mandates, and the number and complexity of the issues and inter-dependencies are high. In the United States – to a greater extent than in Europe – there is also an active threat of legal action for failing to comply with the prescriptions of legislation such as the Clean Air Act Amendments (see for example Garret & Wachs (1996), Waddell, Ulfarsson, Franklin & Lobb (2007)), or the new Californian climate change legislation.[1] Given the litigious tendencies in the United States, perhaps these challenges are more pronounced there than in Europe, although the underlying controversies over land use and transportation investments and projects appear to be no less intense in Europe. Challenges to integrated planning of land use, transportation, and emissions may arise from numerous sources.

Conflicting Institutions have responsibility for different aspects of the domain: transportation investments are coordinated by regional agencies such as Metropolitan planning Organizations in the U.S., but cities, counties, the state and even private institutions in the case of privatized toll facilities, and a myriad of public transit agencies all have mandates and control different aspects of the transport system. Land use plans are controlled by cities and within unincorporated areas by counties, as delegated by individual states. Air quality is controlled by a state air quality board. Other infrastructure investments that interact with transport, such as water and sewer, are generally controlled and managed by special purpose districts, although they can in some cases be controlled by local jurisdictions. While the institutions have different names in the European context, and there may be a greater role for national governments to guide metropolitan planning and investments, it remains the case in Europe that land use planning is heavily controlled by local jurisdictions. So the underlying conflicts among institutions remains very much a relevant issue in the European context.

From even this abbreviated description it is abundantly clear that there is an inherent problem that arises from many institutions having narrow and often competing mandates. Coming to agreement on objectives for integrated

[1] AB32 (CEPA 2006) and SB375 (CEPA 2008)

planning, or even on the need to coordinate plans, is a tough sell to these institutions. Policies that might be optimal for a metropolitan region, arising from close coordination among these infrastructure and land policies, will be all but guaranteed to be sub-optimal for each institution individually, given their narrower scope and constituency. In this regard, it is tantamount to asking institutions to act in ways that are not consistent with their mandates and the interests of their immediate stakeholders. This is a fundamental conflict, and until there is a broader framework that pressures all of these institutions to align their interests with broader metropolitan objectives, it is not likely that implementing integrated planning will advance far – no matter what tools are used. In the European context, there is a much higher public appreciation for the role of government than in the U.S., and the potential for national and even European Union action to coordinate policies is more likely than in the U.S. But, as recent election outcomes in Europe demonstrate, nothing is certain about these differences.

Conflicting Values among both institutional and non-institutional stakeholders, including the citizens of a metropolitan region, and diverse groups of advocates for specific objectives that range from environmental protection to limiting the scope of government, pose fundamental challenges that no technical process will resolve magically. Most metropolitan areas operate within a very imperfect democracy, and governance structures to manage the political process in a democratic way for metropolitan regions are rare indeed. Democratic processes are institutionalized at city, county (to a lesser extent), state and federal levels, leaving metropolitan areas generally with no governance structure for democratic decision-making. MPOs, or the earlier Council of Government structures in the U.S. context, have little authority and no directly elected boards to oversee their actions, and with very few exceptions (e.g. Portland, OR and Minneapolis-St. Paul, MN) have no direct service delivery responsibility and no direct tax revenue to use for services. In short, metropolitan governance has little leverage in the U.S. This means that the political will to develop aggressive management and integration of planning objectives must come from the state or federal level, along with means to revise the governance structure to delegate these responsibilities to a metropolitan level. California may be a test case for a movement in this direction, although the process is very early in its implementation at this time. The absence of metropolitan governance structures may pose both a challenge and an opportunity, however, as we explore later in this paper, within this vaccuum there may be opportunites for experimentation, particularly in the use of social interaction technologies over the web, to facilitate democratic discourse and planning. In contrast with the U.S. context, metropolitan scale coordination and planning in Europe may have more authority, at least in principle, to coordinate development, but this varies widely across European nations.

Conflicting Epistemologies surely also factor into the assessment of the problems of integrating planning. Some individuals, groups, and institutions

are most comfortable using Delphi (Linstone et al. 1975) or other more informal forms of soliciting information and consensus from stakeholders. Others are more comfortable using quantitative methods and models, and feel that this is required to provide a common artifact that can distil theory and common understanding into a coherent and systematic representation that can be examined and to some extent tested for validity or realism. These differences are related to the gap between *implicit knowledge* and *explicit knowledge* addressed by (Te Brömmelstroet & Bertolini 2010). In practice, these differences have led to some bifurcation of planning practices into what may be called visioning or sketch-planning, that rely more on participants knowledge and assumptions, on the one hand, compared to integrated models that rely more on scientific knowledge and empirical testing, on the other. Often these divergent approaches will lead to different conclusions about the problem or the solution, as is to be expected from quite different approaches. The real issue may be less whether qualitative or quantitative epistemologies are more or less correct, than whether some combination of them provides sufficient behavioral and empirical validity to become useful as an artifact for facilitating democratic deliberation in a contested domain. It would seem that these epistemological conflicts are fairly common in both the U.S. and European contexts owing, in part, to common cultural and philosophical roots.

Conflicting Policies at the federal level and below do not make the integration of land use and transportation planning any easier in the United States. Although federal transportation legislation indicates that there is a need to assess the impact of transport infrastructure on land use patterns and subsequent induced travel demand and air emissions, there are other requirements that are in direct conflict. For example, the alternative analysis for highway projects that seek federal funding are required to hold land use fixed across the alternatives for a specific highway corridor – directly denying any possible impact of the alternative corridor transportation infrastructure on land use. Presumably this restriction is imposed to reduce the burden of monitoring whether local planning agencies are 'cooking the books' to shift land uses in ways that would be more supportive of the need for the transport project they desire to implement. But clearly this contradicts other federal policy and undermines the capacity of local and metropolitan activity to promote integrated land use and transportation planning. A similar scenario applies to transit planning, where the Federal Transit Administration imposes extremely restrictive rules on how the analysis of alternatives in a transit corridor are to be evaluated – and again undermining the ability to adapt land use to support the proposed transit alternatives. Surely we will need to find ways to open up the planning process on the project evaluation level to be consistent with broader federal guidance on integrating planning efforts. In the European context, there is considerably less federal legislation that directly addresses the need to integrate land use and transportation planning, so in this sense, the U.S. is more progressive on the advancement of policy to encourage or enforce this level of

coordination, and recognizing its importance. On the other hand, there seems to be no evidence that policy conflicts among different governmental agencies are any less likely in the European context.

1.2.2 Challenges for Integrated Modeling

Transportation models have been used for evaluating alternative transportation plans for many decades in both the European and U.S. contexts, despite of their well-known limitations (see for example Beimborn et al. 1996). Integrated land use and transport models are more recent in their emergence, but their development and use has been riddled with challenges. Douglas Lee (1973) wrote a scathing critique of the early efforts to develop large-scale integrated urban models developed up to the early 1970s, some of which had been widely applied in both the U.S. and in Europe. The criticisms outlined "seven deadly sins" of urban models, and largely focused on problems encountered in the models of the late 1960s and early 1970s, including what Lee saw as excessive complexity, excessive data requirements (for the time), and design flaws that he characterized as wrongheaded and overly mechanical. Lee's critique has been widely cited as contributing to a decline in activity on integrated models in the 1970s and 1980s. During this period, only two integrated modeling frameworks found much application in practice: the spatial interaction model system DRAM/EMPAL (Putman 1983), and the spatial input-output model systems TRANUS (de la Barra 1995) and MEPLAN (Echenique et al. 1990). At a conference sponsored by the Travel Model Improvement Program (TMIP) of the Federal Highway Administration in 1995, drawing together practitioners, academics, and consultants involved in the development and use of integrated models, a consensus emerged that many of the problems identified by (Lee 1973) and later updated (Lee 1994) still described the state of affairs in 1995. The challenge was to develop new approaches and tools that would overcome these critiques and find a more widespread adoption.

In the years following the 1995 TMIP conference, there has been a resurgence in interest in integrated modeling in the U.S., bolstered by the mandate of the ISTEA and CAAA legislation that emphasized the need to coordinate land use, transportation, and air quality planning. In short, there was a compelling need for integrated models, an emerging consensus on how to approach the science of developing such models, and a new mandate to pursue their development and adoption. The key challenges still facing integrated modeling and their application may be summarized below.

Transparency - Models will not have credibility as tools for decision support in complex, conflict-laden domains such as land use, transportation, and environmental planning, unless they can be explained with a sufficient degree of transparency. The term 'Black Box' has been used to criticize the lack of transparency in earlier models (Beimborn et al. 1996). This is not to

say that models cannot be sophisticated, or must be trivialized to a very low common denominator that virtually anyone can understand because it is overly simplified. Albert Einstein has been often cited as stating that theory should be as simple as possible, and no simpler. This principle has been called Einstein's Razor, in reference to Occam's Razor, which can lead to oversimplification. In the context of integrated modeling, there are many emerging sketch planning tools that provide simplicity and transparency, but little in the way of validity in representing how the world works. In this sense, there is a conflict between the principle of transparency and that of validity. These issues are appear to be equally relevant in the U.S. and in Europe.

Behavioral Validity - For a model to be credible for use in a contested domain, it must have sufficient *behavioral validity* to be believable as an independent artifact, within some clearly defined scope of applicability. Behavioral validity is a somewhat broader term than theoretical validity, which would be a more purely academic term, and includes more common sense or intuitive understandings of how the world works. There may be conflicts between theoretical validity as assessed by different academic disciplines – for example between economics and sociology or psychology or political science, which use different theoretical frameworks to explain observed phenomena. Further, there may be conflict between local knowledge or intuition of residents or other stakeholders, and theoretical or empirical research. But the most compelling, credible, behavioral theories will tend to be confirmed across multiple dimensions of disciplinary research and local knowledge. Models that lack any form of behavioral and theoretical foundations cannot pass the credibility test, and are not ultimately useful in supporting integrated land use and transportation planning. To make this case clearer, suppose one convenes a set of stakeholders who have only a superficial understanding of the way the real estate market works, and solicit from them a 'model' of a local real estate market, and 'predictions' about the future based on some proposed intervention. It is possible that the reasoning and predictions of such an exercise will be valid, but the odds are low that it would be able to withstand critical scrutiny. This criticism applies also to planning tools that do not have a theoretical or behavioral basis, and which are now increasingly being used to facilitate visioning or sketch planning workshops. Experience to date suggests that behavioral validity is of equal concern in both U.S. and European contexts.

Empirical Validity - Models must be tested against observed data in order to assess their empirical validity. That is, no matter how much or little behavioral validity a model might have, it is not useful unless it can respond to input assumptions and make predictions that reasonably well correspond to observed reality. This is the process that some refer to as *model validation*. While there are many alternative means of undertaking such an exercise, it should be clear that if a model is intended to be used to predict outcomes well into the future, it should be able to capture the essential trends in outcomes

over some period of history. Many models do not do any form of validation, and leave it to the user to simply believe the outputs. This is a dangerous activity, in that many users of such models place undue confidence in the results, simply because they are produced by a model and perhaps by an academic. Such faith may be ill-founded where models have not been adequately validated against observed data over time. Further, the models must be able to withstand scrutiny from the perspective of being free from bias. If they are perceived as having significant biases in their empirical validation, especially over time, they lose credibility. Again, this concern seems to be equally applicable in the U.S. and European contexts.

Ease of Use - Beyond the behavioral and empirical validity conditions outlined above, if a model is inscrutable and too complex to explain, it also will ultimately not succeed in practice. If it requires the developer to provide extensive and ongoing support, with no assimilation of capacity to use and modify the model system by its users, it will be far less compelling than a model that accomplishes this aim. In short, a model system must strive to achieve a threshold of usability that makes it possible for staff within planning agencies to become capable users. The easier the better, since complexity invites errors. But this priority obviously must be traded-off against behavioral and empirical validity. One could make a stunningly easy to use model with absolutely no behavioral or empirical validity. In fact, there are many such tools now available. It seems that this concern has been given more attention in the U.S., but certainly emerged as a consideration in the progress of the SustainCity project in the European context. Access to data, and familiarity with the tasks involved in generating usable data for modeling purposes contributes to the sense of difficulty in modeling.

Computational Performance - If we find a model system that meets the above criteria we are doing well. But it is possible that such a model will have such poor computational performance that it is rendered irrelevant for most applications. Current generation travel models are veering close to and often over this threshold, with run times for a single year in some cases longer than a full day to simulate a day of travel. On the other hand, approaches that emphasize interactive speed for use in stakeholder meetings and other similar workshop settings, are often of the sketch planning type with no behavioral or empirical validity. They are fast, but at the cost of realism. These concerns appear to be equally relevant in the U.S. and in Europe.

Flexibility - It is a truism to admit that we do not yet have in hand the ultimate models that will satisfy users in all cases and for all their desired applications. We continue to see improvement in theory, in empirical methods, in software development techniques, in user interface techniques, in data, and so forth. As these continue to change, they alter the scope of what is feasible, and what is the best available option to develop and use a model to support

decision-making. As a result, models and software platforms that are too rigid become a serious constraint, and limit applicability. Different users will have different data and needs, and it is clear that models need to be adaptable to these conditions if they are to be widely used. This issue emerged in the U.S. over a sustained period of experimentation in research and in practice, and it also became clear that it was essential to the SustainCity project, and for applied work in Europe.

Data Availability and Quality - In practice, probably the most daunting problem in implementing a model is developing the input data for it. In some cases the data requirements are intimidating to even the most intrepid users. But the real concern is that the science and tools to develop data that is usable in modeling are far from addressing the needs of users. Input data such as parcels, employment records, building and price data are notoriously incomplete and error prone. Further, it is quite difficult to integrate them into a coherent database that is internally consistent. A rough approximation is that 75% of the effort and an even higher percentage of the time involved in developing model applications is due to the difficulty of developing the data for the model system. This is a very important obstacle, and must be addressed. Data availability and quality remain substantial concerns in the U.S., especially for fine-grained land use data on parcels and buildings. In Europe, the problems seem to be less about the existence or quality of such data, but about the obstacles in obtaining such data from public and semi-public entities in order to use it for research and planning efforts. Nevertheless, it remains an important obstacle.

Uncertainty - Uncertainty is a topic that has only recently come into the lexicon of integrated modeling, but is increasingly important in making risk assessments of alternative policy choices or infrastructure investments, and to do this effectively, new methods have been recently developed to analyze uncertainty within stochastic simulation models such as UrbanSim (Ševčíková et al. 2007). An example application for this is the assessment the potential replacement of the Alaskan Way Viaduct in Seattle, Washington, which was damaged by an earthquake and needed to be torn down or replaced due to the risk it presented. The alternatives that have been considered to date, ranging from retro-fitting the existing facility, to digging a tunnel to replace it, to using a lower capacity surface boulevard with transit as a less auto-oriented option, all face uncertainty in terms of their costs and benefits. In such circumstances, it is valuable to inform the debates by assessing alternatives in terms of not only the expected values of their impacts, but also the uncertainty of those impacts, so it is possible to determine that there is in fact significant difference between them (Ševčíková et al. 2011). The public is rarely provided information on these aspects of uncertainty, despite being germane for the choice between risky alternatives. In neither the U.S. nor the European context have policy-makers come to fully appreciate the need to examine uncertainty, although this

is beginning to change.

1.2.3 Addressing the Challenges

To address these multi-faceted challenges for integrated planning and modeling, both in the U.S. and in Europe, much has already been done (see Waddell, 2011 for more on prior work on addressing these challenges), and much remains to be done. The SustainCity project is a major international collaborative effort to address a number of these challenges in the European context, with the aim of developing tools that could be used to inform metropolitan planning in terms of its impacts on sustainability.

1.3 Organization of the book

The organization of this book reflects the overall structure of the SustainCity project, and is presented in a way that is intended to facilitate the book's use as a reference. There are four parts to the book:

1. Introduction
2. Modeling and methodological contributions
3. Integration of transport and land use models
4. Case studies

The first part largely explores the state of the art and state of the practice in integrated modeling. The work in this part addresses several of the challenges described above, focusing particularly on specific data challenges that arise in the context of limited access to household level data.

The second part presents a diverse set of contributions that seek to address one or more of the challenges to integrated modeling, particularly econometric and theoretical improvements that make integrated models more behaviorally and empirically valid.

The third part of the book focuses on a set of methodological issues that more directly focus on the integration of land use and transportation modeling, and provide detailed experience with the integration of UrbanSim with METROPOLIS and MATSim.

The fourth and final part of the book examines the three case studies in which the integrated land use and transport models were applied, and attempts to draw insights from these case studies.

The SustainCity project was an ambitious effort, and even if it did not meet all of the high aspirations its collaborators intended, it has made significant contributions to research in integrated modeling, and the contributions will in varying degrees have potential to impact practice. The motivations that led to

the project are substantial and persistent, and we hope will lead to much more research and application of tools of this kind, to address challenges of urban sustainability in Europe and elsewhere. We hope that the experience of the SustainCity project, largely described in this book, will aid others to extend these methodologies and tools to other parts of the world, where the need to understand the complex relationship between transportation systems and land use is critical for efficient urban planning.

Part I

Introduction

Chapter 2

Agent-based land use transport interaction modeling: state of the art

Christof Zöllig Renner, Thomas W. Nicolai, and Kai Nagel

2.1 Introduction

Land use transport interaction (LUTI) models try to capture the feedback between land use schemes and transport to improve urban planning. Transport and land use are influencing each other in a dynamic and complex manner (Alonso 1964, Mills 1967, Muth 1969, Anas 1982, Borzacchiello et al. 2010). Accessibility, provided by transport, influences location decisions of developers, firms and households (Hansen 1959, Moeckel 2006). Conversely, land use determines the need for spatial interaction (Wegener 2004, Strauch et al. 2005, Wegener 2011). Hence, modeling processes of the urban system needs an integrated view of the interactions between land use and transport (Wagner & Wegener 2007). LUTI models are taking those interactions into account. They are increasingly used in impact analysis and the evaluation of transport, land use and environmental policies (Lee 2009). They help planners and decision-makers assessing the impact of infrastructure measures and to improve policy and investment decisions.

The aim of this chapter is to discuss strengths and weaknesses of such integrated models. To do this, first (i) general aspects of integrated models are discussed and (ii) important evolution steps of land use and transport models are summarized. In order to characterize LUTI models (iii) important underlying subsystems are identified and discussed. Next, (iv) based on the review by Wegener (2004) an overview of existing LUTI models is provided. Finally, (v) a selection of disaggregated and agent-based models are discussed in greater detail. These are ILUTE, ILUMASS and UrbanSim. The chapter finishes with the conclusions.

2.2 A framework of urban systems

We can define LUTI models as simplifications of the urban system to do impact analysis and evaluation of policies and planning with respect to defined targets. The main purpose is to understand how land use and transport influence each other and consequently to use the knowledge to explain effects of modifications to the urban system. If we know the sensitivity of households towards a cordon toll, we have a better idea about the consequences on the city center's housing market. The effects can then be evaluated with respect to defined targets to improve planning. To establish a common understanding we describe the object for which the model is built, discuss general aspects in modeling and define the most important terms. This allows us to put LUTI models within this context. The review papers of Iacono & Levinson (2008), Hunt et al. (2005), Wegener (2004), Verburg et al. (2004), Briassoulis (2000) are important references therefore.

The object being modeled with LUTI models is the urban system. It can be framed in seven subsystems depicted in Figure 2.1 (gray boxes). The natural environment, legal regulations, and infrastructures are the stage on which economic actors (persons) or actor groups (households, firms) perform their activities. The natural environment comprises all elements which would also exist without human beings. It is the basis for human activities and in itself an extremely complex subsystem. The anthropogenic parts — which result from human activity — of the urban system are embedded in the natural environment. The legal regulations and infrastructures are the result of societal decisions, e.g. voting on infrastructure projects or approval of land use plans. The actors' activities constitute the system of land use. Their patterns in space and time emerge from the decisions of the socio-economic actors when

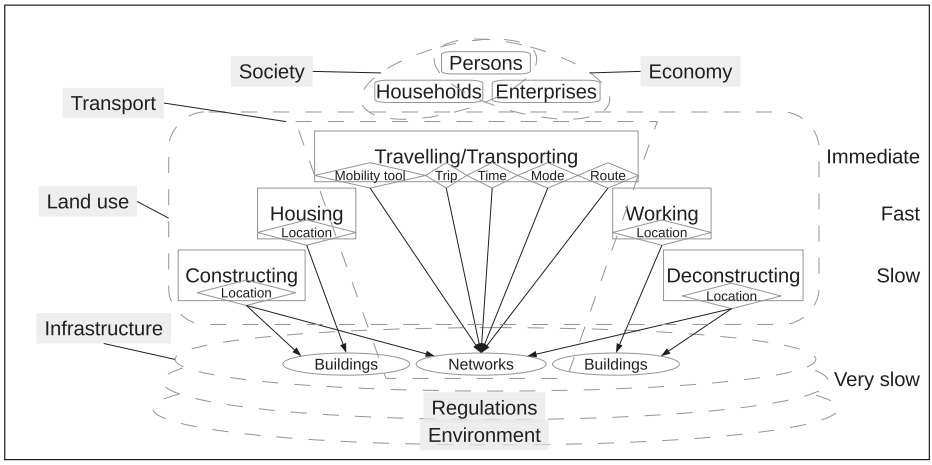


Figure 2.1 A framework of urban systems.

and where to perform an activity. The actors, or agents, consider the conditions of the underlying subsystems for their decisions. These modify again the conditions of the subsystems which means that feedback cycles are in place. The one between land use and transport is discussed in more detail below, depicted in Figure 2.2.

The activities can be seen as decision chains (depicted as rhomboids). Not all decisions are taken equally often. Decisions related to travel are taken rather spontaneous every day whereas the decision on constructing a house is taken much less frequently. Even less frequent are group decisions, e.g. democratic decision processes. Hence, affected subsystems evolve with different speed. This has implications for the modeling and is discussed later. We can distinguish:

- Very slow processes like building infrastructures or modifying a land use plans.
- Slow processes like construction of buildings.
- Fast processes like location choice of households and firms.
- Immediate processes like route choice decisions made on the road.

The activities determine the evolution of the urban system. The resulting processes are related to each other, e.g. the construction of a new residential building gives households new possibilities to reside. However, not all processes are of the same kind. Some are decision processes while others are of a physical nature. Also, the decision processes can be different. One example is the distinction of individual and collective decision making. An example of the latter are discussed in the Chapters 8 and 9 of this book.

In Figure 2.1 we can recognize two main parts of the anthropogenic urban system considered in LUTI models. The spatial distribution of activity places constitute the land use system. This system consists of specialized places which facilitate certain activities like housing, working or traveling. Since specialization often leads to spatial separation, traveling arises as necessity. From this point of view the transport system is the subsystem of land use which relates activity places with each other. Traveling persons and transported goods on transport networks constitute the transport system. The relevance of the transport system within the land use system is the feedback cycle shown in Figure 2.2.

A location's attractiveness depends on how easy it is to reach. Investors anticipate the attractiveness for potential users and are consequently producing new facilities in places with high accessibility. The newly provided facilities get eventually used by the inhabitants of the urban system. Since the performance of transport infrastructures depends on capacity and usage, the attractiveness of a location may decrease if it is more frequently visited.

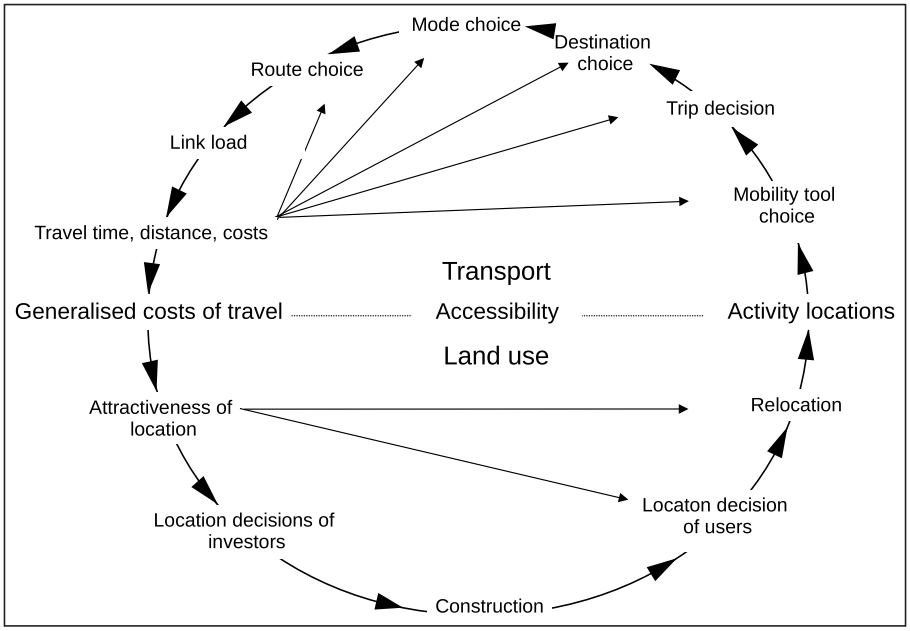


Figure 2.2 Feedback cycle between land use and transport systems adapted from (Wegener & Fürst 1999).

Accessibility is a central measure in this context. Generally, it is a measure of attractiveness of a location. However, there are various definitions in a number of fields (Geurs & van Wee 2004) which shows the generality of this concept. The important aspect in the context of LUTI models is that accessibility measures how opportunities / facilities are related via the transport network to one another. The accessibility measure used should thus be sensitive to variations in generalized travel costs and spatial distribution of facilities.

2.3 Modeling the urban system

A *model* is basically a simplification of reality (Gilbert & Troitzsch 1999). The researcher is the subject who perceives an object and thus creates a model of some form to get a hold of it. The term model is quite general and often more precise terms could be used. The researcher uses different tools for modeling. First of all his brain, but also language, pencils and computers. In analogy we find models in different formulations such as ideas, concepts, theories, physical replications, mathematical formulas or computer programs. These terms are associated to different levels of detail in the formulation of a model.

According to the subsystems and processes discussed before one can distinguish land use transport models and land use transport environment models.

Most LUTI models presented here belong to the first category and focus on the feedback cycle between the two main components. LUTI models are land use change[2] models, which try to explain for what purpose a piece of land is *used*. The terms *interaction* and *change* show that the models do not aim at describing a solid state, but interdependent dynamics, i.e. evolution over time.

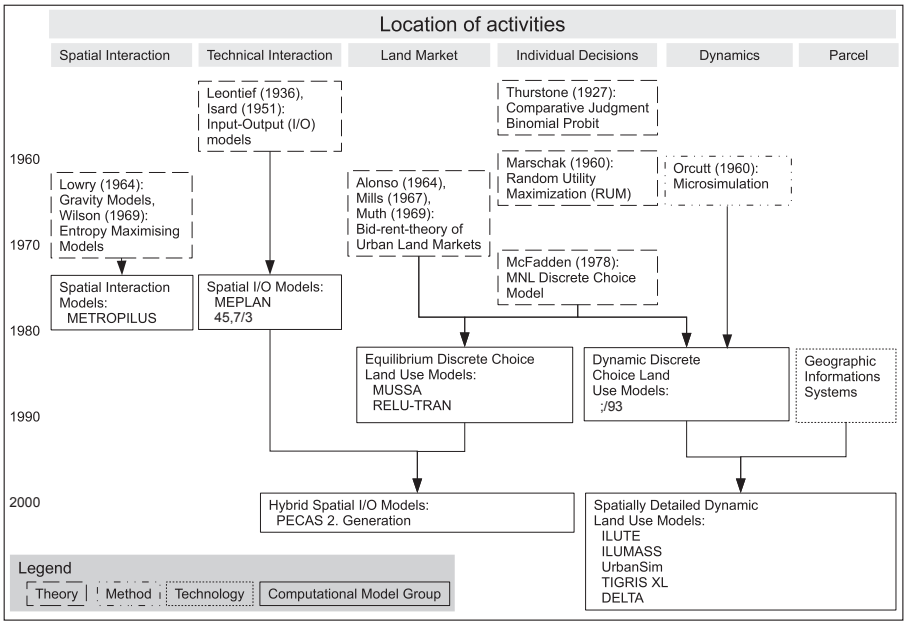


Figure 2.3 Systematics of LUTI models adapted from (Waddell 2005).

Figure 2.3 shows a systematics of LUTI models. The vertical axis is the time line. Aspects of reality which are reflected in the theories and models are distributed horizontally. The rectangles represent elements of different nature as indicated in the legend. An arrow symbolizes *influence*. Representative models of the respective group are shown with capital letters. Predecessors are not included to keep the figure concise.

We recognize four strands of theory which formulate the basic principles underlying LUTI models. The dynamics of the urban system was accounted for by using the method of microsimulation as "a general approach to the study and use of models" (Orcutt 1960, p. 893). The use of parcels became possible through availability of geographic information systems and the according data. Parcels are a meaningful geographic reference because they are the legal entities with highest spatial and regulatory detail.

We can also recognize three waves of development like they have been identified by Iacono et al. (2008): (I) Spatial interaction and spatial input-output

[2] Following the terminology of Briassoulis (2000)

models came up in the sixties, (II) econometric models in the eighties and (III) spatially detailed microsimulation models at the end of the nineties. Each of the three waves has been triggered by theoretical and/or technological developments. The developments in economic theory in parallel with the increasing computational power and data availability lead to this development path.

The first LUTI model was implemented by Lowry (1964) on the basis of gravity theory. The theory was borrowed from physics and stated that the interaction of two regions is proportional to their size (in terms of jobs) and inversely proportional to the distance in between. This model was a start with more inductive than deductive strength.

These first models have been succeeded by econometric models in the eighties (second wave). This is the group of Spatial Computational General Equilibrium (SCGE) models. These work with representative agents (Table 2.2) and explain the interaction of regions based on markets where economic actors make their decisions. Hence, the empirical work required the analysis of more detailed data than aggregates per considered region. Observations of economic actors had to be collected therefore.

The next step followed with the development of discrete choice modeling by McFadden (1978, 1981) and the development of information technologies, which made microsimulation very attractive. HUDS (Harvard Urban Development Simulation) was the first large scale model using microsimulation (Kain 1985). HUDS has been followed by other projects like TLUMIP (Transportation and Land Use Model Integration Project), which further developed TRANUS (Weidner et al. 2007). From the project resulted the Land Use Scenario DevelopeR (LUSDR) (Gregor 2007). Further efforts lead to the development of PECAS (Hunt & Abraham 2003) and UrbanSim (Waddell 2002). This third wave was triggered by the combination of theoretical developments in discrete choice modeling, technical feasibility of microsimulation due to increasing computation power and social relevance, because of pollution and climate debates. The model requirements include ever more ecological indicators like CO_2, land consumption, energy use and air pollution. Up to now, LUTI models treat effects of the anthropogenic system on the natural environment to some extent. The feedback from the natural environment on the anthropogenic system is widely neglected. Coupling a climate model and considering reactions to climate change would be an example of this.

2.4 Overview of LUTI models

This section provides a general overview of operational LUTI models built on Wegener (2004) and Zöllig et al. (2011). We will only look at *operational* models. Operational LUTI models are calibrated for a specific region and ready to use for policy analysis. A non-operational model might be existing

as software or mathematical formulation but it is not yet applied to a specific area. The operationalisation alone can be a work-extensive project (Iacono & Levinson 2008, Gruber et al. 2000) as well as keeping it operational. The number of applications is a proxy for transferability to other locations, which also shows usability and maturity to some extent.

Currently, a number of operational LUTI frameworks exist (Wegener 2004, Zöllig et al. 2011). We can identify two main groups with respect to the aggregation level. The first group, which is of main interest here, operates on the level of individual agents (Table 2.1). These models are also referred to as disaggregate/microsimulation models, multi agent systems (MAS) or agent-based models (ABM). The second group uses representative agents (Table 2.2). Therefore they are also denominated as aggregate models.

The reasons to use disaggregate models are (i) the ability to explain macro level phenomena from a micro level, (ii) the grounding in microeconomic theory, (iii) the capability to represent complex systems, (iv) flexibility with respect to result evaluation (aggregation levels) and (v) flexibility in accommodating different modeling approaches. However, the appropriateness of the model will always depend on the purpose it is used for.

The origins of the models are listed in the tables because the context of development is important given the data dependency of the models. The footnotes of the tables show the references starting with the principal one. Additional references are given for convenience and to show recent activity. Therefore, maintained websites are also included.

2.5 A comparative overview of ILUTE, ILUMASS, and UrbanSim

This section will give a comparative overview of the three model systems UrbanSim, ILUTE, and ILUMASS. The characteristics considered for the selection of these three model systems are their (i) microscopic nature, (ii) capability to interact with a micro simulated transport model, (iii) explicit representation of time (dynamics) and (iv) the representation of social and economic development. The frameworks are *microsimulation* models, which are able to model disaggregated entities such as parcels or persons.

A characterization of the three selected MAS is given in Table 2.3 for easier comparison. The characteristics also serve as structure which is followed in the subsequent descriptions of the three selected models. A general discussion of the characteristics follows hereafter. The notations are derived from Wegener (2004) and Zöllig et al. (2011).

Study regions The number of applications gives an idea on the ease of application of the model. Some of them are applied once, i.e. for one study

Model Name	Place of Origin	Bibliography	Nb. of applications
ILUMASS	Dortmund	Wagner & Wegener (2007), Beckmann et al. (2007), Strauch et al. (2005)	1
ILUTE	Toronto	Salvini & Miller (2005), Miller et al. (2004), Miller & Salvini (2001)	1
LUSDR/TLUMIP	Oregon	Gregor (2007), Weidner et al. (2007)	1
PUMA	The Northern Dutch Randstad	Ettema et al. (2007)	1
STASA	Stuttgart	Haag (1990), Pumain & Haag (1991), STASA (2013)	1
TIGRIS XL	The Netherlands	Zondag (2007)	1
TRESIS	Sidney	Hensher & Ton (2002), Institute of Transport and Logistics Studies (2009)	> 1
UrbanSim	Eugene, Oregon	Waddell (2000, 2002), Waddell & Ulfarsson (2004), ?), OPUS User Guide (2011), UrbanSim Developers (2013)	> 1

Table 2.1 Overview of land use transport interaction models with multiple agents.

Model Name	Place Of Origin	Bibliography	Nb. of applications
BOYCE	Chicago	Boyce & Zhang (1997), Boyce & Bar–Gera (2003)	1
CUFM	California	Landis (1994), Landis & Zhang (1998*b*,*a*)	1
DELTA	London	Simmonds (1999, 2001), Simmonds & Feldman (2005), Bosredon et al. (2009)	> 1
IMREL	Stockholm	Anderstig & Mattsson (1991)	1
KIM	Urbana, Illinois	Kim et al. (1989), Rho & Kim (1989)	1
LILT	Leeds	Mackett (1991)	1
MEPLAN	Cambridge	Echenique et al. (1990), Abraham & Hunt (1999)	> 1
METROPILUS	Ohio	Putman (1996)	> 1
METROSCOPE	Portland, Oregon	Metro Regional Government (2013)	1
MUSSA	Santiago	Martinez (1997*a*), Martínez (1992), Martinez (1997*b*, 2000), Martínez & Donoso (2010)	> 1
PECAS	Oregon	Hunt & Abraham (2003), Abraham & Hunt (2007), Abraham et al. (2005)	> 1
POLIS	San Francisco	Caindec & Prastacos (1995)	1
RELU-TRAN	Chicago	Anas & Liu (2007)	> 1
RURBAN	Tokyo	Miyamoto et al. (1996)	> 1
TRANUS	Carracas	Barra et al. (1984), MODELISTICA (2013)	> 1

Table 2.2 Overview of land use transport interaction models with representative agents.

Criteria	UrbanSim	ILUTE	ILUMASS
Purpose	Assist regional planning at local and state level	Experimental tool for investigating practicability of microsimulation	Develop fully model of microscopic land use, transport and environment
Study regions	world wide	Toronto (Canada)	Dortmund (Germany)
Theoretical Foundation			
Decision rule	RUM	RUM	RUM
Speed of equilibration	lag possible	lag possible	lag possible
Perception	full	asymetric	asymetric
Resolution			
Spatial resolution	grid cells, zones, parcels	grid cells, zones, parcels	grid cells
Temporal resolution	annual	flexible, depending on simulated subsystem	annual
Functioning			
Scope of equilibrium	partial	partial	partial
Simulation of time	iterative	iterative	iterative
Model structure	composite	composite	composite
Transport model	microsimulation	microsimulation	microsimulation
Modelling concept	hybrid	hybrid	hybrid
Usability			
Calibration technique	statistical	statistical	statistical
Calibration scope	model	sub-model	sub-model
Data requirements	micro objects, observed behaviour	micro objects, observed behaviour	micro objects, observed behaviour

Table 2.3 Characterization and comparison of selected LUTI models. This table is built on Wegener (2004) and Zöllig et al. (2011).

region such as ILUTE and ILUMASS. In contrast, UrbanSim has been applied several times to metropolitan areas world wide. UrbanSim is probably the most frequently used microsimulation model. However, the ease of application is also very dependent on data availability and data requirements.

Comprehensiveness The comprehensiveness is shown separately in Table 2.4. The first part of the table shows the comprehensiveness with respect to real world sub-systems, as presented in the framework in Section 2.2. The second part represents the detail with which processes and their interactions are considered. The possible detail in both components is related. For example, preferences of households for upper level living units can only be considered if this information is available in the representation of the building stock.

Criteria	UrbanSim	ILUTE	ILUMASS
Sub-systems			
Persons, households, cliques	yes, yes, no	yes, yes, yes	yes, yes, no
Jobs, firms	yes, yes	yes, yes	yes, yes
Land use	yes	yes	yes
Network, buildings	exogenous, yes	yes, yes	yes, yes
Regulations	yes	yes	yes
Environment	no	no	yes
Processes			
Demography	yes	yes	yes
Firmography	yes	yes	yes
Housing	yes	yes	yes
Working	yes	yes	yes
Travelling	exogenous	yes	yes
Transporting	exogenous	yes	yes
Constructing	yes	yes	yes

Table 2.4 Comprehensiveness of selected LUTI models. The table builds on Wegener (2004), Zöllig et al. (2011).

Modeled Sub-systems What we denominated as society[3] in our framework is modeled as a population of persons. The models also exhibit the structure in the population to some extent. Gender, age, income, and household structure are often considered. The endowment with mobility tools and social networks are less frequently considered despite their importance regarding travel behavior. The economy is represented together with jobs and firms. Again the models differ in the degree of detail in the sense that some represent the structuring of jobs in firms and others do not. In the models the land use sub-system is the outcome of location choices for activities. The available alternatives are

[3] The meaning of the term is very broad and it is thus evident that the models' representation are minimalistic compared to the object in the real world.

buildings/facilities. Thus it is evident that buildings are prerequisites with respect to final activity locations. The transport network is another important infrastructure in the context of transport planning. It is used to calculate more realistic impedances between activity locations which avoids using approximations such as euclidean distances. Regulations are introduced to the model as constraints. An example on the land use side are development constraints which make sure that no residential building is located on a site dedicated for agricultural use. The representation of the environment is reduced to aspects which are found to be influential with respect to modeled decisions. An example is the representation of green spaces or lake view which increase the attractiveness of a location for housing.

Processes The selected models consider all processes distinguished in our framework. An exception is UrbanSim which relies on external transport models. The LUTI models differ in the modeling of the processes constituting the development of the urban system. Some processes are captured by modeling discrete decisions, others are transition models which update certain quantities on the basis of assumed rates. Discrete choice models capture the decision behavior (decision rules and preferences) of the represented actors. Choice models are especially suitable to model markets of discrete goods. Consequently, the theory can be applied to location choice of activities, to travel and transport demand, construction and some parts of demography. In the last case, we think of choice of partners.

There are currently no models which model the modification of regulations and ecosystems. Given the complexity of these processes it is reasonable to work with assumptions.

Theoretical foundations The selected urban simulation models are based on discrete choice theory (Domencich & McFadden 1975), which is used to model the demand side of markets in which goods such as jobs, land, housing or transport services are traded. The usual assumption for the decision rule is *random utility maximization* (RUM). We also find different assumptions about the speed of equilibration in the modeled markets. The models presented in Table 2.3 also allow independent model variables, such as supply and demand, to adjust to equilibrium with some delay. The equilibration process takes multiple time steps in such cases. Microsimulation models can vary in the assumptions about the agents' perception of their environment. In UrbanSim it is assumed that market participants have full information. In ILUTE and ILUMASS the information is assumed to be asymmetric in the markets, that is agents have individual knowledge and search spaces.

Resolution The table shows the supported spatial units of the respective model and the temporal resolution in terms of a typical simulation period. The temporal resolution is one aspect of discrete simulation of time. The temporal

resolution is given on a yearly basis in UrbanSim and ILUMASS. ILUTE allows to specify the scale of a simulation period.

Functioning In all three models a *partial* equilibrium is calculated, i.e. equilibrium is computed separately for each sub-market. This is a fundamental difference to models calculating a general equilibrium, such as the aforementioned SCGE models. Lagged equilibration is also possible if partial equilibrium is calculated.

Dynamic models represent time explicitly, permitting investigation into the speed of effects. Thus dynamic models show the development of the system over time. A dynamic model gives, for example, evidence on the time frame which has to be expected for a desired transformation. Such analysis is not possible with cross-sectional models. The selected models are simulating the evolution of the urban system over time by calculating a sequence of time steps. Modelers discretize the evolution and calculate cross sections of the system in an *iterative* way.

The *model structure* of each framework is classified as *composite*. This means that they consist of loosely coupled sub models, where each submodel has its own independent internal structure (Wegener 2004). A composite structure allows to integrate different model types.

All selected models include or are able to be coupled with a *microsimulation* transport model.

The *modeling concept* distinguishes between input-output-Models (I/O-Models) or multi-agent systems. If both concepts are combined, the model is called *hybrid* (Zöllig et al. 2011). I/O-Models formulate relations between areal units on an aggregate level. MAS explicitly simulate these relations via the behavior of agents in space.

Usability All models are *calibrated* with *statistical* methods on the level of sub models. An approach to calibrate the overall model is only shown for UrbanSim to the authors knowledge (Ševčíková et al. 2007). A possible drawback of microsimulation models is the extensive data requirements. The modeler needs data for the base year, which describe the starting point, and data about the behavior of modeled agents. The latter can be derived from surveys or observations of real world behavior, e.g. route choices can be inferred from GPS data. In many situations, however, it is possible to get a first model from reduced and partially heuristic data sets, which can be extended later if the need arises.

The following sections provide a comprehensive description for each selected model. UrbanSim is selected as reference since it is used in the case studies presented in Chapters 19, 20, and 21.

2.6 ILUTE

ILUTE (Integrated Land Use, Transportation and Environment; Miller & Salvini 2001, Miller et al. 2004, Salvini & Miller 2005) is an integrated, activity-based, microsimulation urban modeling system developed by a consortium of Canadian researchers. ILUTE is built as an experimental tool with the main research objective to investigate to which extent microsimulation can be implemented within a practical model. ILUTE has been applied for the greater Toronto and Hamilton area.

2.6.1 Main components and structure

Four contiguous main components represent the heart of ILUTE: a land use, location choice, auto ownership, and an activity and travel component.

The conceptional design of ILUTE is strongly influenced by the object-oriented software development paradigm. It aims to map objects ("agents") from the real world into the microsimulation model and simulates their activities and interaction. Objects in ILUTE include persons, transportation networks, the built environment, firms, the economy, and the job market.

Market demand and supply interactions are simulated in a disaggregate fashion by a disequilibrium based microsimulation modeling framework for the built space markets (Farooq & Miller 2012). The purchase of a house or car, but also the choice of a spouse or the decision to select a job are considered as market interactions. They have in common that consumers and suppliers interact in a market and try to maximize their individual utility and profit levels. As in the real world consumers and suppliers are assumed to have limited information about the market, e.g. a buyer will not be able to evaluate the entire house market.

The system has been extended for emission modeling to calculate and assess transport introduced air pollution (Hatzopoulou & Miller 2010). A recent effort integrates MATSim to obtain a finer resolution for emission modeling (Hao et al. 2010). A comprehensive description of MATSim is given in Chapter 17.

Multiple temporal and spatial resolutions are supported. In real urban systems time and decision processes are continuous processes that need to be discretized in the modeling system. To simulate change processes of different urban subsystems with different speeds, such as the housing market or transportation infrastructure changes, timestamps are used that are attached to each object. They allow to update and execute an object with its optimal temporal frequency at a flexible time. That is, several levels of temporal aggregation can coexist in ILUTE. Furthermore, different spatial aggregations such as buildings (as a container for activities), parcels, zones (census tracts, traffic zones), planning districts, and grid squares are supported.

The Travel Activity Scheduler for Household Agents (TASHA) represents the conceptual core of ILUTE (Miller & Roorda 2003, Roorda et al. 2009). It is a microsimulation model that generates activity schedules and the resulting travel pattern for every person of a household. The implementation of TASHA is based on three assumptions (Miller & Roorda 2003, pp. 4):

- "[...] scheduling is an event-driven, sequential process [...]"
- "[...] activity/travel scheduling is not an optimizing procedure." This means that the resulting schedule might be sub optimal.
- "[...] travel mode choice (and the associated allocation of household vehicles for individual person travel, as required) is inherent in the activity scheduling process."

Activities are often connected and require the coordination of multiple participants. In ILUTE activities are called projects. A project is defined as a coordinated set of activities that belong together to achieve a common goal (Axhausen 1998). They are used as containers to schedule coordinated joint activities such as work, school, shopping or home-based activities, such as when parents are taking care of their children.

An activity scheduling and mode choice component of TASHA creates activity schedules and travel pattern for each person in a household (Miller & Roorda 2003). The activity scheduling follows a set of rules, according to a predefined priority order, to organize activities into projects. Based on this, schedules for interacting household members are built. This includes: (i) the generation of activities based on observed joint probability distributions for all persons, (ii) the organization of these activities into project agendas, i.e. into a pool of activities, and (iii) the creation of individual schedules built from the project agenda, while taking spatio-temporal constraints into account (Roorda et al. 2009). The mode choice component assigns a transport mode by using a tour based random utility mode choice model that maximizes the overall household utility (Roorda et al. 2009); e.g. in case of overlapping tours of at least two household members, the vehicle is assigned to the person that would get the highest disutility by using another mode.

2.6.2 Key features

A key feature of ILUTE is the representation of compound groups such as (i) households and families, including family relationships as father, mother, spouse and ex-spouse, children and siblings, (ii) businesses and establishments as well as (iii) any collaborating persons that do not belong to the same household.

As in the real-world, in ILUTE decisions can be made by individual persons, or by collectives in compound groups. Individual persons (agents) can take several roles: a person can be a worker, a parent and a property owner at

the same time, and has to take a variety of decisions. In compound groups, location choice and activity scheduling decisions can be traced back to this entity level that provides revealing insights for researchers. An abstraction from real-world entities has been made for businesses (firms), which can take their own decisions and not as a collaborative decision process of employees or firm members. To capture object behaviors, modeling methods such as state transition, random utility, rule-based, learning, exploration and hybrid models are implemented.

Complex decision processes in ILUTE are managed by a flexible and extensible mechanism (Salvini & Miller 2005):

- A time-proxy decision process allows objects to handle future events, such as the purchase of an uncompleted dwelling. This allows anticipatory behavior. This means to make decisions based on speculations. For instance, a family might move into a larger dwelling when they expect offspring.

- ILUTE provides mechanisms to trigger events that are not based on a single stimulus but rather on multiple state changes as a whole. In this context the term stress is used to characterize changes of the current state of a person. Stress occurs if, for instance, the state deviates from a desired, expected or optimal state. The concept of stress is based on a utility based framework. If the stress level reaches a certain level, the stress manager tries to resolve stress by using multiple mechanisms. If, for instance, a long commute causes a serious stress level, this can be resolved by moving into a new home, changing the work place or switching between available transport modes. The stress manager permits the handling of triggered events, joint decisions, accumulated stress, and household interactions.

ILUTE is capable of capturing and tracking complex interactions of the urban system such as activities and behaviors of individual objects, allowing detailed insights for transportation, housing, and urban policy analysis. Therefore, each object in ILUTE has its own, unique representation (perception) of reality. In order to simulate this, an object carries (stores) its knowledge about itself and its environment. The absence of perfect information, e.g. of the road network of a city, is realized by artificially adjusting the access to the true system state, which for performance reasons is only stored once.

2.6.3 Data requirements and preparation

ILUTE provides extensive data synthesis procedures to synthesize households and persons, buildings and dwelling units, the mapping of households to dwellings, and the mapping to work. Moreover, it enables researchers to trace individuals

and to analyze their cumulative effects. To capture complex urban change processes, resulting spatio-temporal data can be visualized in 3D by Houdini 3D from Side Effects Software (Salvini & Miller 2005, Side Effects Software Inc. 2013), a professional animation tool.

2.7 ILUMASS

The ILUMASS project (Integrated Land Use Modeling and Transportation System Simulation) (Strauch et al. 2005, Beckmann et al. 2007, Wagner & Wegener 2007) was under development between 2002 and 2006 by an interdisciplinary consortium of German research institutions. The main purpose of the ILUMASS project was to develop a fully microscopic integrated land use, transport, and environment (LTE) model that helps to explore feasible and successful policies and to achieve sustainable urban transport. ILUMASS has been developed and applied for the urban region of Dortmund, Germany, including the city of Dortmund and 25 surrounding municipalities.

2.7.1 Main components and structure

The three microscopic LTE modules are the main components of ILUMASS. Each module consists of several loosely coupled microscopic sub models with their own independent internal structure. This allows us to capture changes in land use, its impact on activity behavior and transport demand, and the effects of transport and land use on the environment. A comprehensive description is given in Strauch et al. (2005) and Wagner & Wegener (2007). To give a brief overview:

The *land use module* is built on an existing macroscopic urban simulation model of the Dortmund region developed at the University of Dortmund. This is the IRPUD model (Wegener 1982*b*,*a*, 1994, 2004). Its macroscopic modules were re-implemented for ILUMASS in microscopic form. The microscopic land use model consists of the following sub models (Strauch et al. 2005):

- Population: The population model simulates the demographic development of households and persons. This includes the aging of persons, the establishment of new household and their growth, decrease and dissolution over time, e.g. when children are born or a member dies or separates. Also changes in employment are modeled.

- Firms: The firmography, describing the demography of firms, works analogously to the population model. Firms can be founded, they can grow or decline and can be closed.

- Residential mobility: The residential mobility model accounts for decisions of households such as immigration and emigration or just relocation within a region. Relocation decisions are based on the attractiveness of

a dwelling. This includes the attractiveness of a location, dwelling quality, and rent relative to the housing income. A household accepts a new dwelling if it is significantly more attractive than the current dwelling.

- Firm location: Location choices of firms are based on accessibility, size, price, quality, and image. A firm that is unsatisfied with its current location checks out up to ten alternative locations and moves if a selected location provides significant improvements.
- Residential buildings: The residential development model demolishes, upgrades, or builds new buildings for rent or sale.
- Non-Residential buildings: The non-residential development model examines the demand of floor space per zone. If one zone has a low vacancy rate, new floor space is developed. Floor space is subject to land use constraints by municipal land use plans.

Updated locations are provided to the transport model.

The transport module is a detailed agent-based model. It models daily activity pattern, the resulting travel demand and goods movement. For each hour of the day a separate origin destination (OD) matrix is computed. Traffic flows, link loads, and travel times are computed by a dynamic traffic assignment model. A psychological actor model computes a weekly activity plan for each person by considering 29 different activities. These are grouped in 4 main categories: personal, job and school, social activities, and leisure. If an activity is placed outside home, a place, a travel mode and the departure time are selected. When picking a place as well as the capacity limit of the activity location is taken into account. Routes are computed on shortest paths, iterating through all persons until an equilibrium is reached. The goods transport module assigns good flows between companies in the study region. Therefore it takes a macroscopic input-output matrix of the German economy and a German survey on goods transport as input. The generated demand is used in the microscopic travel simulation described above.

The environment module uses travel times, flows, and speeds from the transport model as input, and determines greenhouse gas emissions, air pollution, traffic noise, barrier effects of traffic, and also visual impairments caused by transport and emissions (Strauch et al. 2005).

Each sub module is independent and self executable. The coordination of the simulation procedure and the interaction among the various sub models is managed by a dedicated software component. It executes the modules sequentially and waits with the execution of the next model until required results are calculated by previous modules. The data exchange among the sub models is realized via an integrated database. The data is disaggregated into raster cells of 100 m $\times$ 100 m size. The evolution from the base year data is simulated in annual steps.

2.7.2 Key features

The project contributed to the state of the art in integrated LTE modeling by incorporating three fully microscopic modules for urban land use, transport, and the environment. Distinct features of ILUMASS are the fully microscopic implementation and the consideration of the environment feedback to land use. Shortcomings were found in previous LUTI models in terms of too aggregate temporal and spatial resolutions in order to model sustainable urban transport (Strauch et al. 2005). The latter issue is accounted for by including environmental factors such as clean air or noise in location choices of households and firms (Strauch et al. 2005).

The designers introduce activity patterns in the course of a microscopic implementation of a travel demand model. This is found to be necessary to consider trip chains (multipurpose uni- and intermodal), trips for each hour of a day, the linkage between activity and mobility patterns of household members, emerging lifestyles and work patterns, the interdependencies between travel demand and car ownership, as well as residential and firm location, and also between land use and built form and mobility behavior. This is found to be necessary to overcome issues originating in too aggregate temporal and spatial resolution.

2.7.3 Data requirements and preparation

The base year data incorporates synthetic "populations" of households, businesses, buildings for residential, commercial and public use, vehicles as well as the road and public transport networks (Moeckel et al. 2003, Strauch et al. 2005). Moeckel et al. (2003) present the approach, used in ILUMASS, to generate a synthetic population that is statistically equivalent to a real population.

Households: The synthetic population includes households with individual household members. Households are characterized by household size, income, number of cars, ownership of a monthly season ticket, a car sharing membership and address. Each person is described by age, gender, education, employment status, driving licence ownership, and income.

Firms: Businesses also include public facilities such as schools, hospitals, or museums. They represent the employers and are specified by industry, number and qualification of employers, their capacity to serve customers, the number and type of vehicles and their location.

Buildings: Attributes of residential buildings are the building type, e.g. single or multi family residential housing, size, quality, tenure, and price. The buildings for commercial and public use are described by available and used floorspace for industry, retail, office and public use as well as the price.

2.8 UrbanSim

UrbanSim (Waddell 2000, 2002, Waddell & Ulfarsson 2004, Waddell et al. 2005, OPUS User Guide 2011) is an extensible, agent-based urban simulation model developed by Paul Waddell and his team, first at the University of Washington, Seattle, later at the University of California, Berkeley. UrbanSim was initially developed in 1996 as part of the Transportation and Land Use Model Integration Project (TLUMIP) initiated by the Oregon Department of Transportation (Waddell 2002). In 2005 UrbanSim was reimplemented as part of the open platform for urban simulation (OPUS). In the following, no distinction between UrbanSim and OPUS is made for simplification.

UrbanSim aims at simulating interactions between land use, transportation, economy, and the environment at large-scale metropolitan areas and over a long time span (typically 20–30 years). The motivation for UrbanSim is to assist integrated land use and transportation planning at the regional level within the context of growth management policies carried out at both the state and local level (Waddell 2002). It is designed to explore and analyze the effects of policies at a disaggregated level as a scenario evaluation system (Waddell 2011). It is intended to support modelers and decision makers in government. UrbanSim has been applied for several metropolitan areas such as the Puget Sound Region, the San Fransisco Bay Area and currently, as part of the SustainCity project, the canton of Zürich, the greater Brussels area, and the Île de France.

2.8.1 Main components and structure

UrbanSim is not a single model, rather it is a tool for the integration of several models aimed at the simulation of urban development. UrbanSim mainly consists of six models reflecting the decisions of households, businesses, developers, and governments (as policy inputs) as well as their interactions in the real estate market (Waddell 2002). The responsible models are the Econometric and Demographic Transition Models, the Household and Employment Mobility Models, the Household and Employment Location Models, the Real Estate Development Model, and the Real Estate Price Model. Econometric and Demographic Transition Models, as well as Household and Employment models are independent models: in this chapter these are presented coherently for simplification.

UrbanSim does not model transport itself. To update traffic conditions it relies on the interaction with external transport models (Wegener 2004). As part of the SustainCity project MATSim, an agent-based travel model, and METROPOLIS, a dynamic transport model, are integrated with UrbanSim – detailed descriptions are given in Chapter 17 and 18 respectively. Moreover, external macroeconomic models can be integrated. The scheduling and imple-

mentation of events, meaning read and write access to the database, of these individual model components is managed by a coordinator module.

The UrbanSim models are described in the following according to their processing sequence during the simulation. The sequence of model calls does not necessarily indicate an interaction between contiguous models. A comprehensive model description is provided in (Waddell 2002, 2000):

Accessibility Model: The Accessibility Model is the linkage between land use and transport. It takes the output data provided by the external travel model and maintains accessibility pattern for the internal UrbanSim models. Models that make use of travel model output are the Household and Employment Location Models as well as the Real Estate Price Model.

Econometric and Demographic Transition Models: The Demographic Transition Model simulates births and deaths in the population. These can be specified by providing population control totals, e.g. by income groups or age. In an analogous way, the Econometric Transition Model simulates the creations and losses of jobs. Newly created households and jobs have no location. The location assignment follows later by the household and employment location choice models.

Household and Employment Mobility Models (relocation models): These models simulate whether households or jobs relocate. Such households or jobs are placed in a queue and receive a new location from the location choice models which are described next. If a household or job decides to move their current location becomes vacant. Thus, they change the real estate vacancies conditions, which are used in the real estate development and price model.

Household and Employment Location Models: These models select, in a three step process, a location for each household and job that have no current location. For households, first a random sample of vacant residential units is selected. In the second step the selected units are evaluated for their desirability by a multinominal logit (MNL) model based on the variables and estimated coefficients included in Household Location Choice Model (HLCM). Finally, households pick their most desired location. The Employment Location Choice Model (ELCM) approach is very similar; only the new location of a job is selected randomly among the alternative locations of the random sample.

Real Estate Development Model (developer model): Developer decisions such as new construction, renovation, and reconstruction of existing structures as well as the type of development is simulated by the Real Estate Development Model (REDM). It considers all geographical units of analysis (GUA), e.g. grid cells, for which development is allowed. Each GUA is evaluated by a multinominal logit model for possible development types, including the alternative of no development.

Real Estate Price Model: The Real Estate Price Model (REPM) predicts the prices of each property or GUA based on location characteristics, such as neighborhood accessibility and policy effects. The resulting land values are used as input in the next UrbanSim iteration in the Household and Employment Location Models and the Real Estate Development Model.

2.8.2 Key features

OPUS is a framework for urban land use, transport and environmental modeling and aims to provide a shared platform that can be easily extended by developers or users and adapted for different applications (Waddell et al. 2005). Therefore, the software is released as open source software under the GNU General Public License (GPL) (Waddell et al. 2005). The implementation and maintenance burden of a model infrastructure is taken by OPUS. This approach enables developers and users to focus on experimenting with and applying models (Waddell et al. 2005). Furthermore, OPUS provides an integration of model estimation that allows to keep the model specification consistent between estimation and simulation runs.

The system is easily extensible either by creating an individual OPUS package or by coupling external models via dedicated interfaces (Waddell et al. 2005). This approach eliminates several sources of inefficiencies and inconsistencies such as implementing complex data exchange methods, handling incompatible data formats and software languages, facing problems to access internal algorithms when coupling external models, or adding new OPUS packages (Waddell et al. 2005).

A particular focus of the OPUS framework lies on the computational performance. It is implemented in Python and takes advantage of high performance C and C++ libraries (Waddell 2011, Waddell et al. 2005). Another important aspect of the OPUS software is the usability for a wide group of users and modelers without a profound expertise in software development due to its graphical user interface (GUI) (Waddell 2011). UrbanSim provides various visualization techniques to present model inputs, processes and simulation results as charts and colored static or animated 2D maps (Waddell et al. 2005, Vanegas et al. 2009).

Currently, UrbanSim supports three different geographic units of analysis or GUAs (OPUS User Guide 2011). These are parcels, zones, and grid cells with a configurable resolution. The UrbanSim models simulate the evolution of the data store in annual steps (Waddell 2002).

2.8.3 Data requirements and preparation

The input to the UrbanSim models includes the base year data, access indicators from the external travel model, and control totals derived from external macroeconomic forecast models. In UrbanSim the base year data store contains

the initial state of a scenario (Waddell 2002). It represents chosen attributes of persons, jobs, real estate and locations, and the mapping among these attributes (Waddell 2002). Typically the database includes (i) geographies, initial (ii) household and (iii) job information for a given base year. The geographic layer represents administrative boundaries. Households are represented as individual objects including requisite attributes in order to model location choices. Persons are attached to households and exhibit attributes relevant in terms of travel behavior. Finally, the database includes job entries, incorporating the employment sector, representing employment (Waddell 2002). The primary sources of the base year data are usually surveys or censuses.

If disaggregate information is not available, the population synthesizer in OPUS can be used. The included synthesizer is a snapshot of the PopGen algorithm developed under the SimTRAVEL research initiative (Ye et al. 2009).

2.9 Conclusions

The goal of this chapter is to assess the state of the art of land use transport interaction models. To achieve this goal we presented an introduction to the subject and an overview of operational models on the basis of original articles and review papers. The overview includes aggregated models, as well, for completeness and reference. After, we focused on disaggregated models because of (i) the ability to explain macro level phenomena from a micro level, (ii) the grounding in microeconomic theory and (iii) the capability to represent the evolution of complex systems like cities. A more detailed discussion about added value and challenges of microsimulation is in Chapter 3.

The comparison of ILUTE, ILUMASS, and UrbanSim reveals that even though the purpose of all systems is generally the same, they have different emphases and strengths.

ILUTE is the most detailed system available today, especially in considering interdependencies and structures within society. The aim of the developers to research the practicability of microsimulation approach is recognized.

ILUMASS considers the role of the natural environmental most. The system incorporates explicitly some feedback between land use and the natural environment.

UrbanSim is designed as an open platform for urban simulation. This makes the model flexible, expandable and easily available. UrbanSim includes important tools with respect to practical application such as a graphical user interface, estimation capabilities, data synthesizer, calibration and evaluation tools and interfaces to various transport models as well as data storages. This approach makes UrbanSim attractive for the application in other regions. Therefore, it is today probably the most frequently used microsimulation land use transport interaction model.

Chapter 3

Microsimulation for land use modeling: implementation challenges

Ricardo Hurtubia and Michel Bierlaire

3.1 Introduction

Implementing integrated land use and transport simulators is a complex task that requires to model several groups of agents, urban subsystems and markets, together with their corresponding interactions within the city. Moreover, if the simulation is done at the agent level and with a disaggregate spatial resolution, the complexity is increased due to higher data needs, a larger computational size of the problem, and a bigger set of submodels that need to be estimated.

UrbanSim (Waddell 2002, Waddell et al. 2003) has become one of the most widely used platforms for agent-based land use microsimulation. It has been under development since the late 1990s and it is available as open source software. UrbanSim has a modular structure, with sub-models that can be estimated and run independently and uses a dynamic disequilibrium approach, with all sub-models interacting in a period-based temporal framework.

UrbanSim has been implemented in many cities with diverse results. Most of the applications have been implemented by the developers team, including the cities of Eugene, Oregon; Salt Lake City, Utah; San Francisco, California; Seattle, Washington; and San Antonio, California. Most of these implementations are reported in the literature as producing successful results in terms of accuracy of estimations and market-trend predictions (Waddell, Ulfarsson, Franklin & Lobb (2007), Waddell, Bhat, Eluru, Wang & Pendyala (2007), Waddell, Wang & Charlton (2007)). There is also an increasing number of applications implemented by researchers, urban planning offices, or public authorities outside the UrbanSim core developers team. These include, among others and excluding those reported in this book, the case studies of Paris, France

(de Palma, Motamedi, Picard & Nguyen-Luong 2005); Phoenix, Arizona (Joshi et al. 2006); Volusia County, Florida (Zhao & Chung 2006); Zürich, Switzerland (Buergle et al. 2005, Loechl et al. 2007); Lausanne, Switzerland (Patterson & Hurtubia 2008, Patterson & Bierlaire 2010); Lyon, France (Kryvobokov et al. 2009); Brussels, Belgium (Patterson et al. 2010); Rome, Italy (Zio et al. 2010); and Seoul, Korea (Joo et al. 2011, Hassan & Jun 2011).

The literature reports fewer successful applications in the cases implemented by teams outside of the UrbanSim developers group. The most frequently reported problems are lack of appropriate data for the required disaggregation level, excessive data processing time, partial estimation or implementation of sub-models, and inconsistency of results with observed urban trends. Many of the applications are prototype or proof-of-concept models with simplified datasets and assumptions. Some of the applications report an intensive use of human resources, in some cases excessive considering the quality of the obtained results (Nguyen-Luong 2008).

There is few literature reporting with detail the amount of effort an application of UrbanSim is expected to require. In general, implementation of a fully fledged UrbanSim model can be very expensive in terms of data collection, time and effort for data processing, familiarization with the software platform, model estimation, and implementation of the simulation scenarios. Accounting for the required implementation effort is fundamental for the potential user before deciding which tool, or even which modeling approach, fits best to the objectives and the available resources of a project.

This chapter reports the efforts required for the implementation of UrbanSim for the city of Brussels and provides a description of the steps and processes involved in this task. This is done with the aim of providing guidance to the prospective UrbanSim user, regarding expected costs and development times. The Brussels model was implemented in the context of the *SustainCity* project,[4] financed by the European Research Commission with the objective of adjusting the UrbanSim modeling platform to the context of European cities, focusing on evaluation of sustainability-oriented transport and land use policies. The project considered a simultaneous application of UrbanSim to the cities of Zürich and Paris.

The chapter is organized as follows: Section 3.2 contains an overview of the agents, models, and mechanisms involved in UrbanSim. Section 3.3 describes in the steps that are needed to implement a complete UrbanSim application. Section 3.4 describes the data collection and processing for the Brussels, together with identifying the amount of used resources and implementation difficulties for this case study. Finally, Section 3.5 concludes the chapter.

[4] www.sustaincity.org

3.2 UrbanSim overview

The UrbanSim model implemented for Brussels is a zone-based model, where space is disaggregated in discrete, and relatively large areas. Other types of model consider gridcell or parcel as spatial units, sharing several things in common with the zone version but with some differences on how real estate demand and supply are characterized and simulated. The analysis presented in this chapter applies only to a zone-based model, but it can be easily extended to other versions.

The models in UrbanSim interact in a simulation and data management tool (Open Platform for Urban Simulation, OPUS) that keeps track of the agents' location and real estate evolution through time. Figure 3.1 shows the data structure of UrbanSim and the involved models, that can be classified according to the following models:

- **Transition models** that describe the evolution of the demand agents (households and jobs) for each simulation period.
- **Development models** that describe the generation of new real estate supply (residential and non-residential buildings).
- **Relocation models** where the agents' decision of moving from their current location is simulated.
- **Location Choice models** describing the spatial allocation of new and relocating agents across the different alternatives existing in the city or the spatial distribution of new supply previously defined by the development models.
- **Price models** that compute the prices of real estate good in each period, based on the attributes of the buildings and their locations.

The models interact in each simulation period, generating a new state of the system that is used as a starting point for the simulation in the next period. Many of the attributes that characterize supply and demand are of a dynamic and endogenous nature like, for example, the socioeconomic composition of neighborhoods, the spatial density of types of jobs or the real estate prices. These attributes evolve between periods and are the main input of many the UrbanSim models when predicting the evolution of the city.

Of all the models involved in UrbanSim the most relevant ones are probably the Location Choice models, because they describe the preferences of agents in terms of spatial distribution, therefore giving shape to the city and defining its socioeconomic dynamics and landscape. All location choice models in UrbanSim are based on a "choice" multinomial logit approach, where decision makers (agents) choose from a sample of available alternatives (buildings or locations) selecting the one that provides maximum utility given its attributes and price. Market clearing is treated with a "first come first served" approach

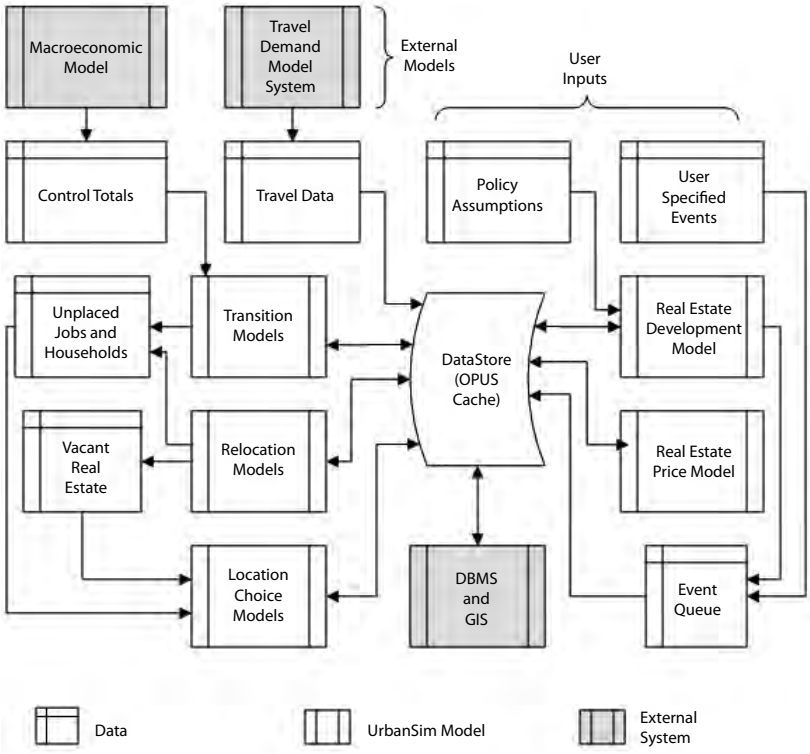


Figure 3.1 UrbanSim data and model structure. Source: UrbanSim Users Guide and Reference Manual (The UrbanSim Project 2011).

(Waddell 2010) meaning that when two agents select the same location conflict is solved by randomly selecting one of them.

Real estate price models are also of high relevance because they describe the market value of the traded goods. In the UrbanSim version used for the Brussels case study, real estate prices are modeled using a hedonic regression of property value per surface unit, on attributes of the building and its environment and market-level vacancy rates (Waddell et al. 2003).

In each period, new households and firms are generated by the Transition models. Simultaneously, new supply is generated by the Development models and distributed within locations in the city. Relocating and new agents enter the market and choose their location following the distribution defined by the Location Choice models. At the end of each period, prices are computed and all location and building attributes are updated to enter as the main input to the next period simulation.

One of the main characteristics of UrbanSim is the independent estimation process for each of the involved sub-models. This is a practical advantage that

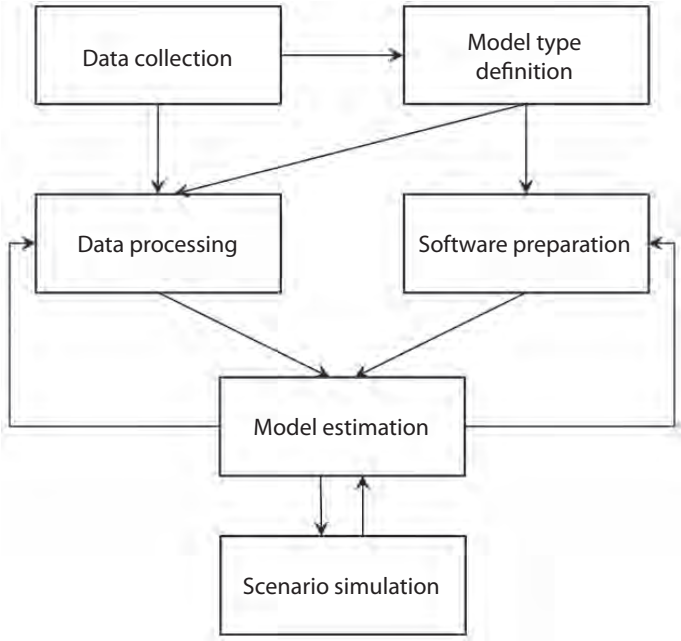


Figure 3.2 Stages of the implementation process.

simplifies the implementation of the model but also implies strong assumptions about the behavior of agents and the interdependence of the decision processes that take place in the city. The modular structure and open source nature of the code makes it feasible to customize UrbanSim for several different circumstances and conditions, although this may require advanced knowledge of the software.

3.3 Implementing UrbanSim

Implementing UrbanSim requires a series of steps to be followed to collect and process data, prepare the software, estimate, and simulate. This section describes the process for the particular case study of Brussels, but it can easily apply to any implementation with similar characteristics. The data requirements analyzed in this section correspond to those required by a zone-based version of UrbanSim coupled with MATSim (Rieser et al. 2007) transport simulations.

Figure 3.2 shows the sequence of steps required to implement an UrbanSim model. Data collection is the initial step where available datasets are identified and gathered. The type of data available will restrict the type of model that can be implemented. For example, if very disaggregate data describing real estate supply is available, a Parcel version can be implemented. If data is of

more aggregate nature, like in the case of Brussels, a zone model is preferable and more feasible. Once the type of model to implement is defined, the data processing can start, following the requirements defined by the type of model. Software preparation and familiarization can be developed in parallel to data processing. After software and data are ready, the model-estimation process can be started. However, it is possible that, during the estimation process, new data requirements are identified or new software issues appear. In this case it will be necessary to re-visit the previous steps several times until all models are estimated and tested. Finally, scenarios have to be defined and run in order to get simulation results.

The steps of the implementation process are described in detail next.

3.3.1 Data collection

UrbanSim requires several datasets to run, the number and complexity of them varying depending on the type of application. The fundamental datasets that are required to estimate and run an UrbanSim application are those describing the basic demand agents (households and jobs) and the basic supply units (buildings). For an extensive list of agents and attributes that may be considered for land use modeling see Hurtubia et al. (2010).

Households must be described at the individual level and exhaustively (all households in the area of study). The basic attributes that should be considered for household characterization are size of the household (number of people), income (or income level), number of cars, and residential location (building type and zone). However, if available, other basic attributes could be considered such as: the age of the head of the household, number of workers, number of children, race and education level, to just mention a few examples. This data is usually found in censal databases but in many cases is not available at the individual level, therefore requiring synthesis of the household population. Population Synthesis procedures like Iterative Proportional Fitting (Beckman et al. 1996) require control totals and aggregate statistics coming from census data and a sample of individual households with detailed attribute information (usually available from OD or household surveys). An alternative way of synthesize populations, using a simulation-based method, has been proposed by Farooq et al. (2013) and is described in Chapter 6 of this book. For a review of different possible population synthesis methods see Müller & Axhausen (2010) and Chapters 5 of this book.

Jobs must be described in terms of economic activity-type and location. This information is usually available at an aggregate spatial level from statistical authorities and tax agencies. It can be easily disaggregated into smaller zones and buildings following some distribution rule like, for example, the availability of workplaces or the built surface by type of economic activity in each zone. UrbanSim also requires the identification of when jobs are home based

or not, and needs information regarding the average built surface consumed by each type of job.

Buildings should be described in terms of type and location (zone), land area used, number of housing units and their average built surface, non-residential surface, and price. Additionally, UrbanSim requires data about residential and non-residential developable capacity (coming from zoning regulations) and historical data of previously built real estate developments.

Besides the main data already mentioned, additional information describing the dynamics of aggregate supply and demand is required. Control totals for households (if possible by categories of attributes) and jobs by type are necessary for all the simulation years. UrbanSim also requires data about structural real estate vacancy by building type, and relocation rates for households and jobs by type. Recent developments incorporate a set of demographic models to UrbanSim in order to simulate the evolution of population within the platform. However, the inclusion of an internal demographic model is very data demanding and was not implemented in the reported application.

Regarding travel data, if an activity-based travel demand model (like MATSim in the case of this application) is to be connected to UrbanSim, additional data about individual persons is required. Persons data must describe, at a minimum level, the residential location and work location of each individual. If a different transport model is available, accessibility data for the area of study at base year must be collected in order to estimate the location choice models using it as explanatory variable.

Finally, additional data describing (non-dynamic) attributes of the locations must be collected in order to characterize neighborhoods or zones of the city. For example the presence of (or distance to) green areas or large sport facilities like stadiums, the distance to the CBD, distance to transportation facilities (airports, train stations) or other specific landmarks.

3.3.2 Model type definition

As mentioned before, the type of model feasible to be implemented will depend on the disaggregation level of the available data. Besides this restriction, the analyst should also consider which type of question will be answered by the model. If the goal is to predict general market trends, a zone or gridcell model should be sufficient. If more detailed and disaggregate analysis is planned, a parcel model could be more appropriate. As a general rule, more disaggregate implementations will be more flexible and allow for a better analysis, but will produce higher cost in terms of data collection, data processing, model estimation, and simulation running time. This trade-off should be taken into account at the time of defining the type of model to implement.

3.3.3 Data processing

UrbanSim models can be adapted to the available data in quite flexible ways. However, if the user only has a basic knowledge of the software and cannot modify the data structure, it is better to process and organize the data in order to fit with the specific format defined by the model templates provided by UrbanSim. In the case study analyzed here, data was adapted to the zone model template (San Antonio) provided with installation of the latest stable release.

The fundamental UrbanSim tables are those related with the basic demand and supply elements (households, jobs, and buildings), these tables concentrate most of the data and require the biggest processing effort (Patterson & Bierlaire 2010).

Synthesis of missing data

Since the level of detail and disaggregation of required data is high, in most cases it won't be directly available from the original data source, either because it was not collected at that level or due to privacy issues. This triggers the need to simulate data, a process that must be done in a way that ensures consistency between all agents and locations.

For example, the already mentioned synthetic population of households must be consistent not only with the real population in term of socioeconomic attributes, but also in terms of location. The households of the synthetic population must be located in one particular residential unit and each unit can't be used by more than one household. This requires to simulate the location distribution of all households for the base year of the simulation.

Another example is the distribution of jobs across buildings. In the reported application, job data was available only at the commune level. Distribution within (smaller) zones and buildings was done through Monte Carlo simulation, following the observed distribution of available non-residential surface by zone and building type.

Adapting data to the platform

In general, land use models require a series of assumptions and simplifications in order to deal with spatial aggregation, and UrbanSim is not the exception. For example, land use regulations and location constraints must be translated into developable capacities by zone and included in the buildings table as available non-residential surface to build and number of possible residential units to build. Similarly, the capacity of non-residential buildings, in terms of the number of jobs they can hold, is defined through the average surface used by each job type in each zone. All these data adaptations require assumptions regarding the average use of land and built space by the agents and the way this

varies across different locations. These assumptions and constraints, however, are less restrictive when using a parcel level model.

Additional data adaptations are required if the implementation is done by adapting a pre-existing model to the new data. Due to inherited model structure, some additional identifiers for agents or supply types and, sometimes, specific variables may be required to run simulations. In the particular case of the Brussels implementation (adapted from the San Antonio model template provided with the UrbanSim stable release), the age of the head of the household had to be included in the households table, even if this variable was not involved in any modeling or simulation process.

Inheriting model structure may also require additional efforts in adapting the travel data. For example, in the San Antonio template, travel data comes in the form of a zone to zone impedance (cost or travel time) matrix. Adapting a different case study to this may be a source of problems, especially if the number of traffic analysis zones of the new transport model is very large, therefore spawning a matrix with a hard-to-manage size. If the transport model does not work with full impedance matrices, as it is the case in some disaggregate transport simulators like MATSim, specific modifications must be introduced to the software. Examples of this can be found in Foti et al. (2012) and in Chapter 17 of this book.

Scenario generation

Sometimes the simulation scenarios that will be run consider specific conditions on supply, demand, or land use regulations. Every pre-set border condition must be explicitly introduced in the corresponding UrbanSim tables that will be considered for scenarios other than the baseline. For example, a particular increase of households in some specific socioeconomic category must be explicitly included in the control totals table while pre-defined (probably large) real estate developments must be included in a "new buildings" table.

3.3.4 Software preparation

After installing the OPUS modeling platform, the analyst needs to familiarize himself with the software and set up the project before any model estimation or simulation is possible. Of all the processes described next, familiarization is the only one that can be performed during (or even before) the data collection and processing steps.

Software familiarization

As any software, UrbanSim requires the user to become familiar with the interface and the coding language used to define modeling elements like, for example, the utility function specifications. This can be easily done following

the tutorials available on-line and by running the example projects provided with the UrbanSim installation package: Eugene (gridcell), Seattle (parcel) and San Antonio (zone). These projects are previously tested and run smoothly, they can be used to practice model estimation and simulation with the latest stable version settings.

Project set up

Setting up a project requires the definition of all the initial settings and relations between sub-models. Every sub-model has a large set of parameters that configures it in terms of sample size for estimation and simulation, tables to consider, dependencies with other models, filtering of alternatives, participating agents, etc. The project settings can be defined either by using a generic template and setting up all parameters from scratch or by adapting a previously existing model. The last option is certainly simpler but it implies inheriting structure and settings from the parent model that can interfere in the estimation and running processes. Additionally, a database can be set up to manage data. UrbanSim supports MySQL and PostgreSQL as platforms for this. For technical details of project and database set up see Gallay (2010).

Database and software debugging

Getting UrbanSim to run is not a straight-forward task. It is very likely that, after initial preparation, many details in the data definitions or structure do not comply with the software requirements. A trial and error process is required before estimation and simulation is possible and some modifications to the original data tables may be needed.

Because of its open source nature, the code behind UrbanSim is constantly evolving in order to deal with new dimensions of the urban modeling problem and to improve its modeling methodologies. For the same reason, troubleshooting can become hard and support from the developers or more experienced users may be needed. Moreover, if modifications to the model structure or simulation algorithms are to be introduced, the developer version should be used, in which case code-writing and software debugging may also be necessary.

3.3.5 Model Estimation and scenario simulation

After setting up the data and the software, model estimation should be straight forward using the OPUS language (based on python) for model specification definition. After all sub-models have been estimated, simulations can be run and, ideally, a validation comparing observed data with forecast results obtained for a baseline scenario should be performed. However, it is likely that both the model estimation and simulation processes will trigger the need to col-

lect and process more data or re-define the way the system is simulated, and this will probably require further software adaptation and debugging. This makes it likely that several steps already described in this section will be, to some extent, simultaneously re-visited during the final implementation of the model. This is important to take into account when scheduling the application project and defining the amount of resources to use in each step.

3.4 Brussels case study

UrbanSim is applied to an extended region around the city of Brussels in the context of the SustainCity project and in parallel with applications for the cities of Paris and Zürich. We focus here on the Brussels application only, but the analysis and conclusion are (keeping proportions) similar for the other case studies. Further details on this case study can be found in Chapters 4, 13, and 19 of this book.

3.4.1 Collected data

The main data sources used for this application were the 2001 Belgium Population Census and the Belgium Land Registry (a cadastre of real estate goods). Both datasets were obtained at an aggregate level from the Belgian Statistical Authority (SPF - Economie[5]). Aggregate data regarding employment by activity type and commune was collected from the ONSS and INASTI databases, from the same source. Additionally, individual level data for households and persons was obtained from the MOBEL travel survey (Hubert & Toint 2002) performed in the area of study during 2002.

Accessibility measures were obtained from an adaptation of the agent-based transport microsimulator MATSim to the OPUS platform (Nicolai & Nagel 2010, Nicolai et al. 2011).[6]

Individual households were generated combining Census aggregate data with individual household level data from MOBEL using a simulation-based synthetic population generation process (Farooq et al. 2013) that ensures consistency with observed marginal statistics by zone. Households are characterized by income level, size, number of children, education level, number of workers, and number of cars categories. For more details on the population synthesis process the reader is referred to Chapters 5 and 6 in this book.

3.4.2 Amount of resources used for model implementation

The development of the model for Brussels, following the steps described in Section 3.3, began in March 2010 with the initial data collection efforts. A si-

[5] http://economie.fgov.be/en/

[6] See also Chapter 17 in this book

multaneous process of data cleaning and preparation took place for a period of approximately 1.5 years. The adaptation of the data into the format required by the platform, model estimation, and simulation preparations took approximately one year until the first simulation run was possible. The process was undertaken by several different PhD students who worked full time on each of the implementation steps in a more or less sequential fashion. This amounts to a total of approximately 2.5 years or 30 persons-month of work. It's important to notice that several processes can't be performed in parallel, meaning that the amount of time required for the full implementation can only be partially reduced by increasing the amount of people working simultaneously on it.

Similarly, in the Paris case study the spent human resources amount to 36 persons-month, mostly by post-doctoral level researchers. This particular case study had a warm start, with a full implementation being previously developed for the same region by the same team. However, extensive data correction, adaptation of new modeling frameworks, model re-estimation and re-calibration was performed. In general, given the size and complexity of this case study, this is a relatively low amount of spent resources. A similar situation took place for the Zürich case study, however using a more disaggregate parcel level model while covering a larger region and, therefore, requiring 40 persons-month in order to achieve a first operational simulation run.

The time and effort required for this application is considerably less that others reported in the literature. For example Nguyen-Luong (2008) reports an effort of four years for a team of four people working full time in a Paris application. Experience is clearly a relevant factor in the amount of time required to implement UrbanSim and implementing a prototype model before the actual application can help to reduce the amount of development time. Patterson & Bierlaire (2010) reports a 5 person-month effort to implement a prototype model, with extremely simplified data and models. It is important to mention that, for the application described here, some of the involved persons already had previous experience implementing basic prototype UrbanSim models and Land Use models in general. It is clear however, that more experienced users (or a larger involvement of the developers team) would have allowed for a much smoother and faster implementation.

3.4.3 Software and modeling difficulties

Despite its rather simple structure of independent sub-models, UrbanSim is complex to implement, basically because it deals with a complex problem. Many details in the way models are estimated and simulations are run can be taken into consideration: sample size, involved agents and datasets, model dependency, definition and sampling of alternatives, data filters, agent group clustering, and data storage options, just to name a few, can be defined by the user. For example, the Household Location Choice Model of the application

described here has more than 50 setting options, with a similar number for other sub-models. Modifying the settings of a model can easily generate errors or internal inconsistencies between submodels that are hard to solve without help from the developers team and, therefore, customizing the model to the analyst's needs can be hard.

Because of its open source nature, the software is constantly and quickly evolving and, therefore, requires more than basic user knowledge from the analyst if the developer version is to be used or if some aspect of the simulation needs to be adapted to particular conditions of the case study. If modifications beyond basic definitions are required like, for example, a change in the behavioral assumptions or type of model for any of the sub-models, the user will certainly need an "almost-developer" level of knowledge or, if possible, intensive and extensive help from the developer team members. Also because of the constant evolution of the software, documentation is usually not up to date. This complicates an overall understanding of the model, sometimes forcing the user to use a "black box" approach when working with it.

Because of these difficulties, the user is likely to end up adapting a pre-existing model to his own case study, meaning that the new application will have to deal with inherited data structures or model settings. This was, to some extent, the case of the Brussels model, which was initially configured using settings from a previous model and was later adapted (with the help of the developers team) to a configuration more specific to the particular elements of the case study.

3.5 Conclusions

This chapter identifies and describes the implementation steps of the a land use model using the UrbanSim modeling platform. The implementation efforts for the Brussels case study, in term of data collection, data processing, software preparation and model estimation are also reported. Implementing UrbanSim for Brussels took approximately 30 persons-month or 2.5 years until a complete simulation was possible. This relatively large amount of effort is basically due to the difficulty to adjust UrbanSim to the existing data structures. This would have required relatively advanced programming skills. Instead, it was decided to prepare the data to pre-existing model implementations and data-structure templates. It is likely that experienced users of UrbanSim, comfortable with modifying the code, would have required a lower level of effort.

Most difficulties during implementation of the model were related to data processing and software preparation. UrbanSim templates requires very detailed data in a specific format that may not be directly available from existing data. Simulation of agent-level data is necessary while data-adaptation, many times using simplifying assumptions, is required to fit the format requirements.

Software preparation is a slow and complex task that requires intensive support from the developers team.

The advantages of using a disaggregate approach could be counterweighted by the fact that data is not always available with the required level of disaggregation to properly estimate all the models. A disaggregate approach allows to simulate very detailed behavior by individual agents but it must be supported by detailed empirical data. Therefore, there is a trade-off between comprehensiveness, level of disaggregation, and implementation cost that should be carefully analyzed before applying this type of model to real case studies, especially when detailed data is difficult to collect.

Chapter 4

Spatial challenges in the estimations of LUTI models: some lessons from the SustainCity project

LUTI *Isabelle Thomas, Camille Cotteels, Jonathan Jones, Alain Pholo Bala and Dominique Peeters*

4.1 Introduction

Land-Use and Transport Interactions (LUTI) models require massive amounts of geo-located data. A model such as UrbanSim can only be efficient if data about the characteristics and behaviour of households, firms (jobs), developers, buildings and infrastructures, but also about transit traffic, visitors, etc., are available and are of sufficient precision on both the statistical and geographical point of view.

Spatial biases (i.e. biases arising from the use of spatial data) are not a new issue in statistical analysis: since Gehlke & Biehl (1934), numerous publications (see Sect. 4.3.1) have shown that estimated parameters of statistical methods are sensitive to both the *size* and *shape* of areal units. Hence, this chapter does not intend to provide a new handbook on quantitative geography, spatial analysis, or spatial econometrics, for which the reader may refer to Fotheringham et al. (2000), Haining (1993), Anselin (1988), Fischer & Getis (2010) and Lesage & Pace (2009). Yet the effects of spatial biases on LUTI models remain largely unexplored and underestimated, not only on the econometric sub models within LUTI models, but also for the results themselves, that is to say on the future situation forecasted by the model. We aim here to remind and illustrate the main challenges arising from the use of spatial data in LUTI models, and to further warn readers and users about their consequences,

such as the dangers of misinterpreting parameters and results and therefore of inadequacy in policy decisions.

One of the main goals of the EU SustainCity research project was to compare the effects of given land-use or transport policies on three case study: Brussels (Chap. 19), Paris (Chap. 20) and Zürich (Chap. 21). To ensure the robustness of such comparison, different challenges related to the geographical space and its representation have to be tackled and controlled for before modelling. An overview of spatial econometrics can be found in Chapter 12 and an application in Chapter 13. Hence, the aim of this chapter is (1) to highlight problems encountered in spatial quantitative analyses when dealing with LUTI models (and UrbanSim in particular), with examples from the project results for the Brussels case study or from a synthetic city specially dedicated for controlling all inputs and their effects (Jones et al. 2013) and (2) to show if (and how far) the three studied cities are, given their differences, comparable in terms of land use and transport modeling. This chapter is divided into sections: Section 4.2 deals with the representation of the urban environment (definition of its limits, size of the basic spatial units, quality of the data). Spatial econometrical biases are then considered in Section 4.3, and Section 4.4 provides a comparison of Paris, Brussels, and Zürich. Section 4.5 concludes the chapter.

4.2 How to represent the studied urban environment?

When modeling spatial realities, one is confronted to the challenge of representing a *continuous space* by a set of *discrete areal units* (e.g. municipalities or census track), automatically leading to statistical artefacts. It impacts several components of the modeling task, including (a) how to delineate the studied urban area, (b) choosing the proper scale of analysis, and (c) handling different types and quality of data. The consequences include the danger of misinterpreting parameters and results of LUTI models, to produce incorrect case study comparisons and potentially wrong planning decisions, because of a spatially biased representation and misunderstanding of the *complex urban interrelationships*. Unfortunately, for most of these challenges, there is no methodological consensus on a solution.

4.2.1 Delineation of the studied area

Defining the spatial extent of an urban agglomeration has led to an impressive amount of scientific literature but also to policy discussions (see for example the cases of Greater London, Barcelona, or Paris) where the rural/urban border is often fuzzy, and there is no consensus about its definition despite it being recognized as politically strategic. *Functional* criteria (e.g. services provided by cities) are commonly used, as well as (separately or simultaneously) *mor-*

phological criteria (extension of the built-up area). See for example Cheshire & Gornostaeva (2002), Parr (2007), Redfearn (2007), Bode (2008), Cörvers et al. (2009), Tannier & Thomas (2013), or the SustainCity report by Pholo Bala et al. (2011). Unfortunately and with some exceptions, *methods* and *criteria* vary from one case study to another, as showed by Dujardin et al. (2007), among others, and so does the *thresholds* and *parameters* values. It implies that comparisons between cities are hazardous (Farmer & Fotheringham 2012). All these choices are often little justified, while it is well known that they influence the number and size of the spatial partitions. Hence, results of urban delineation have to be interpreted carefully, in close relation with the urban context itself. Given the complexity of the urbanization processes, definition and delineation of urban territories may also differ with respect to the goal pursued (transport management, urban design, tax receipts, etc.). Yet, LUTI model are often applied without any questioning about the exact definition of the delineation of the studied city, nor about the link with the "rest of the world" (permeability of the urban border).

The example of Brussels

The capital of Belgium is located in the centre of the country. Since decades the city sprawls out of its official administrative boundary (the Brussels Capital Region or BCR) into two different political regions (Flanders and Wallonia) making the case study interesting scientifically and politically (Riguelle et al. 2007, Thisse & Thomas 2010). Due to the small size of the country and its high built-up densities, the sprawl of Brussels is spatially limited by the proximity to other regional cities and by the existence of a large forest in the south (Van Hecke et al. 2009). Several attempts have been made by geographers for defining the "urban region" of Brussels, i.e. the enlarged area where most of the basic urban function took place: living, working, school, and retail, etc. (Van Hecke et al. 2009). All reports, such as Van Hecke et al. (2009); Dujardin et al. (2007); Vandermotten (2008); and Thomas et al. (2012), lead to a set of communes larger than the BCR (19 communes), but the exact composition depends upon the method and criteria used, and none can be said to be better than the other. Most of these attempts are based on a few justified threshold values (e.g. 650 inhabitants per square kilometre or 15% of the commuters to Brussels) and some combinations of selected variables may lead to major endogeneity problems (in statistical models, a variable is said to be endogenous when this variable is correlated with the error term: this bias arise from a causality loop between dependent and independent variables). Moreover, to be really operational, it should be necessary to regularly update these indicators and thresholds values, but this is far from being the case nowadays; in Belgium, it was - up to now - done after each population census, but with the end of the census, this question remains open.

Innovative methodologies have also been applied (Figure 1). While traditional delineations of the *morphological* agglomeration need a subjective threshold choice (for instance a maximal distance between buildings less than 200 meters), Tannier et al. (2011) use a fractal geometry criteria that has the advantage of not depending upon any threshold choice and to *optimally* define the city delineation. This methodology has been applied to all Belgian cities (Tannier & Thomas 2013), and to Brussels only (Thomas et al. 2012), producing a clear-cut morphological agglomeration different from more traditional partitionings. Another methodological attempt has been proposed by Blondel et al. (2010), with *functional* criteria based on origin-destination matrices: a mathematical method allows networks to be divided into coherent groups in a optimal (modularity criteria) and automatic manner. Figure 4.1 overlays the results obtained for two OD matrices, mobile phone calls (which can be considered as the *"social basin"*) and commuters (*"job basin"*) to the previously defined morphological agglomeration of Thomas et al. (2012). The map shows that there is a well-defined core, but that divergences appear at the outskirts simply because the optimized criterion varies. It is, for instance, well known that rail commuting habits are since decades anchored in the tradition of the western part of the Vlaams Brabant province (Verhetsel et al. 2009), while the phone basin extends more southwards to the French-speaking area.

Finally, the SustainCity research team privileged a study area based on *transport policy criteria*, the RER network (Fig. 4.1). The Brussels Réseau Express Régional, or RER, is a rapid train transport system still under construction. It officially aims at offering fast and frequent public transport connections (to increase rail use by commuters) in an area within 30 km of Brussels. This RER area has been defined by a law (Moniteur belge 2004), without clear justification, nor about the distance threshold, nor the communes considered (sometimes further from the centre, see Fig. 4.1). It constitutes the largest part (126 of 151 municipalities) of the study area used for the SustainCity project (see Chap. 19). Note that the studied area also gathers 25 additional communes (Fig. 4.1), including some belonging to the suburbs of other major urban agglomeration (Antwerpen or Gent) and also incorporating the entire city of Leuven (East).

Hence, there is *no unique delineation* at the exception of the administrative one, which does not anymore correspond to the urban area. If we limit ourselves to the Brussels Capital Region, data collection is simplified (one regional administration involved). But we exclude the periphery of the city, although it is well known that cities need their suburbs and vice-versa (Anas et al. 1998); see also Riguelle et al. (2007) for Brussels. On the contrary, by also studying the suburbs, we open the question of what is the exact delineation of the suburban periphery (where does the city end?) and we complicate the data collection step, as well as the comparability of the variables (3 regional administrations), over time and with other studies.

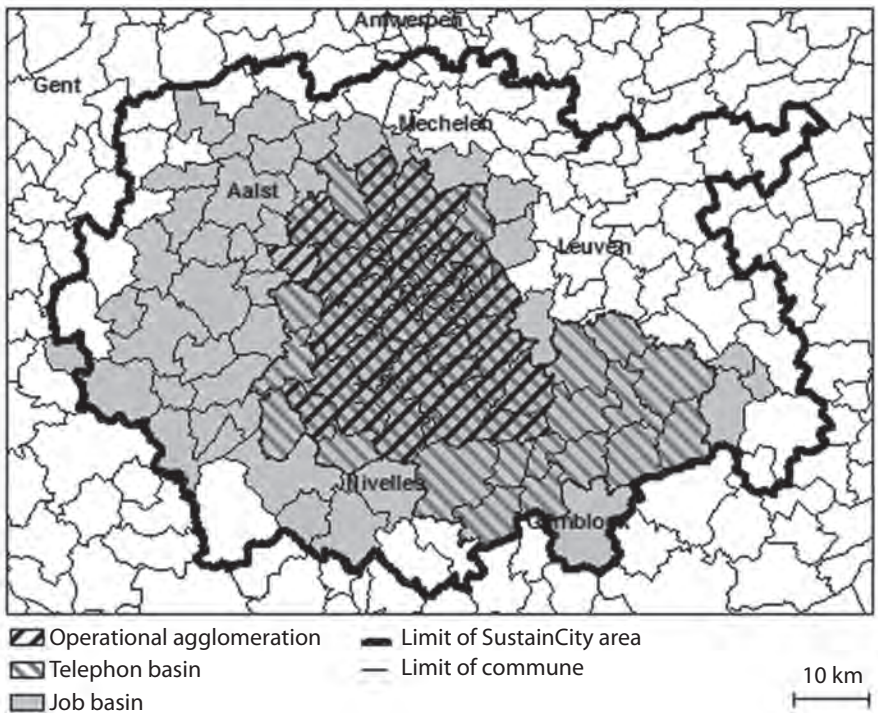


Figure 4.1 Brussels SustainCity area compared to the delineations of Brussels proposed by Blondel et al. (2010) and Thomas et al. (2012).

Consequences for urban modeling

Figure 4.2 shows that the main socioeconomic characteristics of the urban environment dramatically changes with the delineation of Brussels considered. For instance, the BCR concentrates 63% of the population of the fractal morphological agglomeration, 57% of the Brussels' "phone basin", 43% of the "job basin" but only 32% of the RER area. In such a study area, several public transportation modes coexist (train, metro, tram, bus) and are locally anchored, often with different managements. They are run regionally (3 regions) and in some cases are even deserving a city located out of the studied area (Antwerpen, Ghent, Namur). The studied area by the SustainCity project hence has to be considered as a *polycentric urban region* with Brussels on its centre and different secondary cities in the periphery, more than as a genuine urban region.

When extending the size of the study area, the relative importance of the urban core becomes proportionally smaller and – in a dense country such as Belgium – the probability of including secondary urban centers increases, "diluting" the urban reality into a large countryside with loose linkages with the city centre, and where the friction of distance is different. This will affect the

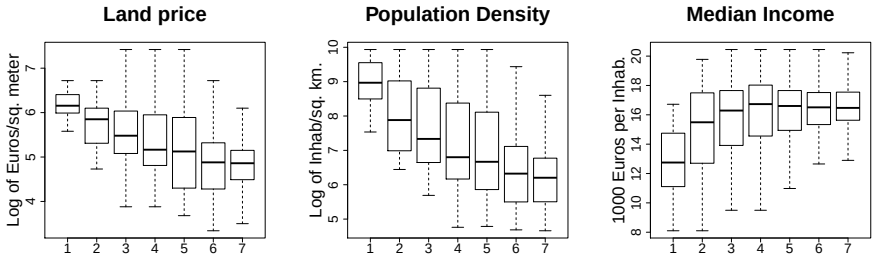


Figure 4.2 Variations of main socio-economic characteristics through delineations of Brussels, where 1 = BCR, 2 = Operational Agglomeration (Van Hecke et al. 2009), 3 = fractal morphological agglomeration (Thomas et al. 2012), 4 = Urban Region (Van Hecke et al. 2009), 5 = Phone Bassin (Blondel et al. 2010), 6 = Job Bassin (Thomas et al. 2012), 7 = SustainCity study area.

nature of the urban model and therefore the UrbanSim parameters estimates. Moreover, by extending the studied area, data collection also becomes trickier because three administrative regions have then to be considered, that is to say three different statistical offices but also three different political instances.

As a corollary, there is the problem of the *border permeability*. If we agree on the definition of a limit to the studied area, we further can question if the city remains open to the "rest of the world". By ignoring relationships with other Belgian cities, we bias the results in terms of all kinds of interactions: social networks, accessibility, agglomeration effects, environmental issues, etc. All these spatial considerations become more critical when comparing cities of different sizes such as Paris, Zürich and Brussels, or when studying very dense urban environment such as Belgium. Brussels, Paris, and Zürich are three cities of different sizes, different administrative constraints, different histories/-geographies, and thus different borders definitions. Urban realities modeled are consequently different, which calls for *cautiousness when comparing results*.

4.2.2 Scale of analysis

Scale is a major issue in geographical analysis and in transportation modeling (Rodrigue et al. 2009). In terms of LUTI application, the scale effect problem can be raised on two methodological issues: First, the three studied cities (Fig. 4.3, Table 4.1) differ in size and, as a consequence, in their position in the urban hierarchy. This means different transportation mode practices, choices and usage, and different development of transportation infrastructures. For instance, trams and metros are more often selected as main transport modes in large cities, while slow modes (on foot, by bike) are naturally privileged in small cities where distances travelled are smaller. The share of the urban center that can be reached in 15 minutes of walk is smaller in Paris than in Zürich,

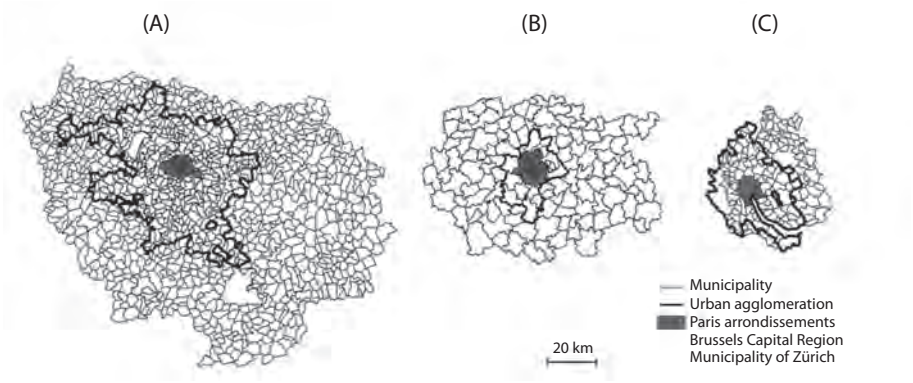


Figure 4.3 Studied areas in the SustainCity project: Paris (A), Brussels (B) and Zürich (C) overlaid by the most common definition of the urban agglomeration.

	Paris	**Brussels**	**Zürich**
Number of communes	1300	151	171
Average surface (sq. km)	*9.3*	*34.2*	*9.7*
Min surface (sq. km)	*0.1*	*1.16*	*1.5*
Max surface (sq. km)	*172.4*	*111.32*	*88.3*
Total Area (sq. km)	12 057	5 170	1 662
% of national territory	*2%*	*17%*	*4%*
Inhabitants (2009)	11 728 000	3 361 000	1 332 000
Jobs (2008)	5 612 000	891 000	812 091

Table 4.1 Comparison of the extension of the studied areas of the three SustainCity case study.

leading automatically to other practices of space. Interurban comparisons of the transport model parameters and outputs are then difficult. The second issue on scale is related to data availability and *basic spatial units*. Ideally, UrbanSim should only use individual (or household) data, measured for each "urban user" and each plot of land. In the real world, however, data are often not available at such a level of spatial detail (privacy reasons) and are approximated by average calculated on *census track* or even *municipalities*, aggregated values replacing individual/household data. Both of these *areal units* vary in *size* and *shape* and the way their boundaries are defined is not controlled for, nor comparable. In the case of Brussels, census data were only made available at the level of communes and ward. Hence actors' behaviours are approximated by averages at the municipalities level which may entails errors in variables and therefore in estimates.

4.2.3 Type, quality, and quantity of spatial data

Geo-located data required by LUTI models are of different nature (*nominal*, *cardinal*, or *ordinal*) and at the level of *points*, *lines*, *surfaces*, or even *volumes* (altitude, height of buildings). As in most econometrical analyses, defining, collecting, and spatially relating data is then amongst the most time consuming tasks in the modeling process. One of the first problems encountered when using UrbanSim for a given city is *data availability*. Let us take the example of Brussels: the 2001 Population and Housing census was the last existing; it has the advantage of covering the entire hinterland of the city for the total population (no sampling). Variables are collected for households, and for individuals and their dwelling. Data definitions are clearly stated with possible comparisons with earlier censuses. However, despite numerous administrative official procedures, it was not made available for the SustainCity project: only aggregated tables by *communes* were obtained. Moreover, after 2001, census data were suppressed and *replaced by independent data files*, not always publically/freely available and often with less information and no relationship between the files. For instance, income or Land Registry data cannot be linked to the characteristics of the households. Some data (especially those on dwellings) are not anymore collected, making an update impossible and reducing the possibility for calibration of the model.

A closely related issue is that, within a given urban area, the same data can have different providers, meaning that managing *data definitions* is a challenge often impossible to overcome. Variables relying on land use planning, for instance, depend on many political institutions and usually cannot be generalized. For the Brussels case study, the BCR, Flanders, and Wallonia have each their own *Regional Master Plan*, defining potential future land uses, but the categories of land uses do not converge, nor do the regulations (maximal number of floors per building for instance) for similar future land use. Definition of land development constraints for the complete study area is thus tricky. Comparison between different data providers is even more difficult when making international comparison or when linguistic differences appear leading sometimes to very subtle differences. Last but not least, each data is expressed in *absolute* but also in *relative* terms. A clear control of both numerator and denominator is hence needed in order to understand the measured relationships. Consider the example of the job's density in a zone i: is the numerator equal to the number of jobs located in i or to the number of inhabitants having a job? And is the denominator the total surface of i or the non-residential area? In the SustainCity project, each team was free to define its own indices for econometrical estimations, and equations are hence not perfectly comparable.

Even if often under-evaluated and constrained by statistical providers, arbitrary choices in terms of aggregation scale, delineation of the study area or the formulation of variables have in common that they generate *measure-*

ment errors that may trigger *"nuisance spatial dependence"*. It refers to the by-product of measurements errors for observations in contiguous spatial units (Anselin 1988). Our experience on the Brussels case study show that errors are almost unavoidable during the data processing step. Hence, prior to running UrbanSim, a consistent and thorough spatial database should be constructed, and a professional high quality *Exploratory Spatial Data Analysis* or ESDA (Anselin et al. 2006) should be conducted. However, due to the time constraints, this was far from being the case here, leading to more "data crunching" process rather than "data modeling".

4.3 What kind of bias can be expected in the econometrical estimations?

Spatial data are known to bias econometric results. In UrbanSim, sampling, regression, and discrete choice models are used. The sensitivity of these methods and their results to spatial issues has already been assessed in the literature (but independently of the LUTIs) and different methods proposed to deal with these biases. However, the use of spatial econometrics has not yet been implemented in LUTI models (Rosenbaum & Koenig 1997, Dowling 2005). Hence, after reviewing the main biases arising from the use of spatial data and the spatial econometrics methods developed to control or reduce these biases, we aim to highlight here their effects, using examples from the SustainCity project.

4.3.1 Main spatial biases and their solutions

Given the huge *number of agents*, random sampling is performed in UrbanSim prior to estimating location choices sub-models. Many alternative sampling techniques could be used (Cochran 2007, Ardilly et al. 2006). In particular, in an urban context where agents are divided among zones, and where the covariates are likely to show a strong autocorrelation pattern, a *spatially stratified sampling* may be a more adequate strategy than a *random sampling*. As this issue influences the data selection on which UrbanSim estimations are done, results may be biased by the sampling procedure.

Spatial autocorrelation denotes spatial dependencies of covariates. In case of autocorrelation, the assumption of independence of observations does not hold anymore; this can lead to unstable parameters estimates and/or to unreliable statistical tests. In linear regression analysis, Spatial Auto Regressive models (SAR) or a Spatial Error Model (SEM) are often used to take into account spatial dependency in hedonic regression (Kim et al. 2003); (Löchl & Axhausen 2010). Note that in most case studies, dwelling values are collected through a continuous variable. When collected by categories, the regression model was adapted: this was the case for the rents in Brussels where a Spatial Auto Regressive Interval Regression (SARIR) model has been developed

in the SustainCity project (see Chap. 13). Choosing the proper model means better estimates of the parameters and hence considering the true mechanisms explaining the dependent variable. Methods for accounting spatial dependencies in *discrete choice models* – such as the multinomial logit (MNL) used by UrbanSim to forecast the location of agents – also exist: let us cite the examples of Guo & Bhat (2004), who have proposed a multi-scale modeling structure to deal with spatial autocorrelation in neighbouring alternatives. For the same problem, Sener et al. (2011) propose a generalized spatially correlated logit model, where adjacent alternatives share common unobserved characteristics. A similar approach can be found in Alamá-Sabater et al. (2011), who compute the utility of each alternative by using a spatially weighted average of the attributes of surroundings locations. However, these methods appear less mature and/or less feasible to implement in LUTI than those for hedonic regressions.

Gehlke & Biehl (1934) outlined that correlation coefficients increase when the size of the Basic Spatial Units (BSU) increases. This *Modifiable Areal Unit Problem* or M.A.U.P., a term introduced by Openshaw & Taylor (1979), outlines that changes in size and shape of the basic spatial units may alter the estimates of the associated statistical analyses. Arbia (1989) and Fotheringham & Wong (1991), among others, have shown a systematic variation of estimated parameters in multivariate regression analysis when the number of zones decrease (some parameters becoming more negative and other more positive). As for spatial dependencies, the effect of the M.A.U.P. on discrete choice model has been far less studied than for linear regression model. Notable exceptions are Carod & Antolín (2004) or Carod (2008). By comparing firms location choice in Catalognia (Spain) for different areal units (provinces, municipalities and functional territorial units), they found significant differences in estimated parameters and conclude that location factors do not act uniformly over broad geographical areas. From a theoretical point of view, Train (2009) outlines that for a linear-in-parameter MNL model, the utility computed for a zone i, created by merging two zones j and k, will not be equal to the sum of the utility of j and k. It should also be noted that it is not straightforward to disentangle the M.A.U.P. and *misspecification issues*. Amrhein (1995) suggests to separate aggregation effects from other types of discrepancies, such as models misspecification in multivariate settings. Briant et al. (2010) investigated the relative importance of size and shape distortions comparatively to misspecification biases in the estimation of spatial concentration, agglomeration economies, and trade determinants. They found that at large scales, the size effect might be important but that at small scales it is rather weak compared to misspecification issues. Solving the M.A.U.P. is often considered intractable, but Openshaw et al. (1984) suggested some initial solutions to control and/or reduce his effects. First, one could simply ignore the M.A.U.P. and hope that the outcome of the research is still robust. Alternatively, the effects of the

M.A.U.P. could be reduced by using either individual data (persons, households), or the "appropriate scale of analysis". The last solution proposed by Openshaw et al. (1984) is to draw the spatial units so that the analysis produces a predicted outcome. However, none of these strategies are advisable in the context of LUTI models: examples hereafter will show that the M.A.U.P. may dramatically influence econometric estimations in UrbanSim. This is especially true when data are only available at aggregated level.

The last spatial bias that we would like to address here is *Ecological Fallacy*, which, formally, denotes that the correlation between individuals can be different from the correlation computed at the level of the group to which the individuals belong (Robinson 1950). On a general point of view, this issue refers to the biases arising when the characteristics of the individuals are deduced from aggregated data. Hence, ecological fallacies do not affect the econometric estimates *per se*, but rather data preparation and interpretation. In particular, UrbanSim requires disaggregated characteristics of households. But such information is almost never available for the entire study area (and particularly for Brussels). The population used in LUTI models will rather be a synthetic population, generated from a sample of individual households (such as the Public Use Microdata Sample in the US) and the marginal distributions of household attributes, per areal units (Ye et al. 2009). If the characteristics of the households are not homogenous in one areal unit, an ecological fallacy problem may occur. For instance, suppose that the average income of the inhabitants of one areal unit is known, but that these inhabitants are divided among a low and a high income group. The average income of the synthetic and observed population will be close, but the distribution of the income will be different, leading to a synthetic population having a different structure and, therefore, a different behavior than the observed one.

To sum up, none of the here mentioned spatial biases have an easy and straightforward solution. However, different spatial econometric tools are now available to control or reduce these biases. The examples presented in the next section will illustrate that their current lack of implementation in LUTI models can dramatically influence the predicted outcome of these models.

4.3.2 Spatial biases within the SustainCity project: some examples

The focus here is on empirical analyses, showing how spatial biases influence the econometric results of the tools embedded in UrbanSim (and in other LUTI models). The examples provided are mostly based on the SustainCity case study (Paris, Brussels, and Zürich), and on a simple, small-scaled, synthetic city (Jones et al. 2013) inspired by Brussels.

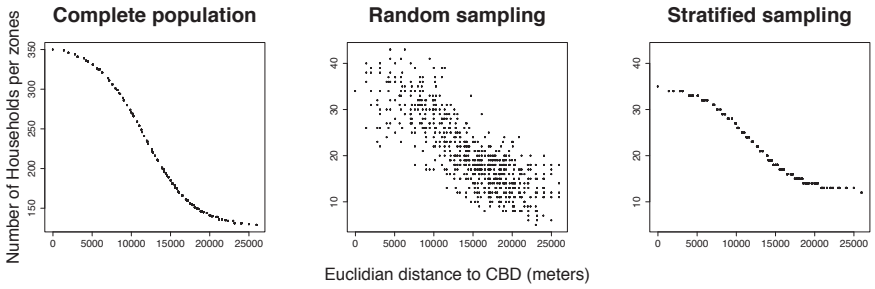


Figure 4.4 Spatial distribution of households within the synthetic city.

Example 1: sampling strategy

Let us start by an illustration of how the *sampling strategy* can influence the estimation of location choices sub-models, based on a synthetic city. The city is monocentric and made out of 729 (27 x 27) gridcells of identical size and shape. Population (Fig. 4.4) is distributed among zones assuming that the car accessibility to jobs increases the probability of a location, and that real estate prices (log) reduce this probability. In order to estimate the household location choices model (HLCM), two samples of 10% have been extracted from the population: a random one and a stratified sample (for each gridcell, 10% of the households are randomly selected). Figure 4.4 shows that there are clear differences in the spatial distribution of observations for these two samples. As a consequence, estimated parameters of the HLCM also changes: the estimate parameters of the (log of) real estate price is equal to 0.06 (+/− 0.02) for the random sample, and to 0.01 (+/− 0.02) for the stratified sample. For the car accessibility, parameters vary from −0.29 (+/− 0.11) to 0,17 (+/− 0.12). Hence, neglecting the spatial heterogeneity of agents in the sampling strategy may lead to misspecifications of the agent's location choice sub-models.

Example 2: spatial autocorrelation

Spatial biases are to be expected in the data used for econometric sub-models, but also in the models themselves. Let us illustrate the effect of spatial autocorrelation on dwelling rents. In the Brussels case study, rents were only available through a categorical variable: for a given areal unit, we know the number of dwellings having a monthly rent below 250 €, between 250 and 500 €, and so on, but not the average rents nor the distribution within each category. Hence, a SARIR model had to be developed (Pholo Bala et al. 2013), and its results were compared to a classical Interval Regression Model (IRM). Details can be found in Chapter 13, but the main findings are highlighted here. Estimations were made on two samples, one in a deprived area of the Brussels-Capital Region and the other in a well-off area. For both, the spatial lag term

is statistically significant. The intercept is by far smaller in the SARIR model than in the benchmark model (IRM), indicating that the omitted variable bias is reduced in the spatial model. Estimated parameters of other independent factors do not differ substantially. Hence, a spatial model allows the inclusion of unmeasured influence of neighboring observations. Furthermore, comparisons between the two samples show that the spatial dependence observed for dwelling rent vary in space. Note that the IRM model has also been used to assess the effect of a change of the size of the BSU. In particular, the estimated parameter of the log of median income is higher when this variable is measured at the municipality level than at the statistical ward level. The same observations can be made for environmental amenities variables. A possible explanation is that the larger the BSU, the lower is the variance of the independent variable.

Example 3: Modifiable Areal Unit Problem

A sensitivity analysis of the Employment Location Choice Model (ELCM) of UrbanSim to the *Modifiable Areal Unit Problem* has been conducted for Brussels (Jones et al. 2012). Location choices of jobs in tertiary activities were estimated for four BSU levels (Table 4.2). The main result is that the population density has a negative impact on utility for small BSUs, while the influence is positive for larger areal units. Hence, it appears that the meaning of the variable changes: depending on the aggregation level, an high population density will either means a residential neighbourhood with no space left for economic activity, or an urban area attractive for tertiary activities. Other estimated parameters of the MNL model also vary in magnitude and significance, and their variations are not monotonous with scale, without easy or unique explanation. Changes in the correlation between independent variables through scales are likely to influence them, as well as the nature of the variables. On an operational point of view, it should be noted that a large BSU having a relatively low probability of location is not composed solely of small BSUs also having

	Small BSU		$\rightarrow$	**Large BSU**
	(1)	**(2)**	**(3)**	**(4)**
Pop. density	−0.11 (*0,02*)		1.04 (*0.03*)	1.06 (*0.04*)
Land prices	−0.35 (*0,11*)	−0.44 (*0.05*)	−0.17 (*0.02*)	−0.18 (*0.03*)
Car accessibility	3.68 (*0,31*)	4.86 (*0.13*)	1.39 (*0,15*)	0.14 (*0.04*)
Dist. to highway	−0.49 (*0,07*)	−0.12 (*0.03*)	−0.85 (*0.04*)	−0.17 (*0.02*)

Table 4.2 Estimate parameters of the ELCM for the Brussels case study (between brackets: *standard deviation*), where 1 = statistical wards (4945 wards, 1 sq. km on average), 2 = sections, 3 = former communes, 4 = communes (151 communes, 34 sq. km on average).

all a relatively low probability but can include an activity park for which the probability of attracting a new or relocating job is relatively high. The scale of analysis is then likely to influence the predicted outcomes of UrbanSim.

Example 4: Ecological fallacy

As mentioned in Section 3.1, ecological fallacy was (in this context) mainly an issue for the data processing step. For instance, in the Brussels case study, the population used by UrbanSim is a synthetic population generated from a sample of individuals and marginal distributions of their characteristics, per zone (Farooq et al. 2013). Data availability was highly limited: the household survey used for individual characteristics only had 1367 records (about 0.1% of the total number of households). Still, the characteristics of the generated population seem to match the observed marginal distribution of these characteristics with a reasonable fit: for the income distribution, at the scale of the municipalities, the R^2 value is 0.799 (Farooq et al. 2013). However, in the Brussels case study, the BSUs correspond to statistical wards. Since the socio-economic level often varies within a given municipality, a good fit at the aggregated scale does not necessarily means a good fit at the disaggregated level. In a similar way, land price data were only available at the scale of the municipalities. Hence, the average municipal land value is allocated to all statistical wards within the commune. Again, given the large socio-economical variations within each communes, this assumption is unlikely to hold.

Last but not least, spatial biases may in some specific cases entail *endogeneity*. Anselin & Lozano-Gracia (2008) investigated the impact of potential error in variables due to interpolated air pollution measures on the estimation of hedonic models of house prices, using a Spatial Two-Stage Least-Squares (Spatial 2SLS) estimation methods. The idea is that interpolated values of error pollution may result in a prediction error correlated with the error term, leading to a simultaneity bias. Their results suggest that the bias from ignoring the endogeneity in interpolated values may be substantial. We may conjecture that using inappropriate spatial scales may further trigger endogeneity of the "error in variables" kind. The use of Spatial 2SLS may mitigate such biases, and therefore is a estimation strategy that deserves to be further explored.

These examples show that spatial biases potentially affect LUTIs results in many different, unpredictable and not tractable ways and that solutions are not unique, nor straightforward. There is a need for controlling the influences of these spatial biases on the predicted outcome of each application. However, little was done in that respect within the SustainCity project, mostly because the time needed to built the database for UrbanSim proved to be far longer than expected. The operational implications of these findings will be highlighted in the conclusion of this chapter.

4.4 Choosing is biasing? Biasing is misinterpreting?

Not only can the representation of space influence the behaviour and the outcomes of LUTI models in different ways, but this is also the case of the choices made by the modelers. This section illustrates two types of selection biases.

4.4.1 "Geographical modeling" versus "data crunching"

Many different explanatory factors can be included in the econometric sub-models of UrbanSim, thanks to the large amount of data required by the model. According to Nguyen-Luong (2008), a compromise between explanative versus predictive model is needed when estimating UrbanSim's sub models. The argument is that a very sophisticated model may be unusable for forecasting, because the variables used to estimate the model cannot be forecasted. Moreover, even if the selected explanatory variables are grounded in recent econometric, transport, and behavioural scientific literature for each sub-topic, one can wonder if a lot of explanatory variables is better than a careful choice of some variables grounded in the economic geography literature.

Let us illustrate this by an exercise conducted on the Brussels case study. The Real Estate Price Model was estimated for two different specifications (Table 4.3). The eight variables included in the specification (see column *a*, Table 4.3) are significant in the OLS estimation equation, and the results are consistent: prices are high where the surface of green areas or income are high, and where there is little industry. But when applying the same model and a second specification (column *b*, with only 3 variables: income, population density, and distance to CBD), the determination coefficient is equal (at the 2-digits level). It suggests that *data modeling* is as efficient in explaining land

	a	**b**
Constant	12.03	12.34
(Log) dist. to CBD	−0.10	−0.12
Population density	0.02	0.01
Income	0.14	0.04
% University degree holder	0.01	
Green area score	0.15	
Density of jobs in education	0.01	
% Jobs in industry	−0.02	
Density of jobs in industry	0.01	
Adj. R-square	0.65	0.65

Table 4.3 Estimated parameters of the Real Estate Price Model of single-family houses for the Brussels case study for different specifications.

prices as the whole set of variables. This also means that (1) the urban structure is well captured by the 3 classical variables of the Alonso-Muth model, that (2) the additional variables simply corroborate the same structure, and that (3) limiting the energy spent on data collection might be useful and will not harm the efficiency of the explanatory model. This confirms that parsimony is important in econometric estimations.

4.4.2 Brussels, Paris, Zürich: do they look alike?

One of the goals of the SustainCity project was to assess and compare the effects of specific land-use and/or transport policies *on three different cities*. Spatial specificities make the comparison of places a difficult task: Paris, Brussels, and Zürich not only have different sizes and hence cope differently with transportation distances, but they also correspond to different urban centre-periphery systems, and more generally to a different "urban model" (polycentricity, location of the better-off households, etc.). Let us add that, practically, each research team has worked independently. The choices made on the study area delineation, basic spatial units, and specifications of the models/-variables/thresholds are thus different and will bias any comparison.

Exploratory Spatial Data Analysis

We shortly present the results of an empirical analysis of the similarities and differences of the urban model of each city. To reduce discussion about data definitions, only three simple variables are used: population and job density and income per capita, with a common scale for each case study (see Fig. 4.5). As expected, population density gradients are more important in Paris, but reveal a clear-cut mono centric situation. At the same geographic scale, Brussels appears more polycentric and Zürich has 2 centres (Zürich and Winterthur). Job densities confirm these trends, but the centre-periphery structure is more marked in Zürich. Income reveals a deprived centre in Brussels, while in Paris better-off households prefer some central wards, confirming former studies, e.g. Goffette-Nagot et al. (2000). Moran's I indices for spatial autocorrelation confirm the differences in spatial structures observed on the maps (Table 4.4). As expected by the literature on urban segregation (Goffette-Nagot et al. 2000), the three variables all show a positive and significant spatial autocorrelation, meaning that similar observations tend to be close to each other. Job and population densities are particularly concentrated in Paris (higher I values). However, one has to be careful with this interpretation, which can be an artefact of the larger size of the Paris study area. Table 4.4 also confirms the maps by showing for each studied area the relationship between distance to the city-centre and the three above-mentioned variables. The correlation with income is negative in Zürich and Paris, while positive in Brussels, where it is well known

	Moran I			Correlation		
	Paris	Brussels	Zürich	Paris	Brussels	Zürich
Population density	0.86	0.60	0.59	−0.53	−0.54	−0.56
Job density	0.79	0.46	0.53	−0.37	−0.44	−0.49
Income	0.58	0.64	0.49	−0.26	0.27	−0.54

Table 4.4 Moran I values and Pearson product-moment correlation coefficients with Euclidian distance to the city centre (all coefficients significant at 5%).

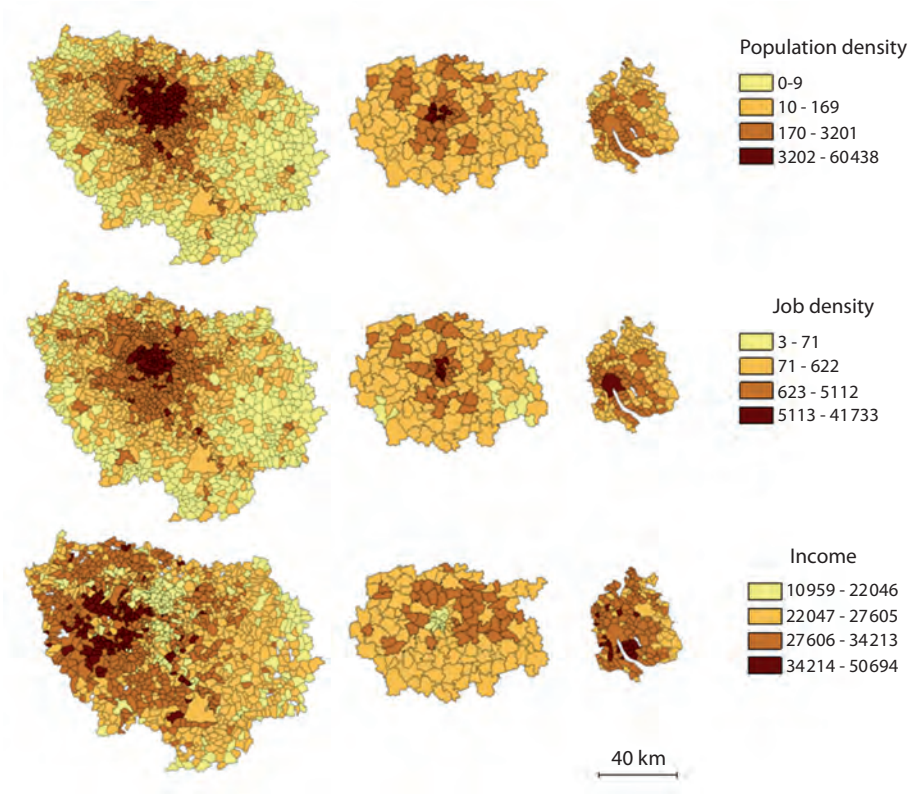


Figure 4.5 Comparison of Paris (left), Brussels (center), and Zürich (right) for population density (top), job density (middle), and income (bottom), by municipality.

that better-off people tend to reside in the periphery, see for example Dujardin et al. (2008) or Goffette-Nagot et al. (2011).

Cluster analysis

We also tested if it is possible to identify groups of municipalities having similar characteristics through the three case studies, by a clustering analysis (Ward)

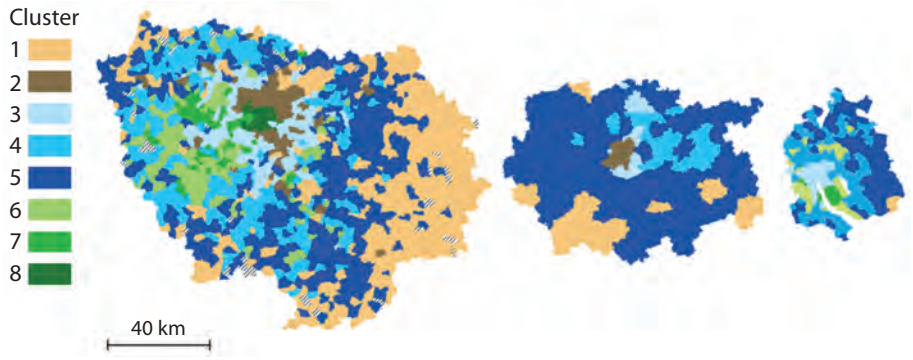


Figure 4.6 Clustering of urban municipalities in Paris, Brussels, and Zürich. CL1: low income and very low densities (population and job); CL2: high densities, very low income; CL3: high densities, medium income; CL4: very low densities, high income; CL5: very low densities, medium income; CL6: low densities, very high income; CL7: low densities, high income; CL8: very high densities, high income.

using the three variables (Fig. 4.6). The Ward density minimises the intergroup variation and maximizes the intragroup homogeneity. The cluster analysis is conducted on the entire set of communes (all 3 cities taken together). Clearly, the spatial structure of the three case studies differs. Larger variations of socio-economic conditions are observed in Paris: it is the only case study where all clusters are represented. However, once again, it is unclear if it reflects larger variations of socio-economic conditions, or if it is an effect of the greater size of its study area, and of the smaller size of its BSU. Compared to Paris and Brussels, Zürich is characterized by higher incomes, and by lower variations of the socio-economics characteristics. The more deprived Brussels Capital Region is clearly identified in the analysis, while peripheral areas are far more homogenous than for Paris. All municipalities surrounding the BCR have medium population densities, and medium income with, on average, higher values in Flanders than in Wallonia. Hence, the urban structure of the three case studies shows clear differences in the centre-periphery structures. These differences may generate discrepancies between model results simply because space is different.

4.5 Conclusion

This chapter highlights the different ways by which spatial data can influence the outcomes of spatial models and specifically of LUTI models. These issues are of different nature and affect the results in different and often not predictable ways. Some of them come from choices made by the modelers themselves: delineation of the study area, selection of the basic spatial units, definition of the variables, and specification of econometrical sub-models. To

solve or, at least, control these issues, one will face different constraints, mostly *external* to the model system: political instances may impose the choice of the study area, and that of the basic spatial units may be constrained by data availability. Some other issues are due to spatial biases in econometric estimations (sampling, spatial autocorrelation, M.A.U.P., ecological fallacy). In this case, one will rather face *internal* constraints: requirements of the model system and the lack of implementation of spatial econometrics tools. Spatial econometric techniques, like the Spatial 2SLS model, may mitigate the biases due to the arbitrary delineation of basic spatial units or to measurement errors in spatial interpolation. However, to our viewpoint, none of the spatial issues presented in this chapter have an easy or unique solution, nor easy and unique consequences.

The "Garbage In, Garbage Out" statement is well known, and clearly applies to LUTI models: such models are very appealing for urban planners, due to their ability to forecast future urban pattern and to evaluate the potential impacts of environmental legislations (Rodrigue et al. 2009). But *"modelling studies almost always present extraordinary results, despite the fact that the data used in the project are barely described, and/or the quality of the data are poor"* (Nguyen-Luong 2008, pp. 3), which raises obvious concerns about the credibility of these results for external observers. In the SustainCity project, data collection and modelers' choices were made independently in each case study, questioning the usability of the results for a comparative point of view. There is a real need of collective and reasonable choices for making LUTIs results understandable and comparable. Results may be credible simply because the urban space is well structured. Simplification and formulation better anchored in spatial theory are needed in order to better grasp the interrelationships.

Eventually, developing a LUTI model for a given city is not only a question of econometrics and computer sciences. As previously outlined, the urban structure of the UrbanSim case studies show clear discrepancies. It is then crucial to take into account the specificities (spatial or not) of each city, and a good knowledge of the study area is vital (Nguyen-Luong 2008). An explicit account of the *geography* of each city and/or case study should help to control or reduces these biases and, therefore, help to improve Land Use and Transport Interactions models.

Part II

Modeling and methodological contributions

Chapter 5

Methodologies for synthesizing populations

Bilal Farooq, Kirill Müller, Michel Bierlaire, and Kay W. Axhausen

5.1 Introduction

Models for microsimulation of urban systems simulate the interactions of synthetic agents. These agents are a discrete representation of the study area's population (or a fraction) that is statistically consistent with known data about the true population. The discretization of the population allows simulating individual decisions at the agent level and aggregating the simulation results only for analysis, while the statistical consistency ensures applicability of the results to the real world scenario. However, due to privacy and other restrictions, detailed data for all agents to be simulated are virtually never available. Synthesis of a population of agents and their attributes is thus one of the first and key steps required for developing an agent based integrated transportation and land use microsimulation.

Notably, the synthesis process is not designed to *improve* the quality of the data. Rather, its purpose is to *prepare* the available data so that it is usable by microsimulation system, while maintaining its statistical properties. It is anticipated that the agent population (a) matches known aggregate totals for all zones in the study area (e.g. the total number of males and females in each zone), and (b) replicates the correlation structure of the true population (e.g. wealthy people tend to be older than average, own several cars and a detached house, etc.). In order to accurately simulate person interactions (e.g. in the household or in a person's social network), the model also requires information on personal relationships. One of the earliest examples of base year agent synthesis can be found in TRANSIMS (Smith et al. 1996). The agent synthesis was formulated as a fitting problem in which weights of microdata were adjusted to fit the aggregate space level control totals (Beckman et al.

1996). Since then various approaches have been developed that are based on the fitting based optimization.

Here we present two methodologies for synthesis and an application in the Brussels case study of the SustainCity project where the required data was not available in disaggregate form. In contrast, for the Zürich and Paris case studies within this project, the complete disaggregate census data was available, and the agent population has been derived directly from this data set without using techniques presented in this chapter. For the purpose of this chapter we only used the Zürich data, but the developed methodologies are applicable to Paris case study as well.

The first approach is based on treating the synthesis as a fitting problem. The fitting of two different types of agents (i.e., household and person) is performed simultaneously. The second approach treats the synthesis as a problem of drawing from a complex joint distribution of agents' attributes. A Markov chain Monte Carlo simulation based procedure is developed. Such an approach is particularly robust in the case of highly limited data availability.

The rest of the chapter is organized as follows. We first present the current state of literature in population synthesis. We then propose two methodologies to synthesize the agent population. The available data sources and input preparation for the two methods are described. An application of the proposed methodologies, and results, are presented. In the end we discuss various aspects of the two approaches.

5.2 Literature review

Agent-based microsimulation model systems for land use and transportation planning have come into widespread use (UrbanSim 2011, MATSim 2013, Bradley et al. 2010, Beckx et al. 2009, Roorda et al. 2008, Bhat et al. 2004, Ben-Akiva et al. 2002, Bowman & Ben-Akiva 2001, de Palma & Marchal 2002, Mahmassani et al. 1995). These models simulate decisions of agents within an urban area, allowing for more detailed simulation and prediction of population distribution, travel demand, and various other indicators than traditional aggregate models. Often, the *agents* represent the individual people living in the study area – grouped in households, as personal decisions are usually influenced by other household members (Jones et al. 1983).

When implementing such a model system, the initial step is the definition of agents and their relationships. This process is called *population synthesis*; also, the term *spatial microsimulation* is sometimes used. Williamson (2013) distinguishes between two practical methods for generating synthetic population: *reweighting* and *imputation*. Reweighting methods construct the population by selecting observations from a microsample. Each record in the sample can be cloned or replicated several times to form agents in the synthetic population. In contrast, imputation methods attempt to generate agents "from

scratch", yielding unique characteristics for each agent. The following two subsections review the literature on the two methods. See also (Harland et al. 2012, Hermes & Poulsen 2012, Müller & Axhausen 2011*b*, Ma 2011, Pritchard 2008) for a comprehensive literature review.

After generating the synthetic population, agent relationships such as household structure (Gargiulo et al. 2010) or social networks (Arentze et al. 2012) can be added to allow simulating interactions between agents. Lenormand & Deffuant (2012) compare the reweighting and imputation methods for generating the household structure.

5.2.1 Reweighting methods

The main idea here is to combine census microdata (the *reference sample*) with aggregate data at various levels (the *control totals*) in order to generate a set of agents for which the distribution and correlation of the agents' attributes are similar to those in the census microsample, and the number of agents within each category matches the aggregate data. Two approaches can be distinguished: The *fitting and generation* (FG) method (Beckman et al. 1996) generates the synthetic population by estimating weights for the reference sample and then drawing with probability proportional to the weight; see also Rahman et al. (2010) for a review of reweighting techniques. In contrast, the *combinatorial optimization* (CO) method (Voas & Williamson 2000) directly expands the reference sample with an integer vector computed using non-linear integer optimization.

Both approaches require a reference sample from which agents are drawn: The generated population contains only realizations that also exist in the reference sample. Furthermore, the weights are generated subject to different optimality criteria, which, depending on the data, sometimes cannot be met. However, reweighting methods usually generate populations with a reasonably good fit to the given control totals.

Fitting and generation

Given a reference sample of households that contains detailed data for all persons, and constraints at both household and person levels, two options are available for the FG method:

1. The weights obey only household-level constraints, person-level constraints are considered when selecting households (*single-level fitting*, cf. Auld & Mohammadian 2010, Ma 2011, Guo & Bhat 2007).

2. The weights obey constraints at both person and household levels, household selection is unconstrained (*multi-level fitting*, cf. Ye et al. 2009, Bar-Gera et al. 2009, Lee & Fu 2011, Müller & Axhausen 2011*a*).

Section 5.3.1 offers more detail on the procedures for estimating the weights.

The FG method requires careful preparation of the data. Especially the *zero-cell* or *missing observation* problem (observations existing in the population but missing in the sample) has been noted early (Beckman et al. 1996). A remedy is to collapse categories, data-driven approaches are discussed in (Auld et al. 2008, Otani et al. 2012). The problem of conflicting control totals has been addressed by Rich & Mulalic (2012). Pendyala et al. (2012) describe the practical application of a multi-level fitting algorithm.

Combinatorial optimization

In this method the sample is replicated zone by zone so as to optimize the weight (where $w = \{0, 1\}$) of each observation under zonal marginal constraints (Voas & Williamson 2000, Ma 2011). Openshaw & Rao (1995), Williamson et al. (1998), and Voas & Williamson (2000) used Simulated Annealing (SA) as an optimization tool to produce the synthetic population. Ryan et al. (2009) compared CO and FG based methods and concluded that CO produced lesser variance. Another advantage is that it has lower memory requirements, though the convergence time of CO based techniques is very slow, especially in the case of SA based optimization.

Barthelemy & Toint (2013) developed a synthesis method that is based on various discrete and continuous optimization procedures. Their method is a step forward in fitting based methods, as it gives more flexibility in terms of data needs. The developed procedure however, involves various hierarchical fitting steps (for each aggregation level where the data is available), maximization of information entropy, tabu search, and various ad hoc matching rules. The method was applied to synthesize Belgian population with limited attributes. The generalization of the methodology, so as it can be used for other cases, is not very clear.

Abraham et al. (2012) have implemented an analogy to multi-level fitting in a CO approach where both household-level and person-level constraints are used for the objective function in the underlying optimization process.

5.2.2 Imputation methods

While imputation methods can make use of a microsample, it is not required. The observed population is assumed to be a realization of a (multivariate) distribution, and this distribution is simulated to generate the synthetic population. In contrast to reweighting approaches, no cloning or replication occurs – when many attributes are generated, the characteristics of each generated agent are unique. Usually, imputation refers to substituting missing values for an existing attribute based on the data. In extreme cases, the attribute is missing for all observations, or the dataset is constructed from scratch. This justifies the usage of the term “imputation” in a very general sense in the context of population synthesis.

One of the early examples of directly drawing from distribution can be found in TORUS (Miller et al. 1987), which microsimulated the households' location choice decision in Toronto area. In TORUS, to generate an agent's attributes, all the available distributions were sampled independently while making sure that there were no logical inconsistencies among the realized attributes (for instance, a 2-year-old child cannot have a university degree). It used the zonal marginals and partial conditionals available from the census.

An example of a multistage approach where attributes are generated sequentially and attributes depend on other attributes generated earlier is described in Williamson et al. (1998); also some earlier studies are referenced therein. In a recent effort, Frazier & Alfons (2012) have adapted this methodology to generate a synthetic population of Ghana. Barthelemy & Toint (2013) and Pritchard & Miller (2012) imputed the associations by sampling the empty household structure from the distribution and fitting the generated persons into it; various rules based on domain knowledge were used to avoid incompatible associations.

5.3 Methodology

Here we describe two different approaches to synthesize the population. For both approaches, the agents are defined by their attributes $X = \{X^1, X^2, \ldots, X^k\}$, where k is the number of attributes to be synthesized. There are n agents in the synthesized population. In the first approach, the synthesis problem is treated as an optimization problem. The second approach treats it as a problem of drawing from a joint distribution which is only partially known and is very hard to directly draw from.

5.3.1 Fitting and generation based method

The fitting and generation (FG) method belongs to the class of reweighting methods and consists of two principal stages. In the *fitting stage*, a fractional *weight* (or *expansion factor*) is estimated for each household in a given *reference sample*. The weighted reference sample corresponds to the full population of the study area and is required to satisfy aggregate constraints (referred to as *control totals* or *controls*). After that, in the *generation stage*, this expansion factor is used to draw from the reference sample; each drawn item becomes an agent in the synthetic population. Correlation between attributes is derived from the reference sample and is altered by the reweighting process. The weight estimators ensure that this change is minimal so that the principal structure of the data is retained. In the case of multidimensional controls, the controls define the joint distributions of the affected attributes in the weighted reference sample.

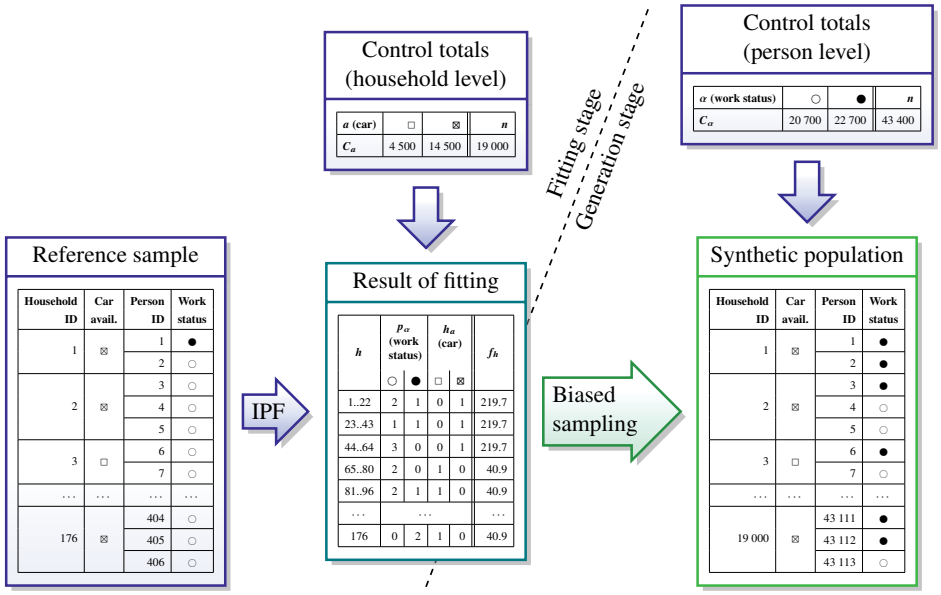
Expansion factors that satisfy the household-level control totals can be estimated using the well-known Iterative Proportional Fitting (IPF) algorithm (Deming & Stephan 1940). In order to also satisfy the person-level controls, the selection procedure must account for this by preferring households with persons in still underrepresented categories; examples can be found in Auld & Mohammadian (2010), Ma (2011), and Guo & Bhat (2007). This approach is sketched in Figure 5.1a. However, apart from the additional bias introduced by the modified selection procedure, the generation stage is more complex and in some cases requires time-consuming computations not suitable for frequent repetition. In addition, it seems to be difficult to specify a mathematical model for the population generated by this procedure.

In contrast, *multi-level algorithms* estimate household-level expansion factors that satisfy the controls at both household and person levels (Ye et al. 2009, Bar-Gera et al. 2009, Lee & Fu 2011, Müller & Axhausen 2011*a*). Here, the weights already define a probability distribution of the households, and construction of the final population is possible using, e.g. weighted random sampling with replacement. Figure 5.1b illustrates this strategy. This simple model allows generating thousands of instances of similar yet different populations, e.g. in the context of multiple imputation (Rubin 1987). In the remainder of the chapter, only the multi-level fitting approach is considered.

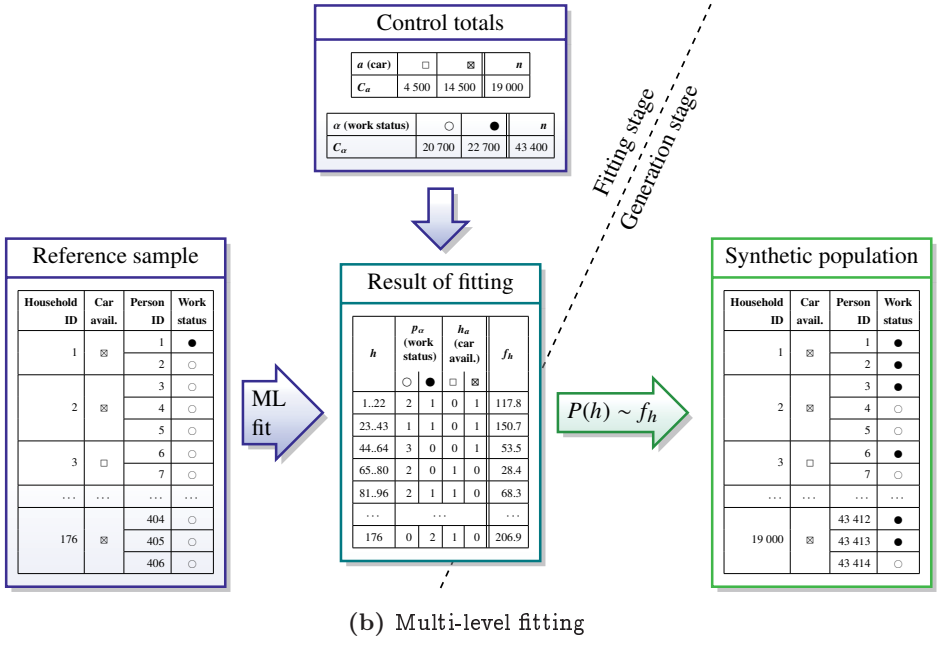
Formally, the reference sample is a $n \times k$ matrix $R = \left(R^1, R^2, \ldots, R^k\right)$, and the control totals are a vector x_j, $j = 1 \ldots c \leqslant k$. Here, n is the number of observations in the reference sample, and k is the number of attributes. For this, a weight vector w_i, $i = 1 \ldots n$ is estimated so that $\sum_{i=1}^{n} (w^\intercal \cdot R)_{ij} = x_j$ for all $1 \leqslant j \leqslant c$. Then, $w^\intercal \cdot R \sim X$ is treated as the multivariate distribution that defines the population of interest, and a realization $\hat{X}$ is sampled from this distribution. The estimation of the weight vector and the format of R and x differs between the algorithms. For single-level algorithms, R contains either households or persons, and w is usually estimated using IPF. Multi-level algorithms always use households in R; in addition, person-level attributes are represented as count variables (indicating the number of persons with a given characteristic in a household). The estimation of w is carried out by repeatedly fitting all constraints in a round-robin fashion, or by maximization of information entropy.

5.3.2 Markov chain Monte Carlo simulation based method

The agents' attributes X can be treated as random variables having a unique joint distribution $\pi_X(x)$ in the real population. In order to synthesize agent population, we are interested in drawing from this underlying joint distribution of agent attributes. This distribution is not completely known and is very hard to directly draw from. We only have partial views of the $\pi_X(x)$ through samples, marginals, and conditional-marginals. Using these partial views, a



(a) Single-level fitting



(b) Multi-level fitting

Figure 5.1 Illustration of single- and multi-level fitting algorithms.

technique is developed here that lets us draw as if we were drawing from the joint distribution $\pi_X(x)$.

Markov Chain Monte Carlo (MCMC) methods are computer based simulation techniques that can be used to simulate a dependent sequence of random draws from very complicated stochastic models/processes/distributions (Hastings 1970). These methods provide flexibility in terms of using various data sources at various spatial scale; bring in prior knowledge in a systemic way; implement assumptions in a coherent manner, where the data is not available; and are, computationally and memory-wise, robust. These techniques have been extensively used in various other domains including physics and image processing.

Here we propose the use of MCMC techniques to simulate drawing from the real population distribution $\pi_X(x)$ and thus to create the synthetic population. Given the fact that we only have partial information about the joint distribution, there can be any number of synthetic populations $\hat{X}$ representing the real population X. The proposed technique lets us draw any number of the possible synthetic populations. The rest of the section describes how a specific MCMC technique can be used to generate synthetic populations.

5.3.3 Gibbs sampling basics

In particular here in this chapter we have used Gibbs sampling (Geman & Geman 1984) procedure to develop the drawing of agents from the joint distribution. We are interested in generating synthetic populations $\hat{X}$ by sampling from the joint distribution of the attributes in the real population X. The problem is thus reduced to first retrieving the best representation $\hat{\pi}_X(x)$ of the joint distribution $\pi_X(x)$ that can then be sampled to generate any number of possible synthetic populations $\hat{x}$ representing X.

Let $X = (X^1, X^2, ..., X^k)$ be a set of random variables with a joint distribution $\pi(x)$ that is hard to retrieve. Instead we have the conditionals $\pi(X^i|X^j = x^j$, for $j = 1...k \ \& \ i \neq j), i = 1, ..., k$ available. The RVs themselves can be discrete or continuous. Then draws from the distribution $\pi(x)$ can be retrieved by using the procedure described in Algorithm 1 and Figure 5.2.

Steady state of a Markov chain is the point at which as if the sampling is done from the stationary distribution representing the true population. To arrive to this state, the Gibbs sampler first has to run for an extended amount of iterations. The draws during this run are ignored. At that point any draw from Algorithm 1 will be as if the draw was from $\pi(x)$ (Train 2009). To avoid the correlation between consecutive draws, few draws between two recorded draws are skipped.

Data:
$\pi(X^i|X^j = x^j$, for $j = 1...k \ \& \ i \neq j), i = 1, ..., k$
iterations (integer): Size of the population pool
interval (integer): Acceptance interval
Result: Draws from $\pi(x)$
initialize X_{prev};
initialize X_pool;
initialize counter;
for size_pool×interval **do**
  Generate a random number from $r = U(1, k)$;
  Generate x^r_{curr} using **Inverse Transform** (Devroye 1986) on
  $\pi(X^r_{curr}|X^j = x^j_{prev}$, for $j = 1...n \ \& \ r \neq j)$;
  $X_{curr} = X_{prev}$ with x^r_{prev} replaced by x^r_{curr};
  if *counter equals interval* **then**
    X_pool.Add(X_{curr});
  end
  $X_{prev} = X_{curr}$;
end

Algorithm 1: Gibbs Sampling.

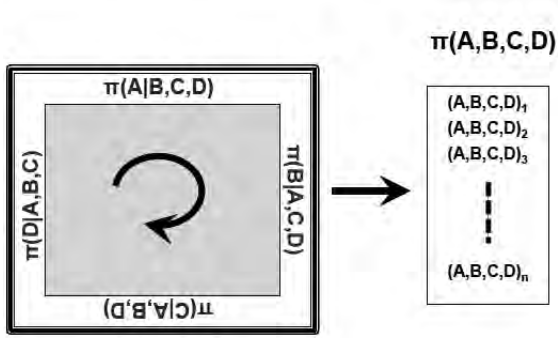


Figure 5.2 Illustration of the working of Gibbs sampler.

5.3.4 Generation of synthetic agents

To generate a population, we simply have to warm up the Gibbs sampler (i.e. ensure that the Markov chain is in a steady state) and then draw the number of agents that are needed. For computational efficiency the state can be saved at the end of the draw. This way the sampler does not need to be warmed up every time a new population is needed. Another option is to pre-fetch a very large pool of agents using Gibbs sampler. This pool can then be used to draw any number of synthetic populations. The procedure is described in Algorithm 2.

Data:
- *n (integer)*: number of synthetic populations
- *size (integer)*: size of a population
- *cond (integer)*: conditions (usually marginals) based on which a population should be drawn. By default the population is drawn from the pool using a uniform distribution

Result: *n* synthetic populations
if *population pool does not exist* **then**
  Call Gibbs Sampler;
end
while k ¡ n **do**
  while l ¡ *size* **do**
    Draw from pool using *cond* ;
  end
end

Algorithm 2: Synthetic population generator.

In Gibbs sampler, the conditional distribution of one attribute over all the other attributes captures the correlation. In an ideal case this correlation will be dictated by the data. If there is no data, then domain knowledge. And if there is no information at all, only then we assume it to be uniform. The important point here is that if there is any information available about the correlation, we can very easily use it to form the input conditional distribution.

5.4 Data

5.4.1 Available data sources

Census and travel surveys have been the primary data sources to construct a synthetic population. Other sources, for instance a household spending survey or a labour force survey, can also be used. However, they are rarely used in the existing literature. The information from these sources is available in two different forms: sample of individual agents and cross-classification tables.

Zoning systems

The data is available at certain aggregations of space that are defined by a zoning system. The aggregation may be based on certain maximum density levels, physical obstacles (river, street, etc.), and political boundaries. There is a hierarchy of aggregations within each zoning system. For instance, in the case the of Canadian Census, the lowest level of zone is called dissemination area where 400 to 700 persons are living/working. One level above is the census tract where the limit is 2,500 to 8,000 persons. Further aggregations are census

sub-division, division, and municipality, respectively. The zoning system also changes with time, i.e., a discrimination area in the year 2001 census may have been divided into two in the 2010 census, so as to satisfy the number of person constraints.

The travel surveys are available at the lowest level of aggregation called Traffic Analysis Zone (TAZ). TAZs are usually defined based on the road network; their size may vary depending on the agency conducting the survey; and they need not overlap with any of the census zoning system. Another zoning system that may be used is the postal code system.

In agent based microsimulations, usually one zoning system is set as base; a mapping between different systems is defined; and all the data is made consistent to that system. An example of such a procedure can be found in Farooq et al. (2008).

Sample of individual agents

The Statistics bureau of a country, among other surveys, also conducts periodic censuses of the entire population. The periodicity of this census ranges from 5 (in the case of Canada) to 10 years (USA, UK, etc.). While the whole dataset is hardly ever available for the research (with some exceptions, like Switzerland), bureaus do provide a representative sample for public use. In North America, this sample of individual agents is called Public Use Micro Sample (PUMS) and in UK and a few other countries, Sample of Anonymised Records (SARs). In this chapter we will use the term PUMS. This sample is available at a large spatial resolution (for instance, City of Toronto, London etc.), so as to make sure that the privacy of individuals is protected. The size of sample may range from 1% to 5% of the total population. The sample may contain a range of demographic and socioeconomic information on households, families, and persons. The exact location, income details, and some other details may be missing due to privacy concerns. Furthermore, the census bureau may hide information on certain individuals, if they deem it exposing the identity of these individuals.

Another source of the sample is the travel survey, usually conducted by the urban regions, municipalities, counties, etc. The focus of this survey is on the travel demand patterns of agents and mode shares, but it also has some information on socioeconomics and demographics of agents. There is more fluctuation here, in terms of the detail, size, and periodicity of travel survey among the regions/countries. Moreover, the periodicity of travel surveys of neighbouring regions may not coincide and it may also not coincide with that of the census.

Cross-classification tables

At various zoning levels (for instance, dissemination area, and census tracts in case of Canada, and sectors, and communes in case of Belgium and France), the Statistics bureau also releases the cross-classification tables for socioeconomics and demographics (for instance, income by age at sector level). These tables are usually restricted to 1 to 3 dimensions due to privacy concerns—the cell values of tables with more dimensions might be small enough to reproduce the population. In the available tables, very small cell values are usually rounded off to zero by the bureau. There is also a random rounding of values done by the Statistics bureau, so as to make it further difficult to reproduce the baseline population.

Details on various techniques used to anonymize the census data can be found (Sweeney 2002, Dalenius & Reiss 1982, Brand 2002). The point that should be noted here is that the available data has already been treated with various such anonymization techniques. Thus, a population synthesized based only on counts data may be not completely representative of the real population.

5.4.2 Data preparation

The two synthesis procedures described in Section 5.3 require the above mentioned data to be preprocessed as input in different forms. Here we first describe the data preparation process for the two processes and then illustrate the results.

Suppose that we are interested in synthesizing the person agents with attributes age A, sex S, household size H, eduction level E, and location in terms of spatial zone Z. In addition, the persons should be grouped into households, and the car availability C is of interest.

Fitting and generation based method

As mentioned earlier, multi-level FG algorithms operate on the following input data:

- a representative reference sample that contains the characteristics of sampled households and all constituent persons, and
- control totals for selected attributes on both household and person levels.

The objective is to estimate a positive weight (or *expansion factor*) for each household so that all control totals are satisfied.

For the given example, the reference sample should contain the household attributes H, C and, for each person in a household, the attributes A, S and E, obtained from a Public Use Microsample or from a large transportation survey

(cf. Sect. 5.4.1). If no dataset with all attributes can be found, statistical matching (D'Orazio et al. 2006) or multiple imputation (Rubin 1987) can be used to create one from multiple datasets.

Cross-classification tables (cf. Sect. 5.4.1) are used as control totals. Here, at least one cross-classification that contains the Z attribute is required to distribute the population over all zones. The attributes in the control totals must be compatible with those in the reference sample. In addition, for each non-zero cell in each cross-classification table there must be at least one corresponding item in the reference sample to avoid the *missing observation problem* (or *zero-cell problem*). One possible remedy to both issues is collapsing categories. This approach has been implemented in a general fashion by Auld et al. (2008) and formalized by Otani et al. (2012). Note that collapsing two categories Y_1 and Y_2 is safe only if they are "similar", i.e. if $P(X|Y_1) \approx P(X|Y_2)$ for all other attributes $X = \{X_1, X_2, \ldots\}$.

Moreover, the cross-classification tables must be consistent with each other (in particular, the grand total and marginal totals for all attributes must match). Rich & Mulalic (2012) present a technique to cope with inconsistent cross-classification tables.

Usually it is assumed that the reference sample is equally representative of all zones in the study area. Hence, even if the reference sample contains precise location information, it is either not used or at best substituted by some higher-order categorization (e.g. a coarser zoning system, a zone classification into urban, periurban or rural, etc.) to avoid the missing observation problem.

MCMC based method

As proposed in Section 5.3.2, we are interested in using a Gibbs sampler to draw from the underlying distribution of the agents attributes. The Gibbs sampler needs the conditional distribution of one attribute over all the other attributes (Fig. 5.2). These conditionals can be generated from the available data sources, i.e., cross-tabulations and microsample. In the simplest case, when the complete crossing is available in the cross-tables, they can be converted into full-conditionals. However, as pointed out in Section 5.4.1, it is rarely the case that the complete crossing is available. So, the conditionals have to be constructed using all the available data, domain knowledge, and bringing in assumptions where there is no information available.

Suppose we are interested in synthesizing person agents (A, S, H, E, Z). This means drawing from $\pi(A, S, H, E, Z)$. For the synthesis process to work, we need to compute the conditionals $\pi(A|S, H, E, Z)$, $\pi(S|A, H, E, Z)$, $\pi(H|A, S, E, Z)$, $\pi(E|A, S, H, Z)$, and $\pi(Z|A, S, H, E)$. In order to generate the population of person agents, suppose that we only have access to a 3% microsample of the region, but it is missing the attributes age and education level. Thus from the sample we only have information about $\pi(A, H, Z)$. At zonal level, the marginals for

sex $\pi(S)$ and household size $\pi(H)$ are available. In addition to that the conditional marginals of age on education level $\pi(A|E)$ are also available. Note that these marginals are available as counts of persons for each category.

Attribute	Required	Available[a]
Z	$\pi(Z\|A,S,H,E)$	$\pi(Z\|A,H)$ $\pi(Z\|A)$, $\pi(Z\|S)$, $\pi(Z\|H)$, $\pi(Z\|E)$
A	$\pi(A\|S,H,E,Z)$	$\pi(A\|H,Z)$ $\pi(A\|Z)$, $\pi(A\|E,Z)$
S	$\pi(S\|A,H,E,Z)$	– $\pi(S\|Z)$
H	$\pi(H\|A,S,E,Z)$	$\pi(H\|A,Z)$ $\pi(H\|Z)$
E	$\pi(E\|A,S,H,Z)$	– $\pi(E\|Z)$, $\pi(E\|A,Z)$

a: The source for first row in this column is the sample and zonal cross-tabs for the second.

Table 5.1 Description of inputs.

Using the available data as described in Table 5.1, we can generate the population for either a) the entire area simultaneously or b) zone by zone. The complete conditionals can thus be prepared as follows:

1. $\pi(Z|A,S,H,E)$
   Here we have two choices. First, we can invert the available marginal counts to develop $\pi(Z|A,E)$, $\pi(Z|H)$, $\pi(Z|E)$. Second option is to construct the person population zone-by-zone. In the first case, we can take one of the conditionals and assume that the value of Z is uniform across other variables, given the selected one. If we do that then we are not using all the information available to us in form of the other marginals. In the case of second option we can construct the other marginals for each zone. This way we will be making the best use of all the information available to us. This decision results in synthesizing population zone by zone.

2. $\pi(A|S,H,E,Z)$
   We have the counts available for $(A|H,Z)$ from the microsample and $(A|Z)$ and $(A|E,Z)$ from the zonal data. In the case of a zone-by-zone synthesis we can estimate an individual level RUM based conditional model for A that uses H at individual level and E at zonal level as explanatory variables. In addition to that, it can use other zonal information, e.g. average age or share of age categories in the zone, or information coming from other zonal statistics, e.g. number of schools, land use in the zone, etc. The quality of this model will depend on the availability of the explanatory variables, but here we want to emphasize that this type of

conditional model is a coherent way of using information from different sources and constructing the conditional marginals for the attributes.

3. $\pi(S|A, H, E, Z)$
   Here we only have $\pi(S|Z)$. If there is no other information we can assume that given Z, S is uniform across A, E, and H. Under this assumption $\pi(S|A, E, H, Z) = \pi(S|Z)$.

4. $\pi(H|A, S, E, Z)$
   Were too, as shown for $\pi(A|S, H, E, Z)$, the conditional models can be estimated by mixing the individual and zonal data.

5. $\pi(E|A, S, H, Z)$
   From the available sources, we only have counts for $(E|A, Z)$. As there is no further information available, we can assume that given Z and A, E is uniform across S, and H. Under this assumption, $\pi(E|A, S, H, Z) = (E|A, Z)$. We can also bring in any possible domain knowledge to justify or improve the conditionals. Here, for instance, we know that the education level is mostly dependent on age rather than sex or household size.

5.5 Application and results

5.5.1 Fitting and generation based method

In this section we summarize the findings of two previous contributions to the FG method (Müller & Axhausen 2011*a*, 2012). For the Zürich case study described in this book, the full year 2000 census was available (Swiss Federal Statistical Office (BFS) 2000*a*), and no synthetic population has been generated.

In Müller & Axhausen (2011*a*), the Hierarchical Iterative Proportional Fitting Algorithm (HIPF) has been first presented and compared with two other multi-level fitting algorithms: Entropy (Bar-Gera et al. 2009) and Iterative Proportional Updating (IPU, Ye et al. 2009). This was not a real case study but a simulation study. A 5% sample has been drawn from the complete census and used as input for the three algorithms. Each canton's population has been synthesized separately. The results of the comparison (see Figure 5.3) indicate an advantage of the new HIPF algorithm over the other two, especially for the less populated cantons.

Müller & Axhausen (2012) describe the preparation process of a household sample with detailed person information from the Swiss PUMS (Swiss Federal Statistical Office (BFS) 2000*b*), a freely available 5% sample of the Swiss population census that has been anonymized (cf. Sect. 5.4.1). Due to this anonymization, the data preparation was considerably more difficult than

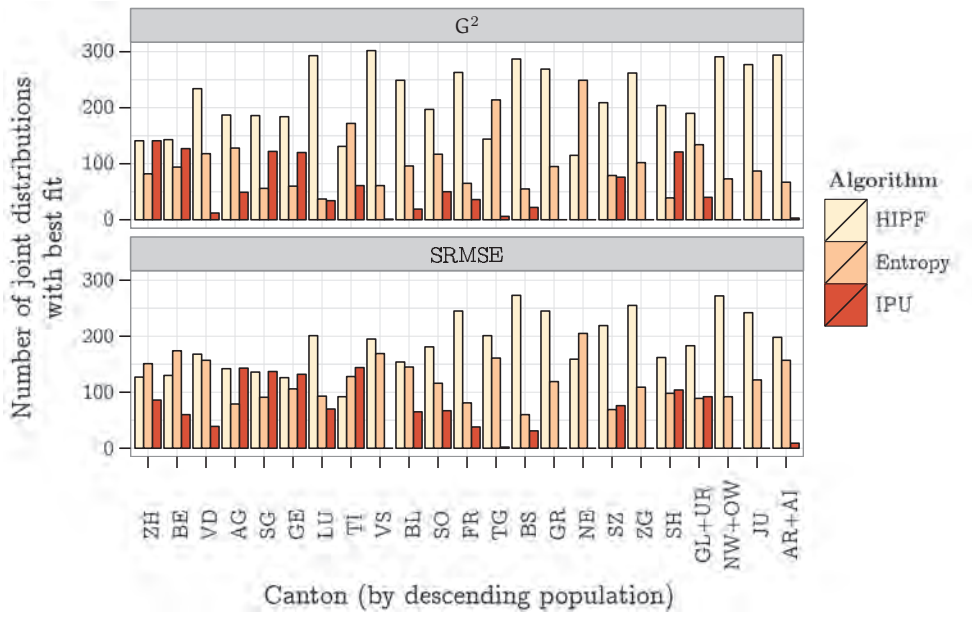


Figure 5.3 Results of the comparison between HIPF, Entropy, and IPU. All three-dimensional joint distributions of the attributes in the synthetic populations generated by the three algorithms have been compared to the true distribution (from the complete census) using two distance metrics.

initially anticipated. However, in this data preparation process, almost all attributes in the PUMS were treated, without a specific application in mind. This suggests that the effort required for the data preparation should not be underestimated, and should focus only on attributes really required in the target population. The data preparation has been carried out entirely in the R programming language (R Development Core Team 2014) following the paradigm of reproducible research.

Figure 5.4 shows an excerpt of preliminary results from synthesizing a population using the PUMS and freely available aggregated totals at both person and household levels. The mean number of full-time workers is similar for all household sizes and independent of the presence of children, although work status has not been controlled in the fitting stage.

5.5.2 MCMC based method

In the Brussels case study the data availability was highly limited. The size of the available micro-sample was 0.1% only, and the zonal level conditionals were also only partially available. The simulation based procedure was used to synthesize a population out of this information. Details on the application of simulation based methodology to Brussels case study can be found in Chapter 6 of this book.

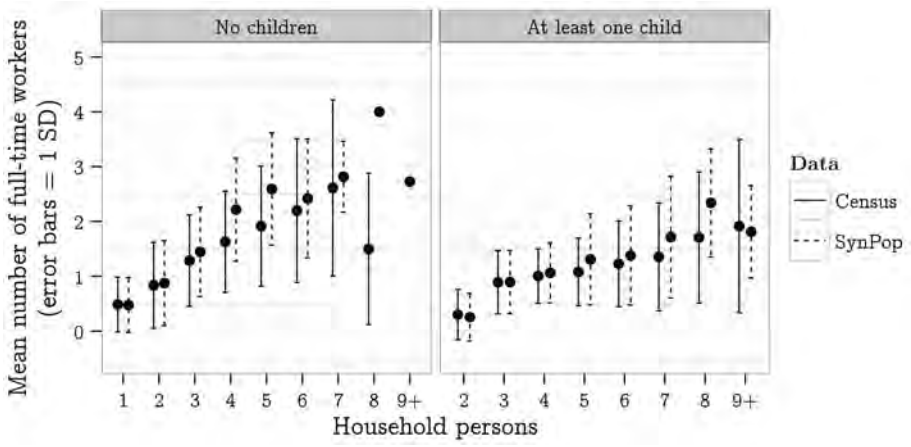


Figure 5.4 Excerpt of preliminary results for the generation of a synthetic population from the Swiss PUMS. The mean and standard deviation of the number of full-time workers in a household for households with a given size is compared between a population synthesized from the PUMS (dashed line) and the full census (solid line). Households with and without children are compared separately.

5.6 Discussion

In this chapter, two different strategies for generating synthetic populations have been presented: the fitting and generation (FG) and the Markov Chain Monte Carlo (MCMC) methods. The FG method requires a sample of the target population which is reweighted and sampled from. In contrast, the MCMC method constructs the target population from scratch by repeatedly sampling from conditional probabilities at the attribute level. While the FG method has been widely adopted and enhanced for almost two decades, the MCMC method is a new and entirely different approach to the problem.

Usually, the limiting factor is data availability: Large samples of individual data are difficult or impossible to obtain, and the anonymization process further complicates their usage especially for FG methods. However, once the data has been obtained and prepared, recent FG methods allow generating synthetic populations controlled at different aggregation levels (households and persons). On the other hand, the MCMC method overcomes these restrictions by employing a simple and robust method that allows using data and even models of different kinds as input for the synthetic population.

As an illustration, simulation based methodology described how a Gibbs sampler can be used for sampling from the joint distribution. Using an example, it was shown how a mix of models/counts, assumptions, and domain knowledge can be used to prepare the input. It was also shown how data from various sources can be coherently merged using discrete conditional models. Such an approach is particularly useful when the data from source (e.g. microsample)

is very limited. The MCMC approach can synthesize a mix of discrete (e.g. marital status) and continuous (e.g. income) attributes. Other samplers (e.g. Metropolis Hastings) or any combination of them can also be used.

In the context of large scale models for European cities in general and SustainCity project in particular, this research has played a critical role in developing methodologies for generating a representative base line population. A large part of the available literature focuses on data from the US. This research provides solutions that are more generic and are easily applicable in the European context.

Two further research directions are identified: First, extending the simulation based approach to accommodate the synthesis of associations among agents (e.g. households and persons). Second, the development of a hybrid approach in which concepts from FG are used to post-process the population generated from MCMC. This way a better fit with available marginals and conditionals can be maintained.

Chapter 6

Simulation based generation of a synthetic population for Brussels

Bilal Farooq, Ricardo Hurtubia, and Michel Bierlaire

6.1 Introduction

The microsimulation of urban systems evolution using UrbanSim and MATSim requires an agent population for the base year. As a first step in Brussels case study a set of household agents with rich attributes was synthesized. The main challenge here was to overcome constraint of very limited data availability. At the sector and commune level, we had access to up to three-way cross-classification tables from 2001 Census. The Census statistics contained aggregate information for the 1.2 million households of the area of study. While on micro-level, we only had access to an approximately 0.1% sample of the households (1367 records) in the study area. This sample came from the 1999 household survey conducted by Hubert & Toint (2002). The area of study consists of 151 communes that are further divided in 4945 sectors.

In the initial attempt we tried to use the Iterative Proportional Fitting (IPF) approach (Beckman et al. 1996) to synthesize the household population, but given the small size of the sample, it was a challenge. In our further investigation on another case study where we had access to the real population, we realized that to have a reasonable fit with the real population a sample size of at least 5% was required. We thus focused our attention to developing a new methodology for synthesis that can overcome the above mentioned challenges. This methodology involved using Markov Chain Monte Carlo (MCMC) simulation to draw agents directly from the joint distribution of their attributes rather than searching of an optimal weighting scheme using some fitting process. In particular, we used Gibbs sampler to simulate the draws. In each iteration, the potential new agent is generated by i) copying the agent generated in the last iteration, ii) selecting one attribute (X_i) of the agent using a uniform draw, and iii) using antithetic draw on the conditional-marginal distribution of X_i over the values of all the other attributes of the agent, the new value x_i is

generated. This process is repeatedly run to generate the agent population. Key contributions of this approach include: a) efficient use of all the available datasets, b) ability to generate all the possible synthetic populations that are based on the available data and are representative of the real population, c) generation of a truly heterogeneous population rather than cloning of the sample, d) ability to synthesize both discrete and continuous attributes of agents e) more flexible data preparation process. The details of the methodology and algorithm can be found in Chapter 5 and Farooq et al. (2013).

The major advantage of using a simulation based approach here was that data from various sources at various spatial and agent aggregation, domain knowledge, and assumptions can be consistently and seamlessly used to construct the conditional-marginal distributions required as input for the Gibbs sampler. MCMC based procedures allow for generation of a mix of discrete and continuous attributes. The proposed methodology does not suffer from the scalability issues with respect of the number of attributes. The only information that is required to generated the next agent is the previous agents synthesized and the input conditionals.

In the next section we first describe the attributes that were needed to be synthesized for the Brussels case study. The available data sources and input preparation process is described. The implementation details of the proposed methodology are discussed. We analyze the generated population and compare it to the available dataset. In the end, a concluding discussion is presented.

6.2 Data

The models estimated in UrbanSim and MATSim (e.g. dwelling or job location models) needed attributes of household and its members to evaluate their decisions during the simulation period. For this, as a starting point, the base year (2001) population with these attributes had to be synthesized. The agent type synthesized was *Household* that contained attributes representing resources (dwelling type/sector, and number of cars) and information on the person agents associated to it. The complete list of these attributes is described in Table 6.1, while details on models that used these attributes as explanatory variables, the parameter values, and other statistics can be found in Efthymiou et al. (2013).

The study area consists of 151 communes that are further divided in 4945 sectors. Data from three different sources were available for the study area. From 2001 Census sector and commune level statistics were available in form of cross-tabulations. Belgium Land Registry data provided a cadastre of real estate goods. These datasets were obtained at sector or commune level from the Belgian Statistical Authority (SPF - Economie, http://www.economie.fgov.be).

Attribute	Levels
Income level of the household (inc_h)	1 (0-1859 Euros) 2 (745-1859 Euros) 3 (1860-3099 Euros) 4 (3100-4958 Euros) 5 (>4959 Euros)
Household size ($size_h$)	1,2,3,4,5+
Number of children ($children_h$)	0,1,2+
Number of workers ($workers_h$)	0,1,2+
Number of cars ($cars_h$)	0,1,2,3+
Number of people with university degree ($univ_h$)	0,1,2+
Dwelling type (v_h)	house (3 types), apartment
Zone[a] (i)	4945

a We used Belgian Census Sectors as zones

Table 6.1 Household attributes synthesized for Greater Brussels Areas in 2001.

While on micro level, we had access to an approximately 0.1% sample of the household (1367) in the study area. This sample came from the 1999 household survey of Belgium, called MOBEL Hubert & Toint (2002).

To generate the new state in a Markov chain, the Gibbs sampler needs access to conditional distribution of one attribute on all the remaining attributes. This means that in total we needed 8 conditionals of as input. Table 6.2 lists out the exact details of the all these conditional-marginal distributions.

No.	Conditional
1	$\pi(inc_h \mid size_h, children_h, workers_h, cars_h, univ_h, v_h, i)$
2	$\pi(size_h \mid inc_h, children_h, workers_h, cars_h, univ_h, v_h, i)$
3	$\pi(children_h \mid inc_h, size_h, workers_h, cars_h, univ_h, v_h, i)$
4	$\pi(workers_h \mid inc_h, size_h, children_h, cars_h, univ_h, v_h, i)$
5	$\pi(cars_h \mid inc_h, size_h, children_h, workers_h, univ_h, v_h, i)$
6	$\pi(univ_h \mid inc_h, size_h, children_h, workers_h, cars_h, v_h, i)$
7	$\pi(v_h \mid inc_h, size_h, children_h, workers_h, cars_h, univ_h, i)$
8	$\pi(i \mid inc_h, size_h, children_h, workers_h, cars_h, univ_h, v_h)$

Table 6.2 Conditionals needed for synthesizing population of Great Brussels Area in 2001.

We chose to synthesize the population zone-by-zone. This meant that the first seven conditionals in Table 6.2 were required for each zone and last one was not needed. The full-conditionals were constructed using assumptions and discrete conditional models. The process is explained in the following sections.

6.2.1 Estimated models

Four models accounting for conditional relationships between characteristics and attributes were estimated, as described by Table 6.3. Of all these models, only the dwelling type model describes a choice, where the alternative (the type of unit) has attributes. The rest of the models (income, car ownership and education) describe the membership of the household to a particular level of the characteristic/variable that can not be described itself in terms of additional attributes. The estimated models do not pretend to provide a behavioral or causal explanation of the choices and characteristics of the households. Instead, they try to find structural (and significant) correlations between characteristics and attributes in order to build the conditional relationships needed for the simulation process.

attribute	household variables	spatial variables
inc_h	$cars_h, workers_h, univ_h, v_h$	$income_i$
$cars_h$	$inc_h, workers_h, univ_h, v_h, children_h$	$income_i$, $car_ownership_i$
$univ_h$	$inc_h, workers_h$	$univ_i$
v_h	$cars_h$, $size_h$	$surface_{vi}$

Table 6.3 Modeled conditionals for Great Brussels Area.

All models consider alternative-specific or level-specific utility functions where the membership to a specific level is described as a function of other socioeconomic characteristic or zonal attributes:

$$V_{in} = \sum_k \beta_k^i X_{nz_{(n)}k} \tag{6.1}$$

where i is the alternative or the "level" or the attribute (e.g. dwelling type or income level) and $X_{nz_{(n)}k}$ is the k-th attribute of household n, or zone $z_{(n)}$ (the residential location of household n). In some cases the attributes are related only to the household (X_n), only to the zone ($X_{z_{(n)}}$) or are an interaction of both dimensions ($X_{nz_{(n)}} = X_n \cdot X_{z_{(n)}}$). β_k^i is a parameter to be estimated which is specific to alternative or attribute-level i. The introduction of spatial information (z) produces a richer set of conditional distributions, allowing to have a different conditional relationship for each zone.

Dwelling type model

Households can live in four types of dwellings: Isolated houses, semi-attached houses, attached houses and apartments. The choice of type of dwelling depends on characteristics of the household like size and car ownership, and zonal attributes like the average size of each type of dwelling in each zone. Estimation results are shown in Table 6.4

Parameter	Variable	Value	Std err	t-test
ASC^2	constant for dwelling type 2	0.423	0.297	1.42
ASC^3	constant for dwelling type 3	0.87	0.305	2.86
ASC^4	constant for dwelling type 4	1.2	0.327	3.68
$\beta_{surf \times h2}$	dummy for hh size=2 x zonal avg surface of dwelling[b]	0.0146	0.00533	2.74
$\beta_{surf \times h3}$	dummy for hh size=3 x zonal avg surface of dwelling[b]	0.0194	0.00597	3.25
$\beta_{surf \times h4+}$	dummy for hh size>3 x zonal avg surface of dwelling[b]	0.0249	0.00299	8.31
β^2_{cars}	number of cars in the household	−0.279	0.182	−1.53
β^3_{cars}	number of cars in the household	−0.593	0.207	−2.86
β^4_{cars}	number of cars in the household	−0.948	0.233	−4.07
A^1	ln of number of dwellings of type 1 in zone[b c]	1	—	—
A^2	ln of number of dwellings of type 2 in zone[b c]	1	—	—
A^3	ln of number of dwellings of type 3 in zone[b c]	1	—	—
A^4	ln of number of dwellings of type 4 in zone[b c]	1	—	—

* dwelling types are: isolated house (1), semi-attached house (2), attached house (3) and apartment(4)
a The alternative of isolated house is used as a reference ($\beta^1 = 0$ in all cases with the exception of the surface parameters)
b spatial variable
c expansion factor used to account for the (un)availability of different types of dwelling in different zones. Not an estimated parameter

Table 6.4 Dwelling type model[*a].

All parameters have the expected sign and most of them are significant at the 95% level with the exception of β^2_{cars}, which is significant at the 88% level. To avoid under identification, all parameters are set to zero for the isolated house type, with the exception of the surface-related parameters ($\beta_{surf \times h2}$, $\beta_{surf \times h3}$ and $\beta_{surf \times h4+}$) which are generic and can be fully identified for all alternatives since the average surface of the dwelling varies between different types of units in the same zone. We observe that larger households tend to prefer units with a bigger surface, meaning that we should observe a concentration of large households in zones where the average size of units is big and also in unit types with bigger average surface. The number of cars in the household has a negative effect for (semi)attached houses and apartments, which is expected since this type of housing tends to be located in denser areas of the city, where parking spaces are scarcer. The utility function is corrected with the expansion factor A^i in order to take into account the fact that, for each zone, households are distributed across types of dwelling according to their availability. This also makes sure that the probability of a household locating in a type of dwelling that does not exist in a zone will be zero.

Income level model

Membership to an income level is modeled as a function of household characteristics and the average zonal income of the residential location. There are five possible income groups: from low (level 1) to high (level 5). Results for this model are shown in table 6.5.

Parameter	Variable	Value	Std err	t-test
ASC^2	constant for income level 2	−0.86	0.789	−1.09
ASC^3	constant for income level 3	−4.64	0.901	−5.14
ASC^4	constant for income level 4	−8.31	1.12	−7.39
ASC^5	constant for income level 5	−10.6	1.55	−6.82
β^3_{educ}	dummy for presence of people with higher educ in the hh	0.831	0.177	4.69
β^4_{educ}	dummy for presence of people with higher educ in the hh	1.72	0.314	5.49
β^5_{educ}	dummy for presence of people with higher educ in the hh	1.92	0.656	2.93
$\beta^2_{zonal_inc}$	average zonal income[b]	0.0008	0.0004	1.84
$\beta^3_{zonal_inc}$	average zonal income[b]	0.0012	0.0005	2.55
$\beta^4_{zonal_inc}$	average zonal income[b]	0.0016	0.0005	3.09
$\beta^5_{zonal_inc}$	average zonal income[b]	0.0016	0.0006	2.47
β^2_{cars}	number of cars in the household	1.16	0.265	4.39
β^3_{cars}	number of cars in the household	1.92	0.299	6.41
β^4_{cars}	number of cars in the household	2.33	0.341	6.83
β^5_{cars}	number of cars in the household	3.2	0.466	6.87
β^3_{house}	dummy for dwelling being a house	0.45	0.193	2.34
β^4_{house}	dummy for dwelling being a house	0.485	0.294	1.65
β^5_{house}	dummy for dwelling being a house	0.485	0.294	1.65
$\beta^2_{workers}$	number of workers in the household	1.14	0.277	4.11
$\beta^3_{workers}$	number of workers in the household	2.22	0.295	7.53
$\beta^4_{workers}$	number of workers in the household	2.46	0.345	7.13
$\beta^5_{workers}$	number of workers in the household	1.74	0.428	4.07

a the super-index in the parameter indicates to which income level (1,2,3,4,5) it is specific. Income level 1 is used as a reference ($\beta^1_* = 0$)
b spatial variable

Table 6.5 Income level model[a].

All parameters show the expected sign and are significant at the 90% level. Because income levels do not have attributes to identify them, and to avoid under-identification, all parameters are set to zero for the lowest income category (level 1). Households that have members with university degree and that are located in high income zones are more likely to belong to higher income

levels, as expected due to observed socioeconomic agglomeration. The number of cars in the household also has a positive effect on the income that is stronger for higher income levels. Households living in a house have a higher probability of belonging to the higher income levels than those living in apartments. Finally, the number of workers in the household also explain a higher income level, although it has a lower effect for the highest level than for mid and mid-high income level, this is probably due to the fact that the highest income levels depend more on the type of jobs than on the number of active persons in the households.

Car ownership model

Car ownership is modeled as a function of household socioeconomics, such as income, education level, presence of children and type of dwelling, and zonal attributes like the average income and the number of households by car ownership-level by zone. Four possible levels of car ownership are modeled: no cars (level 0), one car (level 1), two cars (level 2) and three or more cars per household (level 3). Results are shown in Table 6.6.

All parameters show the expected sign and are significant at the 94% level. Household with members having a university diploma and with higher income levels are more likely to have a larger number of vehicles. Households located in zones where the average income is high and where households tend to have more cars are more likely to own a larger number of vehicles. Households with children living in houses and with working members have a higher chance of owning more cars.

Education level model

The number of individuals with university degree in a household is modeled as a function of socioeconomics like income, car ownership, and number of workers. The presence of other highly educated households in the residential location is also considered as a spatial explanatory variable. Three household education levels are considered: no one in the household has a diploma (level 0), only one person in the household has a diploma (level 1) and 2 or more people in the household have an university diploma (level 2). Results are shown in Table 6.7.

All parameters show the expected sign and are significant at the 93% level. Households with more cars, higher income, and more workers are more likely to have more individuals with a high education level. Households located in zones with a higher percentage of households with highly educated members are more likely to have a higher education level.

Parameter	Variable	Value	Std err	t-test
ASC^1	constant for 1 car	−2.75	0.611	−4.5
ASC^2	constant for 2 cars	−7.02	0.812	−8.65
ASC^3	constant for 3+ cars	−10.1	1.23	−8.2
β^1_{educ}	dummy for presence of people with higher educ in the hh	0.504	0.196	2.57
β^2_{educ}	dummy for presence of people with higher educ in the hh	0.933	0.267	3.49
β^3_{educ}	dummy for presence of people with higher educ in the hh	1.07	0.552	1.94
$\beta^1_{\text{high_inc}}$	dummy for households with high income (>3)	0.977	0.499	1.96
$\beta^2_{\text{high_inc}}$	dummy for households with high income (>3)	2.43	0.569	4.28
$\beta^3_{\text{high_inc}}$	dummy for households with high income (>3)	3.24	0.801	4.05
$\beta^1_{\text{mid_inc}}$	dummy for households with mid income (=3)	0.858	0.267	3.22
$\beta^2_{\text{mid_inc}}$	dummy for households with mid income (=3)	1.87	0.342	5.46
$\beta^3_{\text{mid_inc}}$	dummy for households with mid income (=3)	1.42	0.676	2.11
$\beta^1_{\text{zonal_inc}}$	average zonal income[b]	0.001	0.0003	3.51
$\beta^2_{\text{zonal_inc}}$	average zonal income[b]	0.0013	0.0004	3.29
$\beta^3_{\text{zonal_inc}}$	average zonal income[b]	0.0013	0.0004	3.29
$\beta^1_{\text{car1_zone}}$	percentage of hh's with 1 car in the zone[b]	0.498	0.172	2.89
$\beta^2_{\text{car2_zone}}$	percentage of hh's with 2 cars in the zone[b]	2.13	0.842	2.53
$\beta^3_{\text{car3_zone}}$	percentage of hh's with 3+ cars in the zone[b]	14.1	7.57	1.86
$\beta^1_{\text{children}}$	dummy for presence of children in the household	0.457	0.24	1.9
$\beta^2_{\text{children}}$	dummy for presence of children in the household	0.8	0.276	2.9
β^1_{house}	dummy for dwelling being a house	0.841	0.191	4.4
β^2_{house}	dummy for dwelling being a house	1.86	0.289	6.42
β^3_{house}	dummy for dwelling being a house	2.66	0.776	3.43
β^1_{workers}	number of workers in the household	0.437	0.139	3.15
β^2_{workers}	number of workers in the household	1.24	0.193	6.42
β^3_{workers}	number of workers in the household	1.6	0.358	4.46

a the super-index in the parameter indicates to which car ownership level (0,1,2,3+) it is specific. The alternative of 0 cars is used as a reference ($\beta^0_* = 0$)
b spatial variable

Table 6.6 Car ownership model[a].

6.2.2 Assumptions-based conditionals

For the rest of the attributes (i.e. household size, number of children, and number of workers), we did not have enough information to estimate the discrete-conditional models. Here we had to rely on the assumptions and domain knowledge to construct the full-conditionals. The only dataset that has information

Parameter	Variable	Value	Std err	t-test
ASC^1	constant for 1 person with high educ in hh	−2.96	0.34	−8.72
ASC^2	constant for 2+ persons with high educ in hh	−7.19	0.547	−13.14
β^1_{cars}	number of cars in the household	0.238	0.133	1.79
β^2_{cars}	number of cars in the household	0.701	0.156	4.51
$\beta^1_{\text{educ_zone}}$	percentage of hh's with educ level 1 in zone[b]	3.34	0.566	5.91
$\beta^2_{\text{educ_zone}}$	percentage of hh's with educ level in zone[b]	4.34	0.708	6.13
β^1_{income}	income level of the household	0.24	0.129	1.87
β^2_{income}	income level of the household	1.09	0.152	7.18
β^1_{workers}	number of workers in the household	0.393	0.113	3.47
β^2_{workers}	number of workers in the household	0.851	0.154	5.52

a: the super-index in the parameter indicates to which household education level (0,1,2+) it is specific. The alternative of 0 persons is used as a reference ($\beta^0_* = 0$)
b: spatial variable

Table 6.7 Household education level model[a].

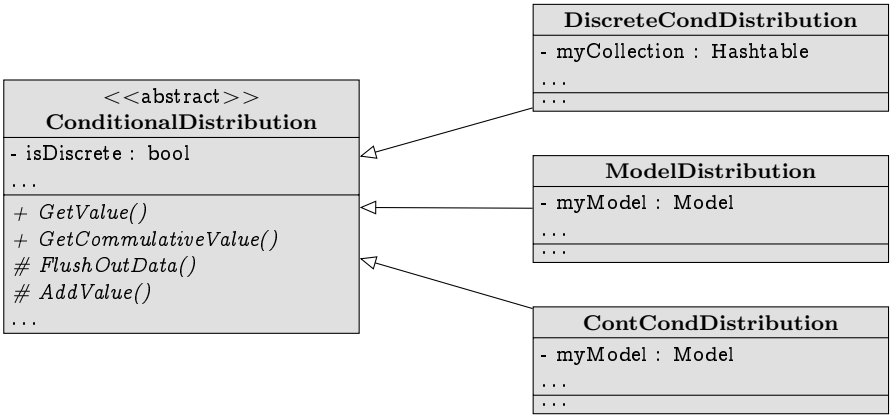
on these three attributes was MOBEL. As mentioned earlier, this dataset is very small and various zones (Census sectors) are either not represented or have very few observations. We thus decided to pool the dataset. The actual conditionals were generated by counting each category of an attribute for rest of the attributes combinations. For example, counting the number of households with size equal 1, 2, 3, 4, and 5+ for all the possible category combinations of income level, children, workers, cars, education, and dwelling type. We assumed that the household structure for the household size, number of children, and number of workers follow the same conditional distributions in all the sectors of the study area. We understand that in general there are differences in the structure among households living in city centres, uptown, suburbs, etc. But in the absence of any other information, we are not able to make the conditionals sensitive to the spatial heterogeneity for these three attributes.

6.3 Population generation

To operationalize the simulation based methodology for the generation of synthetic population, we developed a generic object oriented software called **Sim**ulation based **P**opulation **Syn**thesizer (**SimPSynz**). This software can i) pre-process the data to generate full-conditionals of the attributes, ii) generate a large pool of synthetic agents using different types of samplers, and iii) produce any number of realization of synthetic populations from that large pool. These realizations can be based on user defined marginals or just uniform sampling. The implementation is available upon request in $C^{\#}$ and Java programming languages.

6.3.1 Software architecture

An object oriented class structure was designed.[1] The software architecture was divided into three interacting group of class libraries: *Distributions*, *Samplers*, and *SimulationObjects*. A top level handler class (PopSyn) was implemented in order to integrate the functionality from these libraries into a coherent software. Here we will only describe the main classes of these three libraries. The first library maintains the conditional distributions in form of counts (DiscreteCondDistribution), discrete-conditional models (ModelDistribution), and continuous functions (ContCondDistribution). The structure and association of this library is described in Figure 6.1. Note that at the moment, functionality of the ContCondDistribution class is not completely implemented.



a : "..." denotes that there are additional attributes and functionalities of a class that we did not list here for clarity reasons.

Figure 6.1 Class association for *Distributions* objects[a]

.

The second set of classes represents the four different types of samplers. The main class used to synthesize the population for the Brussels study area was the Gibbs sampler (*GibbsSampler*). However, for further options, we also implemented Metropolis Hasting (*MHSampler*), importance (*ImportanceSampler*), and antithetic (*AntitheticSampler*) samplers. All these classes inherit basic functionality from the main abstract class called *Sampler*. The structure and association of this library is described in Figure 6.2. *GibbsSampler* operates on various conditionals and with the help of a contained *AntitheticSampler* object, it generates the agents iteratively. The iteration parameters (e.g. warm up time) are set in the parent class. The *MHSampler* needs an object of type *StateGenerator* and an evaluator (*StateEvaluator*) to gener-

[1] Here, it is assumed that the reader is familiar with the basics of Unified Modeling Language (UML)

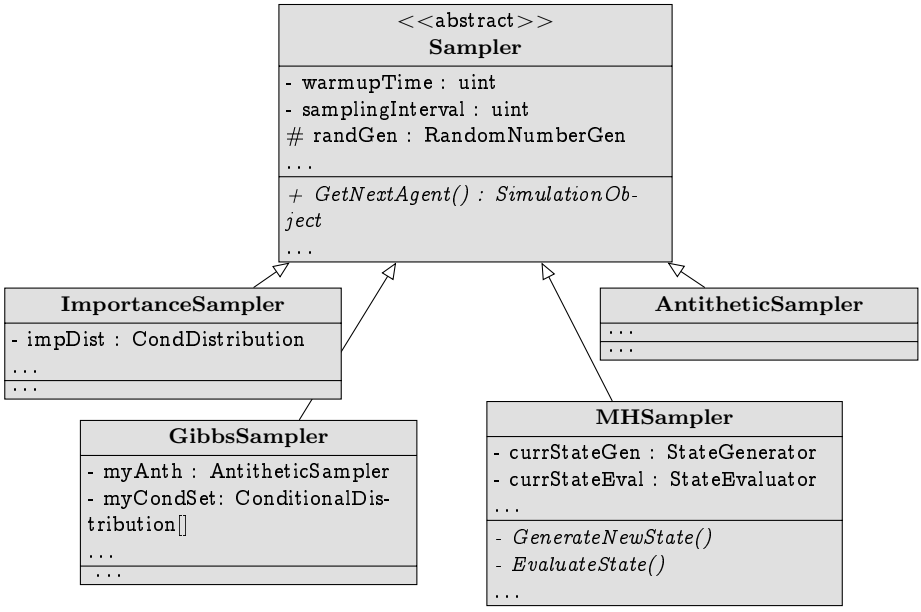


Figure 6.2 Class association for *Samplers* objects.

ate an agent. It also work in iterative manner and inherits paramters from the parent class, *Sampler*.

The classes that are involved in the actual simulation are defined in *SimulationObjects* class library. The basic functionality is defined in the *SimulationObject* abstract class, while all the other classes inherit this class. Figure 6.3 represents the class diagram of the library. *World* maintains and manages the states of various simulation objects. Furthermore, it also controls the execution of different samplers and the state of available information. In essence it represents the functionality of the actual World in virtuality. Space is represented by *SpatialZone* class. It contains the information and operations relevant to specific space. The class is designed such that it can be extended in hierarchy to accommodate different types of zoning systems (sectors, communes, etc.). Currently there are two different types of agent classes, *Person* and *Household*. In the Brussels case study, only the *Household* class was utilized. This library also maintains supporting enumerations and structures that represents the agents' attributes, types, and associations.

In addition to the three main libraries, *Utilities* class library was defined that contained support-classes. They covered the functionality related to ran-

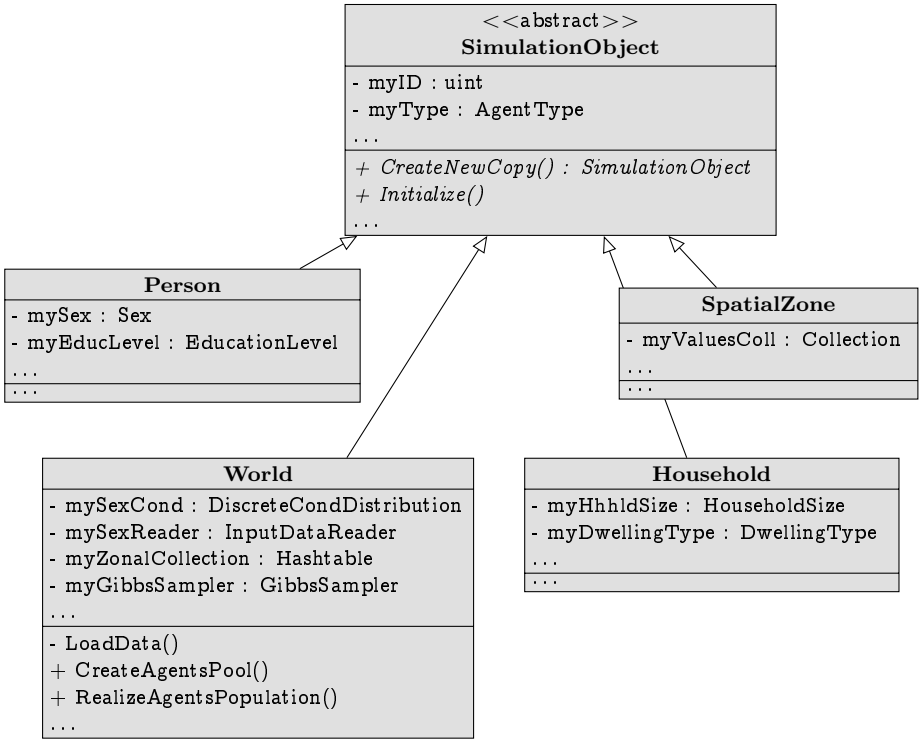


Figure 6.3 Class association for *Simulation* objects.

dom number generations (based on various distributions), file reading/writing, generation of the conditionals from the raw data, and various conversion utilities.

6.3.2 Simulation runs

The actual synthesis of the population for Brussels study area involved first preprocessing the available datasets using the utilities developed in SimPSynz. Once the conditionals were available via preprocessing process, the Gibbs sampler in SimPSynz was used to generate a household agent pool. It contained around 100 million agents with 10,000 or more agents in one sector. Synthesis was done zone-by-zone. UrbanSim requires the number of dwellings to match exactly with the number of households in the zones. The sector level dwelling type marginals were thus used to sample the synthetic population from the pool of agents. The run-time to generate a pool of 100 million agents on an Intel Core 2 Duo processor with 4 gigabytes of memory was approximately 8 hours.

6.4 Results

After the simulation-based synthesis process is finalized, results are validated by comparing them with observed commune level statistics obtained from the Population Census and the Belgian tax authority.

To analyze the goodness of the synthetic population in terms of distribution of household size, Figure 6.4 shows the simulated population (number of people) aggregated at the commune level in the Y-axis and the census statistic in the X-axis. The plot also shows the tendency line (solution to the equation $Y = \alpha X$) and the fit (R^2) for the data. The fit between the simulated and observed data is good (R^2=0.98). However, the synthetic population overall tends to underestimate the number of people by commune ($\alpha = 0.86$), especially for large communes with more than 70,000 people. This difference is expected since the census data accounts for a slightly larger total number of households in the whole region because it considers households located in dwelling of type "other" that was not included in our model.

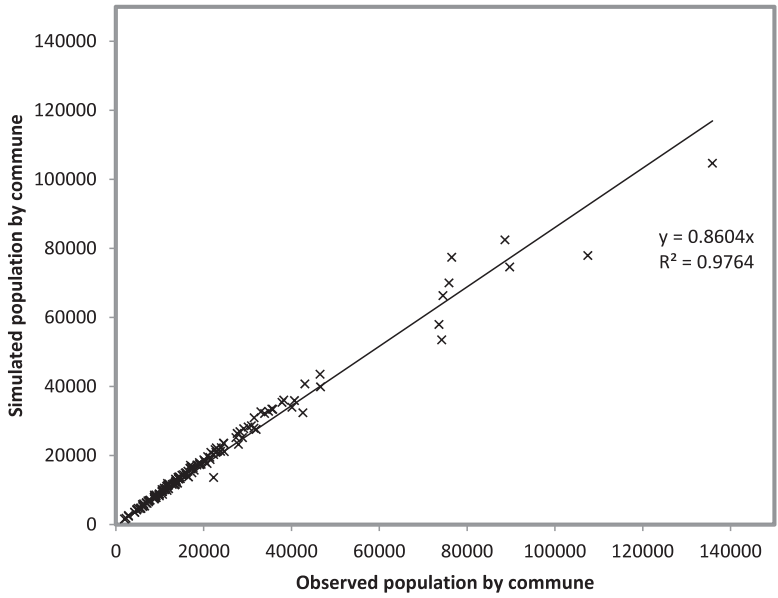


Figure 6.4 Population by commune.

Figures 6.5, 6.6 and 6.7 compare the synthetic population with the observed number of households with no car, one car, and two or more cars respectively. The fit for all three car-ownership levels is considerably good, with a R^2 value of 0.98, 0.99, and 0.94 respectively. While households with no cars tend to be underestimated in the synthetic population, household with one, two, and more cars are only slightly overestimated.

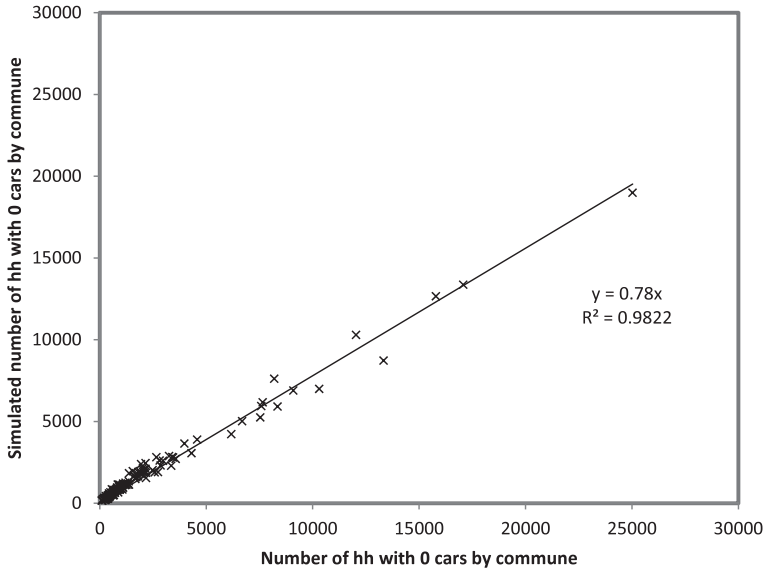


Figure 6.5 Households with 0 cars by commune.

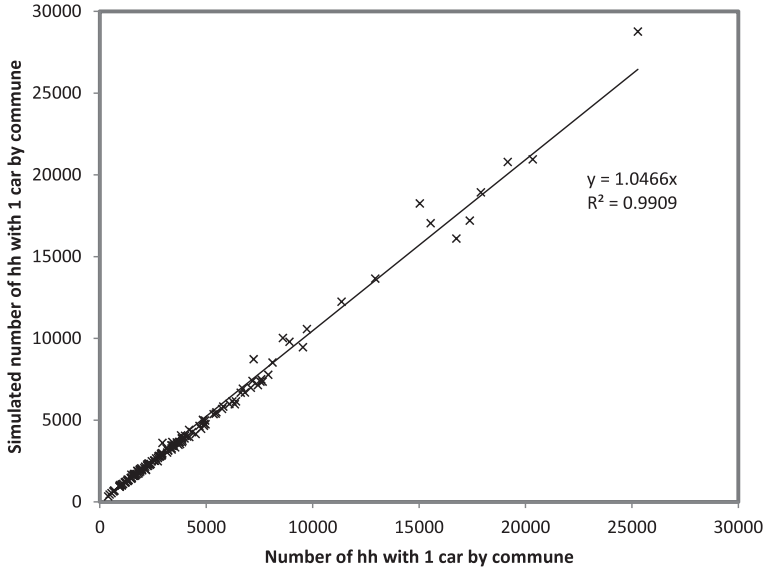


Figure 6.6 Households with 1 car by commune.

Figure 6.8 shows the number of households with one or more people having a university degree by commune. While the fit is good, the synthetic population tends to overestimate the number of university degrees. The trend is, however, correctly followed.

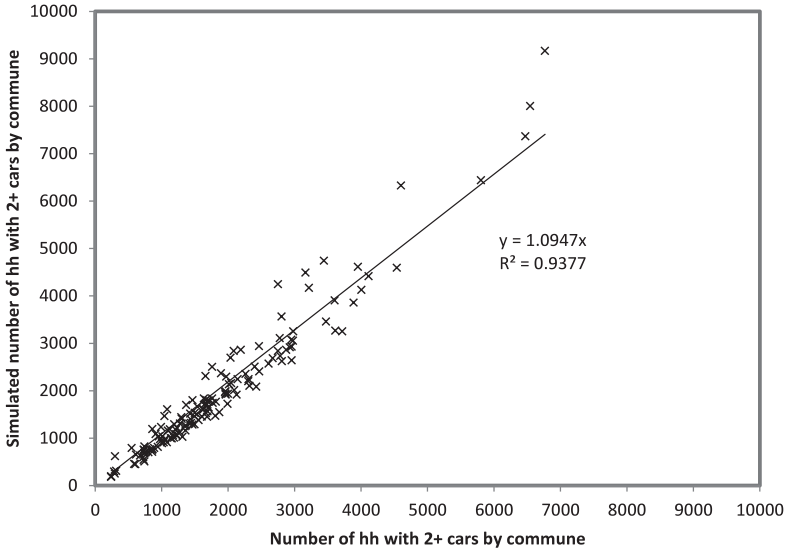


Figure 6.7 Households with 2+ cars by commune.

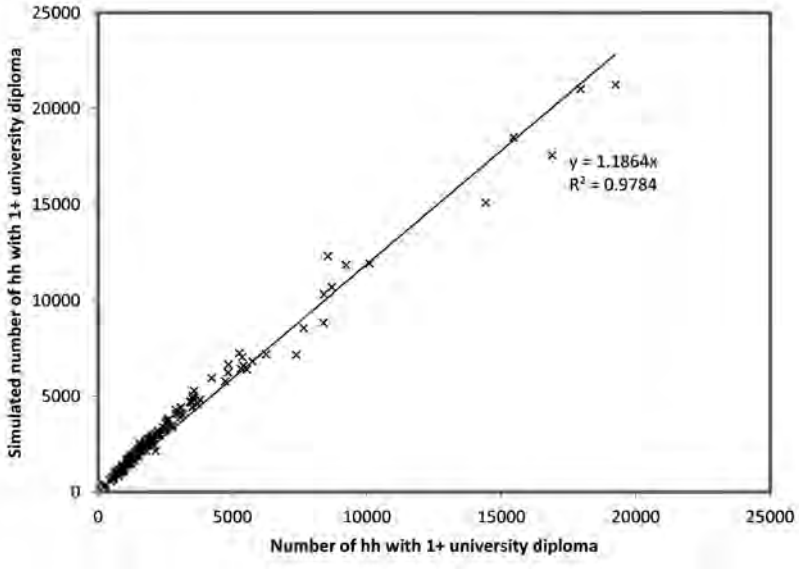


Figure 6.8 Households with 1+ university degree by commune.

Finally, Figure 6.9 shows the observed average household-income by commune, obtained from the Belgian tax authority, compared with the simulated income. The average income from the synthetic population is computed as the number of households by income level multiplied by the median value of each income level. Although this is a much harder variable to fit, results are reasonably good, with a R^2 value of 0.76. The plot shows that the synthetic

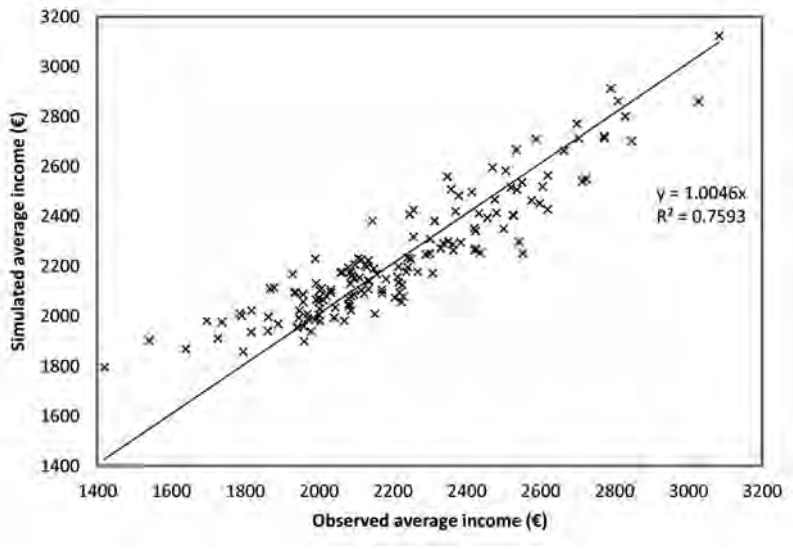


Figure 6.9 Average income by commune.

population tends to overestimate the average income for the poorer communes, however the overall trend is correctly followed.

6.5 Concluding remarks

Most of the effort in terms of methodologies for population synthesis is concentrated on fitting based approaches. Farooq et al. (2013) has pointed out various shortcomings of such approaches, including i) fitting of one contingency table, while there may be other solutions matching the available data, ii) due to cloning rather than true synthesis of the population, losing the heterogeneity that may not have been captured in the microdata, iii) over reliance on the accuracy of the data to determine the cloning weights, and iv) poor scalability with respect to the increase in number of attributes of the synthesized agents. In this regard, the simulation based approach was proposed. Here we showed an application of the methodology by synthesizing a population for the Brussels case study. The available dataset was highly limited. We had access to only a 0.1% microsample for the area. Conventional fitting based approaches were not very feasible. We thus exploited the flexible input requirements of the simulation based approach to make the best out of the available datasets.

A software tool called *SimPSynz* was developed for the operationalization of the methodology. Household agents with eight attributes, including location, dwelling type, car ownership, income level, household size, number of children, number of worker, and education level. A mix of conditional-choice models, assumptions, and domain knowledge were used to prepare the input.

The population was synthesized zone-by-zone. The analysis of the generated population indicated a good fit with the available data.

The literature on methodology for population synthesis is predominantly based on North American datasets (Pritchard 2008, Müller & Axhausen 2011*c*). Extending these methodologies in order to use census formats from other geographic regions (including Europe) is not trivial. In the context of the SustainCity project, the simulation based methodology and its application to the Brussels case study provides a step forward towards approaches that are more flexible to adapt and are directly applicable in the European context.

In the future we intend to extend the simulation based methodology to other cities in Europe and other parts of the world. We intend to develop SimPSynz as a generic tool for the synthesis of agent population and their associations. Moreover, due to its simulation based nature, SimPSynz is planned to be embedded directly in the microsimulation frameworks for modeling the urban systems dynamics.

Chapter 7

Agent-based microsimulation of population dynamics

Lorenzo Turci, Sophie Pennec, Laurent Toulemon, Arnaud Bringé, Rodolfo Baggio, and Elisabeth Morand

7.1 Introduction

Demographic microsimulation models differ from traditional population projection models in two ways. Firstly, they operate at the individual level rather than at aggregate macro level; secondly, they produce results based on the sum of repeated stochastic-driven transitions between states at individual level rather than expectancies of occurrence of events at aggregate level.

During the 1990s, many microsimulation models have been developed in many countries in order to study population changes and demographic and economic consequences of specific policy measures, not only at aggregate level but also at individual level and for specific subgroups of the population.

For the SustainCity project, a demographic microsimulation model was developed using a cross-sectional, closed, and discrete approach. Its aim is to analyze demographic changes at both individual and household levels in "a local area". Each individual belongs to one and only one household. One reference person is defined for each household. Besides the traditional family household (a single person or a couple with or without children), this model considers and simulates complex households that include most of the other forms of cohabitation (e.g. multi-generation households, home-sharing, etc.). It also considers non-private households. To ensure the quality of the simulation, alignment and calibration against some aggregate data had to be performed.

The main types of models which simulate individual behaviour are Microsimulation Models (MSMs) and Agent Based Models (ABMs). The rationale of both types of dynamic individual simulation models is presented in the first section. In the second section, we describe the demographic module

developed for the SustainCity project. The third section deals with calibration of the model and some results based on the Paris region case study.

7.2 State of the art

7.2.1 Introduction

In the 1950s, to avoid some drawbacks of simulation by aggregates and to better capture the complexity of real situations, Orcutt (1957) proposed a model "of various sorts of interacting units which receive inputs and generate outputs" , inputs being "anything which enters into, acts upon, or is taken into account of, by the unit" and outputs "anything which stems from, or is generated by, the unit". Its principal aim was to better simulate variants of policies and in particular their impact taking into account the heterogeneity of individuals cases. With this kind of model it is possible to check whether a policy affects the individuals targeted by the legislation and to see if there is no counterproductive effect of the policy (e.g. a benefit that makes the targeted individual worse off than without the benefit), in order to determine the winners and losers of policy/taxes changes. An important feature of this approach is the possibility to look at the results using the distribution of individuals rather than to rely on average situations produced by aggregate models, like component population projections in demography.

The subsequent implementation of this idea, first carried out by the same author with the creation of DYNASIM (Orcutt et al. 1961) and DYNASIM2 (Orcutt et al. 1976) and later followed by many researchers and institutions (see Sect. 7.2.3 for examples) defined a set of features which today keep characterising MSMs. First of all, these models strongly rely on individual data, on a large amount of them; their main aims have often been predictions and analysis of the impact of policies.

MSMs usually take a sample of individuals as initial population with all characteristics needed for transition. The characteristics evolved over time using transition probabilities.

MSMs have been extensively used to model nation-wide demographic changes which acts as a starting point for more complex modules (especially tax-benefit, pension, health, and transport policies). It has also been used to study kinship (Le Bras 1973, Ruggles 1987, Pennec 1997, Wachter et al. 1997, Tomassini & Wolf 2000). For a review of MSMs benefits over aggregate projections see Van Imhoff & Post (1998).

Nevertheless, these examples do not entirely satisfy Orcutt's initial goal: firstly, because they often consider model and behavioural responses to macro changes as "passive", and secondly, because they often do not take into account the interactions between these agents behaviours.

It may be interesting to analyse the impact of a macro-level change (e.g. policy change) on individual decisions and then the impact of individual decisions on the macro level. Two main streams of research deal with this idea of feed back between macro and micro levels. The top-down approach consists in a combination between MSMs and macro-oriented Computable General Equilibrium (Hérault 2005). A second way, the bottom-up approach, consists in simulating interactions between autonomous agents in order to study the emergence of global features. This idea joins another approach, the Agent Based Models (ABMs) or Multi Agent Systems (MAS) (Blanchet 2014). Origin ABMs can be traced back to Von Neumann (1966). Later, well-known pioneering examples of ABMs are John Conway's Game of Life (Gardner 1970), a grid of cells which can be dead or alive depending on the local alive cell density (they can die both of overcrowding or of isolation) and the famous segregation model by Schelling (1971), which demonstrated that urban segregation can arise even in a rather tolerant society.

MSMs and ABMs share many theoretical foundations, while they differ in their practical applications. ABMs are first and foremost used to develop and explore different models and theories, while MSMs usually intend to empirically evaluate the consequences of given estimated actions to be performed on the system under study or to perform elaborated population projections based on interactions between individuals. For these reasons, ABM techniques appear to be a complement to MSMs.

7.2.2 Microsimulation models

Continuous versus discrete time

Similar to the difference between statistical models of regression and event history analysis, the two types of dynamic microsimulation models differ in their time framework.

In continuous time models, a random process generates the durations to all potential events for each member of the population. The first event, i.e. the event occurring closer to the beginning point, is taking place. The time of occurrence of this first event is then used as the new starting point and the whole procedure, to define the next event, is repeated until an event of exit from the population (migration or death) occurs.

In discrete time models, time periods are treated in a sequential way. For each time period, for all individuals, all possible transitions are determined (without taking into account the exact time of the event within the time period). Each event is considered to occur only once during the period. When different events may occur during the same period of time (competing events), the simultaneous events are considered explicitly by the order chosen for the different events.

Continuous time models have certain theoretical advantages (Willekens 2006), but their implementation is more complex and model operations are less transparent for the user than for discrete time model (Scott et al. 2003, p.5). Calibration and alignement are also much more difficult to perform (Cassells et al. 2006). If the time period unit is very small, discrete-time models can be consider as "pseudo-continuous" models, and in practice they lead to very similar results.

Open vs. closed microsimulation models

Microsimulation models can be open or closed. The distinction does not consider here potential relations with the rest of the world, but the relations within the population itself: in closed models, individuals are explicitly interrelated, while in open models each individual is simulated without reference to the rest of the population (Harding 1993, Cassells et al. 2006).

In open microsimulation models, the events that an individual experiences during each time period are simulated independently from the other individuals: the individuals enter a union, have children, etc., without any explicit relationship with any other individuals of the population under study. For instance, having a partner is an individual event; the partner and children are not explicit individuals of the population, they are created for the sake of simulating the individual lifepath and are not members of the population.

On the contrary, closed models consider the characteristics of units of the population including links between the individuals when appropriate (most often family/household ties). In closed models, individuals can enter the population by birth or immigration and exit by death or emigration. These individuals who enter are new members of the population under study and "full" members in the sense that they are simulated like the other ones for the future time periods of family ties. Moreover, as closed models keep track of the ties between the individuals and in particular family ties, some events occurring to one individual may affect other individuals of the population. For example, when a married person dies, his/her spouse becomes a widow or a widower; if he/she had children, they become orphans following this event. In practice each individual is related to his/her father and mother, and his/her successive partners. Other family links (children, step-children, grand-parents) can be reconstructed from these basic links. This allows describing all family bonds at each period of time, such as parity (number of siblings), or to know whether the father and mother of an individual are still alive, etc.

A main difference between open and closed models relates to how the matching process between partners is handled when a union formation occurs. In an open model, when an individual forms a union, a partner is created with suitable characteristics. This partner is not a full member of the population, he/she only exists by his/her relationship to the full member and never

becomes a full member (all his/her characteristics will be removed from the model when the full member dies or emigrates). In a closed model, a "marriage market" is performed to match those entitled to partner. This procedure matches the partners according to some characteristics (i.e. age, education level, socio-economic status). Maximisation techniques are used and some behavioural rules are applied for those who can't find a perfect partner, they may relax their wishes in terms of partner's characteristics or remain in the marriage market for another time period. Closed models can include migrations, assuming that the "rest of the world" is a tank of individuals where immigrants are taken from and where emigrants are sent to, or where immigrants are "cloned" from the current population.

7.2.3 Five examples of dynamic microsimulation models

For the sake of brevity, we only present here some European models; for a more complete survey see Morand et al. (2010), Li & O'Donoghue (2013), and Blanchet (2014). We consider the following five Dynamic microsimulation models: MOSART (Norway), SESIM (Sweden), SAGE (UK), DESTINIE (France) and SVERIGE (Sweden).

All these models simulate social policies (e.g. pensions, family benefits, support for aged people). For each model, we focus, in this section, on the following aspects:

- initial population
- events simulated and the modeling sequence
- type of data used for estimations

All these models were built using a similar structure: an initial base population, a framework of simulation of all events and their order of occurrence, and simulation outputs. Within the simulation cycle a set of functions calculates the probability of occurrence of the events. Results are given either according to predefined tables or through a "survey type" output of individuals, the user being able to produce tables or other results of interest.

MOSART (Fredriksen & Stølen 2007) is designed for the projection of population, human capital, labour supply, and public pension benefits. The model is used by the Norwegian Ministry of Finance and the Ministry of Labour and Social Inclusion, as well as for research projects at Statistics Norway.

SESIM (Sundberg 2007) was developed by the Swedish Ministry of Finance to study the pension system and the impact of ageing on public finance.

SAGE (Evandrou et al. 2007) is the result of the work of ESRC Research Group "Simulating Social policy for an Ageing Society". Its aim is to

study the impact of the population ageing in terms of pensions and health issues and long-term needs.

DESTINIE (Blanchet & Le Minez 2007) is designed for studying the French pension scheme and related ageing issues. The model computes social security contributions, benefits and taxes and simulates the socio-economic changes.

SVERIGE (Einar et al. 2002) "System for Visualizing Economic and Regional Influences Governing the Environment" is a spatial microsimulation model based on a database covering the whole Swedish population.

Initial population

All the above models are discrete-time models and their initial population is a representative sample of the whole population. For MOSART and SVERIGE, the main data source are administrative population registers.

None of the initial population of the five models are synthetic population based, i.e. created by imputing all the needed variables gathered from different information sources, but almost no country can rely on one data source to create the initial population, therefore some of them, if not all of them, have some imputed data for variables available in the main datasource. Some survey/register data are at household/family level and therefore the ties between the different individuals of a household/family are known but in some cases, like DESTINIE, these ties had been statistically implemented within the survey sample, when a link between two different household is needed (e.g. parents household with their adult-children).

Even If it is not necessary to have detailed and complete individual histories before the starting time, nevertheless a precise situation of the individual/-household has to be known (or imputed, if necessary) according to the events to be simulated. For instance, to simulate a good distribution by number of children in the household, fertility must be varying by number of children ever born (parity) and couple situation, the information on women's parity and couple's situation must be known and those variables included in the initial file. If fertility is also varying by union duration, then the date of union (or union duration at initial time must also be known or imputed). The parameters chosen for the transitions must use only variables available in the initial population and there is thus a trade-off between imputation of variables in the population that weakens the population representativeness and the need for covariates for the transitions.

The initial population must include a record for each individual: these records contain all the individual and household identification numbers and characteristics needed for the simulation at both levels. This implies finding data sources with all or most of the variables needed: those missing need to

be imputed. If no such data sources are available, the initial population has to be created by imputing the variables. If there is a lot of imputation, then it is sometimes called synthetic population.

Some authors consider that a synthetic initial population (Wachter et al. 1997) has some advantages. Firstly, it allows the creation of a population as large as needed and, secondly, there are no confidentiality conflicts or privacy problems, because everything is imputed and the information does not come from a survey sample (where some individuals could possibly be identified).

As it has been underlined by Zaidi & Rake (2001) in their Lesson 8, "The choice of base data is an early, important decision in the model building process. The representativeness of base data is a principal concern here, with the use of matching techniques to be one way of compensating for the shortage of variables in such datasets."

All the microsimulation models presented here start with a cross-sectional population. The initial population consists in a large sample of population coming from survey, census, or register data. The initial "population" size is varying between 54,000 and 480,000 individuals (see Table 7.1) except for SVERIGE which covers all the Swedish population.

Model (Country/date)	Initial database	Sample size
MOSART (Norway, 1988)	12 per cent random sample of population administrative data from 1993	about 480,000 individuals
SESIM (Sweden, 1997)	LINDA longitudinal database, created from administrative data, typical sample about 1% of population	about 110,000 individuals
SAGE (UK, 1999)	0.1% sample of 1991 Census	about 54,000 individuals
DESTINIE (France,1999)	1999 Family History Survey and 2003 Assets Survey collected by INSEE	about 65,000 individuals
SVERIGE (Sweden,1999)	Administrative data from 1985 to 1995	9 million individuals

source : Cassells et al. (2006)

Table 7.1 Initial database for five dynamic microsimulation models.

Microdata used

Most models use, as much as they can, longitudinal surveys to estimate transition probabilities and otherwise cross-sectional survey. In some cases, when

the information is not available in the country, an estimation based on another country is used. For example, in the French model DESTINIE, the USA survey Longitudinal Studies of Aging (about 12,000 individuals studied in 1984, 1986, 1988 and 1990) was used to estimate transition probabilities to disability (see Table 7.2).

Model	Microdata
MOSART	1988 Family and Occupation Survey Population administrative data
SESIM	Initial data base (LINDA) Longitudinal Individual Data for Sweden HINK/HEK survey is based on an interview, merged with register information, and therefore have a much better possibility of creating a household according to an economic definition. This is in contrast to LINDA, which essentially is based on a household according to a tax definition. Regional mobility and tenure choice is based on GEOSWEDE5. This database is constructed from Louise and RTB: GEOSWEDE have been used is in models for regional mobility and health.
SAGE	1991 Household Samples of Anonymised Records (SARs) ONS Longitudinal Study General Household survey 1992-8 Family and Working Lives Survey (1994) British Household Panel Survey, Waves 1-9
DESTINIE	1999 Family History Survey 2004 French Census 2003 Assets Survey 1984, 86, 88, 90 USA Longitudinal Studies of Ageing
SVERIGE	ASTRID longitudinal individual database Administrative registers

Table 7.2 Microdata used in the Microsimulation models.

Events simulated and their order

Microsimulation models are based on Markov chains and the transition probabilities are therefore based only on the current situation. This potential drawback has been ruled out by using biographical information and time-based information as additional covariates. This is the way these models take into

account individuals' past behaviour and some of the covariates of interest (e.g. in order to model the occurrence of a birth, we may use the information on how many children the woman already have).

In all the models under study, demographic events can be classified in 3 groups:

- population growth: migration, birth, death;
- household formation or dissolution: children leaving home, cohabitation and marriage, divorce and separation;
- socio eco-status: education (enrolment and level of education), health status, working status.

Many authors have underlined the necessity to subdivide the model into smaller components: "the subdivision of the model into smaller modules is important. This aspect of dynamic microsimulation helps to make the working of the model more systematic and makes it easier to check problems at different stages in the running of the model" (Zaidi & Rake 2001, p. 8).

Even if the events studied are similar, differences between models exist firstly in the events sequence chosen and secondly in the information used to simulate each event. For instance, if union disruption takes place before union formation, it is possible to experience both events during the same period, i.e. moving directly from one union to another, but it is not possible to enter a union and break it in the same year.

7.3 Demographic model for the SustainCity project

7.3.1 Data structure, software implementation

The model is written in the Modgen language, which is developed and maintained by Statistics Canada. As required by the SustainCity project, this language is free of charge. It has been used for 15 years and different types of models have been developed using this language (Gribble et al. 2003, Morrison 2008). It includes a user-friendly interface for both developers and users. Thanks to this interface, users can easily change the simulation parameters, perform sensitivity analysis, etc.

The demographic module proposed for the Sutaincity project is to be used within the UrbanSim program, but has not been embeded in the UrbanSim/OPUS python environment. The demographic module can be run and tested on its own: this eases the calibration of the module, even for those who are not expert in UrbanSim. Similarly, the demographic module can be used for research questions not related to those answered by UrbanSim, as well as localisation modules within UrbanSim can use other simulated populations and households systems

The proposed model is a closed cross-sectional discrete-time demographic model. The demographic model starts with a population of individuals and households at an initial point in time t_0, and every member of the population ages from one year to the next, from t_0 to the end of the simulation period.

The output of the model is a yearly file of individuals and households whose characteristics are derived from the family status of individuals: children, couples, "isolated" adults who live alone; in the same household of a family; or in a non-family household. For each individual, sex, age, and all the covariates used in the microsimulation are available, and they can be used to build specific variables at the household level. Complex households hosting "isolated adults" living together or with a family, as well as non-private households, are also produced as outputs from the demographic module.

As in most microsimulation models, Demo4 runs with two different types of data: the initial population and the transition probabilities leading to the events faced by the population over time. The technicalities of the model building, the data structure and how the transitions are calculated for the Paris case study as an example are available in Turci et al. (2012)

7.3.2 Individuals and households

Two types of "actors" are modeled: individuals and households. Each individual belongs to one and only one household. One reference person is defined for each household and updated each year. For the time being, only the French statistical office definition of the reference person[7] is used and the rule is the following one:

- if there is a couple in the household, the reference person is the man;
- otherwise, if there is no couple, the reference person is the oldest active person;
- otherwise, if there is no active individual the reference person is the oldest person.

The demographic module does not include any information on dwellings. The location of households within dwellings is part of another module of UrbanSim, the Household Location Choice module.

Individuals are linked with others through 3 possible links: partner, mother and father.

Each person is also related to one household (unless he/she lives in a non-private households, in which case he/she is not linked to any household).

[7] http://www.insee.fr/en/methodes/default.asp? page=definitions/pers-refmenage-exp-prin-rrp.htm

A "simple household" is defined as a household comprising of either:

1. a single person;
2. a couple;
3. a couple with children;
4. a single parent with children.

A "complex household", instead, is a household where, one or more, not partnered individuals add to one of the four possible type of "simple households" e.g. not related adults living in the same dwelling or an old parent living with the family of his/her son/daughter. To keep the model quite simple, there were only one couple in each household. Therefore, a household with the couple of parents living with child also in couple is not possible.

7.3.3 Annual probabilities

The ageing process depends on annual transition probabilities. Ageing here means, all the events individuals/households can face over time. These probabilities vary with several covariates (sometimes only sex and age). Some probabilities may vary with time, i.e. the transition probabilities are estimated for each year based on existing projections. Users can choose to let the transition probabilities constant or to produce a set of probabilities for each year. Moreover, some events may depend not only on their own probability of occurrence, but also on the features of other members of the population (e.g. an agent-based model logic for the mating process is implemented).

Individual probabilities are estimated as sex- and age- or duration-specific rates, and parameters indicating relative risks for some subgroups. Interactions can be included, but the model has to be as simple as possible: events already experienced define the current state of each individual. Transition probabilities (for each event) are defined for each individual, depending on his/her sex, age, and other covariates. A demographic covariate is included in the microsimulation model if it is relevant for other UrbanSim modules or if it is correlated with some of the transition probabilities within the demographic module. Some events may concern groups of individuals. Union formation, separation, and fertility processes are women driven. Migrations involve all the members of a household (migration is considered, in fact, a household-level event). For each transition, a set of basic covariates is defined, within a minimal model. Additional covariates can be included in order to increase the accuracy of the estimation.

As a matter of simplicity, age as a covariate used to estimate transition probabilities is increased every year, and remains constant during the year (period-cohort parallelograms in the Lexis diagram[8]).

[8] An explanation of the Lexis diagram can be found in Keiding (1990).

Social heterogeneity is accounted for with two covariates, both divided in three groups. *Level of education*: low (primary), middle (secondary) and high (university), and *Place of birth*: born in the region, born in the country, born abroad. For immigrants, the duration since arrival can be used as an additional covariate. Additionally, *Labour force participation* covariate can take two values: "in the labour force" (working or unemployed) and "not in the labour force" (student, inactive, or retired).

7.3.4 Events simulated

The events modeled are death, emigration (exits from the region simulated), immigration (entries into the simulated region), education attainment and enrolment, professional status, leaving from the parental home, union dissolution (dissolution of an unmarried union and divorce), union formation (cohabitation and marriage), birth of a child, and entry into and exit from non-private households. The demographic module simulates the evolution of the population size through entries (births, immigration) and exits (death, emigration). Interactions between agents are also taken into account: some events occurring to an individual may have an effect on some other individuals. For instance, when a person dies, his/her spouse, if any, becomes widow(er) and the children, if any, become orphans; moreover the size and composition of the household change.

The events are run in the order presented on Figure 7.1: a person who dies during the year will not face any other event (e.g. the birth of a child can occur only to women alive and living as a couple at the beginning of the year). The situation of each individual is recorded at the end of each year in a "census-like" file containing all the individual characteristics, as well as the links with related individuals (identity of relatives such as partner, parents and children) and the household number (allowing the production of outputs at the individual, or at the household level).

All the events that involve a couple/household decision are woman driven. Hence, the probabilities of these events are estimated only for women. For instance, the probabilities of entering a union are applied to women not living in union; when a woman enters a union, a process of mate-matching to available (i.e. "not in union") men takes place. A set of "convenient partners" is formed through an optimisation technique under constraints, and the partner is chosen within this set. The status of the new partner then changes to "in a union", and a link between the partners is created.

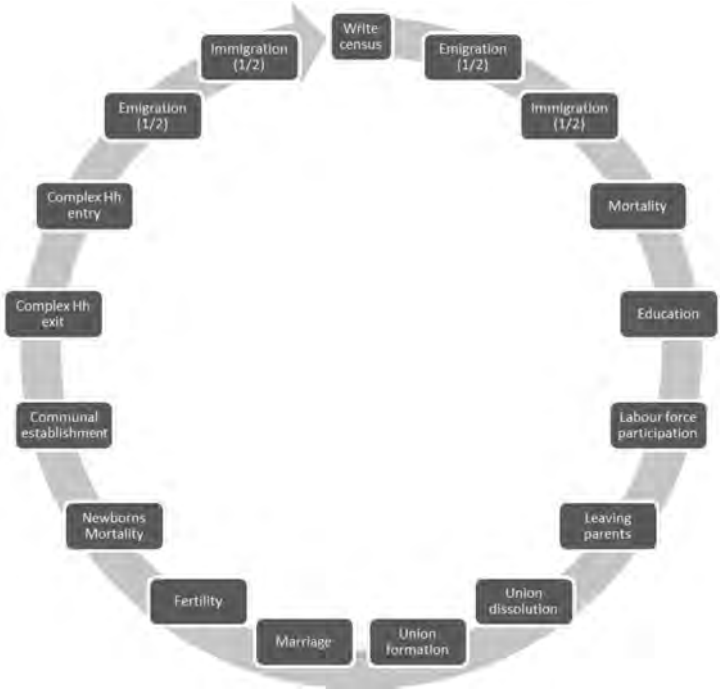


Figure 7.1 Events order.

7.4 Parameters estimation and calibration for the Paris case study

7.4.1 Microdata needed for the demographic module

Initial population

In the Paris case study, Paris region inhabitants was selected from the French survey *Étude de l'histoire familiale*(EHF) based on a sample from 1999 census and variables on education were imputed using census data (Table 7.3). The starting year is set to 1st January 1999.

Event and transition probabilities

For the sake of consistency, we used as much as possible the same data source to estimate most of the transitions probabilities. As for the initial population the main data source is the 1999 Survey on Family History, *Étude de l'histoire familiale*, EHF (Lefèvre & Filon 2005, Cassan et al. 2000). Table 7.4 gives a description of all the transitions included in this model and how they are determined.

	Minimal	Optional
Sex/age	Sex: Male or Female Age: Continuous (0 to 100)	
Fertility	Number of children in the household Parity Age of youngest child	Age of oldest child
Family formation and dissolution	Union status: in union or not in union De jure status: Married or not married Former marital status Year of union	If not in union never live in union or ever been in union If in union/married • union duration • union order If not married never been married/ divorced/widowed
Migrations	Place of birth: «city»/country/abroad Immigration data: residence some years before («city», country, abroad)	
Education and professional activity	Level of education Enrolment Professional activity: Working or non working Age at end of initial studies for those who ended their studies Parent's level of education	Detailed professional activity • working • unemployed • non working • retired
Others	Type of household: private or communal establishment (non-private dwelling)	Type of communal establishment
Identification	Household ID (anonymous) Partner, father and mother ID if they are in the household	Are the parents alive? Do they live in the city?

Table 7.3 Covariates needed in the initial population.

	Basic covariates	Additional covariates
Mortality	Death probabilities by sex and age (country-level)	
Fertility	Female fertility by age Female by birth order (and age)	Fertility among women living in couple First births probability by union duration High order births by duration since last birth (age of the last child).
Emigration	Emigration probabilities to abroad and to another region of the country by type of household, number of children and age of the last child (if applicable) or number of household members and age of the oldest man (or women if applicable)	Place of birth Level of education of the reference person
Immigration	Number of immigrants by sex and age	
Family formation and dissolution	Probability to leave the parental home by sex, age Female union probability by age (separate women never in union and separated women) Some information on homogamy: • New unions by both partners'age • New unions by both partners'education • New unions by both partners'country of birth • New unions (or relative union probability) by both women'parity (number of children ever born) Marriage probabilities by sex and age Marriage by previous cohabitation Female union disruption probability by age	Probability to leave the parental home by sex, age and level of education (and enrolment) Female union disruption probability (Additional variables such as marital status, parity, union duration, level of education are welcome)
Education and professional activity	Age at the end of studies by sex Probabilities of entry and exit into/out of the labor force by sex and age	Unemployment rates by sex and age Retirement rates by sex and age
Others	Estimates of population in communal establishment	Entries into and exits from communal establishments by sex and age

Table 7.4 Covariates used for probabilities estimation.

7.4.2 Transition probabilities

Transition probabilities were estimated from several data sources (see Turci et al. (2012) for more details on the choice of parameters and on the data sources used for the implementation of the initial population and simulated events). For each transition, a balance was made between simplicity and accuracy. On the one hand, it is useful to take into account population heterogeneity in order to estimate the variety of individual transition probabilities: for instance, fertility rates vary not only with age, but also with parity, which allowing accurate simulation not only the mean number of children, but also the distribution of women by ultimate parity (Pennec et al. 2009). On the other hand, it did not appear necessary to take into account the fact that fertility probabilities vary, ceteris paribus, by age of last child: fertility usually reaches a peak around 3-4 years after last birth, but the decline of fertility with age takes sufficiently into account this heterogeneity, and the age of last child did not result sufficiently significant and was thus omitted. Age of last child could have been added as a covariate, in order to improve the simulation of families by age difference between children. Similarly, for a matter of simplicity, we restrained ourselves to woman's age as the main covariate for union formation and dissolution, and we did not include duration of stay or event order (e.g. previous number of unions for union formation, or union duration for union dissolution).

Deaths

The mortality profiles were taken from life tables by sex, age, and year, at the national level (Vallin & Meslé 2001). Probabilities of death vary over time, contrary to all other transitions, which are held constant. When death occurs, the status of the deceased's relatives (children, partner) are changed; an orphan left alone (with no other adult) at age less than 18 is moved to the household of the surviving parent, if any, or otherwise to another household (see subsection "entry into a complex household").

Emigration

Emigration (exit from the region) was estimated as a function of the household type. Among couples with children, the yearly probability of exit was set at 2.6% when the last child was aged less than 5 and at 1.6% if s/he was older. The same probabilities were used for one-parent families. Among couples without children, exit probabilities were set at 10% when the male partner was aged less than 30, 1.6% for those aged 30 years and over, and 2.5% between age 60 and 70. Finally, for households with no couple and no children, estimates were 10% when the oldest working adult was aged less than 30, 1.6% for those aged 30 years and over, and 2.5% between ages 60 and 70.

Immigration

An estimate of the probability to be a "recent migrant" in the region was first estimated among those who were not present at the last census. Immigrants were taken from a pool estimated from the EHF 1999 survey, based on individuals who were not present at the last census. Then, two "pools" of immigrant households were built, from where the immigrants are taken (one composed of people coming from other regions of France, the other of people coming from abroad). Individuals in these "pools" were initialized in the same way as inhabitants. Immigrants come as members of households (i.e. immigration is considered, like emigration, as a household-level event). Immigrant profile of each pool is held constant over time.

Educational attainment

The final level of education was first imputed for all new-borns, as a function of sex, and crossed father's and mother's level of education. Then, the age at the end of studies was imputed from the final level of education.

Labour force participation

The model distinguishes between being in and out of the labour force (unemployed and working individuals are thus considered as in the labour force). Transition probabilities by sex and age were first estimated from the Labour Force Survey. An additional parameter, the presence of child aged 0-2 in the household, was added for women's entries into and exits from the labour force. In our first implementation we did not use the other covariates such as enrolment, level of education, or union status, which appeared to be of second-order when sex, age, and the presence of a young child were considered.

Leaving the parental home

Based on the proportion of young adults living with their parents, probabilities to leave the parental home to live alone were estimated by sex and age. Like the labour force participation, we limited the number of covariates: we did not consider enrolment and working status, as age was a major determinant of leaving home. Union status was not considered because unions are supposed to take place out of the parental home. In order to avoid simulating households with more than one couple, people entering a union are forced to leave the parental home if they are living with their parents (see below).

Union dissolution

Union dissolution probabilities were taken from an estimate based on the family transitions during the five years before the 1999 EHF survey, in 1994-8 (Beaujouan & Toulemon 2009). Union dissolution is woman-driven and it depends

on woman's age and marital status. An additional parameter was added, the presence of a child aged 0-2 in the household. Level of education was not considered. The contrasts by level of education have diminished, if not reversed, in recent years (Buisson & Daguet 2012) and it appeared more careful to simulate no contrast in union formation and dissolution by level of education.

Union formation

Union formation is a two-step process. First, women living by themselves or with their parents are given a probability to enter a union, by age and household status. Second, a random sample of 50 single men are chosen as potential partners; a similarity index is computed, based on difference in age, education, place of birth, and duration since arrival in the region for immigrants. The probabilities of each potential partner being chosen are then computed as a monotone function of each similarity index with the woman's characteristics entering a union. When a union is formed, the couple settles in the woman's household if she is living on her own, with or without children, and, in other cases, in the man's household. If both partners are still living in their respective parental household, a new household is created and they are both moved there.

Marriage

The probability for a woman living in an unmarried couple to get married with her partner is a function of age. Recent couples (created during the current year) have twice the probability to get married (to simulate direct marriages without cohabitaton before the mariage). Woman's level of education and parity, the presence of a newborn and union duration (greater than 0) were not considered, as they did not appear to change significantly the marriage probabilities.

Childbearing

The probability to give birth is estimated by mother's age and parity (number of previous children born) for women in union. In the simulation process, we assume that only women in union can have a child. Fertility was allowed to change between 1999 and 2008, based on estimates civil registration. Sex ratio at birth was fixed at 1.05 boy per girl.

Communal establishments (non-private households)

The model considers all communal establishments (CE) as one type of household. It was not possible to make any distinction between the numerous types of communal establishments (i.e. nursing homes for elders, students' halls, etc.) due to the lack of detailed data at regional level. Individuals living in CEs are not considered to live in a household and are thus not linked to any household.

In fact, when a person is selected to enter a CE, the link with his/her previous household was severed. If the person is in a union, the union is dissolved.

The events that people living in CE may face are death, change of level of education, change of labour market status, and leaving the CE. They cannot form a union, or have a child. When they leave a CE, they create a new one-person household, but they can immediately emigrate, or move to a complex household.

The average probability to enter into a communal establishment was estimated from the census. And then, based on a sex and age profile, people living as a couple have a probability decreased by a ratio of 0.3, while others were doubled. People who enter in a communal establishment are mainly young adults, between 18 and 25 years of age, and elderlies (aged 80+). Enrolment and level of education were not considered as covariates. The probability of exit was set to 0.5, due to the scarcity of data, the same individual could "exit" and "re-enter" in the same time loop, thus staying in a communal establishment (this is common at old ages where the entry probabilities reach 0.2). Those who did not re-enter in a CE or un-enter a household (see below, "complex households"), a new household is created for him/her. It would have been more accurate to assume lower exit probabilities from nursing homes at old ages (and to correspondingly estimate lower entry probabilities), but the high mortality at these ages leads to short sojourn durations, and few moves "in" and "out".

Complex households

In order to simulate the occurrence of complex households, a probability to enter is set by sex and age for men and women living alone, based on the prevalence by sex and age derived from the census. The probability is higher for lone parents having children living in the simulated area. For those chosen to join a complex household, a set of 50 potential households is chosen; a similarity index is computed, based on the age difference between the selected agent and the reference person of each potential household (preference with difference less than 5, or between 25 and 35) and on the size of each household (bigger households having lower probabilities to be chosen).

7.5 Results

One result is the description of the trends in the distribution of the households by size and family structure in a consistent and explicit way, depending on assumptions on individual demographic behaviours.

In the Paris region, between 1999 and 2049, the number of households increases much faster (see Fig. 7.2) than the population size (7.3 ‰per year vs. 3.3 ‰), due to a decline in household size (Fig. 7.3). This is consistent

with the decline in union formation, the increase in union dissolution, and the population ageing, but the trend could be slower due to housing shortage. The convergence to a mean household size of 2 is similar, although somewhat more rapid, than in other projections (e.g. Louchart 2010).

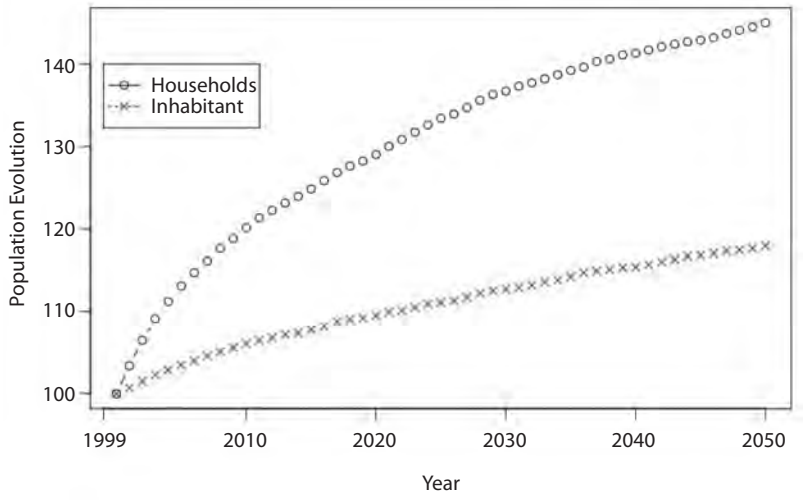


Figure 7.2 Evolution of population based on the population of 1999 (100%).

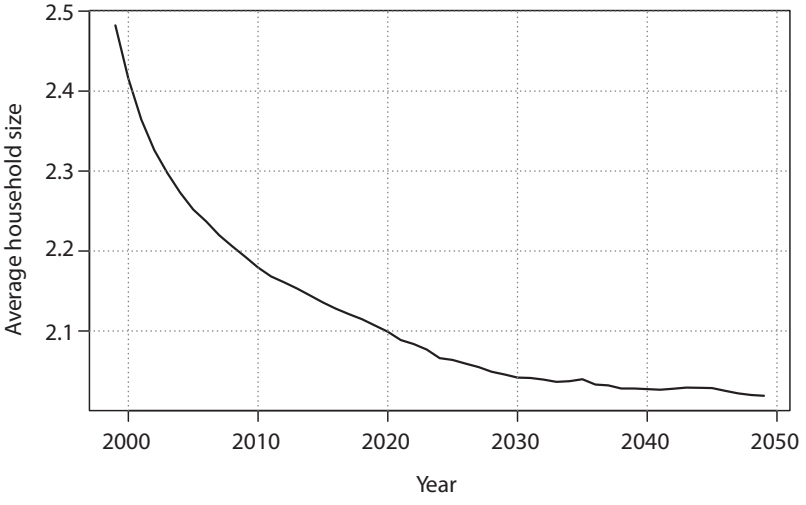


Figure 7.3 Evolution of the average size of the household.

Furthermore, the increase in the number of households is entirely due to a rise in the households with no couple and no child (Fig. 7.4). The first category on the increase is made of households where a person lives alone: a young adult (man or woman), a man of middle age (living alone after a couple disruption, the children living with their mother), more often than a woman at this age, old adults (mainly women) living by themselves after the death of their partner (because men die earlier).

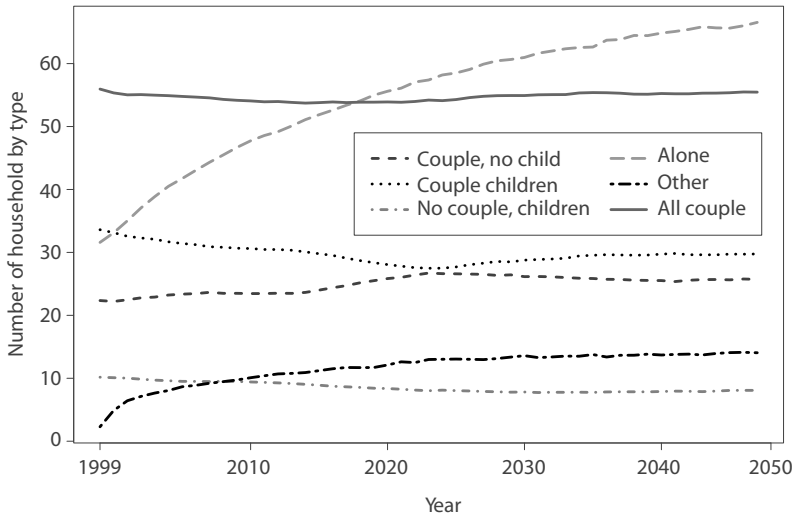


Figure 7.4 Evolution of household type, Base: 100=total number of households in 1999.

The second category on the rise is complex households, made of unrelated housemates or individuals living with (but not within) a family (mostly young adults). Complex households are badly identified in the initial population, as some households include more than one family (a couple with or without children, a person with children) and were not considered as complex. The increase during the year 1999 is thus mainly due to an artefact, but the number converges rapidly, due to the high exit rates. The increase of complex households during the period is a direct consequence of the increase in one-person households, as people leaving alone are those who join existing households without entering a union and thus make them "become" complex.

On the contrary, the number of households with a couple is stable, with a decline of couples with children on the short term, due to the current delay in the birth of the first child. After the year 2020, the equilibrium is reached and the number of couples becomes stable, the ageing of the population becoming the main driver of the increase in the number of households, mainly one-person households.

In the Demographic model, communal establishments are treated as a single household. Figure 7.5 shows the trend in the proportion of inhabitants living in a communal establishment. It remains stable between 1999 and 2009, because of the "hole" in the number of elderlies due to the shortage on births during World War I, and increases slowly afterwards, from 2.0% in 2010 to 2.5% in 2050, due to ongoing population ageing. Men more often than women live in a communal establishment, at all ages before 80. The 2010 census confirms these trends: the proportion of men and women living in a communal establishment in the Paris Region reaches 5.2% for men vs. 3.1% for women at ages 18-24, 2.4% vs. 0.7% at ages 25-64.

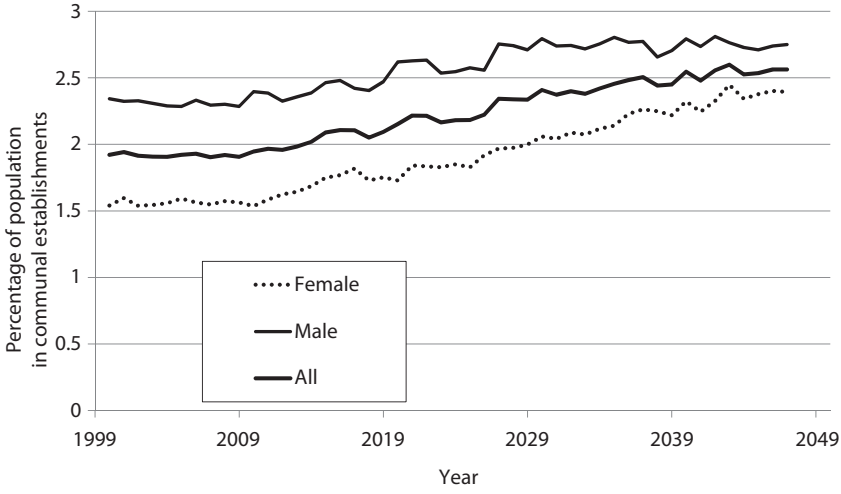


Figure 7.5 Proportion of men and women living in a communal establishment, per cent.

7.6 Conclusion

A specific demographic model has been developed starting from an analysis of the state-of-the-art demographic microsimulations (Turci et al. 2010).

Many individual demographic events were simulated, based on annual transition probabilities (Turci et al. 2012): exits from the population by death or emigration; entries by immigration or birth; exits from the parental household, union formation, marriage, and dissolution; entries into a household (home mates) and exits from complex households; entries in and exits from communal establishments. All these transitions are dependent on sex, age, and specific additional covariates for which parameters estimates are produced; educational attainment and labour force participation are also simulated, allowing the use

of enrolment and working status as optional covariates for demographic behaviour; micro simulation methods allowed to use specific mate-matching rules.

The model produces as output a yearly file of individuals and households with their characteristics derived from the family status of individuals: children, couples, "isolated" adults living alone or living in the same household as a family or living in a non-family household. For each individual, sex, age, and all the covariates used in the microsimulation are available, and can be used to build specific variables at the household level. Complex households hosting "isolated adults" living together or with a family, as well as communal establishments, are also produced as outputs from the demographic module. Yearly estimates of population and households can then also be used by the other modules.

Chapter 8

Modeling real estate investment decisions in households

André de Palma, Matthieu de Lapparent, and Nathalie Picard

8.1 Introduction

There is substantial literature on modeling individual residential location choices. We notice that most of the existing approaches are static. As we think it is more sensible, our proposed framework of analysis is dynamic. We here restrict our attention to some key aspects of such modeling. They remain overlooked in the literature.

First, the set of decisions related to residential location in a multiperiod setting typically involves a very large number of alternatives. At a given date, a choice entails several dimensions: where to locate, which dwelling type, and which tenure status. A fine tuned demand analysis of housing choices consider dozens of small geographical units for potential locations, at least two tenure types (own or rent) and at least two dwelling types (apartment or house).

Another aspect is about who makes residential location related decisions. Recent literature (Chiappori, de Palma, Inoa & Picard 2013, de Palma et al. 2013) shows that accounting for intra-household negotiation provides a new understanding about how residential location choices may result from consensus reached by household members. Most research is based on unitary household approaches that consider there is a single decision maker in the family.

Regarding the dynamics of these choices, one may reasonably assume that a household optimizes its life cycle path of decisions that pertain to labor supply, consumption of goods/services, consumption of floor space, dwelling location choices, dwelling types, tenure statuses (Di Salvo & Ermisch 1997), whether housing is used as a financial asset and whether it is left for bequest to heirs (Arrondel & Lefebvre 2001). All these decisions are closely related and subject to per-period financial constraints.

Also, housing markets are segmented but interacting with each other and with other markets such as the labor market, transport activity and land development. Prices are not always sufficient to clear the markets. Most models do not question equilibrium, or consider only partial equilibria. There are relatively few spatial general equilibrium or micro-simulation models with both heterogeneous agents and markets. de Palma, Picard & Waddell (2007) also showed that accounting for capacity constraints in housing supply considerably changes location choices.

Data requirements are very stringent if one wants to carry out a complete analysis at a very fine level of analysis. In general, one uses several statistical sources and data merging procedures to prepare samples. The types of model and their focus are often adjusted to available data.

The importance of transportation costs has been pointed out by Weisbrod et al. (1980), Anas & Chu (1984), Waddell, Ulfarsson, Franklin & Lobb (2007), Lee & Waddell (2010), to cite a few. They focused on the induced effects of the transportation market on residential location choices. Accessibility to and from a residential location greatly affects households choices.

Quigley (1985), Nechyba & Strauss (1998), Brueckner et al. (1999) show how extrinsic attributes of dwellings, i.e. local amenities and neighborhood play a significant role in spatial distribution of housing demand and on the corresponding market prices. Bureau & Glachant (2010) also find that market prices of owned dwellings are sensitive to their surrounding environment.

de Palma & Lefevre (1985), Ben-Akiva & de Palma (1986) recognized that transaction costs and moving costs affect the dynamics of location choices since they increase the duration of stay at one location. Any fiscal distortion that increases market price (either temporarily or permanently) needs to be compensated by either a longer stay period to return on investment or larger streams of income. Duration of stay at one location is proportional to the levels of transaction prices and moving costs effective in the housing market. As a consequence, households that rent a dwelling are more mobile than households that purchase a dwelling, because of smaller transaction costs.

Analysis of choices of dwelling and tenure types have also been subject to several analysis, e.g. Ioannides & Rosenthal (1974), Henderson & Ioannides (1986), Mills (1990), Cho (1997), Skaburskis (1999). They discussed the effects on individual demand functions of the attributes of dwellings (such as number of rooms, presence of garden/balcony, age of the building, etc.). They mainly discuss the effects of intrinsic characteristics and how they may differ across socioeconomic and demographic groups. McFadden (1978), Weisbrod et al. (1980), Thisse (2010), Dantan et al. (2010), Dantan & Picard (2013) also discussed in a more general way existing trade-offs that may have consequence

on location choices, including differences across individuals with different socioeconomic characteristics.

Linneman & Wachter (1989), Zorn (1989), Brueckner (1997), Gobillon & Le Blanc (2004), discussed the dynamics of housing expenditures of homeowners in presence of credit rationing. They show how the latter may not only affect the demand for floor space but also the choice of a tenure type (for a given demand of floor space) and less directly the choice of a dwelling type (since, on average, consumed floor space is larger for houses than for apartments).

In this chapter, we propose a theoretical microeconomic model to analyze residential choices of households in a dynamic context with perfect information. To keep exposition simple, the decision maker is a household. We do not account for intra-household negotiation between members. We consider that the household lives two periods. We simultaneously address economic choices of residential location, dwelling type and tenure status, and their dynamics. We account for interaction with transportation market, with demand for local amenities, and with financial investment constraints.

At the beginning of each period, household is endowed with a per-period utility function that depends on the level of local amenities, the level of floor space, and the level of consumption of a composite good. It is faced with continuous and discrete decisions: choices of optimal quantity of floor space, consumption level of an outside composite good, and the choices of residential location, tenure status and dwelling type. These choices are subject to budget and technical constraints. In our model, borrowing is allowed in the first period, but not in the second one. No debt can be left at the end of the second period. We also assume that the interest rate is higher when borrowing for a dwelling than when saving/borrowing on the money market (this is a form of credit imperfection). We do not consider any selection mechanism: every household can borrow money for housing purpose. Transaction costs apply to real estate (acts made by notaries or lawyers but also local taxes). We consider existence of moving costs whenever changing home location in the second period. Of course, transportation costs affect the choices of households. Another feature of our model is the introduction of a bequest motive. The household leaves a bequest to heirs at the end of its life cycle for altruistic reasons. It increases its utility level. The bequest consists of money and/or of real estate.

The household program is to maximize the sum of discounted flows of utility over its life cycle subject to an inter-temporal budget constraint. The problem can be solved in two preliminary steps, then by backward induction. First, we compute for each combination of locations, tenure and dwelling types, optimal demands for floor space, and consumption of a composite good. Second, we derive the associated indirect utility functions. The household chooses the combination of locations (hence levels of amenities), tenure, and dwelling

types that maximizes its utility. The levels of optimal (continuous) demands adjust to these optimal (discrete) choices.

We detail our theoretical model in Section 8.2. We discuss the mathematical formulation and related solutions of the household optimization program. We choose functional forms which lead to analytically tractable theoretical model involving inter-temporal utility maximization. We study the effects of the most relevant determinants of residential location choices: demand for local amenities, financial constraints (pay-down requirement, borrowings, and savings), housing and goods/services prices, income, transaction costs, transportation, and moving costs.

In Section 8.3, we then turn to an empirical application. We present data in the first subsection. Our main data source is the 2006 French Housing Survey, which comprises a short retrospective survey back to 2002. We focus on the population living in Paris region during these years. Due to lack of statistical information about precise locations of dwellings, we restrict our analysis to the dynamics of tenure and dwelling type choices given location choices. Also, since the survey is based on revealed preferences, we need to build counterfactuals and to impute values of the attributes of unchosen alternatives. To this extent, we use data on credit from the survey and on prices (values and rents) from the Côtes Callon. The latter provides average observed market prices by tenure type for 2002 and 2006. We also address the problem of missing information about precise locations of dwellings: we only have information about the *département* (large French administrative unit) where households live. Therefore we cannot compute any reliable transportation costs. The moving costs are accounted for through a dummy variable indicating whether the household moved between 2002 and 2006. In a second subsection, we propose an econometric specification of the theoretical model accounting for data availability and identification constraints. Our approach is based on random utility maximization. It takes the form of a Nested Logit choice model, see for example McFadden (1978), Train (2009). We discuss estimation results in the third subsection.

We conclude in a last section and argue that our theoretical model can be used as a building block for detailed analysis of residential location choices.

8.2 A microeconomic model of residential choices

We consider a household living 2 periods $j \in \{1, 2\}$. A bequest B is left to heirs at period 3. At the beginning of each period, the household has to choose its dwelling location, its dwelling type, its tenure type, and the levels of consumption of floor space, composite good, and local amenities. The latter are actually not really chosen but instead determined by residential location choices.

8.2.1 Household preferences

At the beginning of every period $j \in \{1, 2\}$, a household obtains utility u from consumption of floor space Q, local amenities x, and a composite good C (excluding dwelling). We define by $u_j(x_j, C_j, Q_j)$ the household utility in period j. We assume that $u_j(\cdot)$ is strictly increasing and quasi-concave in its arguments.

The utility of a household over its life cycle is measured at the beginning of the first period. It is defined as the discounted sum period-specific utilities plus the level of bequest B left to heirs at period 3 (bequest is transmitted once the household disappears). We assume an altruistic behavior of households. Being altruistic produces (personal) welfare in leaving financial or real estate assets to heirs. We also assume that there is perfect foresight. Stated shortly, perfect foresight corresponds to full information about future (market prices and socioeconomic and demographic personal characteristics): there is neither risk nor uncertainty about any future outcomes. Households being presumably rational, they optimally use all this information to produce their sequences of choices once for all at the beginning of their life cycle. Mixing both features explain in our approach why they leave either savings or real estate but not both, as only the most utility-making choices are made, and how they adapt to be sure to avoid leaving debts to heirs. We agree that it is a limiting case that may diverge from reality and that there is further research to develop on this point. We do however think that it is a baseline approach to be tested in a first step.

We define $\beta \in [0, 1]$ as the discount factor. The resulting inter-temporal utility function $U(\cdot)$ is

$$\begin{aligned} &U(x_1, C_1, Q_1, x_2, C_2, Q_2, B) = \\ &\quad u_1(x_1, C_1, Q_1) + \beta u_2(x_2, C_2, Q_2) + \beta^2 \gamma_2 \ln(B), \end{aligned} \tag{8.1}$$

and we further assume that $u_j(\cdot)$ has a Cobb-Douglas formulation

$$u_j(x_j, C_j, Q_j, B) = \psi_j(x_j) + \alpha_j \ln(C_j) + (1 - \alpha_j) \ln(Q_j), j \in \{1, 2\}. \tag{8.2}$$

If the household does not change dwelling in period 2, floor space consumption is determined once for all in period 1, i.e. $Q_2 = Q_1$. We assume that the level of consumed amenities may change over time whenever location is the same. The inter-temporal utility function simplifies to

$$\begin{aligned} &U(x_1, C_1, Q_1, x_2, C_2, Q_2, B) = \\ &\quad u_1(x_1, C_1, Q_1) + \beta u_2(x_2, C_2, Q_1) + \beta^2 \gamma_2 \ln(B). \end{aligned} \tag{8.3}$$

By assumption, we assume a bequest motive: household's utility depends on what is left to heirs. We are aware that it is not rare to inherit from debt. We here exclude this possibility by forcing a strictly positive bequest.

8.2.2 Budget constraints

Choices of household are made under finite budget constraints. Before writing explicitly the inter temporal budget constraint of the household, we need first to define further notations. For each period $j \in \{1,2\}$, we assume that tenure type (k_j) corresponds to either owning (o_j) or renting (r_j), dwelling type (d_j) corresponds to either a house (h_j) or a flat (f_j), and location corresponds to a predetermined geographical zone l_j in a region that is divided into L mutually exclusive units.

Let $\pi_t^{k_j,d_j,l_j}\left(x_{j,l_j}\right)$ be the price (per unit of surface) in period $t \in \{1,2,3\}, t > j$, of a dwelling located at l_j that is occupied by the household in period j. Whatever d_j, k_j, or l_j, the household is faced with additional transaction costs when considering dwelling consumption. These are of two types. The first applies only once to the transaction itself when purchasing or renting a dwelling, often under the form of a proportional tax rather than a lump sum. Let $\mu_j^{k_j,d_j,l_j} > 0$ be this unit tax level. The second applies to the occupation of the dwelling and is recurrent over time as long as the dwelling is occupied. It mainly models local taxes. Let $\kappa_j^{k_j,d_j,l_j}$ be the associated unit tax level. Prices of the composite good are defined as $p_j, j \in \{1,2\}$.[9]

The household is endowed a strictly positive level of (exogenous) income R_j at the beginning of each period $j \in \{1,2\}$. As household members have to participate in out-of-home activities, e.g. work or education, $D_j\left(x_{j,l_j}, y_j\right) > 0$ is a transportation cost in period j for a residential location l and a set of locations to be reached $y_j, j \in \{1,2\}$.[10]

Let also S_j model monetary savings at period $j \in \{1,2\}$. For $j = 2$, it corresponds to the level of monetary bequest left to heirs, and we assume that $S_2 \geqslant 0$. When $j = 1$, it can be either positive or negative. If positive, the return on money savings is $i_j > 0, j \in \{1,2\}$. We assume that the household can borrow money only during the first period (out of two) of its life cycle. It is excluded to leave debts to heirs. If money is borrowed in period $j = 1$, the household contracts a loan that has to be reimbursed during the second period $j = 2$. The interest rate of the loan is τ_1. We assume that $\tau_1 > i_1$. If money is borrowed for housing purpose, as the household must pay down a fraction $\rho \in [0,1]$ of the total dwelling price, only the remaining portion has to be reimbursed in period 2. By convention, $\rho = 1$ when dwelling is rented in period 1 (no borrowing is allowed for renting), and $\mu_j^{r_j,d_j,l_j} = 0$.

[9] Note that it means that prices of the composite good do not differ across zones in the considered region.

[10] Note that y_j is a vector when many household members work and/or when considering other locations than workplaces. Note also that x is an argument of the transportation cost function as presence of public transport or any special transport infrastructure may be considered as amenities for location l_j and may have effect on the level of transportation costs.

There is yet one component to consider. When the household makes the choice to change dwelling in period 2, moving costs affecting the budget constraint may occur. We model them by $\Delta\left(x_{1,l_1}, x_{2,l_2}\right) \geqslant 0$. By convention, these costs are equal to zero when not changing dwelling.

The budget constraint for the first period is then

$$\begin{aligned} & p_1 C_1 + \left[\left(1+\mu_1^{k_1,d_1,l_1}\right)\rho + \kappa_1^{k_1,d_1,l_1}\left(x_1\right)\right]\pi_1^{k_1,d_1,l_1}\left(x_{1,l_1}\right) Q_1 + S_1 \\ & \quad = R_1 - D_1\left(x_{1,l_1}, y_1\right). \end{aligned} \tag{8.4}$$

The budget constraint in period 2 is more involved:

$$\begin{aligned} & p_2 C_2 + \left[(1+\tau_1)\left(1+\mu_1^{k_1,d_1,l_1}\right)(1-\rho)\right]\pi_1^{k_1,d_1,l_1}\left(x_{1,l_1}\right) Q_1 \\ & + \left[1+\mu_2^{k_2,d_2,l_2} + \kappa_2^{k_2,d_2,l_2}\left(x_{2,l_2}\right)\right]\pi_2^{k_2,d_2,l_2}\left(x_{2,l_2}\right) Q_2 + S_2 \\ & = (1+i_1) S_1 + R_2 - D_2\left(x_{2,l_2}, y_2\right) - \Delta\left(x_{1,l_1}, x_{2,l_2}\right) \\ & + \pi_2^{o_1,d_1,l_1}\left(x_{1,l_1}\right) Q_1 \mathbb{I}_{1\curvearrowright 2}, \end{aligned} \tag{8.5}$$

where $\mathbb{I}_{1\curvearrowright 2}$ is a dummy variable indicating whether dwelling is bought in period 1 and sold in period 2.

We can combine equations 8.4 and 8.5 to obtain the inter-temporal budget constraint:

$$\begin{aligned} & (1+i_1)\, p_1 C_1 + p_2 C_2 + S_2 \\ & + \left\{ \begin{array}{l} \left(1+\mu_1^{k_1,d_1,l_1}\right)\left[\rho + \frac{\left(1+\tau_1^{k_1,d_1,l_1}\right)(1-\rho)}{1+i_1}\right] \\ + \kappa_1^{k_1,d_1,l_1}\left(x_{1,l_1}\right) \end{array} \right\} (1+i_1)\,\pi_1^{k_1,d_1,l_1}\left(x_{1,l_1}\right) Q_1 \\ & + \left(\left(1+\mu_2^{k_2,d_2,l_2}\right) + \kappa_2^{k_2,d_2,l_2}\left(x_{2,l_2}\right)\right)\pi_2^{k_2,d_2,l_2}\left(x_{2,l_2}\right) Q_2 \\ & = (1+i_1)\left(R_1 - D_1\left(x_{1,l_1}, y_1\right)\right) + R_2 - D_2\left(x_{2,l_2}, y_2\right) \\ & - \Delta\left(x_{1,l_1}, x_{2,l_2}\right) + \pi_2^{o_1,d_1,l_1}\left(x_{1,l_1}\right) Q_1 \mathbb{I}_{1\curvearrowright 2}. \end{aligned} \tag{8.6}$$

This general constraint further simplifies when considering specific series of discrete residential choices. For instance:

- if $k_j = r_j$ then $\rho = 1$, $\tau_1 = 0$, $\mu_j^{r_j,d_j,l_j} = 0$: if household is renting, there is obligation to pay the full amount of rent in each period, there is no loan, and there is no purchasing transaction cost;
- if $l_1 = l_2$ then $d_1 = d_2$, $Q_2 = Q_1$, and $\Delta\left(x_{1,l_1}, x_{2,l_1}\right) = 0$ by convention: if household does not change dwelling then location and dwelling type are the same, demand for floor space is the same, and there is no moving cost.

8.2.3 Bequest function

The dwelling(s) transmitted at the beginning of period 3, if any, were necessarily bought previously. For $j = 1, 2$, we denote by $\mathbb{I}_{j\curvearrowright 3}$ the dummy variable

indicating whether a dwelling is bought in period j and transmitted to heirs at period 3. The total value of the bequest evaluated in period 3 is therefore:

$$B = \pi_3^{o_1,d_1,l_1}(x_{1,,l_1})\,Q_1\mathbb{I}_{1\frown 3} + \pi_3^{o_2,d_2,l_2}(x_{2,l_2})\,Q_2\mathbb{I}_{2\frown 3} + (1+i_2)\,S_2. \quad (8.7)$$

8.2.4 Characterization of the solution

Consider now that the household is able to compute optimal demands and derived indirect utility function for each possible trajectory of discrete decisions. It then compares the levels of utility of every possible series of decisions and it selects the one that maximizes its utility. As modelers, we want to define more precisely these optimal demands and indirect utility functions so as to develop a structural framework.

We proceed in two steps to solve the problem of the household. We consider in a first step that the household chooses one of the possible series of discrete residential choices (location, tenure status and dwelling type). Given prices and budget resources, its problem is then to determine its optimal continuous demands for floor spaces Q_1 and Q_2 and other consumption expenditures C_1, C_2 plus a potential level of savings S_2 due to bequest behavior so as to maximize its inter-temporal utility function subject to budget constraints. We however have to account for two key decisions in solving this program. The first regards whether to move from one dwelling to another in between period 1 and period 2. The second regards whether to leave housing or savings as bequest to heirs. Combination of both does not give the same maximization program. These give different solutions.

In a second step, once obtained optimal demands and savings for every possible series of discrete choice, we derive their related indirect utility functions. To summarize, a household first determines its demands in every discrete choice situations. It then derives the related indirect utility function. It finally chooses the series of discrete residential choices that maximizes its level of (indirect) utility.

On bequest composition

As already stated, the solution of the problem depends on the composition of the bequest, which in turn depends on whether markets are perfect or not. We make some proposition accounting only for tenure status but extension to other dimensions of the problem just introduces more complexity without improving understanding of our baseline assumptions.

Definition 1. *Full perfection holds if*

$$\begin{cases} i_1 = i_2 = \tau_1 \equiv i \\ \mu_1^{o_1} = \mu_2^{o_2} = 0 \\ \kappa_1^{o_1}(x_1) = \kappa_2^{o_2}(x_2) \equiv \kappa \\ \pi_j^{b_j}(x_j) = \pi_j^{r_j}(x_j) \equiv \pi_j(x_j), \ j = 1,2 \\ (1+i_t)\,\pi_t^{b_j}(x_j) = \pi_{t+1}^{b_j}(x_j), \ j = 1,2, \ t = 1,2,3, t \geqslant j. \end{cases} \tag{8.8}$$

We remind the reader that perfect foresight corresponds to full information about future (market prices and socioeconomic and demographic personal characteristics): there is neither risk nor uncertainty about any future outcomes.

In order to simplify notations, we define implicit prices :

$$\begin{cases} \Pi_{Q_1}^0 = \{(1+\mu_1)[(1+i_1)\rho + (1+\tau_1)(1-\rho)] + (1+i_1)\kappa_1(x_1)\}\pi_1(x_1) \\ \Pi_{Q_2}^0 = [(1+\mu_2) + \kappa_2(x_2)]\pi_2(x_2) \\ \Pi_{C_1}^0 = (1+i_1)p_1 \\ \Pi_{C_2}^0 = p_2 \\ \Pi_{S_2}^0 = 1. \end{cases} \tag{8.9}$$

Lemma 1. *Full perfection implies that the relative implicit prices in the bequest equation are equal to the relative prices in the inter-temporal budget constraint:*

$$\frac{\Pi_{Q_1}^0}{\pi_3^{b_1}(x_1)} = \frac{\Pi_{Q_2}^0}{\pi_3^{b_2}(x_2)} = \frac{1}{(1+i_2)}. \tag{8.10}$$

Proof. The result is straightforward by applying the simplifications implied by full perfection in Equations (8.6) and (8.7). □

Lemma 2. *Under full perfection, and with no budget constraints ($S_1 \geqslant 0$), at the optimal solution, the household is indifferent between renting and buying, and the bequests are transmitted indifferently as real estate or money. The problem reduces to the maximization of a standard Cobb-Douglas function*

$$C_1^{\alpha_1} Q_1^{1-\alpha_1} C_2^{\beta\alpha_2} Q_2^{\beta(1-\alpha_2)} S_2^{\beta^2\gamma_2}, \tag{8.11}$$

under the inter-temporal budget constraint:

$$\begin{aligned} &(1+i)\,p_1 C_1 + p_2 C_2 + S_2 + (1+i)(1+\kappa)\pi_1(x_1)\,Q_1 + (1+\kappa)\pi_2(x_2)\,Q_2 \\ &= (1+i)[R_1 - D_1(x_1, y_1)] + [R_2 - D_2(x_2, y_2) - \Delta(x_1, x_2)]. \end{aligned} \tag{8.12}$$

Proof. Indifference between renting and buying results from the absence of budget constraints associated with the conditions $\pi_j^{b_j}(x_j) = \pi_j^{r_j}(x_j)$, $j = 1,2$

and $\mu_1^b = \mu_2^b = 0$. Lemma 1 implies that real estate has exactly the same value whether is transmitted as real estate or sold at period 2 (or at period 1 when relevant), and the corresponding value is transmitted to heirs as a monetary bequest S_2. We can therefore assume that dwellings are rented at both periods, and only a monetary bequest is transmitted. □

Proposition 1. *There exists exactly one form of bequest, entailing either real estate (bought either in first or in second period), or money if and only if the following condition holds*

$$\frac{\Pi^0_{Q_1}}{\pi_3^{b_1}(x_1)} \neq \frac{\Pi^0_{Q_2}}{\pi_3^{b_2}(x_2)} \neq \frac{1}{(1+i_2)}. \tag{8.13}$$

Proof. According to Lemma 1, when the three ratios are equal, interior solutions with composite bequest (entailing both real estate and money) are optimal. The reciprocal part proceeds by contradiction. Consider an optimal solution denoted by $\{C_1^*, Q_1^*, C_2^*, Q_2^*, S_2^*\}$. Assume, without loss of generality, that

$$\frac{\pi_3^{b_1}(x_1)}{\Pi^0_{Q_1}} > 1 + i_2.$$

Assume further $\mathbb{I}_{1\curvearrowright 3} = 1$ and $S_2 > 0$. Consider an infinitesimal change $\Delta > 0$ such that Q_1^* is replaced with

$$Q_1^* - \frac{\Delta}{\Pi^0_{Q_1}} < Q_1^*$$

and C_1^* is replaced with

$$C_1^* + \frac{\Delta}{(1+i_1)\,p_1}.$$

The total cost of this infinitesimal change is zero, so inter-temporal utility remains unchanged. Consider an additional change such that S_2^* is replaced with

$$S_2^* + \frac{\Delta \pi_3^{b_1}(x_1)}{(1+i_2)\,\Pi^0_{Q_1}}$$

and first period consumption of composite good becomes

$$\begin{aligned} &C_1^* + \tfrac{\Delta}{(1+i_1)p_1} - \tfrac{1}{(1+i_1)p_1}\tfrac{\Delta \pi_3^{b_1}(x_1)}{(1+i_2)\Pi^0_{Q_1}} \\ &\quad = C_1^* + \underbrace{\left(1 - \frac{\pi_3^{b_1}(x_1)}{(1+i_2)\,\Pi^0_{Q_1}}\right)}_{<0 \text{ since } \frac{\pi_3^{b_1}(x_1)}{\Pi^0_{Q_1}} > 1+i_2} \tfrac{\Delta}{(1+i_1)p_1} < C_1^*. \end{aligned} \tag{8.14}$$

The total cost of this additional infinitesimal change is zero, so inter-temporal utility remains unchanged. The second change moves the total bequest back

to its initial value (here we use $\mathbb{I}_{1\frown 3} = 1$). The combination of these two infinitesimal changes leaves inter-temporal utility, total bequest value, and second-period utility unchanged. As a result, it leaves first-period utility unchanged. This is in contradiction with the decrease in both C_1 and Q_1, since $u_1(\cdot)$ is increasing in its arguments. Similarly, considering

$$Q_2^* \to Q_2^* - \frac{\Delta}{\Pi_{Q_1}^0} < Q_2^*$$

and

$$C_2^* \to C_2^* + \frac{\Delta}{p_2}$$

prove that

$$\frac{\pi_3^{b_2}(x_2)}{\Pi_{Q_2}^0} > 1 + i_2,$$

$\mathbb{I}_{2\frown 3} = 1$ and $S_2 > 0$ cannot hold simultaneously. Finally, a similar infinitesimal change in Q_1^*, C_1^*, Q_2^* and C_2^* leaving both total bequest and total cost unchanged prove that

$$\frac{\pi_3^{b_2}(x_2)}{\Pi_{Q_2}^0} \neq \frac{\pi_3^{b_1}(x_1)}{\Pi_{Q_1}^0},$$

$\mathbb{I}_{1\frown 3} = 1$, and $\mathbb{I}_{2\frown 3} = 1$, cannot hold simultaneously. □

Proposition 1 implies that, under perfect foresight, the case in which the household owns and transmits to heirs both the dwelling where it lives in period 1 and the dwelling where it lives in period 2 cannot be optimal when the relative (implicit) price of these two assets is not the same in period 2 and in period 3, which constitutes a systematic imperfection in real estate markets in the Paris region. It happens that, in our data set, this case represents less than 1% of the sample. This low percentage is consistent with unanticipated changes in personal or professional situation over the life cycle. Such unanticipated changes are ignored in our model.

Optimal series of discrete choices

Another result of our model is that the indirect utility functions for every series of tenure and dwelling types write using only two different mathematical formulations:

$$\begin{aligned}&\bar{V}_{l_1,l_2,d_1,d_2,k_1,k_2} = \\ &\Omega_{l_1,l_2,d_1,d_2,k_1,k_2} - \lambda_{Q_1}^0 \ln\left(\Pi_1^{l_1,d_1,k_1}\right) - \lambda_{Q_2}^0 \ln\left(\Pi_2^{l_2,d_2,k_2}\right) + \lambda \ln\left(W(x_1,x_2)\right)\end{aligned} \tag{8.15}$$

for movers, and

$$\begin{aligned}&\bar{V}_{l_1,l_1,d_1,d_1,k_1,k_2} = \\ &\Omega_{l_1,l_1,d_1,d_1,k_1,k_2} - \lambda_{Q_1}^1 \ln\left(\Pi_1^{l_1,d_1,k_1} + \Pi_2^{l_1,d_1,k_2}\right) + \lambda \ln\left(W(x_1,x_1)\right)\end{aligned} \tag{8.16}$$

for non movers, where:

$$\begin{cases} \lambda^0_{Q_1} = (1-\alpha_1) + \beta^2\gamma_2 \mathbb{I}_{1\cap 3} \\ \lambda^0_{Q_2} = \beta\,(1-\alpha_2) + \beta^2\gamma_2 \mathbb{I}_{2\cap 3} \\ \lambda^1_{Q_1} = (1-\alpha_1) + \beta\,(1-\alpha_2) + \beta^2\gamma_2 \mathbb{I}_{1\cap 3} \\ \lambda = 1 + \beta + \beta^2\gamma_2 \end{cases} \tag{8.17}$$

and Ω's are functions of the possible choices, the exogenous variables, and the parameters of U. The optimal series of discrete choices is then solution of

$$\max\left\{ \max_{l_1,l_2,d_1,d_2,k_1,k_2} \left(\bar{V}_{l_1,l_2,d_1,d_2,k_1,k_2}\right), \max_{l_1,l_1,d_1,d_1,k_1,k_2} \left(\bar{V}_{l_1,l_1,d_1,d_1,k_1,k_2}\right) \right\}. \tag{8.18}$$

8.3 Application

We present in this section a probabilistic discrete choice to analyze dynamics of tenure and dwelling types *given locations*. We focus on the population of inhabitants that lived and may have moved within the Paris region in 2002 and 2006.

8.3.1 Data

Empirical implementation of our theoretical model is data demanding in that we need more than longitudinal disaggregate data . Our main data source is the 2006 French National Housing Survey (FNHS). It reports observed housing choices of French households during this year and it briefly reports those made in 2002. For these two years, we also observe socioeconomic and demographic characteristics of these decision makers. We however point out that available information is somewhat limited for year 2002.

As the 2006 FNHS is a revealed preference survey, we don't have any information about unchosen alternatives. We therefore need to complement this survey by drawing statistical information from other data sources, especially regarding housing prices. To this extent, we use 2003 and 2007 Côtes Callon. Côtes Callon is a database that reports average housing prices by tenure status and dwelling types at the city level.

It also appears that some of the attributes of the alternatives, especially those pertaining to local amenities, are not described.

We thus see two demanding requirements: information about prices by location, dwelling, and tenure types, and information about local amenities by location.

Sample formation

We have initially 6988 observations. When we look at combinations of tenure and dwelling types at these dates, we observe that only 5 of them appear significant. They are written in italics in Table 8.1.

	Dwelling types (2002, 2006)				
Tenure types (2002, 2006)	(h,h)	(h,f)	(f,h)	(f,f)	Total
Move, (o,o), keep	23	9	10	16	58
Move, (o,o), resell	*53*	*14*	*44*	*46*	*157*
Move, (o,r), keep	3	17	1	14	35
Move, (o,r), resell	7	10	4	17	38
No move, (o,o)	*1440*	*0*	*0*	*967*	*2407*
No move, (o,r)	33	0	0	19	52
Move, (r,o)	*25*	*6*	*133*	*170*	*334*
Move, (r,r)	*18*	*41*	*60*	*809*	*928*
No move, (r,o)	25	0	0	35	60
No move, (r,r)	*171*	*0*	*0*	*2748*	*2919*
Total	1798	97	252	4841	6988

Move, No move: change home or not; Keep, resell: keep formerly owned home or resell it
(,) is defined as (type in 2002 , type in 2006); o: own; r: rent; h: house; f: flat

Table 8.1 Matrix of tenure and dwelling types in 2002 and 2006.

For the rest of our application, we do not consider choice situations where a household moves between 2002 and 2006 and keeps its owned dwelling that served as a residence in 2002 whatever its new tenure type. We also do not consider the choice situation where a household changes tenure type from owner to renter. Finally, we do not account for the situation where a household may become owner of the dwelling it was renting in the former period: there is an obligation to change dwelling.

We virtually have 5 sequences of tenure types by 4 sequences of dwelling types but, by natural constraint, we have to remove 4 of them. We thus have 16 choice situations. We observe that our 6745 remaining observations greatly favor non-moving choices. Only 1419 observations (21.04%) concern moving households. We also observe that shift in tenure type from rental to ownership is closely related to households that move from an apartment to a house. The reciprocal exists, but it is not that important. Move and ownership at both periods is often related to a status quo in dwelling type. Data also show that house renting is not a very usual choice as compared to house owning or flat renting.

Attributes of unchosen alternatives

As already stated, the 2006 FNHS is a revealed preference survey. We only have a description of what has been done but not what may have been done. As we want to parametrize and estimate a probabilistic choice model, an important task is to reconstruct pertaining variables that describe attributes of unchosen alternatives. For other reasons, we also have missing data that describe characteristics of some households.

To this extent, we use auxiliary tables of the 2006 FNHS and draw additional information from other data sources. We develop a set of auxiliary regression models. We discuss and present them below. We use them to impute the missing values of our relevant variables that we in turn use in our discrete choice model. Missing values are defined as the expected values as produced by the models. It is a series of single imputation methods. We are aware that single imputation methods have shortcomings in that it does not reflect the full uncertainty created by missing data. Given the goodness-of-fit, the present auxiliary models capture most of the effects describing the dependent variables. We however keep in mind that testing different imputation methods and compare the results in our empirical approach would be of some interest.

First of all, we remind the reader that we don't have accurate information about locations of dwellings. We only have locations at the *département* level, a very large administrative geographical unit (there are only 8 *département* for the region we consider). Even though we stated that we present a model of dwelling and tenure types given locations, we would have preferred to have these variables, at least to be able to compute less roughly housing prices for unchosen alternatives. This is not the case: housing prices of unchosen alternatives are computed at the *département* level. It also not feasible to compute transportation costs or accessibility indicators.

In order to reconstruct housing prices, we need some external data. We use 2003 and 2007 Côtes Callon (data collected during years 2002 and 2006). They provide average housing prices for apartments and houses in several locations of the region and for different levels of comfort and quality of the dwellings. We compute weighted average housing prices for apartments and houses at the *département* level. Weights are drawn from the 2006 census of population and consider only age of the building. We also assume that these weights did not change between 2002 and 2006. We use them to compute housing prices for year 2002.

Secondly, we also need to make purchasing prices comparable to renting prices. To this extent, we convert purchasing prices to pseudo renting prices. We base our analysis on the auxiliary "mortgage" table of the 2006 FNHS that regards housing loans that were contracted by French inhabitants of the Paris region (disregarding our sample selection) for the dwelling they purchased

earlier and currently live in. Using an auxiliary regression model, a system of simultaneous equations, we impute missing information about mortgage level, pay-down fraction and interest rate for household that actually chose to rent.

Thirdly, we make additional assumptions about the attributes of unchosen alternatives for non moving households. Data show that 71.20% of moving households stay in the same *département*. The 2006 Census of population also shows the same. For our model, we therefore assume a non moving household would stay in the same *département* if it had chosen to change dwelling in between 2002 and 2006.

To summarize, non moving households would have located in the same *département* if they had chosen to change dwelling. Values (purchasing prices) and rents per square meter are also computed at the *département* level. They are made comparable as follows:

1. We draw for each dwelling type the average purchasing price per square meter at the level of *département*.

2. We impute how much, how long, and at which interest rate a mortgage would be to buy the dwelling. To impute the values of these variables, we adjust a hierarchical simultaneous equation system by full information likelihood method. We use the auxiliary "mortgage" table of the 2006 FNHS on housing loans that were taken out by households of the Paris region for the dwellings they purchased earlier and currently live in. Our specification is rather simple: we assume that the mortgage amount is a function of income per capita, age of the household head, and household size. Mortgage duration is function of the ratio between mortgage amount and total household income. Mortgage interest rate is a function of mortgage duration and a year-specific variable. Quality of fit is rather good for a such simple system. We obtain that mortgage amount is an increasing function of income per capita and follows a U-shape with respect to age of household head. It is also an increasing function of household size. Duration of a mortgage is an increasing function of the ratio between amount and income. Naturally, interest rate decreases in between 2002 and 2006 and is a decreasing function of the duration. Estimates are reported in Table 8.2.

3. We then impute how much would be the expected monthly reimbursement amount for renters in our sample if they had chosen to buy. To this extent, we use simple financial mathematics. We assume that household i buys a dwelling of type k in location l by means of a mortgage. An amount equal to $(1-\rho_i)\,\pi$ per square meter is borrowed over a period of T_i years at an annual interest rate of τ_i. ρ_i is the fraction of down payment and $\pi_{k,l}$ is the purchasing price per square meter. The monthly expenditure in housing per meter square is defined as the fixed monthly

Label of variable	Estimate	T-stat.
Log. of amount equation		
Intercept	6.66	6.68
Log. of income per capita	0.62	6.84
Age of household head	−0.10	−6.67
Age of household head squared	0.001	5.81
Log. of household size	0.27	4.03
Log. of duration equation		
Intercept	2.59	259.80
Log. of ratio between amount and total income	0.16	24.28
Log.of interest rate equation		
Intercept	3.57	7.77
DV[a]: year 2002	0.21	6.27
Log. of duration	−0.91	−5.23
Number of observations	1628	
Overall ρ^2	0.78	

[a]: DV: dummy variable

Table 8.2 Mortgage-related system of equations.

reimbursement $C_{i,k,l}$:

$$C_{i,k,l} = \frac{\tau\pi_{k,l}}{1-(1+\tau)^{-T}}. \tag{8.19}$$

We finally merge these imputed values with these related to the revealed choices of households. We now have available a set of housing prices for every observed and likely housing choice of all households.

Income and wealth

We also miss 2002 income levels. We actually have 2006 income levels. Some observations are yet missing. We proceed, as it is often done in many applications, to impute missing values of this variable for year 2006. We fit a type II Tobit model (Amemiya 1984). The first part of the model describes why some households did not report their income levels. The second part of the model describes the income levels of the subset of households that reported them, correcting for the likely sample selection bias.

We use a simple specification of the model. Goodness-of-fit of the model is not as good as we expected but we use these results to impute both missing values of income R for years 2006 and 2002. We implicitly assumed that the relation between the dependent variables and the exogenous variables stayed constant over the 2002 to 2006 period. Estimates are reported in Table 8.3.

Label of variable	Estimate	T-stat.
Selection equation		
Intercept	0.84	33.02
DV[a]: household head is an active worker	0.33	11.94
DV: household head is unemployed	−0.14	−3.99
DV: household head is retired	0.36	12.11
Log. of 2006 income equation		
Intercept	7.14	148.47
Log. of age of household head	0.48	39.21
Log. of household size	0.48	81.34
DV: household head is an active worker	0.91	68.16
DV: household head is unemployed	0.04	2.54
DV: household head is retired	0.54	34.53
DV: household lives in Paris	0.02	1.50
DV: household lives in Seine-et-Marne	0.14	6.93
DV: household lives in Yvelines	0.23	10.63
DV: household lives in Essonnes	0.16	8.71
DV: household lives in Hauts-de-Seine	0.11	6.15
DV: household lives in Seine-Saint-Denis	−0.13	−7.89
DV: household lives in Val de Marne	0.06	3.22
DV: household lives in Val d'Oise	0.07	3.82
Standard deviation of error term	0.58	272.49
Correlation of error terms	−0.00	−0.00
Overall ρ^2	0.58	

[a]: DV: Dummy Variable

Table 8.3 Imputation of income variable.

We finally computed wealth as $R_{2002}+\frac{R_{2006}}{1+i}$ where i is defined as the interest rate of a risk-free monetary asset. In our application, it is defined as a 4 years investment in the French *Livret A*:

$$(1+i)^4 = \prod_{t=2002}^{2005} (1+i_t)\,, \tag{8.20}$$

where $i_t, t = 2002, \cdots, 2005$, are drawn from data of the *Banque de France*.

Transportation and moving costs

There is no information about travel behaviors of households in the survey. Added to limited knowledge of housing locations, there is no possibility to compute variables that regard transportation costs or moving costs. It is a shortcoming in our empirical application in that accessibility to workplaces, schools, and other amenities, might play important roles on residential choices of households.

8.3.2 Model specification

We now turn to the econometric specification of our probabilistic discrete choice model.

Choice set

The first step is to elaborate a little bit more on the structure of the choice set. We focus on dynamics of dwelling and tenure types at two dates. There exists several ways to nest the different dimensions of our choice set. We assume the following. Dwelling type for year 2002 is first chosen. Then, tenure type for year 2002 is chosen. In between 2002 and 2006, the choice to move or not is made. In year 2006 the choices of a dwelling type and then a tenure type are made. We may have privileged the choice of a tenure type prior to the choice of a dwelling type or other combinations of the series of choices. The search of the most relevant combination is left aside for future research. Figure 8.1 depicts the nesting structure of choices we consider.

Indirect utility functions

As we only observe choices, we are not able to exactly quantify the utility levels of the households. Also, missing statistical information, proxy variables, and measurement errors, force us to set up in a probabilistic framework . We therefore assume that the indirect utility functions of a household n are defined as continuous random variables:

$$V_{n,x_{n,1},x_{n,2},d_{n,1},d_{n,2},t_{n,1},t_{n,2}} = \bar{V}_{n,x_{n,1},x_{n,2},d_{n,1},d_{n,2},t_{n,1},t_{n,2}} + \varepsilon_{n,x_{n,1},x_{n,2},d_{n,1},d_{n,2},t_{n,1},t_{n,2}} \quad (8.21)$$

for movers, and

$$V_{n,x_{n,1},x_{n,2},d_{n,1},d_{n,1},t_{n,1},t_{n,1}} = \bar{V}_{n,x_{n,1},x_{n,2},d_{n,1},d_{n,1},t_{n,1},t_{n,1}} + \varepsilon_{n,x_{n,1},x_{n,2},d_{n,1},d_{n,1},t_{n,1},t_{n,1}} \quad (8.22)$$

for non movers. The components $\bar{V}$ are those as presented in equations 8.15 and 8.16. They model the strict parts of the utility functions. The ε model error terms.

We also assume that parameters of the utility function, α and γ (see equations 8.15, 8.16, and 8.17), are defined as functions of the characteristics of the household (here household size) to control for observable heterogeneity in preferences.

Probabilistic formulation

The error terms ε enter additively and are independent from observed variables. For each observed household, they are distributed with a GEV distribution,

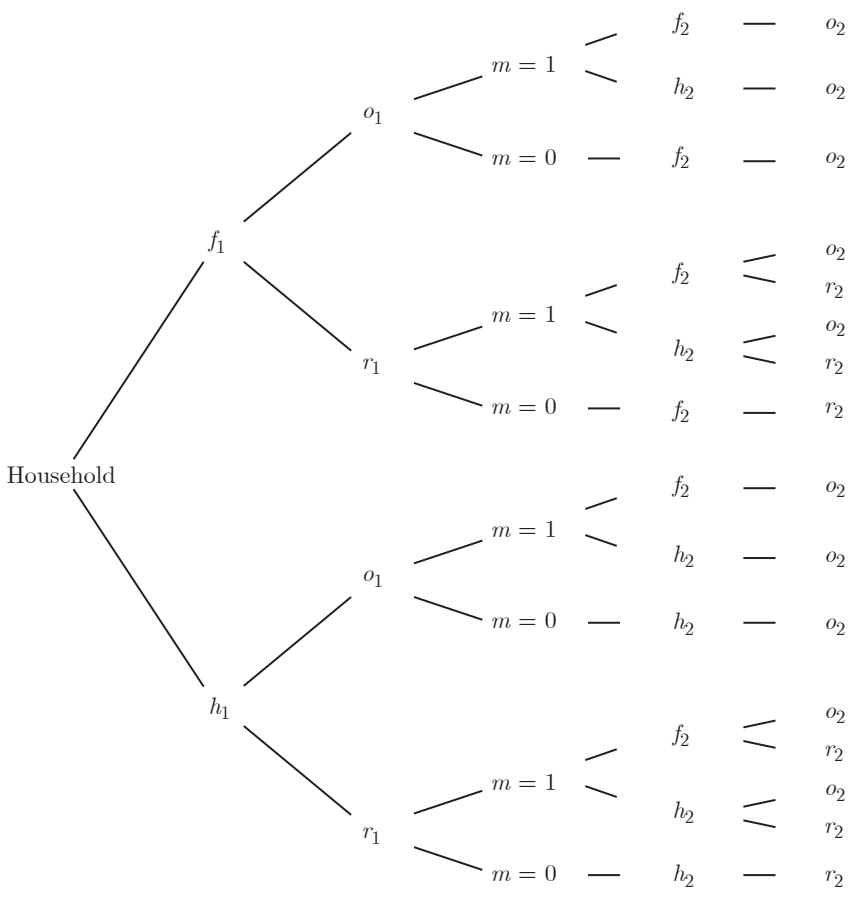


Figure 8.1 Sequence of choices

$\boldsymbol{\varepsilon}_n \to \mathsf{F}(\boldsymbol{\varepsilon}_n) = \exp(-\mathsf{H}(\boldsymbol{\varepsilon}_n; \boldsymbol{\sigma}))$. The choice probabilities are defined as

$$\Pr(\mathsf{x}_{n,1}, \mathsf{x}_{n,2}, \mathsf{d}_{n,1}, \mathsf{d}_{n,2}, \mathsf{t}_{n,1}, \mathsf{t}_{n,2}) = \exp\left(\bar{V}_{n,\mathsf{x}_{n,1},\mathsf{x}_{n,2},\mathsf{d}_{n,1},\mathsf{d}_{n,2},\mathsf{t}_{n,1},\mathsf{t}_{n,2}}\right) \times \frac{\partial \mathsf{H}\left(\mathbf{exp}\left(\bar{\mathbf{V}}_{\mathbf{n}}\right); \boldsymbol{\sigma}\right)}{\partial \exp\left(\bar{V}_{n,\mathsf{x}_{n,1},\mathsf{x}_{n,2},\mathsf{d}_{n,1},\mathsf{d}_{n,2},\mathsf{t}_{n,1},\mathsf{t}_{n,2}}\right)} \mathsf{H}\left(\mathbf{exp}\left(\bar{\mathbf{V}}_{\mathbf{n}}\right); \boldsymbol{\sigma}\right)^{-1}. \quad (8.23)$$

H is chosen to satisfy the nested structure of choices as presented in Figure 8.1. It takes the form of a 5-level Nested Logit model. There are effects of lower choices in the tree on related upper choices ("logsum", i.e. effects of maximum expected utility of a subsequent choice on current choice). We refer the reader to McFadden (1978), Ben-Akiva & Lerman (1985), Anderson et al. (1992), Train (2009) for explicit formulation of these choice probabilities. At each level of the tree, we assume that the nest coefficients are equal.

8.3.3 Results

The model is estimated by maximum likelihood method. Estimates are reported in Tables 8.4 and 8.5. We observe several significant effects, all along with theory and common practice.

Label	Est.	T-stat.
Intercepts:		
dwelling type is house in 2002, same dwelling in 2002 and 2006[b]	−0.94	−3.48
dwelling type is house in 2002, move[c]	−1.26	−3.96
dwelling type is house in 2006, move[d]	−1.30	−3.18
move	−1.71	−4.89
HH head works in 2002 and is unemployed in 2006, effects on[a]:		
own in 2002 and 2006 and move	−2.76	−4.49
own in 2002 and 2006 and stay	−2.00	−9.93
rent in 2002 and own in 2006 and move	−0.94	−2.84
rent in 2002 and 2006 and move	0.16	1.77
HH head works in 2002 and is retired in 2006, effects on[a]:		
own in 2002 and 2006 and move	0.00	0.01
own in 2002 and 2006 and stay	0.72	4.73
rent in 2002 and own in 2006 and move	0.35	1.53
rent in 2002 and 2006 and move	−0.26	−1.29
HH head is unemployed in 2002 and works in 2006, effects on[a]:		
own in 2002 and 2006 and move	−2.67	−3.01
own in 2002 and 2006 and stay	−1.75	−5.16
rent in 2002 and own in 2006 and move	−1.48	−2.20
rent in 2002 and 2006 and move	0.07	0.57
HH head is unemployed in 2002 and in 2006, effects on[a]:		
own in 2002 and 2006 and stay	−0.87	−4.95
rent in 2002 and own in 2006 and move	−1.43	−2.13
rent in 2002 and 2006 and move	−0.12	−1.02
HH head is retired in 2002 and in 2006, effects on[a]:		
own in 2002 and 2006 and move	−1.12	−2.78
own in 2002 and 2006 and stay	0.58	5.50
rent in 2002 and own in 2006 and move	−1.19	−3.36
rent in 2002 and 2006 and move	−0.63	−3.03
HH head is student in 2002 and works in 2006, effects on[a]:		
own in 2002 and 2006 and stay	−2.23	−3.22
rent in 2002 and own in 2006 and move	0.37	1.23
rent in 2002 and 2006 and move	0.61	2.67

[a] reference HH head works in 2002 and 2006 and HH rents the same dwelling over both period; [b] reference is flat, same dwelling in 2002 and 2006; [c] reference is flat in 2002 and move; [d] reference is flat in 2006 and move.

Table 8.4 Estimates of the discrete choice model.

Label	Est.	T-stat.
Income:		
dwelling type is a house in 2002[e]	0.09	5.54
dwelling type is a house in 2006[f]	0.17	3.08
move[g]	0.08	2.44
HH member other than HH head (re)starts working, effect on prob. to move[h]	0.01	0.14
Change in job and or workplace of any HH member, effect on prob. to move[h]	0.37	3.21
HH member other than HH head stops working, effect on prob. to move[h]	-0.16	-2.59
Decrease in HH size in between 2002 and 2006, effect on prob. to move[i]	0.04	0.80
Increase in HH size in between 2002 and 2006, effect on prob. to move[i]	0.73	3.30
$1-\alpha_1$ scaled:		
baseline	0.61	2.69
per additional HH member	0.20	3.18
$\beta\,(1-\alpha_2)$ scaled:		
baseline	0.43	3.56
per additional HH member	-0.06	-2.23
$\beta^2\gamma_2$ scaled:		
baseline	0.29	5.14
per additional HH member	-0.02	-2.48
Nest coefficients (logsums):		
Tenure type choice in 2006 for renters in 2002 who move	0.55	8.77
Dwelling type choice in 2006 for movers	0.51	3.98
Moving	0.69	7.14
Tenure type choice in 2002	0.86	11.21
Number of observations	6745	
Log-lik at convergence	-21073.70	
Pseudo-ρ^2	0.42	

[e] reference is flat in 2002; [f] reference is flat in 2006; [g] reference is stay; [h] reference is no modification in job or workplace for every HH member; [i] reference is no modification of HH structure.

Table 8.5 Estimates of the discrete choice model, cont'd.

It is found that household is sensitive to floor space consumption ($1-\alpha$'s coefficients) and that this sensitivity increases with household size. We also remark that the ratio of periodic sensitivities is larger than 1, as expected, meaning that there exists a strictly positive (and lower than 1) discounting factor. There also exists a significant bequest effect as scaled γ is positive. Accumulating assets, whatever they take the form of money or housing, increases household's level of utility. We are however not sure whether it directly regards heirs or future period of the household as we did not control for life cycle

effects. Anyway, data shows that there exists a positive effect in accumulating assets for future periods of life, whether regarding the household or its heirs.

Results also show that there are different effects of socio-economic and demographic characteristics on dynamics of dwelling and tenure types. Income has a positive effect on the probability to choose a house instead of an apartment. It also has a positive effect on the probability to move, i.e. to change dwelling, over time. This is not surprising as the relative contribution of moving costs is then smaller. The relative larger purchasing power also favors consumption of more space and a more private type of dwelling (no direct up and down neighbors).

We also observe that the structure of a household and its evolution affects the sequence of dwelling and tenure types. It is found that work status of the household head (HH) change the way housing is consumed. It comes from anticipated life cycle effects. If HH retires in the second period, the probability to own a dwelling in the first period and then to live in it in the second period is larger. If HH is unemployed in the second period, the probability to rent in both periods and change dwelling is larger. It is also found that, in this case, the probability to own a dwelling in either the first or the second period is lower. If HH is unemployed in both periods, the probability to rent the same dwelling is larger. When HH is student in 2002 and then works in 2006, the results also show that the probability to rent in the first period and then to buy in the second period is larger.

When there is a change in workplace or job type of any household member, the probability to move is increasing. Such a result is in line with the fact that the household looks at minimizing transportation costs. We also find that a variation of the household size has an asymmetrical effect on the probability to move: an increase of the size favors change of dwelling but a decrease of size does not significantly affect the probability to move.

Finally, everything else being equal, there is a "natural" trend in preferring to consume flat rather than houses.

Looking at the nest coefficients (i.e. "logsum effects") their signs and their magnitudes are all in line with theoretical constraint that ensure that the probabilistic model actually derives from random utility maximization, i.e. these are ordered in ascending order from bottom to top in the tree of choices as depicted in Figure 8.1. Note however that $0.55 > 0.51$ (for 2002 renters) but they are not significantly different from each other. All these nests coefficients are lower and significantly differ from 1, showing "within-nest" substitution patterns. There is no independence from irrelevant alternatives within each nest. It means that removing or adding one choice alternative or one series of choice alternatives alter the ranking of all the others. The chosen one may not anymore be the best option. Given estimates, there are however moderate correlation in unobserved factors within each nest. There is no perfect within-

and between-substitution of the different dimensions of the residential choices that we consider.

8.4 Conclusion

We developed a structural microeconomic framework to analyze simultaneously several aspects of the dynamics of residential choices: location, dwelling, and tenure. We accounted for inter-temporal budget constraints while allowing for bequest motives. Our analytical formulation is tractable for empirical matters. We therefore proposed an econometric formulation of the approach by using a Nested Logit probabilistic choice models. We discussed demanding and stringent data requirements to implement it.

Our results show that there is sensitivity to floor space consumption and that there exists significant discounting and bequest effects. It differs along with household types and the period we consider. Given the results, we speculate that there exist a preference for present consumption of floor space thereby recognizing a strong attachment of households to their housing conditions. The estimated nested structure also show that there exists substitution pattern within the different dimensions of the choice set: tenure, dwelling, and moving choices. There is neither perfect within- and between-substitution nor -perfect within- and between-complementary of the different dimensions of the residential choices that we consider – it lies in between.

Our work is a first step towards integration of a dynamic location choice model in the UrbanSim architecture that is currently developed for the Paris region. It may be further extended in several ways. First, the model may be extended to a multi-period maximization program. Second, the assumption about perfect information and perfect foresight of market variables is questionable. The approach may be formulated as a dynamic discrete choice model with forward-looking economic agents. We also conjecture that using estimated parameters from the econometric model would provide a better basis to perform simulation of an urban system equilibrium.

Chapter 9

Intra-household decision models of residential and job location

Nathalie Picard, André de Palma, and Ignacio A. Inoa

9.1 Introduction

The residential and job location choice literature is dominated by models considering a single decision-maker in each household. In such early models, the analysis of joint decisions (such as residential location choices, involving monetary cost, but also emotional and time dimensions) are usually simplified by assuming away the within-household bargaining process. In fact, it is assumed that there is a single decision maker, who can be seen as a more or less benevolent dictator, to use the terminology introduced more than 30 years ago, by Gary Becker, the father of Economics of the Family.

In the economic literature, models considering a single decision-maker are usually referred to as "unitary model". Many researchers, especially in family economics, tend to criticize this assumption, questioning the hypothesis that the behavior of multi-person households can be described by a single decision-maker model satisfying the basic preference axioms (such as transitivity, for example). Incidentally, we know, from the Arrow impossibility theorem, the difficulty to aggregate preference within small groups.

In the context of residential location, it is highly probable that some intra-household bargaining process plays an important role in dual-earner households. Intra-household bargaining process refers to the negotiation process within the household with respect to joint decisions, such as residential location, use of the car (when there is a single car), or choice of a holiday location. In particular, in the context of residential location, each worker in the household has his own preferences and constraints depending on his own workplace, generally different from the workplaces of the other workers in the same household.

When there is more than one decision-maker in the household, a complex bargaining process usually takes place, and the resulting decisions would appear irrational if they were analyzed using a single decision-maker model. Since the 1980s, economic literature has developed models analyzing the within-family bargaining process in other contexts such as labor supply or consumption. For example, Lundberg & Pollak (1993) developed a specific bargaining process in the presence of multiple interrelated decisions made by multiple decision-makers, whereas Chiappori (1988) developed the so-called "collective model". Collective models take account of the fact that multi-person households face joint decisions involving more or less altruistic members with specific preferences and constraints. In collective models, decisions are assumed Pareto-optimal in the sense that the solution of the bargaining process is on the Pareto frontier, that is, that it would not be possible to increase the well-being of one household member without decreasing the well-being of at least another member. Spouses jointly make decisions leading to a Pareto-optimal outcome, if it is not possible to make one spouse better-off without making the other spouse worse-off. Pareto-optimality can be bested empirically, and is usually verified. By contrast, the assumptions and/or conclusions of unitary models are usually rejected in multi-person households.

Recently, economists and transport researchers have worked on explaining the effect of intra-household variables on the choice of residential location and workplace (Waddell 1996, Abraham & Hunt 1997, Sermons & Koppelman 2001). Their analysis addressed the multiple-member household decisions, but not the multiple decision-maker problem, involving a within-household bargaining process. Overall, intra-household considerations remain limited in this literature; the theoretical and empirical contributions on the effect of intra-household variables on the residential location and workplace choices are scarce. In particular, there are no published contributions on within-household bargaining process in location decisions (with the exception of the ongoing research by Chiappori, de Palma, Picard & Inoa (2013)).

Our objective in this chapter is to make a first step to bridge the gap between two strands of literature which have developed independently up to now. The first one covers location choice models that consider more than one household member or worker, but a single decision maker. The second one focuses on the decision process in other contexts, when several decision makers have diverging preferences and constraints. The ultimate objective is to make a step forward in the analysis of joint mode choice and residential and job location choices in a dual-earner household.

Beyond the simple differentiation of location choice by socio-demographic characteristics, recent models dealing with family location decisions allow for the identification of differences between females and males, and between multiple-worker households and single-worker households. In the context of the Paris

Area, we analyze differences between spouses' commuting times, and we show that spouses' disparities in commuting is a key element in the intra-household decision process. The single-worker household approach might leave aside important intra-household considerations that influence commuting time and accessibility to jobs. As a consequence, the one-worker oriented location choice models may lead to misleading conclusions for dual-earner households.

We review here different models potentially useful to study intra-household decisions. We assess the framework introduced by Chiappori, de Palma, Picard & Inoa (2013), which applies the collective approache of household behavior (Chiappori 1988, 1992) to describe residential location choice of dual-earner households. In the context of a collective model (assuming Pareto-optimality), bargaining powers play an important role to measure the values of time of the two spouses.

We then argue that the framework introduced by Inoa, Picard & de Palma (forthcoming), could also be adapted to analyze residential and job location choices in a two-worker household. By accounting for two accessibility variables (one for each spouse), their three-level nested Logit model could be used to study the interdependency of residential location and workplace, while accounting for variation of preferences for job types across individuals. This could be performed both in single- and in two-worker households.

The outline of this chapter is briefly described below. Section 9.2 provides some background of the studies analyzing household behavior in multi-person households. We focus the literature specifically on residential location and mode choice studied in this chapter. Differences between female and male commuting time in the Paris Region are presented in Section 9.3, which also analyzes the role of spouses respective travel times in a model of residential location that considers intra-household decisions, and explores original considerations in individual-specific accessibility measures. Section 9.4 discuss insights for future work and concludes.

9.2 Literature review

The models we review here present the two literature strands which developed independently up to now. The transportation and location choice literature has mainly modeled the household as a single decision-making unit. Timmermans (2006) offers an extensive review on past research in the transportation literature, whereas Vermeulen (2002*a*,*b*) reviews the work on unitary and collective household models in other contexts.

Different literature reviews and special issues have already been interested in the modeling of intra-household decisions in such contexts. Two special issues on modeling intra-household interactions, edited respectively by Bhat & Pendyala (2005) and by Timmermans & Zhang (2009), have become the

main references in the transportation literature. The former focuses on utility-maximizing models, whereas the latter presents other methods such as micro-simulation approaches and group decision theory.

de Palma, Picard & Inoa (2013) update the previous reviews of intra-household decision models in the transportation literature, and focus on discrete choice models. It differs from such earlier reviews by putting more emphasis on economics of the family and collective models of household behavior.

These reviews reveal that two general research streams have modeled, independently, multi-person households and intra-household decisions. On the one hand, there has been interesting research on within-family interactions on what can be broadly categorized as the transportation, activity-demand, and location literature, which paid no attention to the within-family bargaining process itself. On the other hand, household decision-making processes have been studied in depth in the Collective Model literature, but they paid no attention to residential location, workplace, or other transport-related choices.

These two strands of literature differ not only with respect to their topics, but also with respect to the nature (discrete versus continuous) of the decisions analyzed. As a consequence, these two strands of literature used and developed totally different techniques. This distinction is at the core of new theoretical and empirical developments.

9.2.1 Family decisions in the transportation literature

The transportation and activity-demand literature heavily relies on discrete choice modeling and more occasionally on structural equation modeling and seemingly unrelated regressions (a system of equations, where the error terms are assumed to be correlated across equations), to explain the time invested in activities by the household members, as reported by Srinivasan & Bhat (2005). Furthermore, Timmermans (2006) classifies transportation and activity-demand works that consider intra-household decision-making into three categories: micro-simulation (simulation of a household member daily activity-travel pattern), rule-based (if-then decision tree structures), and utility-maximizing models (RUM and time allocation approaches).

The discrete choice models of intra-household decisions in the transportation, activity-demand, and location literature covers choices from the short run to the long run decision context.

Contributions regarding the long term decision context includes choices of residential and job location, mobility, and car ownership as in Abraham & Hunt (1997), Freedman & Kern (1997), Sermons & Koppelman (2001), Waddell (1996).

Contributions regarding the short term decision context includes modeling (joint) activity and task allocation, travel, car sharing, and mode choice as

in Wen & Koppelman (1999, 2000), Gliebe & Koppelman (2002). Scott & Kanaroglou (2002), Vovsha et al. (2003, 2004*a*,*b*), Bradley & Vovsha (2005) work on discrete choice of maintenance activities allocation in tour-based travel demand modeling systems, as Srinivasan & Bhat (2005, 2006) and others.

More details can be found in the special issues and reviews mentioned before, i.e. Bhat & Pendyala (2005), Timmermans (2006), Timmermans & Zhang (2009), de Palma, Picard & Inoa (2013).

9.2.2 Intra-household decision process in the economic literature

Following the impetus originated from seminal work by Nobel Prize laureate Gary Becker (1965, 1973, 1974*a*, 1974*b*, 1991), recent developments in economics of the family, broadens the classical research field, which initially focused on fertility and labor supply, to new questions such as the marriage decision or choices related to the number of children, their education, and the allocation of tasks and time among household members.

When modeling household decisions, the household was traditionally considered as a single decision-maker, and the study of the decision-making process and of the transactions between family members were neglected, as discussed in de Palma, Picard & Inoa (2013). These so called "unitary models" neglect the differences and potential conflicts between the interests of the different household members, and implicitly assume that household members pursue consensual objectives. This assumption leads to a poor understanding of decision mechanisms within the household, even when individual-specific variables (such as spouses' respective ages or education levels) are introduced in the model.

Collective models and other within-households bargaining models aim at answering the theoretical and empirical criticisms addressed to unitary models of family decision-making. They were developed in two major directions. Strategic models rely directly on the theory of non-cooperative games (see, e.g. Ashworth & Ulph (1981), Leuthold (1968)), while collective models proposed by Chiappori (1988, 1992) rely on the basic assumption that the household decision process leads to Pareto-efficient allocations. The bargaining process may then be either explicit (as in McElroy & Horney (1981), or in Lundberg & Pollak (1993)), or non-specified (as in Chiappori (1988, 1992)). In the latter case, the bargaining process is very general and not restrictive, only assuming Pareto-optimality. Pareto-optimality hypothesis seems natural for analyzing household decisions since family members, who interact over a long period, are probably able to find mechanisms leading to efficient decisions.

As shown by Chiappori (1992), collective models can be used to study the welfare level of each household member, and therefore to analyze and measure

in a consistent way the redistributive effects of any economic policy, not only at the household level but also at the individual level. More specifically, he showed that, under some rather plausible conditions, individual utility functions can be recovered from household behavior (and disentangled from bargaining power effects, whereas bargaining effects induce a bias in the measurement of preference parameters in unitary models).

The collective model introduced by Chiappori (1988, 1992) was originally restricted to continuous decisions, because it relied on the first-order condition resulting from Pareto-optimality. The extension to continuous-discrete applications (Blundell et al. 2007) is very recent because it implies complex technical developments. A few years before, Van Soest (1995) considered a discrete choice model of joint labor supply within the family, but it did not take into account the negotiation process within the household, and it did not test or impose the restrictions implied by the optimality of the joint decision. Vermeulen (2006) made another step to introduce discrete (female) labor supply, but his model does not impose Pareto-optimality of discrete labor supply decision.

9.3 Modeling intra-household decisions in two specific examples

An important issue in modeling residential and job location is the interdependence of these two decisions, both at the individual level and at the household level. It can be argued that the choice of residential location is made conditional on the workplace, or vice versa (Waddell 1996). We can also argue that the choice of residential location is made in different time frames according to the life cycle of individuals. What households then consider in their joint location decisions, with a varying degree of importance along their life cycle, are the chances of each active member in the household to access a good job in each alternative location, which determines the accessibility to jobs of each household member, from any given residential location.

We present here an innovative perspective on the interdependence between residential location and workplace of spouses, taking into account the process of negotiation within the household, the travel time of each spouse, and the job accessibility for each spouse. For this we propose models of residential location to give clear answers to various theoretical and empirical issues concerning: the residential location conditional on workplaces and the joint choice of residential location and workplace of each spouse.

This work has allowed expanding and developing models that correct the bias in the individual value of time of spouses. This is achieved by measuring separately the influence of explanatory variables on (1) the decision-making process within the household and (2) the value of time of spouses. The bar-

gaining power is taken into account to analyze the effect of the commuting time of spouses to the household residential location choice. For more details on the modeling framework of the results developed below see Chiappori, de Palma, Picard & Inoa (2013).

9.3.1 Mode choice and spouses travel times

The differences between male and female observed travel times depend (1) on spouses' endogenous joint mode choices, (2) on the distribution of male and female jobs over the region (marginal distribution of the destination of morning commuting trips), (3) on the distribution of dwellings over the region (marginal distribution of the origin of morning commuting trips), and (4) on the household joint residential and job location choices. The link between the marginal distributions of origins (home) and destinations (work place) and their joint distribution is partially explained by the spouses bargaining powers for location choice in the household. If the women's bargaining power is larger than the men's one, then the joint distribution of household residential location and spouses' job locations results in a reduction in woman's travel time compared to man's travel time.

In this section, we make a first step at disentangling the effects of these different sources, focusing on the relative reduction in actual women's and men's travel times (compared to all potential travel times) resulting from their respective bargaining powers.

Using data form the 1999 Census in the Paris Region, we analyze the distributions of spouses' potential and actual travel times, and interpret their differences in the light of respective bargaining powers. Travel times are computed using the dynamic transport network model METROPOLIS (de Palma & Marchal 2002). The Paris Region is composed of 8 *départements*: Paris, 3 *départements* around Paris ("inner ring"), and 4 *départements* farther away ("outer ring"). The city of Paris accounts for 20 *arrondissements* that correspond to 20 communes. The region contains 10,724,748 inhabitants for a total of 4,510,369 households. Household location and workplace are observed only in a 5% sample. About half of the individuals in this 5% sample work, which represents 242,516 workers, among which 239,499 work in the Paris area. We finally restrict our sample to couples in which both spouses live and work in the Paris region, which leads to 60,798 dual earners households (the difference corresponds to singles and one worker families).

Commuting network distances and travel times in chosen alternatives are computed using individual information on actual residential location and workplace, and reported in Table 9.1, in our sample of 60,798 dual-earner households. These distances and travel times correspond to the actual, or observed commuting trips. The upper (resp. medium) part of the table reports distance and travel time information for women (resp. men), and the lower part of

the table reports the average differences and tests the statistical significance of these differences.

In the chosen alternatives, women's commuting distances and travel times are shorter than men's. The average difference is 2.27 km and 5.67 minutes through the public transportation network, and 1.85 km and 2.23 minutes through the private car network. The standard errors are those of the average difference between female and male actual commuting distances and travel times. The t-Stats are the ratios of the Mean to Std Err columns. Under the null hypothesis that the male and female average distances and travel times are equal, this statistic follows a Student distribution. For the 4 variables considered, the absolute values of the t-Stats are very large, and the null hypothesis is clearly rejected: actual commuting distances and travel times are significantly lower for women than for men.

This difference may come from two different phenomena: (1) the marginal distributions of dwellings (irrespectively of the male and female workplaces) is such that dwellings are on average closer to female jobs than to men jobs; (2) the distribution of dwellings conditional on female and male workplaces is such that, for a specific household, the dwelling is closer to the woman's workplace than to the man's workplace.

In order to quantify the first phenomenon, we randomly select 9 hypothetical residential locations for each household, independently from the husband and wife workplaces. To do so, we consider a uniform distribution over the 5 million dwellings of the region. Equivalently, we perform importance sampling of communes, with weights proportional to the number of dwelling in this commune (see Chapter 12 for more details on importance sampling). Following this strategy, we build a sample of dwelling locations independent from both spouses workplaces, and the distribution of these hypothetical dwellings is statistically identical to the marginal distribution of dwellings in the region. The resulting distances and travel times are reported in Table 9.2, which is built on a sample of 60,798*9=547,182 hypothetical (unchosen) alternatives for residential location. The structure of Table 9.2 is the same as the one of Table 9.1, except that the figures are now those of hypothetical travel times.

In the hypothetical case in which households would randomly choose a residential location independent from husband's and wife's workplaces, the average commuting distances and travel times would be approximately multiplied by 2. Stated differently, the fact that households (partially) adjust their residential location choice to the spouses' workplaces divides by 2 the spouses commuting distances and travel times, compared to the hypothetical case in which they would randomly pick one of the existing dwellings. More interestingly, in this hypothetical situation, women's commuting distances and travel times would be slightly larger than men's travel times. This means that, on average, dwellings are marginally closer to male jobs than to female jobs. The fact that actual commuting distances and travel times are shorter for wives than

Variable	Mean	Std Dev	Min	Max
Women				
Network distance (public, km)	13.151	13.973	0.25	134.73
Network distance (private, km)	10.710	11.716	0.25	107.32
Travel time (public, minute)	42.574	34.936	1.00	384.56
Travel time (private, minute)	15.200	14.248	0.76	151.51
Men				
Network distance (public, km)	15.417	15.047	0.25	134.73
Network distance (private, km)	12.557	12.620	0.25	109.12
Travel time (public, minute)	48.242	36.732	1.00	384.56
Travel time (private, minute)	17.431	15.199	0.76	149.36
Difference (Women - Men)	**Mean**	**Std Err**	**t-Stat**	**p-value**
Network distance (public, km)	−2.266	0.083	−27.21	<0.01%
Network distance (private, km)	−1.847	0.070	−26.45	<0.01%
Travel time (public, minute)	−5.668	0.206	−27.57	<0.01%
Travel time (private, minute)	−2.231	0.084	−26.41	<0.01%

Table 9.1 Actual commuting distances and travel times.

for men according to Table 9.1 cannot come from the marginal distributions of dwellings and jobs because the relationship is reversed in Table 9.2, but it comes from the conditional distribution of dwellings given spouses workplaces. Stated differently, the endogenous household residential location favors wives rather than husbands, which suggests a larger bargaining power for women than men in the residential location choice. This result is also consistent with Abraham & Hunt (1997), who explained that the probability of moving is more strongly related to commuting distance for women than for men, which results in shorter commuting distance for women after a relocation.

Note also that the endogenous residential location reduces more actual travel time of women compared to men by public transit than by private car. Given that women have a stronger tendency than men to commute by public transit, this result suggests that households locate so as to favor woman's commuting by transit.

The fact that both dwellings and jobs are rather located in the center of the region divides by about 2 the distances and by about 2.5 the travel times in comparison with the hypothetical situation in which dwellings and jobs would be uniformly and independently distributed among the communes. In this even more hypothetical case, the average distance between dwellings and jobs would be 66 km both through public and private networks, and it would represent an average travel time of 215 min by public transportation and 80 minutes by car.

Variable	Mean	Std Dev	Min	Max
Women				
Network distance (public, km)	31.712	21.177	0.33	182.33
Network distance (private, km)	26.995	19.399	0.33	155.90
Travel time (public, minute)	86.200	50.193	1.32	564.52
Travel time (private, minute)	35.020	20.778	1.00	233.01
Men				
Network distance (public, km)	31.154	20.813	0.35	182.33
Network distance (private, km)	26.464	19.020	0.35	155.90
Travel time (public, minute)	84.844	48.704	1.40	564.52
Travel time (private, minute)	34.469	20.355	1.05	233.01
Difference (Women - Men)	**Mean**	**Std Err**	**t-Stat**	**p-value**
Network Distance (public, km)	0.558	0.0401	13.90	$<0.01\%$
Network Distance (private, km)	0.531	0.0367	14.46	$<0.01\%$
Travel time (public, minute)	1.356	0.0945	14.34	$<0.01\%$
Travel time (private, minute)	0.551	0.0393	14.01	$<0.01\%$

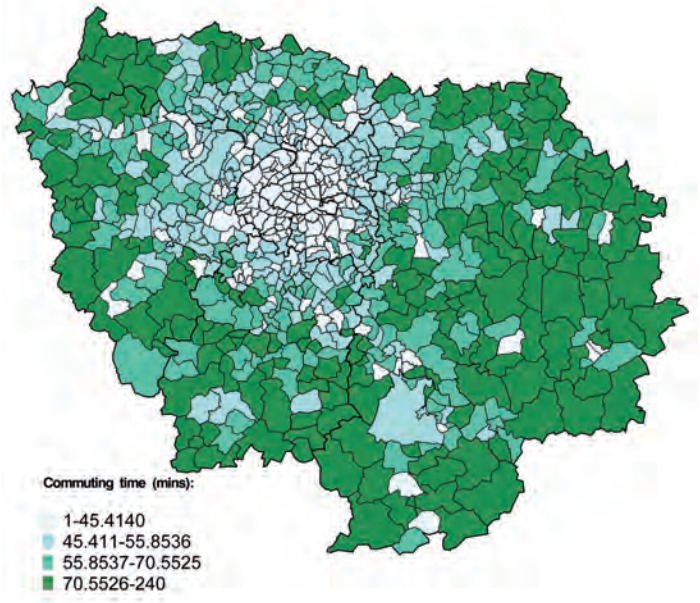
Table 9.2 Hypothetical commuting distances and travel times using the marginal distribution of dwellings.

The geographical distribution of actual commuting time for women and men by public transportation and private care are depicted in Figure 9.1 and Figure 9.2, respectively, by gender. The "Other" category corresponds to communes with no observed trip by the mode considered, for the gender considered. Actual commuting times in public transportation and private car are increasing functions of the distance to the center of the region, in which most of the jobs are concentrated. The difference in travel times between the outer ring and the central part of the region (Paris and inner ring) is more pronounced for public transportation than for private car, which reflects the fact that the public transportation network is concentrated in the center of the region and is poorly adapted to trips from the outer ring to the outer ring.

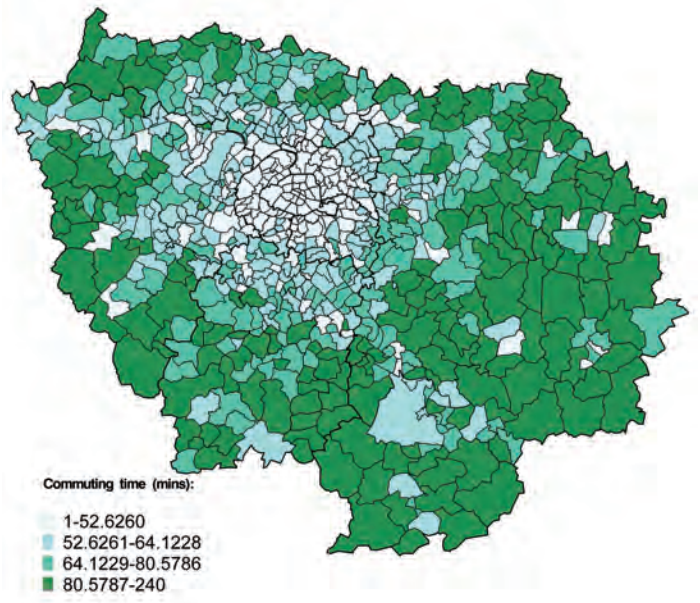
The geographical distribution of relative (woman-to-man) travel times depicted in Figure 9.3 shows that women work closer to their dwelling than men in the outer ring, whereas, for households living in Paris or in the inner ring, the husband's workplace is closer to the dwelling than the wife's workplace. This is consistent with the fact that men's jobs are concentrated in the core of the region, whereas female jobs are more uniformly spread over the region.

9.3.2 Car ownership and mode choice

Using the same sample as in Section 9.3.1, we estimate a multinomial logit model of residential location conditional on husband and wife workplaces, as a

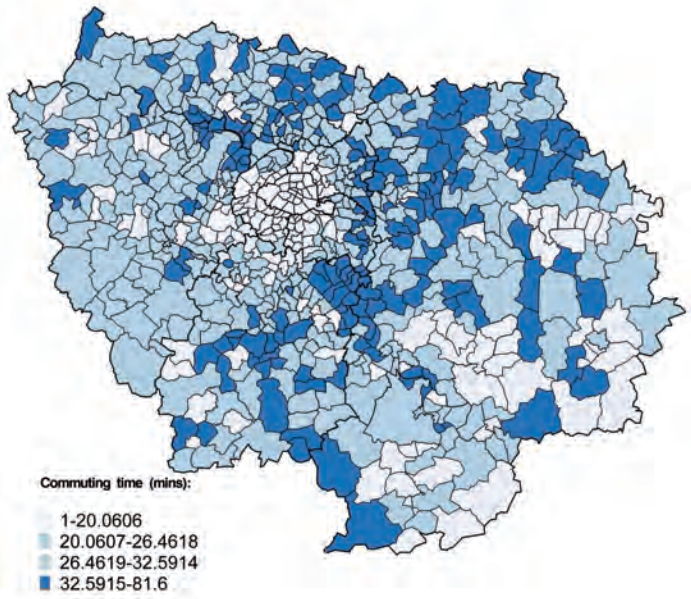


(a) Wife

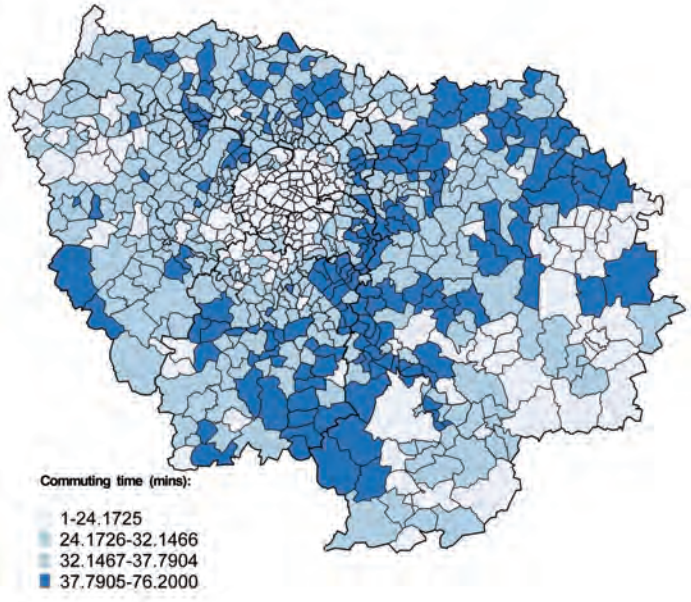


(b) Husband

Figure 9.1 Public transportation observed commuting times.

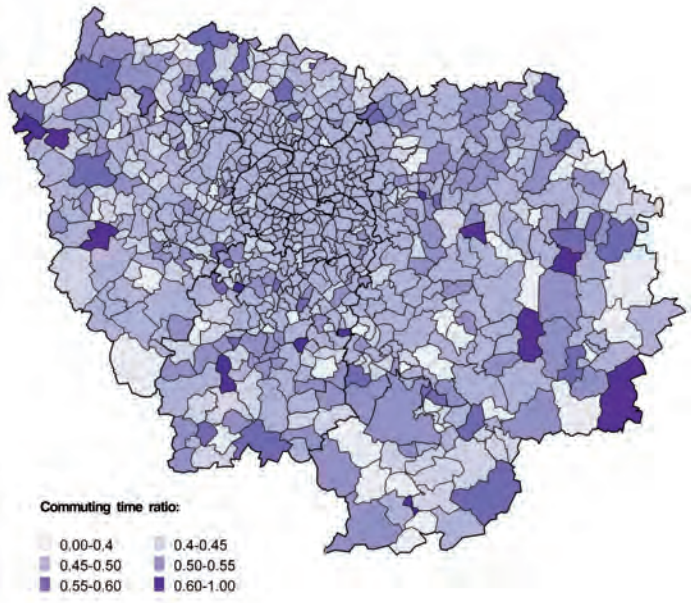


(a) Wife

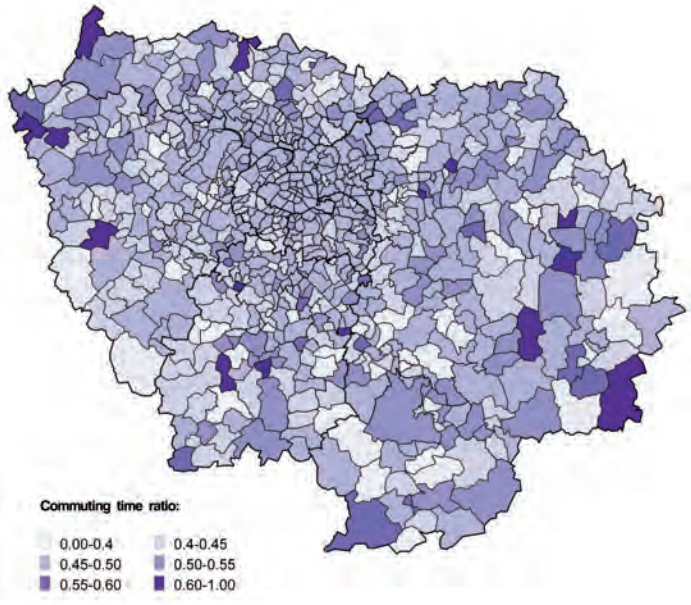


(b) Husband

Figure 9.2 Private vehicle observed commuting times.



(a) Public transportation



(b) Private vehicle

Figure 9.3 Geographical distribution of relative travel times.

Parameter	Coefficient	Std Err	t Stat	Pr>|t|
Travel time by mode and gender				
Log(Price)	–3.7977	0.0314	–120.9	<0.0001
Travel time (public transportation) Wife	–0.0325	0.0005	–69.59	<0.0001
Travel time (public transportation) Husband	–0.0172	0.0005	–36.58	<0.0001
Travel time (private car) Wife	–0.0212	0.0010	–21.68	<0.0001
Travel time (private car) Husband	–0.0183	0.0010	–18.87	<0.0001
Travel times by sex, mode, and number of cars				
Log(Price)	–4.0008	0.0323	–123.87	<0.0001
Travel time (public transportation) Wife, no car	–0.0578	0.0023	–25.57	<0.0001
Travel time (public transportation) Wife, 1 car	–0.0394	0.0007	–56.9	<0.0001
Travel time (public transportation) Wife, 2 cars	–0.0233	0.0006	–36.07	<0.0001
Travel time (public transportation) Husband, no car	–0.0414	0.0023	–17.84	<0.0001
Travel time (public transportation) Husband, 1 car	–0.0232	0.0007	–33.31	<0.0001
Travel time (public transportation) Husband, 2 cars	–0.0094	0.0007	–14.46	<0.0001
Travel time (private car) Wife, no car	–0.0059	0.0048	–1.23	0.2191
Travel time (private car) Wife, 1 car	–0.0114	0.0014	–7.89	<0.0001
Travel time (private car) Wife, 2 cars	–0.0319	0.0014	–22.97	<0.0001
Travel time (private car) Husband, no car	–0.0075	0.0047	–1.58	<0.0001
Travel time (private car) Husband, 1 car	–0.0169	0.0014	–11.89	<0.0001
Travel time (private car) Husband, 2 cars	–0.0191	0.0014	–13.95	<0.0001

Table 9.3 Residential location choice as a function of travel time.

function of local dwelling price (per square meter, in log) and spouses' actual travel time, by mode and gender. Table 9.3 shows results of two different models. In the first model, the Value of Time (coefficient of the travel time variable) only depends on mode and gender. In the second model, the travel time variable is crossed not only with mode and gender, as in the first model, but also with the number of cars owned by the household.

The Price has the usual negative sign and is highly significant in both models. According to the first model, household location is slightly more sensitive to woman's than man's travel time by private car (|–0.0212| is significantly larger than |–0.0183|), but it is twice more sensitive to woman's than man's travel time by transit. The second model helps understanding such differences. Travel time by private car (both for husband and wife) has no influence on household location for households with no car, which is totally consistent. Households do not care how much time husband or wife would spend commuting by car, when they have no access to any car.

For each household, 9 unchosen alternatives are generated using importance sampling, alternatives corresponds to the 1300 communes of the Paris Region.

The difference between the coefficients of travel time by private car for the wife in households with 0 or 1 car is not statistically significant. For the husband it is the difference between 1 and 2 cars which is not significant. This means that the husband's commuting time by private car becomes relevant in residential location choice as soon as there is at least one car in the household.

By contrast, the wife's commuting time by car becomes really relevant only when there is at least 2 cars in the household. This gives a strong indication that the husband has the priority to use the car to commute when there is competition between spouses for the use of the unique car.

The influence of travel time by public transportation on residential location is decreasing with the number of cars in the household, both for the wife and for the husband. This influence nearly disappears for men when there are two cars in the household, suggesting that the husband usually commutes by car and does not care about transit travel times when there are 2 cars in the household. These results are consistent with the fact that, on the one hand, public transportation is a substitute for private car when household members can easily reach a station, which is usually the case in Paris and the inner ring, whereas it is a complement when individuals have to drive to the station in order to use public transportation. The fact that husband's transit commuting time plays virtually no role in residential location of two-car households, whereas the wife's transit commuting time still matters, suggests that when there are two cars in the household men will anyway entirely commute by car while wives go to the station by car.

9.3.3 Couple residential location conditional on spouses workplaces

We now discuss the results obtained by Chiappori, de Palma, Picard & Inoa (2013) in a structural collective model assuming that spouses workplaces are chosen before the residential location. In that case, like in Section 9.3.2, residential location depends on the spouses actual commuting times between the current workplace and the potential residential locations contemplated. However, by contrast with the reduced-form model estimated in Section 9.3.2, we analyze here a structural model explicitly taking into account both spouses individual preferences and respective bargaining powers rather than mixing them in a household utility function which may or may not be consistent with collective rationality. To the best of our knowledge, this the only contribution to the literature that takes into account the role of the within-family decision process and spouses bargaining powers in a residential location. Picard et al. (2013) develop and estimate a one-step structural model to disentangle bargaining powers from spouses' values of time in a joint mode choice model. They find results similar to the results presented in this section concerning the determinants of the bargaining power. The other contributions such as Abraham & Hunt (1997), Beharry-Borg et al. (2009) do consider the influence of individual characteristics on residential location choices, but the associated coefficients mix the influence of individual preferences and bargaining powers.

The objective here is to analyze the within-family decision process involving bargaining between members with diverging preferences, objectives, and con-

straints. Spouses may or may not have diverging preferences concerning local amenities, and these preferences probably differ from the preferences of singles. The same individual will tend to enjoy bars and discos when he is single, and open spaces when he is married with children. Given this change in preferences of the same individual when he marries, it is not possible to disentangle respective bargaining powers and spouses' preferences for local amenities, and we will not try to do so. Instead, we consider a joint preference of the household for local amenities.

On the opposite, it is obvious that the husband preference for his own commuting time (let's call it Value of Time (VOT), although it is more complex when utility is not linear in travel time because then VOT is given by the local derivative of utility with respect to commuting time, and it varies with commuting time) is different from the wife's preference for the husband commuting time. The influence of the husband's commuting time on household residential location mixes (1) the role of the husband's VOT, which a priori depends only on the husband's individual characteristics and (2) the role of respective spouses bargaining powers, which a priori depends both on the husband and wife individual characteristics. Similarly, the influence of the wife's commuting time on household residential location mixes the role of the wife's VOT and of the respective spouses bargaining powers. As a result, neglecting spouses respective bargaining powers leads to biased estimates of the values of time of the household members.

Chiappori, de Palma, Picard & Inoa (2013) have developed a method to disentangle the bargaining power and the spouses VOTs, and to measure separately the influence of explanatory variables on the bargaining power and the value of time. They applied this method to the same data as in Sections 9.3.1 and 9.3.2. Each spouse's bargaining power is normalized to 1/2 in the reference case (the two spouses are French and 20 years old), and the husband and wife bargaining powers always sum to 1, so that bargaining powers can be interpreted as percentages. Any increase in the woman's bargaining power corresponds to a decrease of the same percentage for the husband's bargaining power. The econometric results show that the woman's age and man's nationality play a crucial role in determining bargaining power. The magnitude of the effects depends on the covariates considered, but some general patterns emerge. Each spouse's bargaining power increases with age, but the increase is faster for the wife than for the man. Consider two equally-old men. The wife of the first man is 10 years older than the wife of the second. Our estimates show that, in this case, the bargaining power of the first wife is 4.28% larger than that of the second woman. Consider now a third couple, in which the wife has the same age as the wife in the first couple, but the husband is 10 years older than the husband in the first couple. Then, the bargaining power of the husband in this third couple is 0.78% larger than that of the first husband. Consider now the first couple 10 years later. Each spouse is 10 years older

and, as a result, the wife's bargaining power has increased more the the man's. As a consequence, the wife's bargaining power tends to increase over time (by 4.28% – 0.78%= 3.5% each 10 years).

The bargaining power of the husband is reduced by about 4.5% when he is foreign, whereas the wife's bargaining power does not significantly depend on her nationality.

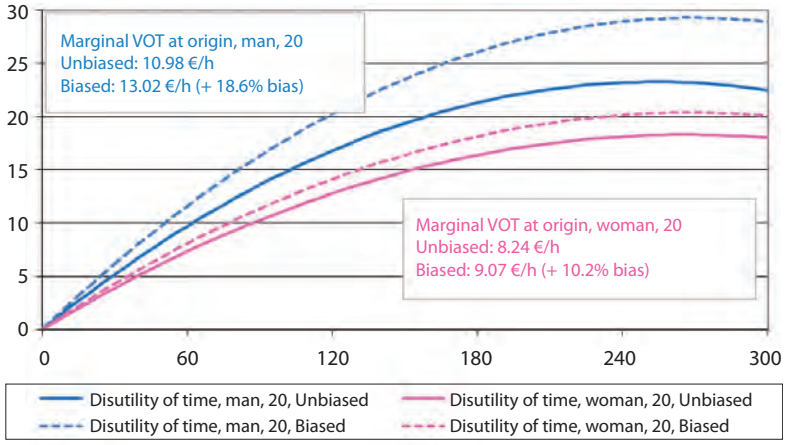
The econometric results of this model also show a large bias in the measurement of spouses' VOT when the bargaining power is not taken into account in a residential location model. For example, there is a 20% underestimation of the value of time for a 40-year-old French man, which becomes a 18% overestimation for a 40-year-old French man. For the wife, the VOT is always overestimated when the bargaining power is omitted, and the bias is 15% for a 40-year-old French woman and 10% for a 20-year-old French woman. There is an additional bias upwards in the measurement of the VOT of foreign men, when the bargaining power is omitted.

The commuting cost estimated in the model is a concave function of commuting time for each spouse. The value of time for a man (compared to a woman) of 20 years old is about €11 (respectively €8) per hour at the origin (i.e. when the travel time tends to 0). The value of time is a decreasing function of age.

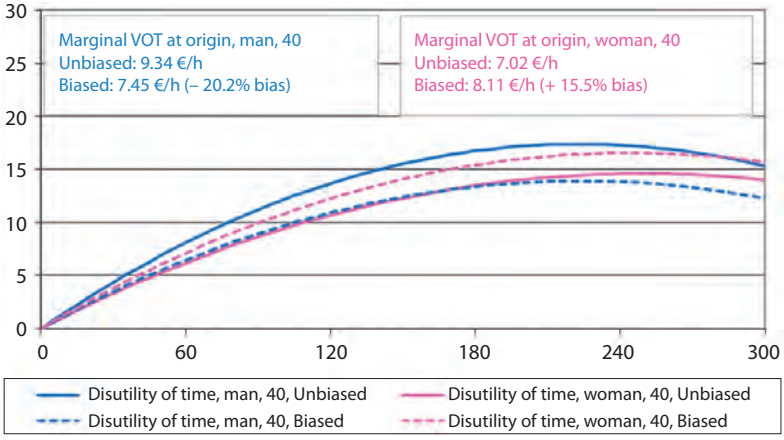
9.3.4 Accessibility in a residential location, workplace, and job type model in one- and two-worker households

We now discuss the extension to dual-earner households of the single or one-worker household structural model developed by Inoa et al. (forthcoming). By contrast to Sections 9.3.2 and 9.3.3, this model assumes that residential location is chosen before individual workplaces. This model analyzes a three-level nested Logit model of residential location, workplace, and job type choice. Residential location is the upper-level choice, workplace the middle-level choice, and job type is chosen at the lower level.

The decision tree is fully consistent and rational: on one hand, workplace choice is anticipated at the upper-level residential location choice; on the other hand, the middle-level workplace choice is made conditional on the residential location actually chosen. This model allows defining and computing an individual-specific measure of attractiveness of geographical units, which takes into account the local distribution of job types, and individual preferences for job types. This measure of attractiveness of geographical units is fully consistent with the model, from the theoretical and econometric points of view. It corresponds to a log-sum term going up the decision tree from the lower level to the middle level. Going one step up in the decision tree, we obtain an individual-specific measure of accessibility to jobs, which takes into account



(a) Age 20



(b) Age 40

Figure 9.4 Biased and unbiased value of time.

the heterogeneity in VOT and preferences for different job types captured in the measure of attractiveness of potential workplaces.

The model is estimated on single or one-worker households, using the same data as in Sections 9.3.1 to 9.3.3. At the lower level of the decision tree, preferences for job types depend on gender, education, and number of children (for women only). The same variables also explain observed heterogeneity in the Values of Time at the middle level of the decision tree. At the upper level, observed heterogeneity in preferences for local amenities is even richer, with the introduction of profession of household head, and household composition and income, crossed with the local distribution of the same characteristics, in order to reflect the tendency of households to locate close to similar households.

9.4 Conclusion

We have discussed in this article different models describing mode choice, residential location and workplace of the various members of a family, mainly the husband and wife, but more sophisticated models should also take into account the children. The key concept introduced in this chapter is the negotiation within the couple, that is to say, the decision-making process related to joint choices, such as the residential location, which is usually, but not always, unique in couples.

The tools we have discussed in this chapter are based on the work developed during recent decades in the field of economics of the family. They will revolutionize discrete choice models used in transport and urban economics, in the sense that the decisions are no longer described as the outcome of a unique decision maker. Instead, in collective discrete choice models, the decisions depend on the preferences and constraints of two individuals tied by economic, but also emotional ties. Discrete choice theory still has to be adapted to this new way of approaching the individual decisions. Current literature in transportation (Abraham & Hunt 1997) has developed a few family decision models. However these models ignore the bargaining process within the family, as studied by family economists (Chiappori 1988, 1992).

An objective of the research is to provide new tools for researchers in urban and transportation economics, as well as in engineering. In particular, we plan to provide collective choice models to the integrated transport and land use tool, UrbanSim. This article begins a search path that may be still very long and hopefully fruitful.

Chapter 10

A real estate development model with heterogeneous agents

Christof Zöllig Renner, and Kay W. Axhausen

10.1 Introduction

In a spatially and dynamically explicit simulation of spatial development[11] real estate supply models have to determine when, where, and how much of which type of real estate is built. The events constituting to the evolution of the building stock include construction, alteration, deconstruction and replacement of buildings. These events are the consequences of decisions made by the owners of the respective land and buildings. This suggests that real estate development can be modeled by applying behavioral modeling. Other events can be the consequences of physical processes (e.g. a building might be destroyed by an earthquake) but they are not further considered here. The model would ideally answer all questions above at once since in reality all these decision facets are considered simultaneously. The decision facets of time and quantity are continuous whereas location and type of real estate are discrete. It is not only the options which are relevant for decision making but also the decision maker and the circumstances. A complete model thus comprises explanatory variables describing the alternatives, decision makers, and decision situations to determine the decisions on time, location, quantity and type of real estate.

The aim of this research is to analyze the influence of developer characteristics on their development decisions. Accordingly, the following research questions are formulated to guide our work:

- Can behavioral differences be expected?
- How does the behavior of real estate developers differ?
- Is it possible to capture their behavioral differences in quantitative models?
- Can the models be improved by considering developer types?

[11] A framework of the modeled object is shown in Figure 2.1 (p. 18).

These questions are approached in two main steps. Firstly, we conducted eleven in-depth interviews with developers active in the Canton of Zürich in Switzerland. The qualitative work allowed us to get access to the subject and develop some intuition about the data at hand. Secondly, we estimate deterministically segmented location choice models according to developer types we find in the data. We concentrate on the question *where* new development projects occur and if there is distinguishable decision behavior. The basis for both steps is a dataset of almost 60 thousand records of real estate development project applications in the study area from 2000 until 2010. The records contain contact details of the real estate developers as well as the addresses of the development projects. The dataset is bought from the firm DOCUMEDIA (Docu Media Schweiz GmbH 2013) which collects the information to facilitate the formation of construction consortia. The study area and the observations finally used for model estimation are shown in Figure 10.1.

The remainder of this chapter is organized as follows: Section 10.2 gives a brief overview on the literature. In Section 10.3 the main idea of the chapter is derived from classical theories. Section 10.4 reports on the qualitative study and its conclusions, the quantitative models are presented and discussed in Section 10.5 which is followed by conclusions (Sect. 10.6).

10.2 Literature review

The literature review has two parts. The first part summarizes qualitative studies whereas the second part focuses on real estate supply modeling in the context of land use transport interaction (LUTI) models. The focus of the first part is on categorization of real estate developers and emphasizes the situation in Switzerland.

10.2.1 Developer types

A study in eastern Australia identifies six developer types according to their behavior (Coiacetto 2001). The research used semi-structured, flexible, and interactive interviews. The following developer types were identified: *passive local property owning developers*, *"Means to a mission" developers*, *specialized client developers*, *showpiece developers*, *"eyes on the street" developers*, and *value adding opportunity developers*. These names show that purpose and endowment of developers potentially influence their behavior.

Three developer types are found by Ball (2003) when investigating the structure of the house building industry in France. His categorization is along the lines of purpose and legal status. The types identified are: *owner-developer*, *construction firms*, and *social and semi-public developers*.

Ruming (2010) finds *small local*, *medium local/regional*, and *large regional* developers. Differentiation is made according to developers' size and

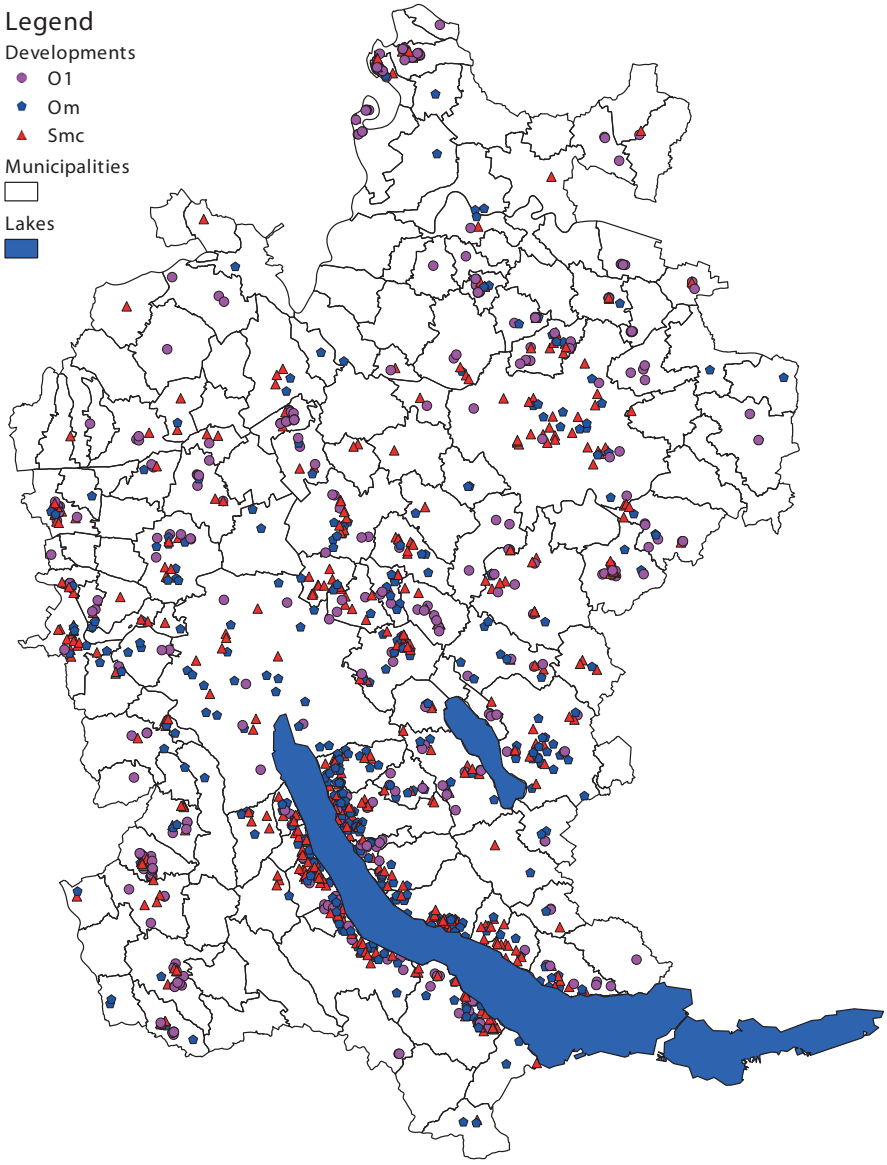


Figure 10.1 Observed real estate development projects in the study area according to developer types © 2011 Vector200 Swisstopo (JA100120).

area of activity. It is also found that informal associations with local administrations are very important to facilitate the approval process.

In Switzerland three studies come up with developer categories. Schüssler & Thalmann (2005) identify two main types, the *promoter* and the *owner-occupier*. The distinction is made according to the main objectives of the

developers which is in the first case to buy and sell property while maximizing profits. The second type has more diverse goals. A first subgroup considers investments into real estate as diversification option. A second subgroup tries to maximize benefits for the community and a third group holds real estate but wants to sell it as soon as the market allows for it.

van Wezemael (2005) identifies developer types according to legal status (commercial or public), strategy (portfolio strategies, object oriented strategies) and size (big, small). It can be expected that developers following a portfolio strategy are also bigger in size. These developers are also likely to have more resources available.

The categorization by Friedrich (2004) is done according to the legal status. She uses the categories of *institutional*, *public*, and *private* developers.

10.2.2 Modeling approaches of real estate supply

Gravity models The first LUTI models did not explicitly model real estate supply. They related spatial distribution of residents directly with employment distribution (Lowry 1964, Putman 1974) following a gravity approach borrowed from physics.

Economic models With the second wave (Iacono et al. 2008) of LUTI models economics provided the principles for land use allocation. From than onwards the land use development processes were understood as resulting from market interaction. We can distinguish *aggregate* and *disaggregate* models among economic models.

Aggregated models A first category of aggregate models focuses on trade relations of regions (*regional economic models*). Real estate supply, in terms of floor area, adapts as a reaction to price change. Examples of such models are TRANUS (de la Barra 1989) and PECAS (Hunt & Abraham 2005). A second category of models elaborate on *land markets* by describing both demand and supply. Anas & Arnott (1993) developed a model predicting construction and demolition probabilities according to real estate related costs and expected market prices. Demand is modeled with a nested logit choice model. A similar approach is taken in DELTA (Simmonds 1999) where the amount of new built space provided is determined calculating a ratio of current rent levels and construction costs. Allocation to zones is based on profitability. Martínez (1992) further developed the bid-rent theory formulated by Alonso (1964) and introduced the bid-choice approach to MUSSA. This framework simulates an auction process determining simultaneously price and allocation of households. A more recent version of the model exhibits a real estate supply sub-model which assumes profit maximizing developers. There aims are to maximize the rents of their stocks and to find combinations of buildings and zones with

maximum profit. The decisions on what combination to choose is modeled with a multinomial logit (MNL) model. The amount of units supplied is derived from demand (Martínez & Donoso 2010).

Disaggregated models With increasing computational power the trend of disaggregation leads to microsimulation models where the behavior of individual actors is modeled (as proposed by McFadden (1978)) and simulated.

Waddell (2002) uses a MNL choice model which determines the probabilities for a set of land use transitions (including no land use change) for each location in each year in early versions of UrbanSim. In a more recent version (Waddell et al. 2010) a development transition model determines the number of real estate units of a certain type on the basis of target vacancies and simulated vacancies. Development projects are sampled from a history of observed projects until the calculated number of units is reached. The sampled projects are again located with a MNL model. The choice set includes randomly drawn parcels with suitable zoning and remaining capacity. A hedonic real estate price model (Rosen 1974) is used to determine prices which are introduced among other variables in the location choice model.

Zhou & Kockelman (2008) micro simulate supply and demand of single-family housing on the basis of bid-rent theory. Real estate is built on that vacant piece of land which maximizes profit. Cost of land purchase and construction are imposed on the developer and selling prices are determined by the highest bid of a household at equilibrium (i.e. all households are allocated to a dwelling). An interesting feature of the model is the treatment of large undeveloped parcels which are subdivided according to an empirically derived size distribution. This is a statistical parceling model.

In ILUTE (Salvini & Miller 2005) real estate supply is determined with an economic and location choice model. The economic model predicts the number of housing starts. Four location choice models predict in which zone the housing start will occur. The location choice models are stratified according to project types (Haider & Miller 2004).

A similar approach is described by Dong & Gliebe (2011). The quantity of new housing supply is determined with a time series model. The type of project to be located is determined with Monte Carlo drawing from a fitted empirical distribution. Three models for spatial distribution of single family housing projects are compared. The first is a MNL model without market segmentation. The second is a MNL model with exogenous market segmentation and the third model is a latent class model, i.e. endogenous market segmentation. The authors find that the models with taste heterogeneity are theoretically more appealing. However, predictive power is not improved due to practical issues. Even though the authors use the term "developer heterogeneity", they are not actually discriminating developments on the basis of

developer attributes. They use project size as discriminating variable, which describes the project rather than the developing actor. The authors presented models with discrimination according to developer type using contractor status and specialization as attributes in a earlier version of the paper (Dong & Gliebe 2010).

10.2.3 Conclusions from literature

The literature on developer types reveals that there are behavioral differences according to purpose, legal form, size, and endowment with properties. None of the studies was explicitly undertaken to inform discrete choice modeling later on.

The conclusion with respect to real estate supply models is that most of them assume a representative agent. Basically, all real estate construction industries are treated likewise. The work which investigates heterogeneity on the supply side is limited to segmentation according to project types. The studies focus almost exclusively on the housing segment due to data availability. There is much more research done on the demand side for that reason too (as found by Hunt et al. (2005)). The household population is much better known than the real estate developer population. The research presented in the following starts to close that knowledge gap.

10.3 A concept for a real estate development model with heterogeneous agents

In this section we embed a model with heterogeneous real estate developers into urban economic theory. Firstly, we describe why we expect real estate developer types to behave differently. Secondly, we discuss expected consequences on a more aggregated level.

Comparative advantages We understand that some combinations of site and developer just fit better. A large scale developer with a lot of overheads needs development opportunities with appropriate returns to scale. On the other side of the spectrum we think of households which develop their customized houses. The location choice for a real estate development project will be in accordance with these prerequisites.

Parcel size Parcel size is the result of the ownership history of the land and of the attached building (e.g. consider countries with inheritance laws allocating an equal share between the heirs). The consolidation of parcels is in itself a costly and lengthy process, which offers chances for developers of different scale.

Bid-rent model of Alonso (1964) The bid-rent theory of Alonso (1964) explains how comparative advantages of a location are capitalized into the land price. For example, less transportation costs allow the budget constraint household to make a higher bid for a central location.[12] On the supply side the developer needs a piece of land as primary resource. This piece of land will be more expensive at central places due to the higher bids, i.e. the primary input factor is more expansive. The increase in land price leads to factor substitution of land into construction technology and know-how. This relation is the core of Muth and Mill's theory (Mills 1967, Muth 1969). On the supply side the requirements with respect to know-how and technology increase. Consequently, a specialization is necessary to build in central locations.

Spatial effects The specialized developer is more likely to find better suitable parcels at central places. The less skilled developer will thus be outbid by the more professional one and ends up with a more remote parcel. Further reasons to expect more professional developers at central locations are *a*) that the market size is larger at central places. Thus the developer, who has limited information about the preferences in the market, has less risk. *b*) When financial agents are present, capital can be borrowed. We can assume lower capital costs for professional large scale developers since they can provide more securities. This favors professional developers at central locations which require more capital input.

Short run vs. long run In the short run we would expect that households who develop their own house simply lack the necessary endowment and thus get pushed away from central places by more efficiently working, professional developers. In the long run we would expect a consolidation of the developer industry such that less professional developers disappear. The assumption for short run equilibrium can be justified by entry barriers.[13]

10.4 Interviews with developers in Zürich

The purpose of the in-depth interviews and the analysis of the resulting qualitative data is to inform the quantitative modeling presented in Section 10.5. The usage of the same dataset for the qualitative work and the modeling allows us to get some intuition about the observations by contacting the actors responsible for the decisions.

[12] Centrality is measured in terms of accessibility in our case.
[13] A list of examples can be found in Coiacetto (2009, 122).

10.4.1 Methodology

The methodology used is described in detail by Gläser & Laudel (2004). The main steps are the preparation of a theoretical framework and an explanatory strategy on the basis of which the interview questions are formulated. The following in-depth personal interviews are the data collection step. The third step is the qualitative analysis of the transcribed interview texts. The work of this step includes the extraction of the relevant information with an extraction raster, the preparation of the extracted content and its analysis. The last step is the interpretation of the results.

In this case the interviews shall help in particular to confirm different developer behavior for the hypothesized typology[14] shown in Table 10.1 which is applicable to the data from DOCUMEDIA. Thus the interviews shall investigate the behavioral differences between these three developer types. We try to relate attributes of the developers with characteristics of their decision making.

Developer type code	Developer type name	Purpose	Number of projects
O1	Self-owning without portfolio strategy	Own use or letting	1
Om	Self-owning with portfolio strategy	Own use and/or letting	Several
Smc	Commercial developer (Promoter)	Sale	Several or 1

Table 10.1 Definition of developer types in DOCUMEDIA data.

Eleven interviews were conducted in spring 2011 either by telephone or face-to-face. Their duration varies from half an hour up to one and a half hours. The interviews followed an interview guideline which was not strictly adhered to, thus the interviews were semi-structured. For the recruitment of the interviewees we used the contact information from the DOCUMEDIA dataset. Stratified sampling is employed according to the a priori defined developer types in order to get at least one candidate per category. In addition we selected the ten most active developers in terms of recorded project applications of which eight were of type *Smc* and two of type *Om*. Of the respondents two were of type *O1*, four of type *Om* and five of type *Smc*. The interviews were recorded on tape and later transcribed. This raw material is then analyzed with respect to behavioral differences according to developer types. A more detailed description can be found in Zöllig & Axhausen (2012*a*).

[14] Public housing corporations are not available in the data and could not be considered explicitly.

10.4.2 Differences in behavior of real estate developers

In the analysis we list the behavioral information according to developer characteristics which gives an idea about the relation between developer characteristics and their behavior. We distinguish the developers according to purpose and professionalism.[15] Along these lines differences are found with respect to decision criteria, underlying information, the treatment of the information for decision making, perception of alternatives (i.e. parcels on the micro level and municipalities on the macro level) and tasks undertaken during the construction process. The findings are summarized in the following.

Decision criteria None of the interviewees relied only on one criteria. Most mentioned criteria are necessary but not sufficient with respect to a development decision. The criteria and attributes are not mentioned on an operational level which indicates that most of these actors do a lot of their evaluation work in a qualitative way.

The responses of the interviewees confirm that developers with purpose *sale* are profit oriented and take more risk. The higher risk is priced in by requiring higher return and shorter pay-off time. This must be balanced over multiple projects. One interviewee noted that the high risk of development units is likely to be compensated by good margins in the construction unit in case of promoters.[16] This was recognized as cross-subsidizing of business units which eventually leads to underpricing of risks. This is probably a reason why promoters tend to separate the development unit from the construction unit.

Underlying information The evaluation methods of professional developers are more data oriented. Professional developers improve their knowledge by applying data mining methods. They also buy processed information from specialized service providers. Such processed data is in most cases compared to other sources for the sake of a more robust assessment. Professional developers must rely on data analysis and more sophisticated tools because they manage large portfolios. This is also true for professional promoters which are usually active in large areas. Unprofessional developers mention service providers[17] as sources of information. This suggests that some actors are specialized in selling development services rather than built space.

Perception of alternatives Two interesting indicators have been mentioned by professional developers. One reported that the opening of super-

[15] A developer is classified as *professional* if it is a company with more than five realized projects and a turnover of more than one million Swiss Francs. The number of realized projects is used as a proxy for professionalism on the firm.

[16] See Section 10.2.1 for a definition.

[17] Service providers are specialist within the real estate development industry, e.g. brokers who know about market prices or banks which help with financing.

markets is a good sign for a location. Not only that the location gets more attractive but it is expected that the retailers make well informed decisions. Another interviewee points out that immigration of young households is of special interest. The comparison of the number of reported attributes shows that professional developers analyze the situation in more detail. The general aspects such as high accessibility, low taxes or suitable shape of the parcel are also covered by the unprofessional developers.

Attributes related to administrative areas are reported as well. This can be explained by two reasons. Firstly, people identify with administrative areas such as quarters or municipalities, if, for example, an area has a media presence and image. Secondly, a lot of data considered by developers is associated with administrative areas.

The interviews confirm that the consideration of location characteristics depends on planned usage, e.g. noise is more important in case of housing projects compared to office projects. Most interviewees answered from the perspective of housing use. However, there were few respondents which distinguished attributes according to planed uses.

The search space of unprofessional developers is found to be local. This confirms that familiarity with a region influences location choice. Beige (2008) found strong attachment to the region of current residence in case of household location choice. In one case the project location was given as an inherited property. This is an example pointing at the path dependency within the spatial development process. Other prerequisites which may result in different personal preferences are social networks or familiarity with the regionally varying education system.

Another finding is that professional developers can profit by offered development opportunities. In such cases they are contacted by property owners who want to sell their property or have it developed. This passive way of obtaining development opportunities is only likely for actors that are known for their development skills.

Tasks undertaken during real estate development Self-owning developers also finance their project. Thus they are representing the demand side in a market where development services are traded. On the opposite side we have the *development service providers* which all undertake the construction management task. This confirms the findings of Healey (1991) who defined the developer as coordinator.

The condition that all tasks have to be carried out for the realization of a development and the finding that none of the developers cover the full task spectrum shows the presence of a network of actors in all cases. The complete network is unknown, but we can assume that these networks would help to explain particular developments. These developer networks are dynamic in

the sense that they change from project to project and sometimes even during a project. However, we can conclude from statements in the interviews that certain parts of such a developer network can be more stable because social and business contacts have been established.

Buying property can be optional for developers with existing portfolio. Typical examples are large industrial firms which have land banks. One interviewee reports that the current business resulted from a decline of a large industrial company. This example shows that endowment especially with an existing portfolio creates a completely different situation in terms of development opportunities. The portfolio serves as well as collateral for a loan.

10.4.3 Conclusions from interviews

The decision making of developers is varying according to the variables *purpose* and *professionalism* with respect to their decision criteria, the information sources, and executed tasks. The a priori defined developer types seem to exhibit behavioral differences. The hypothesis regarding the attributes of location choice are presented in Table 10.3.

Developers with purpose *sale* see their profit either in providing efficient development services or in selling a finished product. Therefore, they expect shorter pay-off periods than developers with *own use* as purpose. Developers with purpose *letting* have reduced expectations for profitability which we explain with their preferences for secure investments.

Professional developers have more resources and know-how which allows them to exploit more information than unprofessional ones. The information advantage allows them to be more strategic in their choices. We also find that professional developers have a wider activity space.

The behavior also varies in terms of tasks carried out within the development process. The business cases are related to the endowment in terms of real estate portfolio and skills. The endowment with a real estate portfolio is found important because the land resources are development opportunities at hand and as well securities to get better financing options.

10.5 Real estate development models considering heterogeneity of developers

In this section we present the quantitative work which investigates behavioral differences between developers estimating discrete choice models. The aim is to confirm the findings from the qualitative study with the data at hand.

10.5.1 Heterogeneity in discrete choice modeling

In the realm of discrete choice modeling (Train 2009, Ben-Akiva & Lerman 1985) we find several notions of heterogeneity. Depending on the aspect of choice situations within which heterogeneity is assumed, different methods can be used to accommodate the assumption. Heterogeneity can be modeled with respect to alternatives and decision makers. In the first case it is about revealing the structure of the set of alternatives. The modeler has to find the right model to accommodate the correlations found in the error terms of a basic MNL model. Possible models are nested logit (NL), cross nested logit (CNL), pairwise cross nested logit (PCL), or ordered general extreme value (OGEV).

In this study we use deterministic segmentation to address heterogeneity among decision makers. The deterministic segmentation has the advantage to be more easily interpretable and ties taste variations to observed attributes which will be helpful for simulation. The segmentation assumed follows from the available data and is confirmed by the qualitative study. The segments, in this case developer types, are shown in Table 10.1. We estimate basic MNL location choice models in order to apply random sampling of alternatives (McFadden 1978). Model specification and estimation with UrbanSim (Waddell 2002) is described in 10.5.3.

10.5.2 Observations and data preparation

The DOCUMEDIA dataset contains real estate development projects localized by address. This allows us to construct a dataset which combines the information of developer type, constructed project, and chosen parcel. The projects can contain multiple buildings. This is preferable to observations of single buildings because of possible correlation precisely due to the fact that they have been realized in the same project. For a discussion see Zöllig Renner & Axhausen (2013).

Another reason to use the DOCUMEDIA dataset is the information on the developer behind it. There is an attribute which contains the purpose (own use, letting, sale) and contact details. From the contact details we can infer the frequency of projects a developer carried out. The frequency is used as a proxy for professionalism. In addition, there is a lot of detail on components of the building which is not useful for our purpose. The dataset is quite extensive with almost 60 thousand records. However, the number of usable observations after data preparation is dramatically lower (Table 10.2). There is no guarantee that the dataset contains all development projects carried out in period from 2000 until 2010.

The data in the register of buildings and dwellings (GWR) contains additional attributes on the buildings constructed. Therefore, we match the DOCUMEDIA and GWR datasets via a spatio-temporal join. We use an ad-

dress matching procedure to match DOCUMEDIA projects to parcels and then match the buildings located on that parcel to the respective project. We only use projects in which new buildings are constructed. The number of observations is reduced from 21384 to 1301 when applying the matching procedure. This is unfortunate since more complex models can be estimated on larger samples.

The parcels, which are the alternatives, are enriched with context data. The preparation process is described in Chapter 21 of this book and in more detail by Schirmer et al. (2011). The result is a database comprising parcels with planning constraints, buildings with dwellings, and households with persons and jobs. All entities are related with each other via their relative spatial position. Households are located in dwellings, dwellings are associated to buildings and buildings are standing on a parcel. This abstraction of spatial reality is the context of the development events, which are the observations used for model estimation.

The sample sizes which remain after data cleaning and integration are shown in Table 10.2. In this study we only consider projects with residential use. The samples of mixed use and non-residential projects are too small to allow an investigation of heterogeneity.

Developer type	*SFH*	*MFH*
O1	281	54
Om	148	139
Smc	150	274
Total	579	467

Table 10.2 Number of observations.

The qualitative study confirmed that location characteristics are considered among other factors. Unfortunately, we find only few of the other potentially valuable explanatory variables in the data at hand. Thus, we focus on different evaluation of location attributes. The hypotheses with respect to the variables are shown in Table 10.3.

10.5.3 Model specification and estimation

Firstly, we develop two reference models which do not use information on real estate developers. There is a model for single family housing (SFH) and multi family housing (MFH) projects. This segmentation is captured in sub-models to allow specific filtering of alternatives. The filters are such that only parcels can be chosen on which the respective building type is allowed by the land use regulations. Therefore, the project type is at least equally important as the developer type. The land use regulations define also floor area densities which limits the capacity of parcels in terms of allowed floor space.

Variable	Hypothesis
Accessibility car	We expect professional developers to have stronger preference for high accessible parcels. ($\beta_{Smc} > \beta_{Om} > \beta_{O1}$)
Accessibility PT	We expect professional developers to have stronger preference for high accessible parcels. ($\beta_{Smc} > \beta_{Om} > \beta_{O1}$)
New neighbouring buildings	We expect no difference.
Share of recreation area in zone	We expect professional developers to have weaker preferences for neighbourhood endowment with recreation area. ($\beta_{Smc} < \beta_{Om} < \beta_{O1}$)

Table 10.3 Hypothesis regarding developer type specific parameter estimates.

Secondly, building on the reference model, we tested for each variable if developer type specific estimation results in significantly different parameters. The significance of parameter difference is tested with a two sample t-test (Cressie & Whitford 1986). The variables which show significant differences between developer specific estimates are introduced as basic deterministic interaction terms in the developer based model.

The models are developed and estimated in UrbanSim. The available template to formulate agent location choice models is used. The agents are in our case development projects which choose to locate on parcels as alternatives. The utility is calculated with a utility function which is linear in parameters. The variables are either directly taken from the database or specified using the domain specific language in UrbanSim (Borning et al. 2008). Each observed location of a project is completed with a random sample of thirty parcels. A parcel is a valid alternative if it has capacity for further development and it is assigned to a residential zone. The models are then estimated with UrbanSim's estimation procedure employing the B-triple-H algorithm for maximum likelihood estimation (Berndt et al. 1974).

We tested various specifications of the model searching for variables which are significant, with expected sign and weakly correlated. A correlation of less than 0.2 is considered weak. The remaining explanatory variables are described in Table 10.4 and respective data sources are given. More detail about the data sources can be found in Chapter 21.

Here we only discuss the variable *Fit of development to parcel constraints*. It interacts the project size with the allowed capacity on the parcel (Eq. 10.1). Utility is drastically reduced if the project is larger than the allowed floor space, i.e. there is no hard capacity constraint assumed. This reflects that in reality negotiations are possible allowing projects to exceed allowed densities. The logarithmic formulation in case of unexploited floor area capacity reflects that there is a decreasing marginal utility for unexploited building capacity.

Variable	Description	Data sources
Accessibility car	Accessibility of jobs by car according to formula (17.9) on page 349. Travel times are calculated with MATSim.	Cadastre, employment census 2000, MATSim road network
Accessibility PT	Accessibility of jobs by public transport according to formula (17.9) in Chapter 17. Travel times are calculated with the cantonal transport model (Vrtic et al. 2005).	Cadastre, employment census 2000, impedance matrix from cantonal travel model
Fit of development to parcel constraints	Step function (Eq. 10.1) of the difference between permitted floor area (FA_p) on the considered parcel and floor area of the development project (FA_d).	Cadastre, zoning plans, DOCUMEDIA, GWR
New neighboring buildings	Number of buildings with year built later than 1995 within 150 m.	Cadastre, GWR
Share of recreation area in zone	Share of land area dedicated to recreational use within the traffic analysis zone of the development.	Cadastre, Zoning plans
Slope	Average slope of parcel.	Cadastre, digital terrain model

Table 10.4 Description of used variables.

The descriptive statistics of the choice sets are given in Table 10.5.

$$\text{fit}(FA_p, FA_d) = \begin{cases} \ln(FA_p - FA_d) & \text{if } FA_p - FA_d > 0 \\ 4 * (FA_p - FA_d) & \text{if } FA_p - FA_d < 0 \end{cases} \quad (10.1)$$

10.5.4 Estimation results

The parameter estimates, t-values and estimation statistics of the reference model are shown in Table 10.6. The only remaining unexpected sign we find in case of SFH and car accessibility. This might be due to the limited number of parameters and resulting endogeneity problems (Guevara & Ben-Akiva 2006). A possible interpretation is that the variable captures preference for remote locations. However, the parameter estimate also remain negative when the variable of recreation area is tested. The adjusted ρ^2 is quite low with 0.091 in case of SFH. For MFH the goodness-of-fit is better with 0.155.

The results of the models with developer type specific parameters are shown in Table 10.7. Since it is not possible to introduce effect coded variables, we can only say if a variable is more or less weighted by a certain developer type.

Variable	*Mean*	*St. dev.*	*Sum*	*Min*	*Max*
Single family housing (ref.)					
Fit of development to parcel constraints	−818.00	2234.90	−14 208 700	−50 127	10.3
New neighbouring buildings	5.08	8.93	88 299	0	107.0
Accessibility car	13.74	0.23	238 707	12.9893	14.4
Slope	5.71	4.39	99 217	0	30.9
Multi family housing (ref.)					
Accessibility PT	12.30	0.66	172 343	9.77058	13.7
Fit of development to parcel constraints	−5144.44	9353.75	−72 073 600	−130 918	10.7
New neighbouring buildings	5.00	8.87	69 996	0	99.0
Share of recreation area in zone	1.11	0.49	15 502	0.668107	14.2
Single family housing					
Fit of development to parcel constraints	−813.25	2250.04	−14 126 100	−50 145.8	10.4
Accessibility car			1		
for O1 developers	6.67	6.87	115 833	0	14.4
for Om developers	3.51	6.00	61 024	0	14.4
for Smc developers	3.56	6.02	61 828	0	14.4
New neighbouring buildings					
for O1	2.52	6.75	43 768	0	86.0
for Om	1.29	4.96	22 357	0	86.0
for Smc	1.33	5.03	23 031	0	106.0
Slope	5.75	4.41	99 898	0	32.3
Multi family housing					
Accessibility PT					
for O1 developers	1.42	3.93	19 884	0	13.8
for Om developers	3.67	5.64	51 357	0	13.7
for Smc developers	7.22	6.08	101 184	0	13.9
Fit of development to parcel constraints	−5130.26	9369.33	−71 875 000	−130 903	10.7
New neighbouring buildings	5.00	8.70	70 102	0	106.0
Share of recreation area					
in zone for O1	0.13	0.39	1779	0	7.0
in zone for Om	0.32	0.56	4541	0	14.2
in zone for Smc	0.64	0.64	9006	0	14.2

Table 10.5 Descriptives of variables in sampled choice sets.

The different estimates of developer specific parameters can be interpreted as follows with respect to single family housing projects:

Accessibility by car Self-owning developers weight accessibility by car most strongly negative. Again, this seems to capture preference for remote, quite places to live. Self-owning developers with multiple projects already care less. Developers selling the project seem to underestimate the importance users attribute to that characteristic.

New neighboring buildings The estimated parameter is insignificant for type Om. There is no significant difference for the significant parameters.

Variable	*Estimate*	*T-value*
Single family housing		
Fit of development to parcel constraints	0.0009	12.06
New neighbouring buildings	0.0329	8.32
Accessibility car	−1.7301	−9.24
Slope	0.0575	6.03
LL(conv.)	−1785.62	
LL(0)	−1969.29	
Adj. ρ^2	0.091	
Observations	579	
Multi family housing		
Accessibility PT	0.3574	3.73
Fit of development to parcel constraints	0.0005	18.58
New neighbouring buildings	0.0180	2.80
Share of recreation are in zone	0.4449	11.24
LL(conv.)	−1337.90	
LL(0)	−1588.36	
Adj. ρ^2	0.155	
Observations	467	

Table 10.6 Estimated parameters of residential project location choice models (reference models).

This confirms the hypothesis that dynamic areas are more attractive for all developer types.

In case of multi family housing projects the developer specific parameters are interpreted as follows:

Accessibility PT Stronger preference for good public transport is found for self-owning developers with multiple projects. The interpretation is that self-owning developers do know about their use of public transport. Consequently, public transport accessibility is valued more. In contrast, selling developers also have to expect car users to buy their estates. The insignificance for O1 developers is probably due to the small sample of 54 observations.

Share of recreation area in zone Self-owning developers with one project value green space most. It seems like more attention is paid to that aspect if just one project is built. This is probably due to the fact that these developers live later on in these building.

The developer specific model shows better goodness-of-fit in case of SFH projects and MFH projects. However, as second estimation showed that the improvement depends on the estimation run and thus probably on the sampling of alternatives. Also, parameter estimates have deviations up to 50%. These findings suggest that the assumptions for an MNL model do not hold and thus random sampling biases parameter estimates. The model should be estimated on the full choice set. We do not find that adding developer information improves the models with the data at hand.

Variable	*Estimate*	*T-value*
Single family housing		
Fit of development to parcel constraints	0.0009	9.61
Accessibility car for O1 developers	−2.3598	−8.76
Accessibility car for Om developers	−1.2414	−2.78
Accessibility car for Smc developers	−1.0018	−2.67
New neighbouring buildings for O1	0.0440	8.86
New neighbouring buildings for Om	−0.0001	−0.01
New neighbouring buildings for Smc	0.0348	3.95
Slope	0.0566	5.98
LL(conv.)	−1769.37	
LL(0)	−1969.29	
Adj. ρ^2	0.097	
Observations	579	
Multi family housing		
Accessibility PT for O1 developers	−0.1500	−0.76
Accessibility PT for Om developers	0.5717	2.85
Accessibility PT for Smc developers	0.3526	2.66
Fit of development to parcel constraints	0.0005	17.63
New neighbouring buildings	0.0203	3.41
Share of recreation are in zone for O1	0.5775	3.69
Share of recreation are in zone for Om	0.4652	7.97
Share of recreation are in zone for Smc	0.4890	8.25
LL(conv.)	−1329.13	
LL(0)	−1588.36	
Adj. ρ^2	0.158	
Observations	467	

Table 10.7 Developer specific parameters of residential project location choice models (developer based models).

Another concern is a potential *panel effect* coming from the data. In this work it is assessed with the "sandwich estimator method" as described by Daly & Hess (2010). We use biogeme (Bierlaire 2003) for estimation because it computes the necessary robust statistics. The analysis shows that the classic and robust estimations differ which suggests that there are issues with the model specification. The geometric means of the robust t-ratios are smaller for both models if the panel information is considered for estimation. We conclude that a panel effect is present. The reduction in the geometric means of the t-ratios is 5.8% and 2.2% for the SFH and MFH model respectively. Compared to the results of Daly & Hess (2010) the effect is small. One reason is that only 9.6% and 8.4% of the observations are of the same respondent for SFH and MFH projects respectively.

10.6 Conclusions

At the beginning of this chapter we hypothesized that characteristics of real estate developers would influence their decisions. The presented work shows

that real estate developers do behave differently by a qualitative analysis of in-depth interviews and a quantitative analysis of real estate developers' location choices in the Canton of Zürich, Switzerland.

The qualitative analysis reveals differences in various aspects. These include decision criteria, type and amount of information considered for decision making and tasks executed "in house". This notion of specialization points to the developers' role as coordinator of construction. The developer is the central actor within a network of specialists. The endowment of real estate developers with own properties, financial resources, and skills influences opportunities and decision behavior. The existence of an own land bank not only provides the real estate developer with collateral for a loan but primarily with the natural resource he needs.

The multinomial location choice models show significant differences between the three investigated developer types for some variables when formulating deterministic basic interactions. The insights in behavioral differences of real estate developers are that *a*) self-owning developers have strongest preference for remote locations and also value recreation areas highest. Viewed from the perspective of a selling developer one can also argue that the importance of these characteristic is underestimated. *b*) Public transport accessibility is higher valued by self-owning developers. These results are potentially biased due to endogeneity.

While the models can reveal the behavioral differences between developer types there is no better fit to the data. We conclude that the considered information is insufficient to assess the heterogeneous preferences among real estate developers with enough detail. This is not only true for the estimation of the models but even more so for the simulation of scenarios. The needed synthesis of an entire population of real estate developers would add additional uncertainty to simulation results. However, the relevance of considering developer heterogeneity depends as well on the scenarios which shall be evaluated.

Goodness-of-fit depends on the estimation run and thus on the sampling of alternatives. This indicates that the multinomial model and the chosen sample size might not be fully appropriate for the given modeling task. A possible reason is that the alternatives (in terms of parcels) available to a developer are limited in the land market.

The negative sign regarding car accessibility is unexpected. This issue is probably due to an omitted variable which biases the parameter estimate (Guevara & Ben-Akiva 2006). One possibility is that effects of competition are not represented.

Further options for model improvement are to consider model formulations which capture correlations of alternatives in space and time. Spatial correlations can be accommodated in "competing destinations formulations"

(Fotheringham & Curtis 1992). While this work is in the context of migration it seems suitable to investigate the activity spaces of real estate developers.

In this study inter temporal decision making is neglected. However, real estate developers consider rather long time horizons compared to other decision makers. Therefore, it seems relevant to estimate models considering dynamic optimization (Train 2009, p. 169), i.e. optimizing choice over multiple time periods.

The quality of the observations in the DOCUMEDIA dataset is found to be limited for the purpose of discrete choice modeling. A lot of interesting information is not included because the data has been collected for another purpose in the first place. The dataset itself comprises little information on real estate developers. Further reasons are incompleteness and missing match with other datasets. The latter issue is important if the observations shall be enriched with information from other sources.

The reconstruction of the decision situation is clearly imperfect. Only one cross-section (2000) has been generated. One can argue that this is less of an issue when observing long term decisions. However, it is still preferable to derive the context of the observations from a more detailed spatio-temporal database. In addition to neglecting the update of context information, all data sources contain errors which influence the results too.

Future research may also apply our findings in simulations of developer type-specific policy scenarios of urban development. Policies of interest can be developer specific zoning regulations (e.g. in Switzerland the share of foreign property owners is limited) or developer type specific subsidies (for example in case of energy policies). Another possibility is to test assumptions about the future structure of the real estate development industry. The explanatory power of the simulated models can be tested with procedures similar to the one proposed by Pinjari & Bhat (2010).

Chapter 11

Modeling the life-cycle of firms and its effect on relocation choice

BR Bodenmann and KW Axhausen

11.1 Introduction

Like people, firms are born, they grow, mature, eventually decline, and may reach the point of death. These stages are the same for all firms, but strategic decisions within these stages are strongly influenced by different circumstances, such as industry and type of products supplied, firm structure, economic environment and, of course, the people involved (owners, managers, employees, suppliers, and customers). For example, small privately owned firms will be perfectly congruent in the location chosen by their owners, whereas subsidiaries of large international companies choose their location based on the holding company's criteria. In fact, the allocation of a firm to a certain life-cycle stage is generally quite uncertain. It depends not only on the firm's internal circumstances, but also on the life-cycle stages of the products and services supplied.

In particular, the life-cycle stage of a firm substantially influences strategic decisions such as relocation choice. Different researchers (e.g. Bodenmann (2011); Hutter (2005); Maoh & Kanaroglou (2004)) show that these decisions are influenced by various company-specific factors: the industry in which a firm operates, the stage of the life-cycle of the goods manufactured the stage of life of the company itself, the firm's organization and structure, and the size of the company. An additional factor is whether the existing location is the base of an existing local and regional network that should not be lost due to a relocation process.

The aim of this chapter is to show the impact of life-cycle status on firm relocation behavior. As the stages in a life-cycle are not directly measurable, researchers generally switch to other indicators that are supposed to represent a certain state in a life-cycle. Of particular interest is the influence of age, size, industry, and location on firm migration behavior. In addition, this chapter will discuss distribution of business formations and closures.

11.2 Conceptual framework and data

11.2.1 Research on firmographics

The demographic processes of firms, in field-related literature also known as firmographics, have increasingly attracted interest, particularly in the context of large-scale micro-analytical models of the interactions between transport and land use, but research on the topic is far from having reached its full extent. Relevant studies include Maoh & Kanaroglou (2007), Maoh & Kanaroglou (2009), de Bok & Bliemer (2005), Moeckel (2005), and Karniouchina et al. (2013).

The conceptual approach to firmographic processes used in the current study is based on Maoh & Kanaroglou (2005) and takes the specific characteristics of Switzerland into consideration. It is assumed that the changes over time (from t to $t+1$) of spatial distribution and of the number of businesses in a region are primarily affected by formations and closures of businesses, and migration of businesses from, to, and within the region (as well as mergers and acquisitions). A useful illustration (Fig. 11.1) of the impact of the above-mentioned factors in urban areas is presented in Maoh & Kanaroglou (2005). Regarding the impact of these factors on the total number of businesses, mergers and acquisitions usually lead to the deletion and subsequent creation of one or more businesses. Newly created companies are simply added to the number of businesses and closures lead to the complete deletion of a company (Federal Statistical Office FSO 2005).

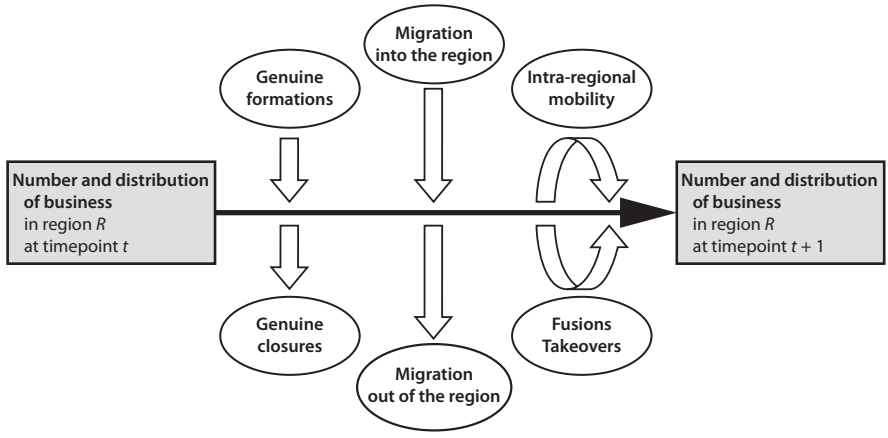


Figure 11.1 Development of firmographics in regions (Source: based on Maoh & Kanaroglou (2005)).

11.2.2 The life-cycle concept

The concept of a life-cycle was initially adopted by the field of business administration for individual products and has been used to explain the development of businesses and industrial sectors (see also Nischalke (2006); Schaetzl (2003); Pfingsten (1998)). In analogy to the biological life-cycle, the product life-cycle is divided to four or five phases. During the *introductory* phase only relatively small amounts of the product are produced. In the *growth* phase, the quantities produced are increasing significantly up until the *maturity* phase when growth rates start to decline again. During the *decline* phase, reduction of the produced quantities continues unless the product is *re-introduced* in a modified form and a new product life-cycle begins.

Although the application of the life-cycle concept to explain the development of economic and social systems is conceptually simple, there are significant differences in comparison to the biological life-cycle. Probably the most important one is that in economic systems the succession of life-cycle phases does not necessarily follow one direction and it can be affected by external factors. Particularly *decline* or death do not necessarily follow *growth* of an economic system, as further development of the product or other relevant actions can lead to re-entering the *growth* phase.

Another important difference with the biological life-cycle is that, in contrast to biological organisms, in socio-economic systems, the life-cycle phases do not follow a certain succession pace. Depending on the product, industry, area or specific measures, a system can remain a shorter or longer time in a life-cycle phase.

Although still a controversial topic, the application of the life-cycle concept to understand and explain the development of economic systems has gained recognition (Nischalke 2006). However, it needs to be noted that the complexity of socio-economic systems does not allow life-cycle models to reliably forecast their development.

11.2.3 Life-cycle of businesses

The life-cycle of a business can be seen as the combination of the life-cycles of the different products produced or sold by the business. As in the case of the product life-cycle concept, in the business life-cycle concept, we assume that the organisational changes in a company depend on its current phase. Lester et al. (2003) provide a number of references on life-cycle models for businesses, as well as on the development and support of four and five-stage models. A business's life-cycle can be divided into the following four or five phases (Nischalke 2006):

1. **Start-up phase**: New companies try to find their position in the first phase as viable organisations and are mainly represented by one person,

i.e. the founder. Decision-making is mainly concerned with satisfying short-term goals.

2. **Growth phase**: The company becomes successful and establishes its position in the market. Typical characteristics in this phase include large growth of sales and adoption of a business strategy that focuses on developing its resources' potential and advantages of scale. Additionally, the number of employees increases.

3. **Maturity phase**: The level of sales and size of the company are stabilised and the organisational structure becomes increasingly bureaucratic, while the company mainly focuses on operational efficiency.

4. **Revitalisation phase**: The business is using specific measures in order to prevent tendencies toward stagnation and decline.

5. **Decline phase**: Unsuccessful innovation and saturated markets increase decline tendencies leading either to prolonged stagnation or closure.

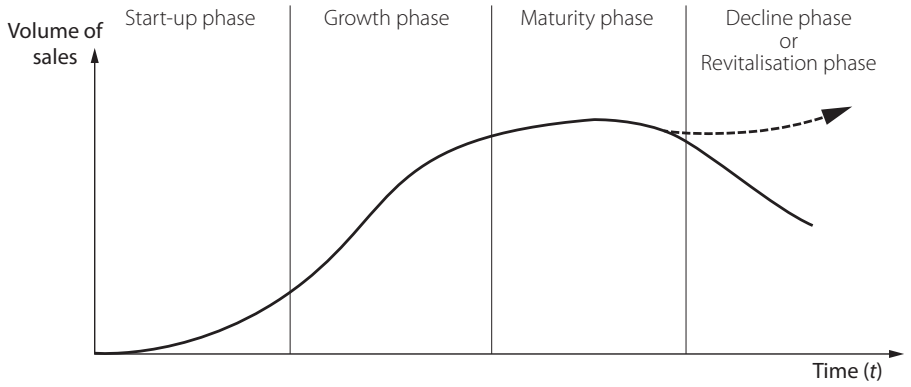


Figure 11.2 Phases of a business life-cycle (Source: based on Schaetzl (2003)).

Crisis management is a key issue for companies in every phase. The most widely known model considering crises in business life-cycles was developed by Greiner (1972). The author suggests that a company's success is almost exclusively determined by its ability to deal with crises. According to Greiner's model, long steady growth phases are followed by short phases where fundamental changes occur and so on. The timing of the different crises depends on the age and size of the company, while the duration of the growth phase varies depending on the industrial sector.

By combining Greiners model with the phases of the life-cycle concept, we conclude that fundamental changes are expected to occur in the *start-up, growth*, or *revitalization* phases (Fig. 11.2). Furthermore, because relocations

can be a measure of crisis management, the following assumptions regarding business relocations are made:

a) Businesses that are in the start-up phase are relocated relatively often and they seem to prefer the central parts of agglomerations (Thompson & Barnbrock 1975).
b) Businesses in the growth phase are under increasing pressure to expand, but their need for more space is difficult to meet in central areas. As a result, it is relatively common to decide to relocate away from agglomeration centres.
c) Among businesses in the maturity phase, only a limited number are expected to relocate.
d) Finally, regardless of their life-cycle phase, older companies are more resistant to crisis as they have more experience. Hence, the probability of survival increases with age.

Maoh & Kanaroglou (2007) investigated the case of Hamilton, Canada with a model that considers key demographic processes such as growth, decline, survival, and failure. They suggested that a firm's desire to move is largely explained by internal factors such as age, size, growth, and industry type, and found that mobility decreases with age while small distance moves are in general more common and most often occur among smaller firms. The impact of growth varies depending on sector. The authors concluded that while these factors do not necessarily translate into spatial policies, the age and growth variables provide an insight into the relationship between the life-cycle of the establishment and its willingness to stay or move.

11.2.4 Life-cycle of industrial sectors

Similar to the business life-cycle, the industry life-cycle is, in principle, the combination of different product life-cycles and as such it is described and interpreted by Porter (1999) and Ernst (1997). Karniouchina et al. (2013) suggest that as businesses go through growth, maturity and decline, business unit effects are more determinative of performance in the growth phase, while industry conditions play a more significant role in latter phases and the importance of industry peaks in the decline phase. Furthermore, over the three phases industry and corporate parent effects explain more variance in performance, while business unit effects decrease (Karniouchina et al. 2013). Various research projects have examined the trend of average company size during the industry life-cycle showing that in order to cover all the life-cycle phases, the study period must cover at least 50 years (Klepper & Graddy (1990); Gort & Klepper (1982)). However, for technological and Internet companies, which had been increasing rapidly in number at the time of the research, this period might be too long. Peltoniemi (2011) examines the theoretical background

and empirical support of the industry life-cycle concept and concludes that the industry life-cycle theory "has spawned well-established theoretical arguments that have been prolifically tested and elaborated upon" and offers a comprehensive review of relevant research studies. The focus of the review of the theory is on the stages of industry emergence and transition to maturity, while the investigation of empirical evidence is organised around entry, exit, survival, innovation, and industry structure.

The model presented was initially developed by Storey & Tether (1996) and has been applied to Switzerland by Meyer (2004). The life-cycle phases are represented by the changes over time of the number of businesses and workers for a particular industrial sector. Figure 3a, the basic model described in Meyer (2004), divides the industry life-cycle into three phases.

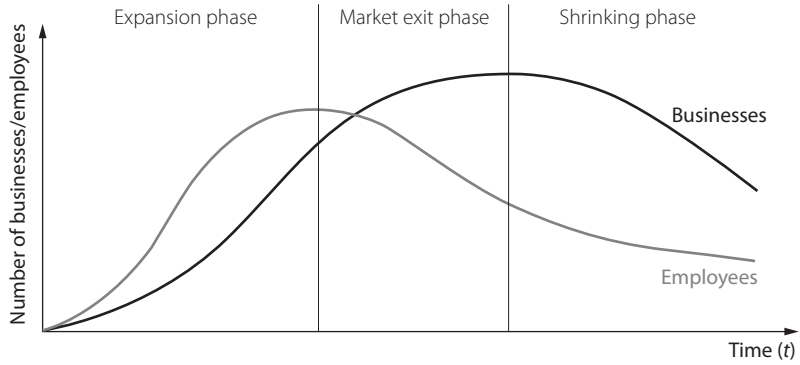
a) In the **expansion phase**, the number of businesses and workers of the specific industrial sector increases. This phase is roughly the analogue of the start-up and growth phases of the product life-cycle.
b) In the **market exit phase**, the number of businesses of the industrial sector decreases and the number of workers grows leading the average size of companies to a sharp increase. Similarities with the maturity phase of the product or business life-cycles can be found.
c) In the **shrinking phase**, the number of businesses decreases, but slower than in the previous phase, and so does the number of workers.

According to the model of Storey & Tether (1996) the number of companies starts to decrease before the number of workers. However, this is not necessarily the case for Switzerland. Meyer (2004) suggests there can be another phase between the expansion and shrinking phases as shown in Figure 3b:

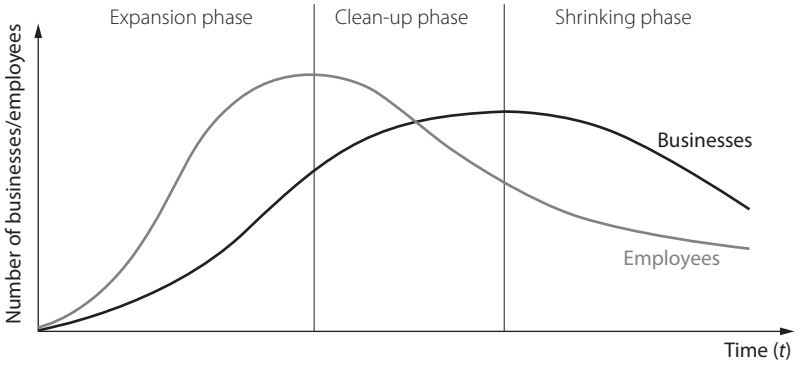
d) During the **clean-up phase** the number of companies continues to grow, while the number of workers decreases leading to a relatively sharp decrease in average business size. The clean-up phase, similar to the market exit phase, has common characteristics with the maturity phase of the product life-cycle.

The different phases of an industry life-cycle can be represented according to the changes in the numbers of businesses and employees on a two-dimensional Cartesian coordinate system (Fig. 11.4):

I) The first quadrant represents the expansion phase. For the industrial sectors under the diagonal line (45 degrees) the number of businesses increases faster than the number of workers.
II) The second quadrant represents the market exit phase.
III) The third quadrant represents the shrinking phase and is divided in two areas by the 45-degree diagonal. Above the line the number of businesses



(a) Base model



(b) Extended model

Figure 11.3 Phases of an industry life-cycle (Source: based on Meyer (2004)).

decreases faster than the number of workers and the opposite below the line.

IV) The fourth quadrant represents the clean-up phase.

In general, for the industrial sectors that lay above the diagonal line, the average size of businesses grows, while it declines for those under the average size.

Considering the relationship of the business life-cycle to location and relocation choices, the following hypotheses can be made:

a) The *expansion phase* is dominated mainly by start-ups, but also includes new arrivals of companies in the observed region.
b) Businesses grow fast in the *market exit phase*, hence an increase in relocations from the centre to surrounding areas is expected.

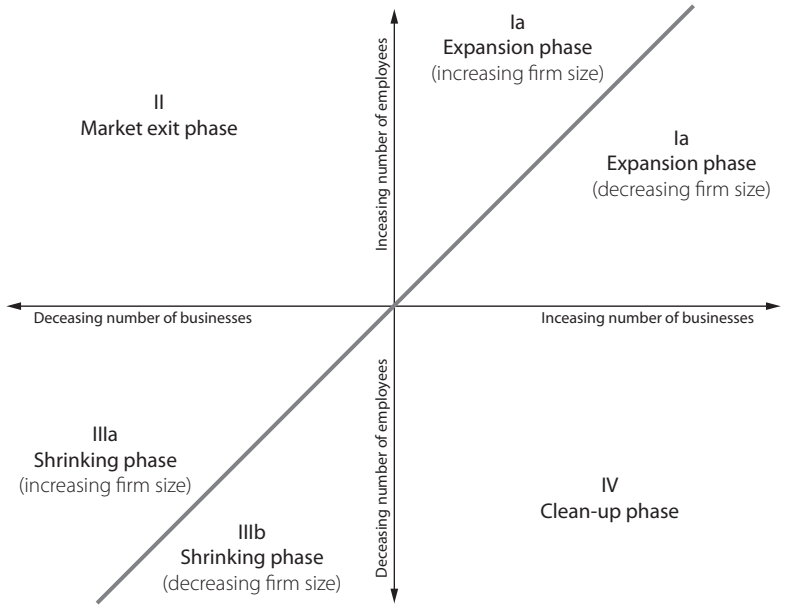


Figure 11.4 Stages of the industry life-cycle in a cartesian coordinate system (Source: based on Meyer (2004)).

c) The *shrinking phase* is dominated by businesses closing down or moving out.
d) In the *clean-up phase*, many small businesses enter the market. They are likely to locate in the centre of agglomeration areas.

11.2.5 Data sources

The results presented in the following chapters partly refer to the whole of Switzerland and partly focus on the cantons of St. Gallen, Appenzell Inner Rhodes, and Appenzell Outer Rhodes. The main reason for focusing on specific cantons has been data availability, or more specifically, a lack of data concerning age and businesses relocations for the whole of Switzerland. Data on business that were used for the Swiss-wide analysis were obtained from Swiss Federal Statistical Office (FSO).

The data regarding companies in the case study area, i.e. the cantons of St. Gallen, Appenzell Inner Rhodes, and Appenzell Outer Rhodes, were exported from the commercial registries of the cantons for the period 1990 to 2005 using software developed for this purpose. The values obtained for the 31st of December of every year and other data obtained include site location at community level, number of registered persons, and founding date. The data allowed us

Characteristics	SG	AR	AI	CH
Number of municipalities (2006)	88	20	6	2,721
Size [km^2]*	1,951	243	173	39,996
Percent of settlement area*	6%	6%	3%	5%
Number of inhabitants (2005)**	459,999	52,561	15,220	7,464,788
Number of employees (2005)***	200.383	17,991	5,589	3,260,796
Largest city / municipality				
Name	St. Gallen	Herisau	Appenzell	Zürich
Number of inhabitants (2005)***	70,316	15,646	5,684	347,517
Number of employees (2005)***	49,669	6,568	3,114	276,091

SG: St. Gallen; AR: Appenzell Ausserrhoden; AI: Appenzell Innerrhoden; CH: Switzerland.
Source: FSO (2009); Databases: *Arealstatistik 92/97, **census of enterprises, ***ESPOP

Table 11.1 Characteristics of the three cantons included for data source.

to examine the development of approximately 51,700 companies over a period of 15 years. The total number of registered companies in the case study area rose from 21,700 at the end of 1990 to more than 30,700 at the end of 2005.

The firmographic events outlined in Figure 11.1 regarding relocation can be simply and effectively captured using the address information of the commercial registry. In contrast the distinction between real start-ups (or closures) and new businesses (or business closures) that are the result of mergers is more difficult. Hence, in many cases the status was determined with the help of notes and comments that appear in the registry. As a result, it was estimated that 9% of all start-ups and 13% of all businesses closures can be attributed to mergers. From the data on mergers and acquisitions, it is not possible to extract safe conclusions about location choice, more specifically, it is not possible to know whether the company remained in the same location or not. However, on average this only affects 0.6% of all companies in the case study area. Data on business size and sector are only available from 2003 to 2005.

The distances between business locations, e.g. between former locations and potential new locations, were calculated at the Institute for Transport planning and Systems (IVT) at ETH Zürich (Froehlich et al. 2004).

11.3 Events in the life-cycle of a company

11.3.1 Business start-ups and close-downs

The balance of the number of business start-ups and closures varies considerably over the different business sectors. Figure 11.5 illustrates the differences between annual average business start-ups and closures for Switzerland from

2001 to 2008. It can be noted that in the sectors of manufacturing, trade and repair of motor vehicles, and education, there are more businesses closing down than opening. Especially for the two former, this difference is particularly high. IT, Telecom appears to be the most dynamic sector with the highest percentage of start-ups and closures; while finance and insurance, real estate, and administration and support services are the fastest growing sectors with significantly more business creations than closures. Health and social services, along with accommodations are the least active sectors, while transport and logistics remained marginally steady in size as there were similar numbers of new businesses and business closures. These results give a first hint on the life-cycle status of the different sectors, as we assume that dynamic sectors are in the expansion phase.

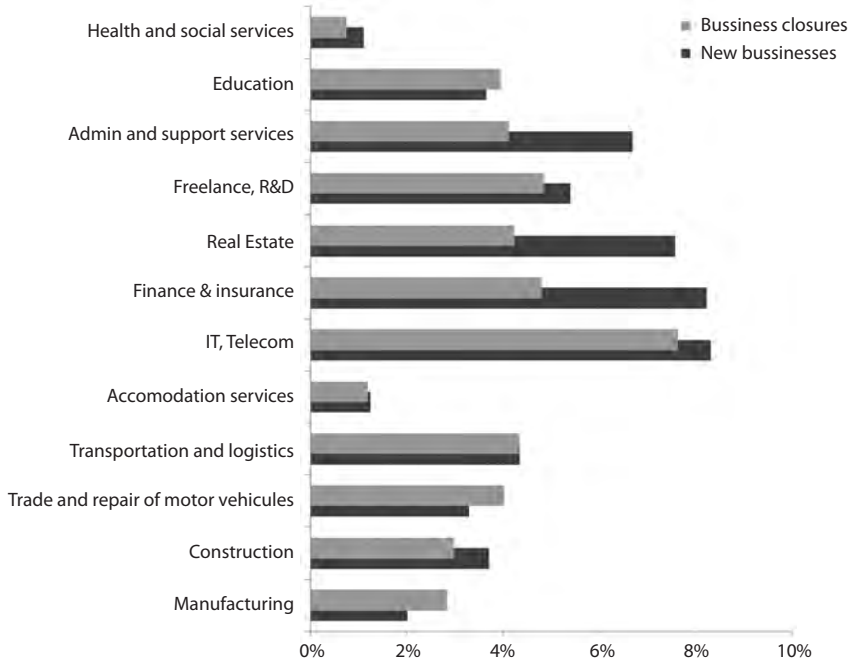


Figure 11.5 Annual average percentage of business start-ups and closures from 2001 to 2008 in Switzerland.

11.3.2 Age

The age of a company is not necessarily an indicator of its life-cycle phase as the duration of the phases vary significantly among different companies and business sectors. Furthermore, age cannot be associated to the revitalization phase. However, age can be an indicator of the company's experience in managing crises related either to the company itself or to the market (Greiner 1972).

Based on the distribution of businesses in Switzerland and the case study area according to size and age, it is clear that start-ups are almost exclusively very small enterprises as they make up 97% of all companies aged between 1 and 10 years old. This is in line with the results of other studies, such as Hutter (2005). The proportion of very small enterprises decreases as the age of the group increases and reaches its lowest value at 66% for the group of companies aged 41–50 years old. Then it rises to 71% for the age group 51–100 years old.

The proportion of small enterprises increases initially and reaches its maximum at 22% for the age group 31–40 and then decreases, while the proportion of medium enterprises increases until it reaches its maximum at 12% for the age group 41–50 and then decreases. The proportion of large companies steadily increases with age reaching its maximum at 3% for the age group 51–100. This small percentage of large enterprises has a large impact on employment as 67% of the employees of companies aged 51–100 years old work in large companies. The percentage of employees in small enterprises reaches a maximum of 40% for the age group 21–30 years old and for medium enterprises, the highest percentage of employees is 37% and occurs in the age group 41–50 years old.

The aforementioned figures seem to support the theory that business size tends to grow with age and this is also indicated when the average ages of the various business size categories are compared: the average age of very small enterprises in the case study area is 13 years old (median: 7 years old), small enterprises 25 years old (median: 22 years old), medium enterprises 38 years old (median: 32 years old) and large enterprises 58 years old (median: 45 years old).

However, what is the impact of age on relocation decisions? Figure 11.6 illustrates the variation in address changes of businesses by age for our case study area. The percentage of businesses relocating within the same municipality starts from 3.3% in the second year, rising to 3.5% in the sixth year and then falling until it reaches 2.3% in the 25th year after which it starts increasing. In the case of relocations beyond the borders of the municipality, young enterprises, approximately 1.9% move annually during the second and third year of their life, much more often than older ones. The relocation rate decreases sharply initially and more smoothly later. Relocations inside and outside the borders of a municipality seem to follow different rules. Although relocations within small distances are little affected by the age of the company, in the case of larger distances, age plays a significant role. The main location decisions are made by businesses during the first years of operation.

11.3.3 Duration of staying in the same location

The duration of the period that a company spends in the same location depends on whether it was founded there or moved there. In our case study area, the probability that a company will still be in the same location where it was founded after one year is 88%, after three years 68%, and after ten years 29%

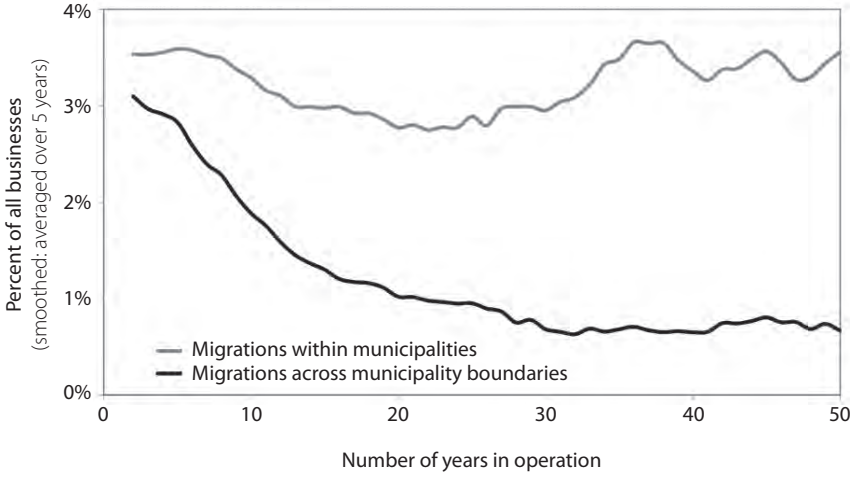


Figure 11.6 Percentage of firms with registered address changes by age of firm (SG/AR/AI, 1991–2005).

(numbers for the region of St. Gallen). For a company that was founded in another location and moved to its current one, the probability of remaining there after one year is 92%, after three years 76% and after 10 years 38%, i.e. significantly larger. In Figure 11.7, both curves become flatter the longer a company stays in one location which indicates that the longer a company stays in the same location, the higher the probability it will remain there in the following year. Both curves almost perfectly follow the form of a negative exponential curve with the coefficient around 0.99.

11.3.4 Migration

In contrast, in the relationship between start-ups and close-downs, the in- and out-migration of businesses is in balance as shown in Figure 11.8, which illustrates the annual in- and out-migration rates of businesses. Between 1991 and 2005, 1,200 businesses moved in and 1,192 moved out. Even considering that 85% of the businesses (1011) move within the case study area, the surplus of incoming companies is very small.

The percentage rates of migrations classified by economic sector and type of migration are presented. In our case study area most relocations occurred in the services and finance sector with 5.8% companies relocating annually between 1991 and 2005. In contrast, in the gastronomy sector, which relies heavily on location, only 2.6% changed their location. As expected, the close-down rates for agriculture and mining (3%) and retail (3.6%) are well below average.

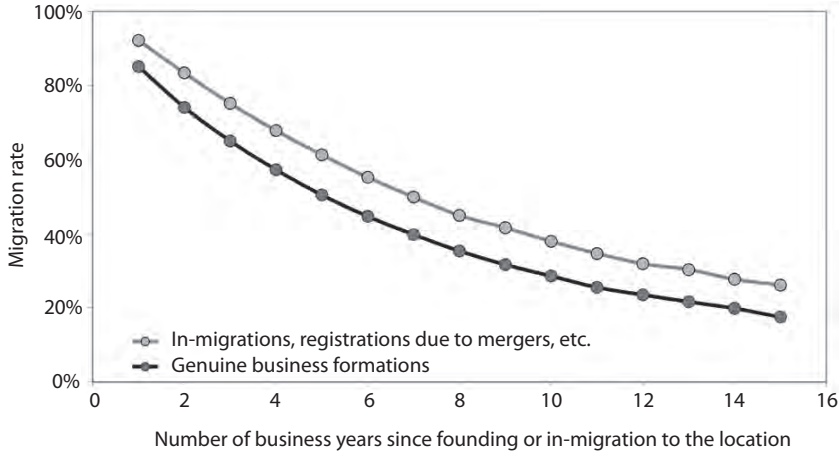


Figure 11.7 Probability of a firm to remain at current location (SG/AR/AI, 1991–2004).

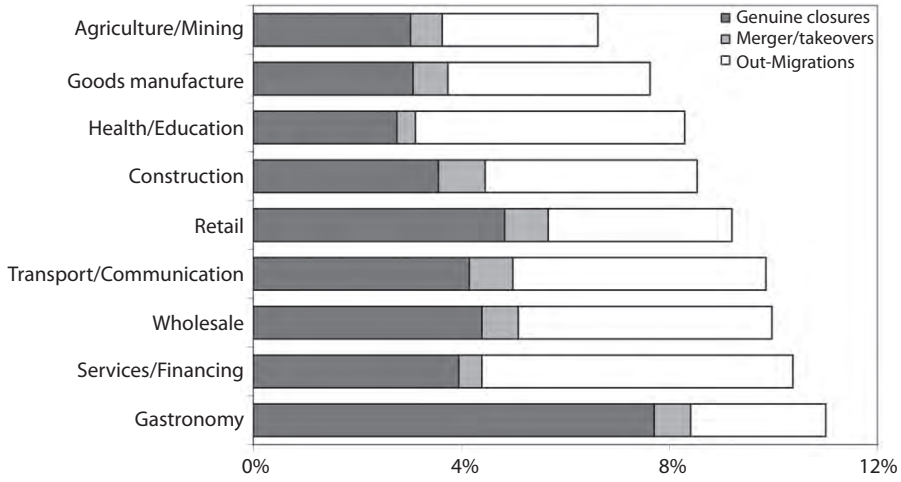


Figure 11.8 Percentage of businesses leaving their location (SG/AR/AI, 1991–2005).

Similar results have been obtained for the Netherlands (Pellenbarg (2005); van Steen (2005)). Van Steen examined 2,000 companies over a period of five years and estimated the annual average relocation rate to be 5.7%, very similar to our case, which is 5.8%. Pellenbarg (2005) examined data from the Chamber of Commerce and estimated the annual relocation rate to be 2% higher. He attributes the difference to the fact that van Steen does not consider new businesses. Relocation rates extracted from the commercial register tend to be low because of two main reasons: 1) Due to the uncertainty resulting from a number of companies that were not actually deleted (0.6% of all companies)

and 2) by not counting companies that moved out of the case study area in the same year they were created (0.25% of the companies). However, there still remains a small difference of 2% to explain. In contrast, the ratio of the relocation rates between individual sectors in Pellenbarg is similar to the results presented earlier.

11.3.5 Influence of the life-cycle of industrial sectors

Following Meyer (2004) and the methodology presented in Section 2.4, the different industrial sectors can be classified into life-cycle stages according to changes in the number of employees and companies. Figure 11.9 is based on the annual average percentage changes of the number of employees and companies in Switzerland between 1995 and 2008 and presents the results for key industrial sectors. The change in the number of employees and companies indicates that the majority of the industrial sectors are in the expansion phase, while mining, manufacturing, water services, and education are in the shrinking phase. Construction, and transportation and logistics are in the clean-up phase and trade and repair of motor vehicles is in the market exit phase. In total, from 1995 to 2008, there were 30,696 more businesses and 578,743 additional employees in the industrial sectors that were in the expansion phase. There were 6,236 fewer businesses and 34,278 more employees in the industrial sectors in the market-exit phase, 4,427 more businesses and 26,449 fewer employees in the industrial sectors in the clean-up phase and 5,810 fewer businesses and 126,640 fewer employees in the industrial sectors in the shrinking phase.

Figure 11.10 illustrates the average annual percentage changes of the number of employees and companies in Switzerland for three different time periods

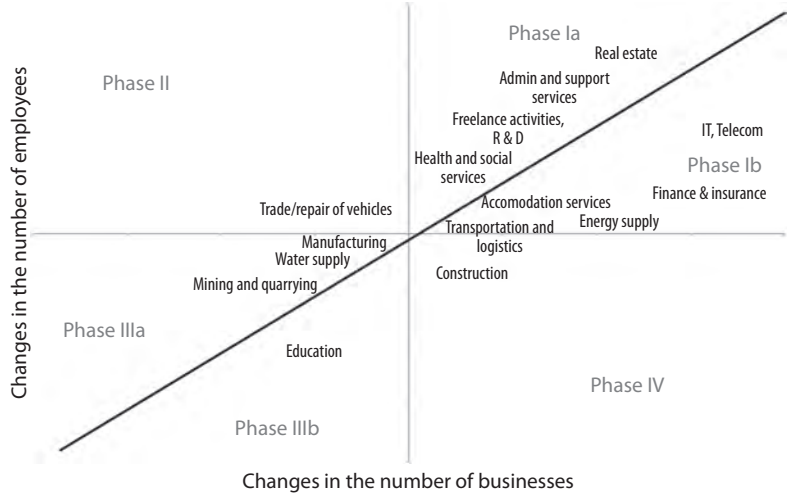


Figure 11.9 Life-cycle phases of industrial sectors in Switzerland 1995 to 2008. The results are based on average annual percentage changes.

(1995–2001, 2001–2005, and 2005–2008). At first, we can see that while the majority of the industrial sectors are in the expansion phase in the period 1995–2001, during the 2001–2005 period, most industries appear in the market exit and shrinking phases and then, in the 2005–2008 period, all but two industrial sectors are in the expansion phase. Only administration and support services and real estate remain in the expansion phase during these three periods and both of them were among the fastest growing sectors during the 2001–2008 period as shown in Figure 11.10. In another change, construction moves from the clean-up phase to the expansion phase during the second period.

The examination period is relatively short to be able to draw conclusions about the life-cycle of industrial sectors, but the results can be indicative of the relationship between industry and the economy. At the beginning of the 2001–2005 period, there was an economic downturn that led to the decline of the GDP and this might be the reason for a general trend towards the market exit and shrinking phases. However, the appearance of most industries in the expansion phase in the 2005–2008 period probably indicates the recovery from the crisis. IT, Telecom is a sector that seems to have been affected by this decline and growth of the economy as shown by its status in the three consecutive periods because it has the most start-ups and closures during the 2001–2008 period as shown in Figure 11.10. In contrast, education shows a constant improvement from period to period, moving from the shrinking phase to the market exit phase and finally to the expansion phase, although during the 2001–2008 period there are more closures than start-ups as shown in Figure 11.10. Interestingly, even manufacturing, which is generally considered a declining sector, appears in the expansion phase during the 2005–2008 period, but, as mentioned earlier, the period is not long enough to allow us to identify trends. The results for finance and insurance services also show interesting behaviour, as the impact during the 2001–2005 period is on the number of employees and not companies. As shown in Figure 11.10, finance and insurance services is among the fastest growing sectors in terms of new businesses, indicating the employment flexibility of the sector, if we assume that companies used redundancies as a measure against the crisis.

Figure 11.11 illustrates the percentage of businesses created and closed-down from 2001 to 2008. As expected, the creations and closures of businesses vary significantly over the life-cycle phases. In the expansion phase, the proportion of both business creations (4.5%) and closures (3.7%) are relatively high. For the industrial sectors in the market exit phase (second phase), the percentages of both start-ups and close-downs drops to 3% and 3.5% respectively. Interestingly, the drop in business closures is relatively smaller than in start-ups. In the shrinking phase (third phase), business creations fall to 2.2% indicating that there are fewer start-ups than closures (3%). Finally, in the clean-up phase, start-ups increase to 3.8%, while slightly more close-downs (3.3%) are observed.

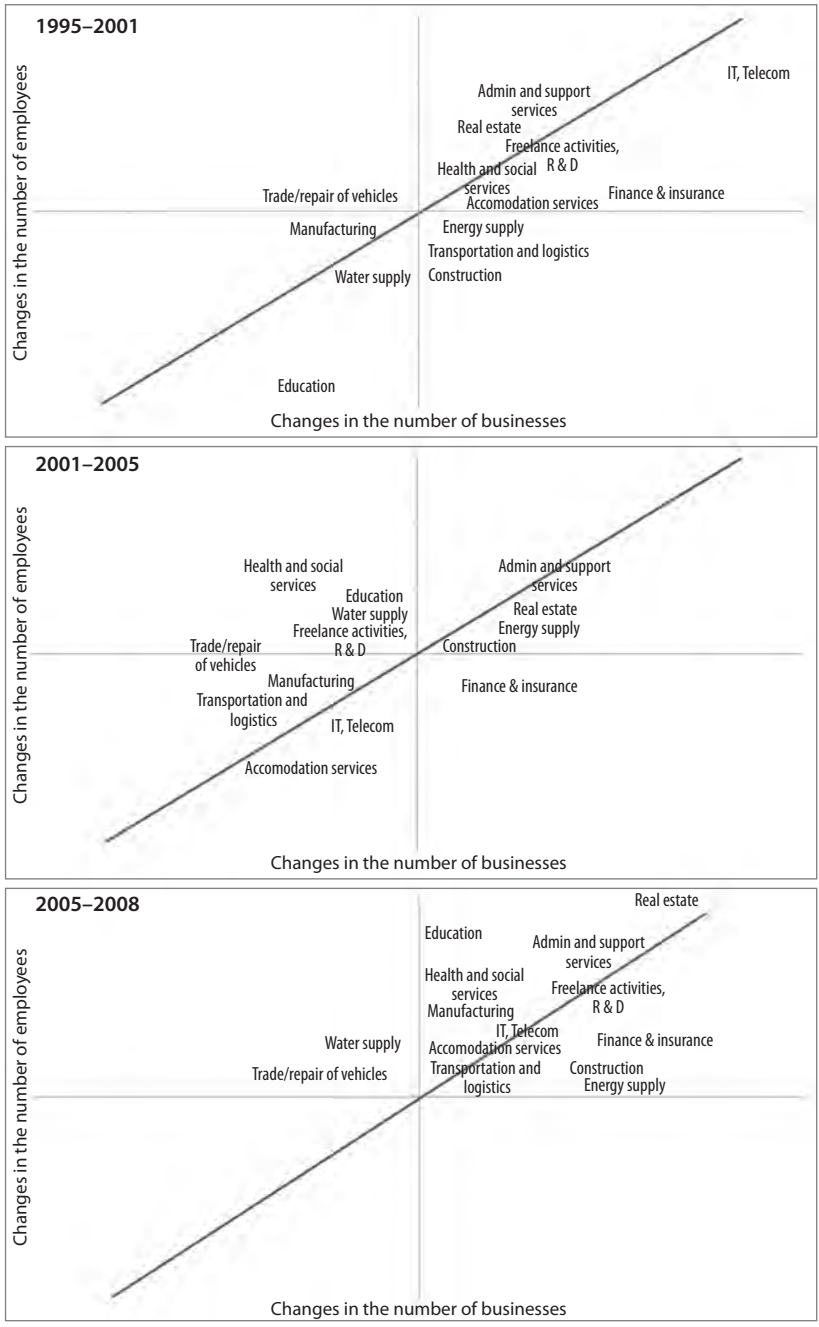


Figure 11.10 Annual average percentage changes of the number of employees and companies in Switzerland for three different time periods.

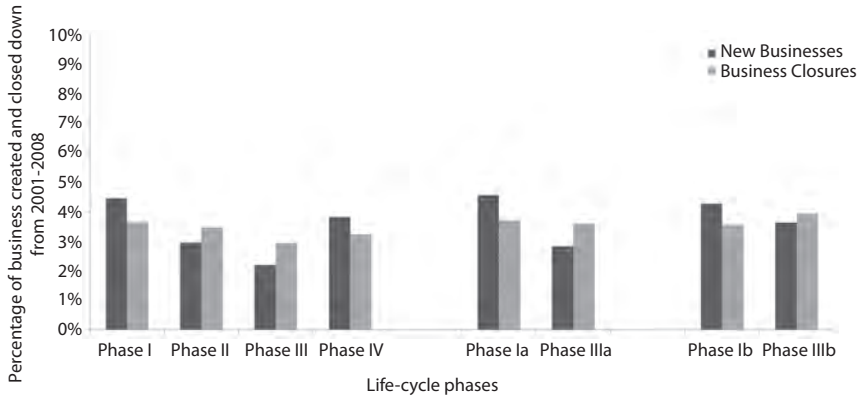


Figure 11.11 Business start-ups and closures in the different industrial sector life-cycle phases.

11.4 Conclusion

Based on the analyses of firmographic data on Switzerland and specifically the Cantons of St. Gallen and the Appenzell Outer Rhodes and Inner Rhodes, different variables related to the life-cycle concept prove to have substantial effects on the behaviour of businesses. Indeed, the implementation of the life-cycle concept also causes problems. In particular, allocation of specific life-cycle phases is quite difficult, if not impossible. Regarding business life-cycles, available data in Switzerland only allows estimating life-cycle phases based on the founding date of the firm and, derived from this information, its age. Understandably, this information is only sufficient to allocate the start-up phase of a business with a limited reliability. In contrast to the level of industries, allocating other phases on the level of individual businesses is impossible. Nevertheless, the life-cycle concept is useful to understand and interpret observed firmographic events.

In line with the life-cycle concept for businesses, young firms in the *start-up phase* focus on internal organisational issues and therefore confine their actions to short-term objectives. As a consequence, this leads to high rates of firm closures, mergers, and relocations. Different analyses in this topic show that new firm establishments take place disproportionately often in urban areas (e.g. Bodenmann (2011)). As expected for businesses in the *growth phase*, new established firms tend to leave the city centre after a few years to settle in suburban areas. In contrast, older firms tend to relocate less frequently. This is in line with the expectations for businesses in the *maturity and decline phases*. Additionally, survival rates of older firms tend to be considerably higher than those of young firms.

In contrast to the business life-cycle, the proposed approach for the life-cycle of industrial sectors enables a quite robust allocation to different phases.

The proposed allocation is based on two indicators: a) the development of employees and b) the development of firms in a certain sector. Interestingly, in Switzerland, firm immigration and emigration on the level of regions tend to be quite balanced (Bodenmann 2011). Therefore, the number of businesses in a certain region is mainly influenced by business foundings and closures. In general, the analyses prove that firms in sectors in the *expansion phase* tend to relocate disproportionately more often. In contrast, businesses in the *shrinking phase* tend to stick to their domicile.

Hence, these results show that life-cycle phases of the industrial sector demonstrably influence the migration behaviour of businesses on two different levels: 1) migration rates are directly influenced by an increasing number of business establishments and this leads to a higher rate of young firms; 2) the high number of growing businesses potentially leads to a shortage of working space and therefore also higher migration rates.

The examples of life-cycles of industrial sectors also show occasionally a very volatile assignment of phases between different time periods. Therefore, we note that a) the life-cycle phases in reality do not follow a constant sequence and b) can change fundamentally within a short time period. This could be a consequence of the fact that life-cycle phases of businesses as well as sectors are decisively influenced by the phases of the services and products provided. The strong impact of (global) economic growth makes allocations even more difficult.

Regarding land use and transport simulation, the life-cycle concept reveals how important it is that firmographic simulations have to be carefully validated and tested in different scenarios. This is due to the fact that comparatively small changes in the economic situation may cause quite large effects on the number of foundings, closures, and migrations of businesses. This has also a large negative impact on the reliability of projections on the future regional economic situation. Today's widely spread estimation of economic forecasts, based on average economic growth, has to be replaced by the use of different scenarios. The analysis used in this project distinctly shows that in this field a so-called average is not useful as an indicator. Scenarios are more coherent, as they provide information about a wider range of possible developments. Indeed, an important task remains: to set up consistent and plausible scenarios.

The results have been used to model firmographic events in the SustainCity Z'urich Use Case. The definition of the modeled sectors is based on these results as well as the relocation probabilities of the firms, which were modeled in UrbanSim directly on the level of individual employees working in a certain sector.

Chapter 12

Econometric methods for land use microsimulation

Constantinos Antoniou, and Nathalie Picard

12.1 Introduction

Typically, urban development models have been based on aggregate principles. UrbanSim (Waddell et al. 2003) is among a new breed of models that use microsimulation in an effort to overcome the limitations of earlier models and provide a more dynamic and detailed paradigm. The advantages and disadvantages of using microsimulation are not within the scope of this chapter, but the main implication is that more data, as well as more detailed ones are required for microsimulation than for aggregate models.

In the context of the SustainCity project (http://www.sustaincity.org), three European cities (Brussels, Paris, and Zürich, described in other chapters of this handbook) have been modelled using the land use microsimulation platform UrbanSim. This platform relies on various models interacting with each other, to predict long–term urban development. The aim of this chapter is to provide some econometric insight into this process.

A common set of notations and assumptions are first defined, and the more common model structures (linear regression, multinomial logit, nested logit, mixed MNL, and latent variable models) are described in a consistent way.

Special treatments and approaches that are required due to the specific nature of the data in this type of applications (i.e. involving very large number of alternatives, and often exhibiting endogeneity, correlation, and (pseudo–)panel data properties) are discussed. For example, importance sampling, spatial econometrics, Geographically Weighted Regression (GWR), and endogeneity issues are covered.

Specific examples of the following models: (i) household location choice model, (ii) jobs location/firmography, (iii) real estate price model, and (iv)

land development model are demonstrated in the context of the case studies in Brussels, Paris, and Zürich. Finally, lessons learned in relation to the econometric models from these case studies are summarized.

12.1.1 Data availability and limitations

The following sections provide an overview of the data available for the three case studies considered within the SustainCity project: Brussels, Paris, and Zürich. Microsimulation models have strong data requirements, making data collection a long and complicated effort. Data collected from various sources need to be processed, matched, and homogenized before they can be used. Besides these practical issues, however, there are further challenges to be dealt with. For example, some of the data collected imply further restrictions (e.g. those related to data protection) or are not public and therefore their use is limited. Finally, there are also privacy issues that can limit the usability of data, as often only aggregated data are available, which is a problem, as such projects rely on microeconomics. Other forms of anonymization might be sought to overcome this limitation. Such restrictions were particularly stringent for the Brussels case study, and had important consequences on data and econometric methods used.

12.1.2 Objectives and policy implications

The policy objectives of the SustainCity project can be summarized in the following three points:

- Define objectives (sustainability and others) of policy makers: what are the components (economic, environmental, social, etc.), what is the horizon (5 years or 50 years), valuation of each component (monetary and/or categorical) as well as the level of aggregation?
- Translate the objectives for policy makers into model outputs: this includes developing output reports for the model and suggesting feedbacks of some elements for the model development (local environmental quality has a clear feedback on housing demand and prices).
- Define alternative sustainability policy packages, translate them into model inputs and discuss expected outcomes.

The policy objectives need to be defined by type, level of aggregation, and level of quantification. The impacts of standard policies have been studied in various projects. For example, much attention has been devoted to the impact of road pricing. Road pricing has a positive aggregate impact, but implementation costs are not trivial, and acceptability is an issue since some agents gain, while others lose. The transfers needed to improve acceptability and preserve equity are well understood, but the land use impacts and the implementation

are still not clear. Parking policies, traffic restrictions, pedestrian areas in city centres, and lanes restricted to bicycles provide other types of "soft" policies, whose short run and long run impacts are likely to be non–negligible.

12.2 Suitable techniques

12.2.1 Notations and assumptions

The majority of the models that have been estimated fall under two general categories: Linear regression models and discrete choice models. The objective of this subsection is to provide a basic set of notations and assumptions, in order to ensure that the model development work can be presented in a consistent manner.

The linear regression model is given by

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \beta_K X_{Ki} + \varepsilon_i \tag{12.1}$$

where i denotes the observation and the error terms ε_i are assumed to be white noise (normally distributed with zero mean and variance σ^2). Linear regression can be applied to quantify the dependent variable Y as a linear combination of a number of explanatory variables (or regressors) X. The regression coefficient β_k denotes the increase of the dependent variable per unit change in the corresponding explanatory variable X_{ki}, ceteris paribus.

In the discrete choice framework, the entity of reference is the individual decision–maker, described by a number of socio–economic characteristics, e.g. age, gender, and income. See de Palma et al. (2013) for a review on joint decisions with multiple decision-makers within the household in the context of transport or location choice models.

These decision-makers choose among a set of available (discrete) alternatives. The identification of the choice set among all available alternatives is one important aspect, which becomes particularly relevant when a huge number of alternatives may be available. A decision–maker n selects one and only one alternative from a choice set $C_n = \{1, 2, \ldots, j, \ldots, J_n\}$ with J_n alternatives.

The specification of a random utility model uses the following utility specification (for a decision-maker n choosing alternative j from a choice set of J_n alternatives)

$$U_{jn} = X_{jn}\boldsymbol{\beta} + \nu_{jn} \tag{12.2}$$

where X_{jn} are observable variables that relate to the alternative j and decision-maker n, $\boldsymbol{\beta}$ is a vector of coefficients of these variables, and ν_{jn} is a zero–mean, random term that is usually i.i.d. (independently and identically distributed) across decision-makers. Several assumptions can be made about the distribution and the variance/covariance structure of the error term. The most common assumptions lead to the multinomial logit model (Gumbel error terms i.i.d.

across alternatives) and probit model (Normal error terms possibly correlated across alternatives for a given decision-maker). We will see in Section 12.2.2 that the i.i.d. assumption of the MNL model can be relaxed in the mixed MNL model.

12.2.2 Model structure

The main types of models that have been considered in the SustainCity project are outlined in this section, starting from the more straightforward and moving to the more advanced.

Linear regression

The linear regression model is an attractive and simple method that is being used extensively. While the linear regression model is simple (to run and interpret), elegant, and efficient, it is subject to the fairly stringent Gauss–Markov assumptions (Washington et al. 2010). If these assumptions hold, it can be shown that the solution obtained by minimizing the sum of squared residuals ('least squares') is BLUE, i.e. Best Linear Unbiased Estimator. In other words, it is unbiased and has the lowest total variance among all unbiased linear estimators.

The basic Gauss–Markov assumptions require linearity (in the parameters; nonlinearity in the variables is acceptable); homoscedasticity; exogeneity of regressors; and uncorrelated disturbances. In addition, performing tests on parameters requires either Normally distributed disturbances or a large enough sample.

It is often the case, especially concerning income or price data, that information is missing on the exact value of the explained variable. Instead, the only available information is that the explained variable lies in some interval. In that case, a maximum likelihood estimator can be used. The likelihood then corresponds to the probability that the explained variable lies in the observed interval. The statistical structure of the model (which is often called *interval regression*) and the assumptions are similar to simple linear regression.

Multinomial Logit (MNL)

The most common discrete choice model is the linear in parameters, utility maximizing, multinomial logit model (MNL), developed by McFadden (1974). One of the most noteworthy aspects of the multinomial logit model is its property known as Independence from Irrelevant Alternatives (or IIA), which is a result of the i.i.d. Gumbel disturbances. The IIA property states that, for a given decision-maker, the ratio of the choice probabilities of any two alternatives is unaffected by other alternatives. This property was first stated by Luce (1959) as the foundation for his probabilistic choice model, and was a catalyst for McFadden's development of the tractable multinomial logit model. There

are some key advantages to IIA, for example the ability to estimate a choice model using a sample of alternatives (McFadden 1978). However, as Debreu (1960) pointed out, IIA also has the distinct disadvantage that the model will perform poorly when there are some alternatives that are very similar to others (for example, the now famous red bus—blue bus problem); this can be a significant concern when dealing with the models in software such as UrbanSim where a large number of rather similar alternatives may be available. Controlling for a large number of regressors helps to reduce the disadvantage of IIA assumption since the similarity of alternatives is a problem only for the residual part.

Nested Logit (NL)

There are many ways to relax the IIA assumption, and many variations of discrete choice models aim at doing just that. Nested Logit (NL), introduced by Ben-Akiva (1973) and derived as a random utility model as a special case of Generalized Extreme Value (GEV) models by McFadden (1978, 1981), partially addresses this issue by explicitly allowing correlation of residual terms within sets of mutually exclusive groups of alternatives. The nested logit is widely used in practice due to its extremely tractable closed form solution.

Multinomial and nested logit are the workhorses of discrete choice modeling, and form the foundation of models in areas such as travel demand modeling and marketing. This is because they are extremely tractable and fairly robust models that are widely described in textbooks (for example, Ben-Akiva & Lerman (1985), Greene (2002), Louviere et al. (2000)) and can be easily estimated by numerous estimation software packages (for example, `biogeme`, Bierlaire (2003)). Nested logit models have been used to estimate extremely complex decision processes, for example, detailed representations of individual activity and travel patterns (Ben-Akiva & Bowman 1998), joint residential location-workplace-employment type choice (Inoa et al. forthcoming), or joint tenure status-dwelling type-residential location (Dantan & Picard 2013).

Beyond MNL and NL, there are many directions for enhancements that are pursued by discrete choice modelers. Two of these categories of models (mixed MNL and latent variable models) are outlined in the next two sections.

Mixed MNL (MMNL)

Mixed logit is a highly flexible model that can approximate any random utility model (McFadden & Train 2000). It obviates the three limitations of standard logit by allowing for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. Unlike probit, it is not restricted to normal distributions. Its derivation is straightforward, and simulation of its choice probabilities is computationally intensive, but simple. Like probit, the mixed logit model has been known for many years but has only become fully applicable since the advent of simulation.

A detailed description of mixed logit is available in Train (2009) and Walker (2001). The specification of a random coefficient mixed logit model uses the following utility specification (for a decision maker n choosing alternative j from a choice set of J_n alternatives) $U_{jn} = X_{jn}\beta_n + \varepsilon_{jn} + \nu_{jn}$ where X_{jn} are observed variables that relate to the alternative j and decision maker n; β_n is a vector of random taste parameters specific to individual n; ε_{jn} is a Gaussian, zero–mean error term, with a standard deviation σ_{jn}; and ν_{jn} is a zero–mean, random term that is i.i.d. extreme value. β_n captures both observed and unobserved taste heterogeneity.

Several other approaches that allow for the explicit modeling of correlation among observations exist and could be applicable to this problem. To name a few: Normal mixing distributions (Abdel-Aty et al. 1997), Generalized Estimation Equation (GEE) models (an extension of generalized linear models) (Abdel-Aty & Abdalla 2004), Heteroscedastic Extreme Value (HEV) model, and the multinomial probit. MMNL has several interesting properties that make it attractive. MMNL is conceptually very close to the MNL, which is arguably the most widely used discrete choice model. Furthermore, the tools to specify and estimate MMNL models have reached a level of maturity that can make them accessible to a wide range of researchers and practitioners. Finally, the MMNL is a fairly flexible model, as the additional error term ε_{jn} may have a normal, uniform, log–normal, or other distribution. The additional term may also capture heteroscedasticity among individuals and allow correlation over alternatives and time. While each of these reasons may be relevant to some other method, the MMNL combines these arguments.

The most widely used model specification for the deterministic part of the utility is the standard linear-in-the-parameters specification, used in the vast majority of such models. The actual choice of variables is determined based on economic theory, data availability, and estimation results of alternative considered models.

Latent class models

The nested Logit model is relevant when the upper level category is observable. This is the case, for example, for dwelling type or tenure type. In some cases, the upper level category is implicit and cannot be observed. This is the case, for example, for budget constraints, which prevent the constrained households to borrow in order to buy their dwelling, and so that they are bounded in the tenant category even though their expected utility may be lower in this category than in the owner category (Dantan & Picard 2013). The modeler cannot know a priori which households are tenant because they chose so, and which households are tenant because they are budget–constrained. The latent class model allows modeling at the upper level of the nest the probability that the household is subject to binding budget constraints.

12.2.3 Data issues in discrete and continuous models

Importance sampling

In a MNL model (or its generalized mixed version) random sampling of alternatives consists in randomly selecting a small number of alternatives for each respondent. Random sampling can prove very useful when the number of alternatives is very large (several hundreds or thousands) for two reasons: to reduce computational burden and to avoid dealing with very small probabilities (with heterogeneous preferences and thousands of alternatives, the probabilities could be of the same order of magnitude as the precision in computations, which is typically 10^{-8} or 10^{-16}).

Ben-Akiva & Lerman (1985) show that efficiency can be improved by using importance sampling, i.e. by letting the probability of inclusion of a given alternative in a given choice set vary. Importance sampling is discussed here in the context of household location, but it could easily be extended to workplace or other choices among a large number of alternatives. In the case of importance sampling, the probability that a zone is included in a choice set is proportional to the "size" of the zone, which may be measured either as the population stock (number of dwellings existing in the zone, number of households living in the zone) or as a flow (number of movers to this zone, number of vacant dwellings in the zone). This is equivalent to uniform sampling of (vacant) dwellings located in that zone. Under the IIA assumption, importance sampling leads to unbiased coefficients.

One specific issue may arise with importance sampling. The same zone should a priori have the opportunity to appear twice or more in the choice set, but some econometric software does prevent such repetition. When the same zone cannot appear twice in a choice set, there is to an under–representation of the largest alternatives. This results in a bias in the coefficients of all variables correlated with zone size. This bias should be corrected. This correction is implemented in UrbanSim and in the SURVEYSELECT procedure under the econometric software SAS.

Note that the under-representations of large alternatives, and the resulting bias, become more and more severe when the number of alternatives in the individual choice sets is increased. As a result, this number should not be increased too much (10 alternatives randomly chosen for each household choice set was a reasonable figure for household location choice in the Paris case study) when the software used for estimating models does not allow for repetitions and does not correct the resulting bias.

Under the IIA assumption with importance sampling of alternatives, when the zones are large enough (e.g. more than 100 households each), aggregate demand can be consistently computed based on the probabilities computed in the individual choice sets. This means that, for computing aggregate demand,

it is not necessary to compute the probability of each of the alternatives for each individual or household, which saves a lot of time when the number of alternatives is large. However, this simplification (working on the alternatives randomly selected in the individual choice set rather than on the full set of alternatives) should not be used for computing segregation effects or, more generally, when focusing on the geographical distribution of population characteristics.

Finally, extending random sampling to nested logit model is not straightforward. In particular, in the nested model, the inclusive value should be computed on the whole set of alternatives rather than only on the alternatives randomly selected in the individual choice set.

(Pseudo–)Panel data

The data that are used in the UrbanSim models come from several time periods. The main example concerns prices. When dealing with such panel data it is often useful to consider the heterogeneity across individuals, often referred to as unobserved heterogeneity or individual effect. In general, pooling data across individuals (or locations in the case of price data) while ignoring heterogeneity (when it is present) will lead to biased and inconsistent estimates of the effects of relevant variables (Hsiao 1986), especially if they are correlated with unobserved individual effect (endogeneity bias). Several approaches have been developed to incorporate these effects in the model formulation.

One such approach is to estimate a constant term for each individual (or each location in the case of price data), which is referred to as a "fixed-effects" approach (Chamberlain 1980). The main advantage of the fixed-effects approach is that it corrects the bias when estimating the effects of explanatory variables correlated to the individual effect. However, the main drawback to this approach is the large number of parameters and, consequently, the large number of observations required per individual or per location. A more tractable approach is to assume that the fixed term varies across individuals according to some probability distribution, which is referred to as a "random-effects" specification (Heckman 1981, Hsiao 1986). The random-effect model suffers from the endogeneity bias, but it is easy to estimate even with very few observations per individual (or per location).

Spatial econometrics

Spatial effects represent some of the main methodological challenges that have to be tackled in price hedonic regression. We may distinguish two kinds of spatial effects: spatial dependence and spatial heterogeneity.

Spatial dependence may be "considered as the existence of a functional relationship between what happens at one point in space and what happens elsewhere" (Anselin 1988). Many recent hedonic price studies suggest that in

a cross-sectional hedonic price analysis, the value of a property in one location may also be affected by the value of other properties located in its neighboring area (Yusuf & Resosudarmo 2009). Two broad causes may lead to spatial dependence. Firstly, there is the by-product of measurements errors for observations in contiguous spatial units. In several cases, data are collected only at aggregate scale. This often implies a poor correspondence between the spatial scope of the phenomenon under scrutiny and the delineation of the spatial units of observations and thus potential measurement errors. Those errors will tend to spill over across the frontiers of spatial entities as one may expect that errors for observations in one spatial unit are likely to be correlated with errors of neighboring geographical entities. A more fundamental cause of spatial dependence is due to varieties of interdependencies across space. Location and distance do matter and formal frameworks proposed by spatial interaction theories, diffusion processes, and spatial hierarchies structure the dependence between phenomena at different locations in space (Anselin 1988). Part of this effect can be captured by explanatory variables.

Spatial heterogeneity is related to the lack of stability over space of the behavioral, or other, relationships under scrutiny. It implies that functional forms and parameters vary with location and are not homogeneous across the dataset. Several factors, such as central place hierarchies, the existence of leading and lagging regions, vintage effects in urban growth, etc., suggest modeling strategies considering the particular characteristics of each location or spatial entity.

It has been amply demonstrated that neglecting spatial considerations in econometric models not only affect the magnitudes of the estimates and their significance, but may also lead to serious errors in the interpretation of standard regression diagnostics such as tests for heteroskedasticity (Kim et al. 2003).

Several contributions have attempted to control for spatial effects in first stage hedonic price estimation. They mostly use two kinds of frameworks: spatial econometrics models or Geographically Weighted Regression (GWR) (Brunsdon et al. 1996). There is no consensus about the variety of solutions proposed in the literature. The best modeling strategy often depends on the specificity of the case study investigated.

Spatial econometrics models capture spatial dependency in econometrics models, avoiding statistical issues such as inconsistent or inefficient parameter estimates. In those models, spatial dependency can be handled in several ways. Indeed, in the spatial econometrics toolbox we distinguish: the Spatial Autoregressive Model (SAR), the Spatial Error Model (SEM), a mix of the SAR and the SEM—the Spatial Mixed Model (SMM)—and the Spatial Durbin Model.

In a SAR model, both the direct and indirect effects of a neighborhood's housing characteristics are captured through a spatial multiplier. This model is particularly appropriate when there is structural spatial interaction in the

market and the modeler is interested in measuring the strength of that relationship. It is also relevant when the modeler is interested in measuring the "true" effect of the explanatory variables, after the spatial autocorrelation has been removed (Kim et al. 2003, p. 29).

A contrario, in a SEM model, spatial autocorrelation is assumed to arise from omitted variables that follow a spatial pattern (Kim et al. 2003). Conversely to the SAR model, the SEM is appropriate when there is no theoretical or apparent spatial interaction and the modeler is interested only in the correction of spatial autocorrelation (Anselin 2001).

The Spatial Durbin Model includes a spatial lag of the dependent variable as well as spatial lags of the explanatory variables. This model is an extension of the SAR that allows the structural characteristics of neighboring houses to influence the price of each house. It also captures how the price of houses in one area depends on the characteristics of neighboring areas (Brasington & Hite 2005).

Besides spatial econometrics models, GWR is a local version of spatial regression that generates parameters disaggregated by the spatial units of analysis. This allows assessment of the spatial heterogeneity in the estimated relationships between the independent and dependent variables.

One field in which spatial econometrics models have been shown to be more effective than the traditional linear regression framework is real-estate price modeling. Efthymiou & Antoniou (2013, 2014) estimated local and global models for Athens and Thessaloniki in Greece, and showed that the local models (spatial and GWR) can successfully address the spatial autocorrelation of the data, since they result in lower Moran's I values (Anselin 1995). Moreover, the lower Akaike Information Criterion (AIC) indicates that their predicted values are closer to the observed ones. Furthermore, spatial econometrics models and GWR have been applied to model land use changes (Wang et al. 2011), optimal location of facilities (Efthymiou et al. 2012), and annual average daily traffic (Zhao & Park 2004).

Most of the contributions using those models assume that the dependent variable, house price or dwelling rent, is continuous. In some cases, the need to handle the following issue might arise: the information about the dependent variable, dwelling rent or price, may be collected through a categorical variable. Each modality of this discrete variable refers to a unique interval of dwelling rent. Therefore, we have to resort on techniques designed to estimate spatially dependent discrete choice models. Lesage & Pace (2009) provide a detailed overview of spatially dependent discrete choice models. From all those models, the ordered spatial probit model is the one that proposes the modeling strategy that is the closest to the one we have to implement. However, there are important differences between the "Spatial Interval Regression" model and the ordered spatial probit model. In the ordered spatial probit model, the cut points separating the intervals of the latent variable are unknown. Therefore,

there is an identification issue and the variance has to be normalized to one so that regression coefficients as well as cut points may be estimated. In this model the vector of boundaries of the dependent variable is known. Hence, regression coefficients as well as the variance may be jointly estimated. A similar analysis has already been undertaken by Goffette-Nagot et al. (2011), who explore the spatial variation of land prices in Belgium. While they also account for spatial autocorrelation, their analysis differs since they consider land prices rather than rents as their dependent variable. Moreover, land price information is collected at the level of the municipality rather than at an individual level.

Endogeneity of variables and selection bias

Endogeneity is a serious problem commonly faced in LUTI models interested in interactions between modules. A typical example is given by the prices in the household location choice model, which is correlated with the error term. This problem is caused either by the simultaneous determination of the supply and the demand for dwelling units, or by omitted attributes that are correlated with price. Indeed, empirical residential location choice models have often reported estimated coefficients of dwelling-unit price that are small, statistically insignificant, or even positive. This would imply that households are insensitive to changes in dwelling unit prices, which is not only counter-intuitive, but also makes the models useless for policy analysis. See de Palma, Motamedi, Picard & Waddell (2005*b*, 2007) or Guevara & Ben-Akiva (2006) for examples and discussions.

When endogeneity results from omitted attributes, the best solution is to include enough explanatory variables in the model of interest. Instrumental variables techniques can be used to correct for endogeneity, provided that at least one instrument is available for each endogenous variable. It often proves to be difficult to find such instruments. In the case of household location, if it can be reasonably assumed that dwellings and offices compete for land, then variables measuring local business tax can be used to instrument dwelling prices. In their application on the Paris case study, de Palma, Motamedi, Picard & Waddell (2005*b*) used such instruments and found that endogeneity bias becomes negligible when the household location choice model is rich enough (i.e. when enough explanatory variables are included). Note that a rich enough model can be estimated precisely enough only when the sample size is large enough. In the Paris case study, with 600,000 households, we could precisely estimate more than 50 coefficients reflecting heterogeneity in the preference for accessibility and local amenities, as well as the tendency to locate close to similar households.

12.2.4 Diagnostics

Model diagnostics are a key tool in developing appropriate models. In general there are two families of diagnostics: statistical and graphical. In order

to ensure that the output of the various case studies within SustainCity are consistent and comparable, it has been suggested to ensure that the same diagnostics are provided. Each table of results should contain, for each explanatory variable, the following four pieces of information: estimated coefficient, standard error, t–statistic, and p–value. For summary tables comparing multiple models, it is sufficient to present the estimated coefficient value and t–statistic. In terms of summary statistics, regression results should report corrected R^2 for linear regression. For MNL/NL/MMNL/latent models that are estimated using maximum likelihood, the null log likelihood and the final log likelihood should be reported, along with the AIC. Degrees of freedom should also be reported. Likelihood ratio test values should be performed to determine whether model restrictions should be retained or whether the more general models should be used. As well as to reporting corrected R^2 for linear regression, it is recommended that corrected likelihood ratio test values be reported. For example, McFadden's LRI (McFadden 1974) adjusts for unbalanced alternatives in its definition of the reference likelihood (the reference is a constant-only model), and Estrella (1998) corrects for sample size and the number of estimated parameters.

The econometriceconometric models described in this chapter have some explicit underlying assumptions that need to be satisfied by the data, in order to be valid. A number of violations may often occur, however, resulting in residuals that are not i.i.d. In order to ensure that these assumptions are not violated (or, to be able to resolve them, or at least consider their implications), it is important to perform a series of tests, e.g. for normality, autocorrelation, endogeneity, and heteroscedasticity.

One of the most effective ways to determine and visualize violations, such as autocorrelation and heteroscedasticity, is through the use of (partial) autocorrelation functions (sometimes called "correllograms") and residual plots (e.g., Box & Jenkins (1976)). Residual plots over time or against the magnitude of the dependent variable can help identify heteroscedasticity. QQ normal scores plots can be used to identify deviations from the normality assumption. These visual tests should also be accompanied and further supported by formal statistical tests. The Shapiro–Wilk test can be used to test the normality assumption. The computation of the skewness and kurtosis also provide additional information.

The Box–Ljung test should be used to check for autocorrelation for various lags. A different way to test this type of lack-of-fit of a model is to consider the first few autocorrelations as a whole, using a so-called "portmanteau" test. It should be noted that the number of autocorrelations to use depends on the data and while a lag of 4 or 5 might be sufficient, using a lower lag might not illustrate the dependency. Larger lags do not add to the inference, but are also rather harmless in this context. A popular test for checking the heteroscedas-

ticity assumption is the Breusch–Pagan test (Breusch & Pagan 1979), while a test for heteroscedasticity for non–linear models is described in Gourieroux et al. (1987). A usual way to test for endogeneity is to use a Hausman test.

12.3 Models to be estimated

Table 12.1 outlines the types of models that are estimated for each model type and case study. These models are explained in the next subsections.

12.3.1 Household Location Choice Model (HLCM)

HLCM has been modeled using MNL with importance sampling. Extensions such as NL, MMNL, or latent variables were also estimated, but are only discussed here because they cannot be implemented yet in the current version of UrbanSim. In the case of NL, stratified sampling is an option to be discussed. Household location choice model could be estimated on the whole sample, irrespectively of tenure type and dwelling type. However, when possible, we recommend that tenure type and dwelling type are considered separately, with coefficients specific to each tenure type and dwelling type, and that the decision to move (relocation choice) is estimated together with location choice. In

Model	Paris	Brussels	Zürich
Household location	Nested: relocation/ dwelling type/ tenure status/ location	Multinomial Logit structure. Besides this, nested structures of choice is tested in order to account for correlation of attributes across alternatives	MNL with explaining variables of domains: life style, dwelling type, location (Household relocation: Probabilities for relocation of HH according to income and age)
Job location	Matching workplace/ business	Multinomial logit; sampling of alternatives	Hierachical NL of firm location choice (Bodenmann & Axhausen 2010)
Real estate price	Simultaneous equation (5 types), spatial correlation, Dwelling level	Hedonic model; estimated using "interval regression". A spatial autoregressive model is considered	Spatial error model (Löchl & Axhausen 2010)
Land development	Matching project location/land use transition	2-step model: Supply by building type per zone: linear regression/Choice of zone: Multinomial logit	NL with explaining variables of domains: project, developer and development constraints

Table 12.1 Models by case study.

this case, we recommend the following nested structure: 1) decision to move; 2) tenure choice; 3) dwelling type; and 4) location.

An extension to latent class was successfully estimated for the Paris case study, but it will probably not be included in UrbanSim any time in the near future. In this experimental latent class model, two cases are considered for step 2 (tenure choice): under credit constraint, the only option available to the household is to rent; other wises unconstrained households are free to choose either renting or buying of a dwelling. The probability of credit constraint is estimated simultaneously with the other parts of the model, as an upper level conditional on moving. Additional extensions are scheduled at the bottom level, for dealing with geographical nests and scalability.

Endogeneity of prices is a serious problem in HLCM. It can be solved by instrumenting dwelling prices. Instruments are not obvious in this context, and the choice of instruments is guided by assumptions concerning the real estate markets. In case dwellings and offices are competing for land, instruments can be found in the list of variables influencing the demand for offices. In the Paris case study, variables related to local business tax (French *Taxe professionnelle*) appeared to be valid instruments (see de Palma 2005).

12.3.2 Job location/firmography

A distinction is operated between firms and plants. The way plants can be related to firms depends mainly on data availability. When the identifier (Id) of the plant is not maintained because of the move, this induces fake deaths and births since the available data does not distinguish between a move and a death and birth when the plant Id changes. All models estimated are sector-specific, since the dynamics of the job market significantly varies across activity sectors. Given the stability of activity sector either from the plant point-of-view or even from the worker point-of-view, no model is estimated for transition between sectors.

Three options may be used to study employment location: job location, either by itself or together with household location, and firmography. Each of these models uses Multinomial Logit (MNL) or Nested Logit (NL). In the simplest option, each job is located independently from the other jobs in the same firm or plant and from Household location, using a Multinomial Logit (MNL) model. This simplest option should be considered as a second best, and less relevant than the other ones.

The second option, relevant from the point-of-view of the worker, builds a more elaborate job location choice model. It is a Nested Logit (NL) for workplace and household location, in either order. In such a model, commuting time is a key variable explaining the location at the lower level of the nest, which happens to be by far more significant than any variable measuring either accessibility or expected time typically used in location choice models.

In the third option, firmography, relevant from the point-of-view of the firm, all workers working in the same plant are located simultaneously, at the same place. In addition to the location of new plants, firmography estimates the "death" of the plants using a binary logit model, as well as growing/shrinking of stable plants, using a Linear Regression model.

Note that the "birth" of plants, which is implemented in UrbanSim is not estimated. In the simulation process, newly born plants are randomly selected from the distribution of existing plants.

12.3.3 Real Estate Price Model

Real Estate Price Models correspond to Hedonic price models, including simultaneous regressions, which are relevant in the case of imperfect real estate markets. Two distinctions should be operated in the model: a first distinction between renting and buying, and a second distinction between houses and flats. The relevance of these distinctions depends on: (i) correlations between various prices in the same location, (ii) market shares of each category, and (iii) turnover in each category.

The Real Estate Price Model (REPM) uses the following methods: Simple linear regression, interval regression, and Spatial correlation (SAR, GWR). These methods can be applied either on individual prices or on average local prices. In UrbanSim, the parcel version uses data on individual buildings. In this case, the determinants of real estate prices include both individual and local attributes. Linear regression models are estimated on the log of total selling price, including surface in the regressors list. Aggregate data are on average prices per m^2. In this case, only the local attributes will be included in the regressors list. Potential determinants of the prices are: (i) individual attributes with building characteristics (surface, age, view), and (ii) local attributes with accessibility (to jobs, households, activities), neighborhood (households, jobs, land use), and distance to stations.

12.3.4 Land Development Model

This is the less advanced model in the 3 case studies (and in UrbanSim). UrbanSim currently proposes two options, which are substitutes for the moment. It is desirable that UrbanSim can evolve so that these two options are complements, and describe respectively the supply and demand for land, in relation to the politicians or stakeholder versus investor points-of-view:

Stakeholder point-of-view: Land Use Type transition model, in which there is choice between the different land use types, and transitions between land use types for a given parcel. The model is formulated as a MNL with a relatively limited choice set (e.g. there are 83 land use types in Paris region, which can

be grouped in 9 homogeneous aggregated types). Trade-off between competing land uses can be considered.

Investor point-of-view: captures the choice between the different locations for a given project. It is formulated as a MNL with importance sampling (trade-off between competing locations for a given project). The list of potential alternatives depends on the land use type attached to the project and on the surface of the project: the project can be located only in parcels (e.g. communes or IRIS) for which the surface available for this land use type is larger than the surface of the project.

A common initial model is used for the project definition. A project is defined as a parcel which changed detailed land use type. It is characterized by location (parcel ID, geocoded), size, former land use type, and new land use type. This is the upper level of a NL model for either option (i.e. either point-of-view).

12.4 Conclusions and recommendations

This section summarizes some of the lessons that have been obtained from the case studies within the SustainCity project and could be useful to other similar applications.

12.4.1 Lessons from case studies

Depending on data availability, find the best econometric strategy for each model

The need for detailed and reliable data has been motivated in this chapter, along with the difficulties in obtaining and using such data. However, data availability, quality and restrictions are in general location- and application-specific. Therefore, it is not practical to try to develop general econometric strategies that would be applicable (let alone effective) in all settings. Instead, the modelers should be able to find the "best" econometric strategy for each model, based on the data availability. When detailed data of good quality are not available, a more parsimonious or aggregate model might provide more reliable information than a more detailed and elaborate model (that may not be adequately supported by the data).

Compare estimation results obtained with an econometric software and with UrbanSim until you get exactly the same results

UrbanSim provides facilities for the estimation of econometric models, the results of which can be then be used for modeling purposes. It is also possible to estimate models outside of UrbanSim and then use the estimated coefficients within UrbanSim for simulation purposes. Consistency between the model

estimation and simulation is of paramount importance for the credibility of the results. As there are multiple secondary reasons that might obfuscate the model estimation process, it is recommended that UrbanSim model estimation results are compared against standard econometric software (that the modeler is familiar with) to make sure that the data and underlying assumptions made by UrbanSim are indeed understood correctly.

One practical way to check that the results obtained by the two models are consistent is to use a systematic test "à la Hausman" (Hausman 1978).

Endogeneity issue and order for running models

UrbanSim involves the running of a sequence of models in cycle. This type of models is known to suffer from endogeneity, which can have significant implications in the model results. In order to minimize the impact of this problem, it is important to ensure that the model coefficient estimates and the order, in which the simulations are run, are consistent.

12.4.2 "Standardized views"

The objective of this subsection is to provide some practical guidance towards the development of uniform and "standardized" views. This is particularly important in order to be able to develop some composite insight from the output of the models developed for the various cities (but also among different model forms for the same type of model within applications). The identified suggestions reflect already identified items and it is expected that further similar suggestions will need to be made to ensure uniformity.

Vocabulary and units

When dealing with buildings/apartments, specific distinctions should be made between floor space and land area (cover), in order to avoid confusion. The unit that should be used is (square) meters. Floor space indicates the total area of the property (in m^2); for example if an apartment is spread over two floors and has $100m^2$ per floor, then the total floor space should be reported as 200 m^2. [The fact that this is a two-floor apartment, which could be considered an advantage e.g. for the real-estate price model, could be e.g. captured by an additional explanatory parameter in the hedonic regression model].

On the other hand, land area (cover) reflects the physical space that a property occupies on the land. For example, a 6-story commercial building with 200 m^2/floor would have floor space of 1200 m^2 and land area of $200m^2$.

The unit for monetary measures (e.g. income/cost/rent) should be Euro in all cases, in order to provide uniformity and more direct comparisons. When the original data is in another currency (e.g. Swiss Franc, CHF), then they should be converted to Euro.

When values are considered in logarithm, the natural logarithm should be used, and denoted by ln, in order to avoid confusion.

Results presentation

As it has been presented earlier in this chapter, UrbanSim applications for different cities within SustainCity involve different types of models, at different granularities. Therefore, it is expected that the results may vary substantially. However, there are some measures that can be taken towards providing coherent results that can be used to perform some meta-analysis. One such aspect that relates to the price levels is the recommendation to use a log–transform (both for the model estimations and presentation of results).

Another price-related aspect that can have a large impact on the reported summary statistics is the way that property prices are used. In particular, using total property prices leads to summary statistics (e.g. R^2) with much higher values (than if prices per m^2 are used). Therefore, the results will be more easily supported if they are accompanied by these statistics and it is recommended that total property prices are used in the model estimations.

Acknowledgements

The authors would like to acknowledge the contribution of the following persons (alphabetical order) in the work that led to this chapter: B. R. Bodenmann, L. Chauveau, A. de Palma, D. Efthymiou, P. Fastre, S. Gayda, R. Hurtubia, J. Jones, K. Motamedi, K. Müller, H. Ouaras, D. Peeters, A. Pholo–Bala, P. Schirmer, I. Thomas, and C. Zöllig.

Chapter 13

Spatial issues in econometric analyses: a hedonic estimation of rents in Brussels

Alain Pholo Bala, Dominique Peeters, and Isabelle Thomas

13.1 Introduction

The increasing concerns about sustainable development and the growth of urban areas have facilitated a renewed enthusiasm for the use of quantitative models in the field of transportation and spatial planning. The SustainCity project proposes to advance the state-of-the-art of urban simulation models, and their underlying methodologies, and to improve their diffusion among planners and decision-makers in government.

To do so the partners of the project propose to extend OPUS/UrbanSim platform (Open Platform for Urban Simulation). Some spatial issues may arise from the implementation of that platform. One of them is that in OPUS/UrbanSim, a massive amount of geographic data is collected from various sources, and often at different spatial scales. Another issue is that the definition of agglomerations or more broadly the delineation of the study area may differ in the different case studies that are used to implement OPUS/UrbanSim. All those problems are likely to influence and bias spatial econometric analyses. Moreover, spatial autocorrelation is also likely to have significant impacts on statistical findings.

In this chapter, we plan to check the magnitude of those "spatial" biases and to propose some suggestions to control or at least limit those biases. To

This study was largely conducted while Alain Pholo Bala was holding a Post-Doc position at CORE, Université Catholique de Louvain from February 2010 to January 2011, funded by the SustainCity Project.

do so we will base our econometric investigation on the first-stage hedonic regression model, which is well represented in the OPUS/UrbanSim platform as the Real Estate Price Model.

From a conceptual point of view, the problem of spatial autocorrelation and the issues of the choice of spatial scales and of the study area boils down to spatial dependence and spatial heterogeneity problems. Spatial dependence is one of the main methodological problems that has to be tackled in first-stage hedonic regression. In general terms, it may be "considered to be the existence of a functional relationship between what happens at one point in space and what happens elsewhere" (Anselin 1988, p. 11).

Two broad causes may lead to spatial dependence: the nuisance and the substantive spatial dependence (Magrini 2004). The nuisance spatial dependence refers to the byproduct of measurement errors for observations in contiguous spatial units. In several cases data are collected only at an aggregate scale. Because it implies a poor correspondence between the spatial scope of the phenomenon under scrutiny and the delineation of the spatial units of observations, it may entail measurement errors. Those errors will tend to spill over across the frontiers of spatial entities as one may expect that errors for observations in one spatial unit are likely to be correlated with errors of neighboring geographical entities (Anselin 1988).

Such measurement errors may be caused by problems of spatial aggregation or by arbitrary delineation of spatial units of observations. The aggregation of spatial data is not benign regarding statistical inference. The question of the sensitivity of statistical results to the choice of a particular zoning system is well known as the Modifiable Areal Unit Problem (henceforth, MAUP). Several contributions have assessed the impact of the MAUP on multivariate statistics (Gehlke & Biehl 1934, Fotheringham & Wong 1991, Amrhein 1995, Briant et al. 2010).

A related but less investigated issue is the one arising from the choice of the delineation of the study area. This issue points more to spatial heterogeneity, i.e. the lack of uniformity of the effects of space. Any structural instability of a given relationship across space would entail different econometric results for distinct study areas. Our contribution focuses on Brussels. For this specific city several delineations may be considered: administrative delineations, morphological delineations (Donnay & Lambinon 1997, Tannier et al. 2011, Van Hecke et al. 2009), and functional delineations (Cheshire 2010, Van Hecke et al. 2009, Vandermotten et al. 1999)[18] Considering administrative definitions may be harmful since administrative borders do not capture the essence of economic phenomena and transportation issues, which often spill over boundaries. In this chapter, we analyze nuisance spatial dependence and spatial

[18]The issues related to the delineation of the study area are more exhaustively described in Chapter 4 of this book.

heterogeneity by investigating the impacts of choices of the aggregation scale and of the delineation of the study area.

The substantive spatial dependence is more fundamental and is due to varieties of interdependencies across space. Location and distance do matter and formal frameworks proposed by spatial interaction theories, diffusion processes, and spatial hierarchies structure the dependence between phenomena at different locations in space (Anselin 1988).

It has been amply demonstrated that the neglect of spatial autocorrelation in econometric models not only affects the magnitudes of the estimates and their significance, but may also lead to serious errors in the interpretation of standard regression diagnostics such as tests for heteroskedasticity (Kim et al. 2003).

In this chapter, we assess substantive spatial dependence by considering two components of the spatial econometrics toolbox: the Spatial AutoRegressive Model (SAR) and the General Spatial Model (SAC). Several contributions investigate the spatial dependence issue in cross-sectional hedonic price analyses through the estimation of Spatial Models (Gawande & Jenkins-Smith 2001, Kim et al. 2003, Brasington & Hite 2005, Löchl & Axhausen 2010).

In most of these contributions, the dependent variable (house price or dwelling rent) is continuous. In this chapter, we have to face an extra problem: the information about the dependent variable (here: dwelling rent) is collected through a categorical variable. Each modality of this discrete variable refers to a unique interval of dwelling rents.

There are two ways to handle that issue. The first approach consists on using a Gibbs Sampling algorithm to design "Spatial Interval Regression" models. The second approach implies the use of a two-step estimation procedure with locational fixed effects. The rest of this chapter is organized as follows. The next section describes the study area and Section 13.3 presents the data used for estimation. Section 13.4 is devoted to a detailed presentation of the estimation strategy. Then, Section 13.5 presents the results of estimations and Section 13.6 concludes the chapter.

13.2 Delineation of the study area and basic spatial unit

13.2.1 Delineation of the study area

We restrict the focus of our analysis to the private renting market of Brussels. Several delineations of the capital of Belgium have been proposed based on different criteria: administrative, morphological (Donnay & Lambinon 1997, Tannier et al. 2011, Van Hecke et al. 2009), and functional (Cheshire 2010, Van Hecke et al. 2009, Vandermotten et al. 1999). Table 13.1 and Figure 13.1

Spatial entity	Description	Nmun	Nobs
Brussels Capital Region (BCR)	Administrative definition	19	177,721
Agglomeration	Morphological agglomeration adjusted to municipalities boundaries	36	208,371
Urban region	Agglomeration+Suburb	62	233,582
Residential urban complex (RUC)	Urban region+Residential Commuting area	122	283,079
Extended Residential urban complex (ERUC)	Brussels RUC+Leuven RUC	134	301,160
RER area	Area served by the RER	135	314,279
Extended RER area	RER area + 12 municipalities STRATEC study area	147	325,135
"Union"	Municipalities pertaining to at least one delineation	157	330,147

Nmun= number of municipalities, Nobs= number of observations

Table 13.1 Number of observations (dwellings) for eight different delineations of Brussels. Source: Van Hecke et al. (2009) and STRATEC.

present macrozones that are consistent with eight delineations of Brussels. One of those delineations, the Brussels Capital Region, corresponds to a purely administrative definition of Brussels. Another delineation, Brussels (operational) agglomeration corresponds to a morphological definition of Brussels.

The other delineations represent merely functional definitions of Brussels. From the Van Hecke et al. (2009) nomenclature, we may consider the Brussels urban region on one hand and the Brussels residential urban complex on the other hand. We may also consider Brussels "extended residential commuting complex" which is the union of the Brussels and Leuven residential commuting complexes. Stratec, an independent consultancy company, has proposed other definitions: "Stratec RER Area" and "Stratec Extended RER Area" on the basis of commuting ties through the railway network transportation system.[19]

The 2001 Belgian census includes the information from 177,721 dwellings pertaining to the Brussels Capital Region and 330,147 observations in the set of municipalities pertaining to at least one of the most extensive delineations of Brussels. This set is labeled "Union" in Table 13.1. Therefore, Brussels Capital Region concentrates more than half of the rented dwellings of the most extensive definitions of Brussels. In Figure 13.1, the "extended residential commuting complex" corresponds to the union of the "Extended residential commuting area" (in lightest blue), the Suburb (in medium blue), and the

[19]Those spatial entities are essentially based on the Official RER Area defined by ministerial decree and composed of 126 municipalities (Moniteur belge 2004).

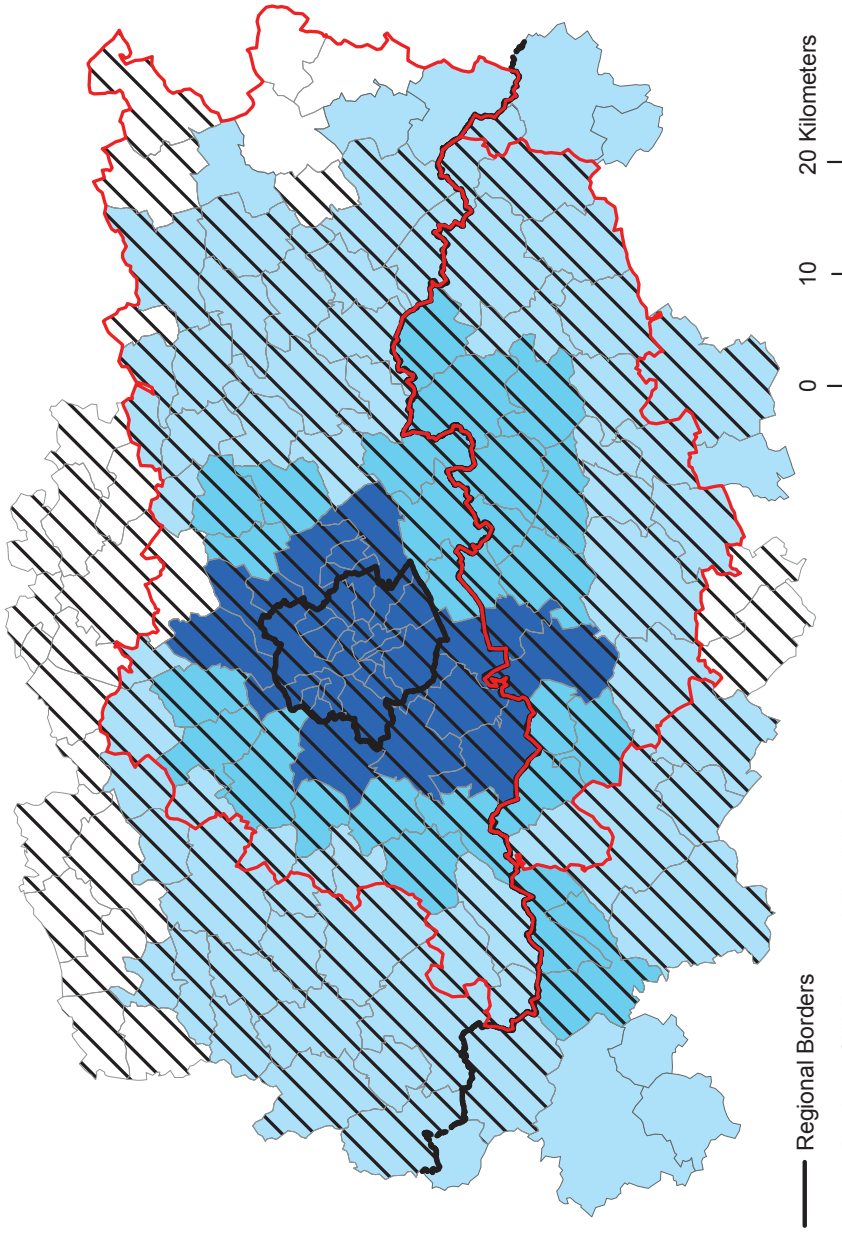


Figure 13.1 Different limits of Brussels agglomeration. Cartography: Alain PHOLO BALA, UCL, 2010.

Agglomeration (in darkest blue). The hatched zone corresponds to "Stratec Extended RER Area" and the small and central area with a black border depicts the Brussels Capital Region.

13.2.2 Basic spatial units

Environment quality attributes and neighborhood characteristics can be measured at different spatial scales. From the least to the most disaggregated level, we may distinguish the following spatial units: the municipality, the former township, and the statistical sector. Those spatial units are nested. The Union area is divided in 147 municipalities, 742 former townships and 5,127 statistical sectors. Variables measured at the municipal level are followed by the label (COM), those measured at the former township level are followed by the label (AC), and those measured at the statistical sector level are followed by the label (SS).

13.3 Data description

In hedonic regression models, dwelling rent is characterized by a bundle of several kinds of characteristics (Kim et al. 2003, Brasington & Hite 2005). The first attribute type refers to the structural characteristics of the dwelling. The second includes neighborhood characteristics such as median income and accessibility. The third type of characteristics relates to environmental quality, such as air pollution and proportion of agricultural areas or forests.

13.3.1 Dwelling structural characteristics

The 2001 Belgian census includes several different housing attributes that can be taken into account into a hedonic regression (see Table 13.8). Most of the variables that can be constructed from those attributes are categorical and qualitative. There is only one potential quantitative variable: the number of rooms.

Information about monthly rents (in euros without charges) has been collected into intervals corresponding to the following categories: 1 for rents below 249.89; 2 for rents between 249.90 and 495.78; 3 for rents between 495.79 and 743.67; 4 for rents between 743.68 and 991,56; 5 for rents larger than or equal to 991,57.

Table 13.7 in Appendix 13.A provides the list of the dwelling attributes used in this chapter.

13.3.2 Environment quality attributes

Land cover information

The Belgian Corine Land Cover database provides land cover information classified according to a legend adapted for the whole European continent. It was obtained in the form of a raster dataset that was used to produce the following synthetic variables at the municipal level: percentage of each municipality and former township covered by Forest (*PERFOR*), and percentage of each municipality and former township covered by Agriculture (*PERAR*).

Pollution indicator

Several hedonic price studies attempt to find out whether air quality is associated with property value. Most of those studies suggest that air pollution affects property value negatively (Boyle & Kiel 2001, Kim et al. 2003, Brasington & Hite 2005, Anselin & Lozano-Gracia 2008). The Belgian Interregional Cell for the Environment (IRCEL–CELINE) provides information on air quality in all 3 Belgian Regions through a raster file. This raster allowed us to compute the average concentration of PM_{10} in every Belgian municipality.[20]

Slope indicator

The average gradient of relief, noted SLOPE, is computed for each statistical sector, each former township, and each municipality from a Digital Elevation Model. It is used as a proxy of the average landscape slope and it will be useful to test the assumption that hilly landscapes are more attractive to residents (Goffette-Nagot et al. 2011).

13.3.3 Neighborhood attributes

Median and average income

Localities where most inhabitants have a high social and economic status are characterized by more expensive dwellings. We use median or average income by tax declaration as indicators of social and economic status. Those data were obtained from Belgian National Statistical Institute for 2001. Median income is available at the level of each statistical sector and each municipality, while average income is available at all spatial scales.

Accessibility indicators

Belgium is a densely populated country with large commuting flows. Its small size and its high population density mean that several employment centers are

[20] PM_{10} denotes particulate matter (suspended particles) with a diameter of $10\mu m$ or less. The use of the concentration of PM_{10} pollutants as a proxy of air quality may be found in other studies as Anselin & Lozano-Gracia (2008) and Montero et al. (2008).

often reachable from a given point (Goffette-Nagot et al. 2011). Greater accessibility implies an increased quality of life for the individual (greater freedom to choose activities and more time to devote to them). Hence, we may expect an influence of accessibility to employment centers on residential land prices and dwelling rents. We use the potential accessibility measure computed by Vandenbulcke (2007). It captures the accessibility of a zone to all the jobs of the other zones.

13.4 Estimation strategy

We estimate the hedonic model by means of interval regression and we analyze spatial effects. Therefore, in this section we present the methodological aspects of our two estimations approaches.

13.4.1 Benchmark model: interval regression

In some databases the information on rent prices is collected through a categorical variable (cf. 13.3.1). Therefore, they do not give the actual value of the rent y_i^*; they just provide the value y_i of a categorical variable from which we can infer the interval where y_i^* lies:

$$y_i = j \text{ if } \alpha_{j-1} < y_i^* \leqslant \alpha_j$$

where $j \in \{1, ..., J\}$ and $\alpha = (\alpha_0, \alpha_1, \cdots, \alpha_J)$ is a given vector of cut points with $\alpha_0 \leqslant \alpha_1 \leqslant \cdots \leqslant \alpha_J$.

To estimate this model, without taking spatial effects into account, we rely on "interval regression". In such a framework, $y^* = (y_1^*, y_2^*, \cdots, y_N^*)'$ is a variable that has a quantitative meaning and not just a latent variable with only an ordinal signification, as in the ordered probit model (Wooldridge 2002). As in Geoghegan et al. (1997), we opt for double-log estimation. Therefore, in this model we are interested in estimating $E(\ln(y_i^*)|x) = x_i'\beta$ where x_i denotes a vector of dummies and of logarithmic transformations of continuous variables.

13.4.2 Spatial Autoregressive Interval Regression Model

A first approach to control for spatial dependence is to extend the basic Interval Regression model by the following specification:

$$\begin{aligned} \tilde{y} &= \rho W \tilde{y} + X\beta + \epsilon \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I_N) \end{aligned} \tag{13.1}$$

where $\tilde{y} = \ln(y^*)$, N is the number of observations, X is a $N \times k$ matrix, ρ is the spatial dependence parameter, W is a $N \times N$ standardized spatial weight matrices.

W tells us whether any pair of observations are neighbors. For example, if dwelling i and dwelling j are neighbors then, $w_{ij} = 1$ and zero otherwise. Whether or not any pair of dwellings are neighbors is based on whether or not they are located in the same geographical entity or in contiguous spatial units. We follow Kim et al. (2003) by considering two spatial units as contiguous when they share a common border.

We assume that there are S spatial entities. Any spatial entity l is populated by N_l individuals, with $\sum_{l=1}^{S} N_l = N$. Therefore, if $y^*_{s_l}$ denotes the $N_l \times 1$ vector of rents paid by the N_l households living in the lth spatial entity, the $N \times 1$ vector of rents paid by all the households of the sample is $y^* = \left(y^*_{s_1}{}', \cdots, y^*_{s_l}{}', \cdots, y^*_{s_S}{}'\right)'$.

While much has been written on the techniques for dealing with spatial dependence in continuous econometric models, the study of spatial dependence in discrete choice models has received less attention in the literature. This is clearly due to the added complexity that spatial dependence introduces into discrete choice models and the subsequent need for more complex estimators.

Several techniques are used to estimate this spatially dependent discrete choice model in a purely cross-sectional setting. In this chapter we opt for the Gibbs Sampling approach since it outperforms the most relevant alternative methods: the Recursive Importance Sampling (RIS) simulator and the Expectation Maximization (EM) algorithm (Flemming 2001).[21]

13.4.3 Two-step estimation with locational fixed effects

We may use an alternative approach to capture spatial dependence based on a two-stage procedure (Ahlfeldt 2011). In a first stage we may obtain estimated values of dwelling rents by carrying out an interval regression on a set of dwelling structural attributes as well as a set of locational-fixed effects that capture location heterogeneity. After this first step, we may retrieve the fixed effects and obtain an estimate of the average log of dwelling rent within former townships, adjusted for dwellings characteristics.

Then in a second stage, we can regress the average log of dwelling rent within former townships on location characteristics. This two-step estimation strategy ensures that there is no endogeneity bias due to the correlation between unobserved location characteristics and observed neighborhood and environment quality attributes. The implementation of this method requires however, the selection of the spatial scale at which the locational fixed effects will be defined. Common sense would suggest to opt for the finest spatial scale, but this is likely to imply huge computational constraints on first step estima-

[21] Further details on the estimation of the Spatial Autoregressive (SAR) model depicted in (13.1) are provided in Appendix 13.B.

tions. In the upcoming estimations, we accommodate for these computational constraints by defining fixed effects at the former township level.[22]

13.5 Results

In this section, we present estimation results for different specifications. We first present results from the benchmark model which further allows us to investigate MAUP and delineation issues. We discuss spatial autocorrelation with the spatial interval regression model. Eventually, we discuss MAUP, delineation issues and spatial autocorrelation with the two-step estimation procedure.

13.5.1 Interval Regression Model

We use the Interval Regression Model (IRM) to serve as a benchmark for comparison with more appropriate models. Since it is less computationally demanding than the spatial interval regression model, the IRM enables estimations with huge databases. This allows for enough variation of environment quality and neighborhood attributes to compare estimation results with different basic spatial units and with different delineations.

Impact of the choice of the delineation of the study area

Table 13.2 displays results of the IRM for different delimitations. Standard errors are clustered on statistical sectors in order to account for variance shifting in the error across space. Most of the results for different delineations of Brussels have the same sign but show differences in magnitude. This suggests that the choice of the limits of the agglomeration has a strong impact on econometric results.

For the dwelling structural characteristics, most of the results are as expected. The value of a dwelling increases with its number of rooms and its surface. Renovated dwellings and dwellings recently built have higher monthly rents. Dwellings with central heating are dearer than those with other heating installations. Dwellings using coal or wood as the source of energy for heating are less expensive. A contrario, dwellings using electricity as the source of energy are more expensive. Dwellings with a garden, bathrooms, toilets, double glazing, and wall isolation are more valuable. Moreover, the more garage places a dwelling has, the more expensive it is.

Most of the results about environmental quality variables and neighborhood attributes are in line with intuition. Rental prices decline with increases of the pollution indicator. This is not surprising since we expect greater demand for dwellings located in less polluted areas. The coefficient of **LNSLOPE (AC)** is positive. This confirms the assumption that hilly landscapes are more attractive to residents. The coefficient obtained for the

[22] More details on this two-step estimation procedure are given in Appendix 13.B.

variable **LNPERFOR (AC)** is positive and significant. This indicates that dwellings located in neighborhoods covered by forests are sought-after.

The sign of the **LNPERAR (AC)** coefficient is more difficult to interpret; it suggests that dwellings located in neighborhoods largely covered by agriculture are less valuable. This may indicate that such neighborhoods are deprived of infrastructures and amenities (schools, shopping centers, etc.) that are required by most households or that agriculture entails some negative neighborhood externalities. Dwelling rents increase with **LNPOTJOBS** the indicator of potential accessibility to jobs, which suggests that there is a greater demand for more accessible residential areas.[23] Finally, the coefficient of the variable **LNREVMED** is positive and significant, indicating that dwellings located in wealthy neighborhoods are more expensive.[24]

To sum up, choices regarding the limits of the study area are not benign regarding econometric results. The reason of those discrepancies may lie on a spatial heterogeneity argument. The different delineations imply distinct geographic structures: the larger the study area, the larger will be the implied proportion of rural hinterland by nature less urbanized and less accessible from the main centers of activities and employment.[25] Therefore, the choice of the study area is a very sensitive issue regarding the precision of estimates. This further stresses the need to delineate the study area in a way that is consistent with the problem under investigation. A failure to do so would entail biases that may mislead statistical inferences and the policy recommendations that they may drive. For instance, the elasticity of the potential accessibility indicator rises from 0.1917 for the Operational Agglomeration macrozone to 0.3663 for the Brussels Capital Region, which means a 91% increase. Such a discrepancy may lead to highly mistaken conclusions in terms of land use and transport policies.

Impacts of the choice of the basic spatial unit

After having evidenced that statistical results are sensitive to the delineation of the study area, we now investigate the impact of the aggregation scale on econometrics findings. Our dependent variable and the dwellings structural

[23] Results for the **LNPERAR (AC)** and the **LNPOTJOBS** will be challenged in the two-stage estimation approach where we control for endogeneity.

[24] Table 13.2 shows the impact of the choice of the limits of the agglomeration on econometric results through estimations of separate samples. The econometricse literature recommends rather to perform an estimation with interaction terms based on the largest sample. This procedure has the advantage of testing whether each coefficient varies significantly, in a statistical viewpoint, across different delineations. In Pholo Bala et al. (2013) we show that several interaction terms are significant which clearly imply that the coefficients of several regressors vary significantly across different definitions of the study area. This further confirms the sensitivity of statistical results to the delineation of the study area.

[25] Further details on the geographic structure of Brussels and its suburb may be found in Chapter 4 of this book.

Variable	Union		ERUC		Urban Region		Agglo.		BCR	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E
Intercept	**-2.727**	0.548	**-2.487**	0.603	**-2.185**	0.596	**-1.987**	0.793	**-4.169**	0.929
APP[1]										
OTHER	**0.101**	0.041	**0.094**	0.043	0.030	0.050	-0.020	0.065	0.010	0.071
SING	**0.074**	0.005	**0.076**	0.006	**0.077**	0.007	**0.072**	0.008	**0.051**	0.009
LOFT	0.019*	0.011	**0.031**	0.011	**0.064**	0.011	**0.058**	0.012	**0.057**	0.012
LNROOMS	**0.096**	0.005	**0.103**	0.005	**0.125**	0.007	**0.118**	0.008	**0.120**	0.009
SURFA[1]										
SURFB	**0.046**	0.006	**0.049**	0.006	**0.050**	0.007	**0.050**	0.007	**0.050**	0.008
SURFC	**0.097**	0.006	**0.103**	0.006	**0.106**	0.007	**0.107**	0.007	**0.108**	0.008
SURFD	**0.183**	0.008	**0.194**	0.008	**0.212**	0.010	**0.224**	0.010	**0.234**	0.011
SURFE	**0.271**	0.010	**0.286**	0.011	**0.328**	0.013	**0.348**	0.013	**0.369**	0.016
SURFF	**0.436**	0.015	**0.454**	0.016	**0.516**	0.019	**0.541**	0.022	**0.544**	0.028
RECENTBUIL	**0.174**	0.006	**0.174**	0.007	**0.200**	0.008	**0.203**	0.009	**0.189**	0.013
RENOVATION	**0.065**	0.006	**0.066**	0.006	**0.083**	0.007	**0.090**	0.008	**0.098**	0.008
FURNISH	**-0.190**	0.009	**-0.188**	0.010	**-0.182**	0.012	**-0.166**	0.012	**-0.159**	0.013
HEATA[1]										
HEATB	**-0.012**	0.005	**-0.014**	0.006	**-0.019**	0.006	**-0.020**	0.007	-0.013*	0.007
HEATC	-0.002	0.019	-0.003	0.019	-0.013	0.021	-0.007	0.024	0.007	0.024
HEATD	**-0.192**	0.004	**-0.187**	0.004	**-0.170**	0.005	**-0.158**	0.005	**-0.143**	0.006
FUELA[1]										
FUELB	**-0.165**	0.013	**-0.170**	0.014	**-0.178**	0.023	**-0.168**	0.024	**-0.127**	0.026
FUELC	**-0.088**	0.024	**-0.093**	0.025	**-0.160**	0.039	**-0.117**	0.048	-0.041	0.069
FUELD	0.107*	0.065	**0.126**	0.062	0.095	0.089	0.050	0.090	0.059	0.097
FUELE	**0.085**	0.008	**0.080**	0.009	**0.056**	0.011	**0.036**	0.012	**0.029**	0.014
FUELF	0.008	0.005	0.004	0.005	**-0.018**	0.006	**-0.028**	0.006	**-0.030**	0.007

[1] Reference case.
Bold: significant at 0.05 level.
* significant at 0.10 level.

Continued on next page

Table 13.2 Interval regression: estimation results for different study areas.

Variable	Union		ERUC		Urban Region		Agglo.		BCR	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E
FUELG	-0.028*	0.015	-0.027	0.017	**-0.056**	0.022	**-0.051**	0.024	-0.034	0.026
FUELH	0.013	0.039	0.015	0.042	-0.057	0.052	-0.090	0.057	-0.084	0.058
DGLAZING	**0.030**	0.004	**0.027**	0.004	**0.020**	0.004	**0.018**	0.005	**0.021**	0.005
WALLISO	**0.041**	0.004	**0.040**	0.004	**0.035**	0.004	**0.032**	0.004	**0.027**	0.005
BATHROOM	**0.252**	0.008	**0.244**	0.008	**0.217**	0.009	**0.209**	0.008	**0.200**	0.008
TOILET	**0.040**	0.008	**0.041**	0.009	**0.040**	0.009	**0.034**	0.010	**0.037**	0.010
PARKA[1]										
PARKB	**0.104**	0.006	**0.108**	0.006	**0.132**	0.006	**0.141**	0.006	**0.151**	0.007
PARKC	**0.151**	0.008	**0.152**	0.009	**0.176**	0.011	**0.174**	0.012	**0.143**	0.014
GARDEN	0.007	0.005	0.008	0.005	**0.015**	0.005	**0.013**	0.006	**0.019**	0.006
LNPM10	**-0.155**	0.042	**-0.221**	0.051	-0.048	0.054	-0.076	0.060	**-0.188**	0.071
LNPOTJOBS	**0.355**	0.027	**0.348**	0.029	**0.234**	0.038	**0.192**	0.054	**0.366**	0.066
LNSLOPE (AC)	**0.132**	0.008	**0.119**	0.010	**0.136**	0.014	**0.116**	0.017	**0.150**	0.027
LNPERFOR (AC)	**0.011**	0.002	**0.010**	0.002	**0.014**	0.003	**0.011**	0.003	0.003	0.005
LNPERAR (AC)	**-0.020**	0.003	**-0.024**	0.003	**-0.028**	0.003	**-0.030**	0.004	**-0.029**	0.005
LNREVMED (SS)	**0.407**	0.028	**0.415**	0.029	**0.469**	0.026	**0.520**	0.025	**0.550**	0.028
scale(σ)	0.343		0.343		0.340		0.334		0.327	
Nobs	70,839		62,695		44,319		37,805		30,315	
Sample size	330,147		301,160		233,582		208,371		177,721	

[1] Reference case.
Bold: significant at 0.05 level.
* significant at 0.10 level.

Table 13.2 Interval regression: estimation results for different study areas (*concluded from the previous page*).

attributes are individual characteristics. Therefore, in our analysis aggregation biases concern only environmental and neighbourhood variables.

Table 13.3 compares estimation results when the average income by tax declaration is measured successively at the statistical sector, the former township and at the municipal levels. It shows that the coefficient of the logarithm of the average income is higher when average income is measured at the municipal level.

In Table 13.3, we also observe substantial changes in the intercept, which is almost multiplied by 2.4 from the first to the third specification, the **GARDEN** coefficient, which is not significant in the first specification, is significant at the 10% significance level in the second specification and is significant at the 5% significance level in the third specification, and the **LNPM10** coefficient, which is not significant in the second specification and is positive — a surprising result — in the third specification. Therefore, the intercept, the garden and the pollution variables' coefficients appear to be very sensitive to changes in the size of the basic spatial units.

Such results are consistent with Gehlke & Biehl (1934) findings which outline that the correlation coefficient tends to increase as the size of spatial units increases. What is the rationale of such findings? A possible explanation may be the following: the larger the BSU, the lower the variance of a variable. As the standard deviations of variables lie in the denominator of the correlation coefficient and the simple regression coefficient, this may explain their increase when the size of a BSU increases. We may conjecture that a similar effect operates on variable coefficients in the interval regression model.[26]

The results just described outline the necessity of using the appropriate spatial scale for the definition of environmental quality and neighborhood attributes. Using inappropriate spatial scales for the definition of those variables may result in biased estimates. Once more, such biases may imply misleading conclusions in terms of the policy recommendations drawn from econometric results. However, due to constraints in data availability, sometimes they are unavoidable as the information of some variables may only be obtained at specific spatial scales. This is precisely the case with the pollution indicator which, because of raster resolution, can only be computed at the municipal level. In such cases of spatial scale constraints, one should be aware of the potential biases.

Using inappropriate spatial scales may trigger endogeneity of the "error in variables" kind. In a case of continuous dependent variable, Anselin & Lozano-Gracia (2008) use a Spatial 2SLS method to handle both spatial dependence and endogeneity generated by spatial interpolation of air quality values. This

[26] We have however to be cautious in confirming Gehlke and Biehl (1934) findings. This effect may not apply in every context. In some cases we may notice variables displaying a higher variability at finer scales, while others may show up higher variance at a more aggregate level.

approach may be relevant for handling biases due to spatial scales in cases where the regressand is continuous, it is tremendously challenging in the interval regression context. However, it may be considered in the second stage of the two-step procedure with locational fixed effects and deserves to be explored in future contributions.

Variable	(1)		(2)		(3)	
	Coeff.	S.E	Coeff.	S.E	Coeff.	S.E
Intercept	**-2.281**	0.529	**-3.439**	0.477	**-5.376**	0.457
APP[1]						
OTHER	**0.122**	0.051	**0.086**	0.042	**0.091**	0.040
SING	**0.070**	0.005	**0.080**	0.005	**0.078**	0.005
LOFT	0.020*	0.011	0.016	0.010	0.015	0.010
LNROOMS	**0.094**	0.005	**0.094**	0.005	**0.097**	0.005
SURFA[1]						
SURFB	**0.047**	0.006	**0.047**	0.006	**0.048**	0.006
SURFC	**0.096**	0.006	**0.099**	0.006	**0.100**	0.006
SURFD	**0.188**	0.008	**0.186**	0.008	**0.185**	0.008
SURFE	**0.278**	0.010	**0.275**	0.010	**0.270**	0.010
SURFF	**0.427**	0.015	**0.438**	0.014	**0.430**	0.014
RECENTBUIL	**0.176**	0.007	**0.177**	0.006	**0.177**	0.006
RENOVATION	**0.067**	0.005	**0.058**	0.005	**0.056**	0.005
FURNISH	**-0.187**	0.010	**-0.189**	0.009	**-0.191**	0.009
HEATA[1]						
HEATB	**-0.013**	0.006	-0.008	0.006	-0.008	0.006
HEATC	0.000	0.019	-0.015	0.017	-0.021	0.017
HEATD	**-0.180**	0.004	**-0.192**	0.004	**-0.191**	0.004
FUELA[1]						
FUELB	**-0.152**	0.014	**-0.162**	0.013	**-0.161**	0.013
FUELC	**-0.073**	0.027	**-0.087**	0.026	**-0.090**	0.023
FUELD	0.066	0.070	0.093	0.062	0.091	0.061
FUELE	**0.078**	0.008	**0.083**	0.008	**0.081**	0.008
FUELF	0.008	0.005	0.006	0.005	0.005	0.004
FUELG	-0.017	0.016	-0.026*	0.014	-0.026*	0.014
FUELH	0.026	0.038	0.012	0.037	-0.001	0.037
DGLAZING	**0.027**	0.004	**0.028**	0.004	**0.029**	0.004
WALLISO	**0.038**	0.004	**0.043**	0.003	**0.044**	0.004
BATHROOM	**0.245**	0.009	**0.256**	0.008	**0.257**	0.008
TOILET	**0.039**	0.008	**0.038**	0.008	**0.039**	0.008
PARKA[1]						
PARKB	**0.100**	0.006	**0.111**	0.006	**0.114**	0.006
PARKC	**0.137**	0.008	**0.156**	0.008	**0.161**	0.008
GARDEN	0.001	0.005	0.009*	0.005	**0.013**	0.005
LNPM10	-0.076*	0.046	0.015	0.042	**0.193**	0.044
LNPOTJOBS	**0.292**	0.028	**0.276**	0.027	**0.266**	0.022
LNSLOPE (AC)	**0.113**	0.007	**0.105**	0.007	**0.096**	0.007
LNPERFOR (AC)	**0.008**	0.002	0.005	0.003	**0.005**	0.002
LNPERAR (AC)	**-0.022**	0.003	**-0.031**	0.003	**-0.034**	0.003
LNREVMOY (SS)	**0.409**	0.021				
LNREVMOY (AC)			**0.512**	0.024		
LNREVMOY (COM)					**0.656**	0.027
scale(σ)	0.334		0.342		0.342	
Nobs	63,426		70,828		72,105	
Sample size			330,147			

[1] Reference case.
Bold: significant at 0.05 level.
*: significant at 0.10 level.
(1): Estimation with **LNREVMOY (SS)**, **(2)**: Estimation with **LNREVMOY (AC)**, **(3)**: Estimation with **LNREVMOY (COM)**.

Table 13.3 Interval regression: impact of the choice of the size of the basic spatial unit for the variable LNREVMOY, Sample=Union.

13.5.2 SAR Interval Regression model

Let us now investigate the substantive spatial dependence issue through a Spatial Autoregressive Interval Regression (SARIR) model and a General Spatial Interval Regression (SACIR) model. We have not been able to run those spatial models algorithms on the full dataset. Indeed, they are very demanding in terms of computational resources. So we ran those algorithms on two subsamples of respectively 2,969 observations and 62 statistical sectors (Sample 1) and 2,565 observations and 81 statistical sectors (Sample 2).

Dwellings of Sample 1 are located in the municipalities of Anderlecht, Berchem-Sainte-Agathe, and Molenbeek-Saint-Jean. The geographical locations of Sample 2 dwellings lie within the municipalities of Auderghem, Woluwe-Saint-Lambert, and Woluwe-Saint-Pierre.[27] Figures 13.2 and 13.3 depict the average income in Brussels Capital Region (BCR) as well as the locations of Sample 1 and Sample 2 dwellings in the BCR.

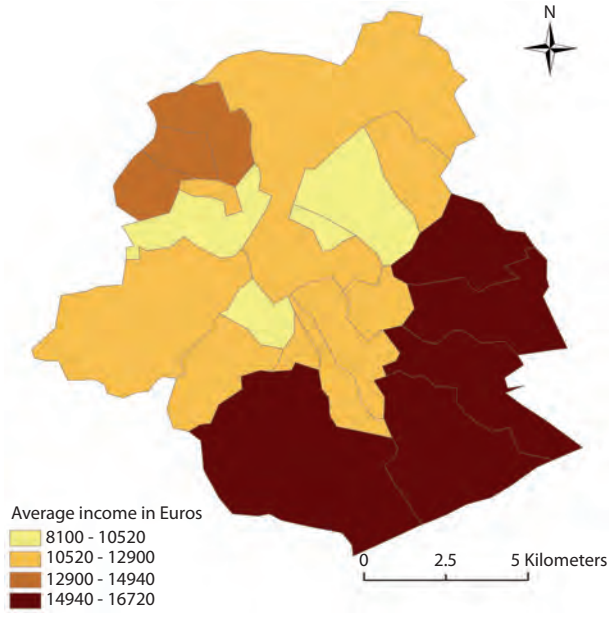


Figure 13.2 Average income in municipalities of Brussels Capital Region. Cartography: Alain PHOLO BALA, UCL, 2011.

[27] Considering an axis dividing Brussels Capital Region from the South-West to the North-East, then the first sample lies in the part of Brussels above that axis and the second sample in the other side. As shown in Figure 13.2, most of the municipalities of Brussels above that axis, like Moleenbeek-Saint-Jean, Anderlecht and Saint-Josse have a lower average income (Ganshoren, Berchem-Sainte-Agathe, and Jette are exceptions characterized by higher incomes). Municipalities below that axis, especially Uccle, Watermael-Boitsfort, Auderghem and Woluwe-Saint-Pierre have a higher average income.

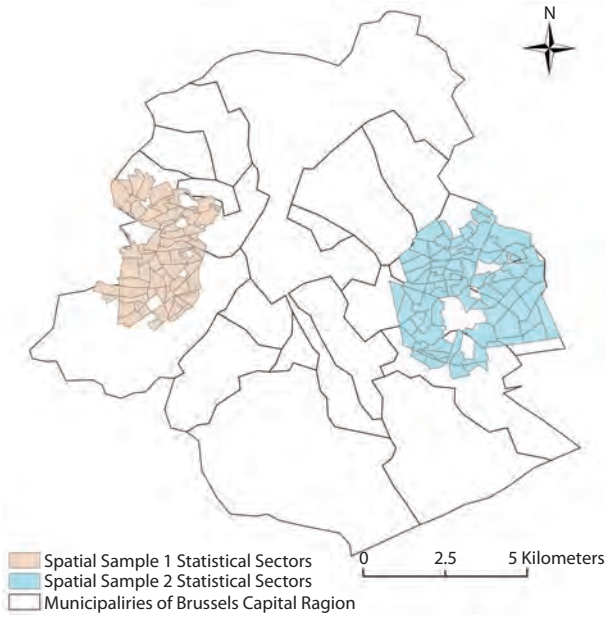


Figure 13.3 Location of Sample 1 and Sample 2 in Brussels Capital Region. Cartography: Alain PHOLO BALA, UCL, 2011.

Unfortunately, there is not enough variation in the environment and neighborhood attributes to assess reliably any potential impact of changes of the basic spatial unit. Table 13.4 displays the results of the basic interval regression model as well as those of the SARIR and the SACIR algorithms. The SACIR model seems less credible than the SARIR model because the disturbance spatial dependence parameter λ is not significant. The results of the IRM and the SARIR model differ substantially only for the spatial dependence parameter and the intercept.

While traditional hedonic estimation does not address omitted variable bias, the spatial autoregressive model mitigates that issue. Indeed, the spatial lag term $W\tilde{y}$ picks up unobserved influences that affect dwelling rent (Brasington & Hite 2005). It relies on a linear combination of dwelling rents nearby in space. Unmeasured influences help determine the rent of neighboring dwellings and the rent of a given dwelling is related to the rents of neighboring dwellings as well. Hence, any dwelling rent is affected by the unmeasured influences of neighboring observations. Therefore, the $W\tilde{y}$ term incorporates the influence of omitted variables on the rent of a dwelling.

Omitting the spatial lag term, as in the non-spatial IRM, would definitely entail biased estimates. Such biases are perceptible in the **LNREVMED (SS)** coefficient which is 18% higher in the non-spatial IRM estimation based on Sample 1 and 7% higher in the non-spatial IRM estimation based on Sample 2.

By acknowledging the impact of dwelling rents nearby in space, the spatial model implies a lower impact of the median income. Another noteworthy observation may be made about the **LNREVMED (SS)** coefficient: it is the only coefficient of an environment or a neighborhood attribute that is significant. As explained earlier, since the other environment and neighborhood variables are measured at a more aggregated level, they do not have sufficient variation to allow enough precision in the measure of their estimates.

Another interesting result lies in the discrepancy between Sample 1 and Sample 2 SARIR model results. For instance, the spatial dependence parameter and the **LNREVMED (SS)** coefficient are higher for Sample 2 than for Sample 1. Those results indicate that the neighborhood has a stronger impact in the determination of Sample 2 dwelling rents than in Sample 1.

The difference between Sample 1 and Sample 2 results yields some evidence of spatial heterogeneity in the estimation. Spatial heterogeneity is related to the lack of stability over space of the relationship under study. It implies that functional forms and parameters vary with location (Anselin 1988). While the analysis of this issue is beyond the scope of the chapter, it may trigger interesting perspectives in terms of spatial issues investigation.

Variable	Sample1			Sample2		
	Inter.reg.	SAR	SAC	Inter.reg.	SAR	SAC
ρ		**0.145**	0.200*		**0.180**	0.187*
		(0.084)	(0.144)		(0.090)	(0.137)
λ			-0.066			-0.053
			(0.285)			(0.276)
Intercept	16.099	7.453	9.706	-6.867	-8.145	-0.905
	(13.283)	(13.541)	(18.221)	(7.284)	(7.005)	(9.864)
SING	**0.072**	**0.073**	**0.077**	**0.124**	**0.120**	**0.128**
	(0.020)	(0.021)	(0.021)	(0.025)	(0.026)	(0.027)
OTHER	**0.572**	**0.570**	**0.609**	**-0.529**	**-0.556**	**-0.561**
	(0.158)	(0.156)	(0.180)	(0.256)	(0.273)	(0.270)
APP[1]						
LOFT	**0.185**	**0.182**	**0.177**	**0.118**	**0.113**	**0.120**
	(0.040)	(0.040)	(0.041)	(0.053)	(0.053)	(0.055)
LNROOMS	**0.231**	**0.230**	**0.227**	**0.296**	**0.297**	**0.300**
	(0.021)	(0.021)	(0.022)	(0.026)	(0.027)	(0.028)
RECENTBUILT	**0.245**	**0.243**	**0.244**	**0.245**	**0.244**	**0.250**
	(0.022)	(0.022)	(0.024)	(0.031)	(0.032)	(0.034)
RENOVATION	**0.071**	**0.072**	**0.066**	**0.093**	**0.098**	**0.102**
	(0.025)	(0.026)	(0.027)	(0.030)	(0.032)	(0.032)
FURNISH	**-0.132**	**-0.132**	**-0.146**	**-0.120**	**-0.121**	**-0.126**
	(0.033)	(0.033)	(0.037)	(0.038)	(0.039)	(0.041)
QUALITY2	**0.148**	**0.149**	**0.149**	**0.253**	**0.253**	**0.247**
	(0.021)	(0.021)	(0.023)	(0.038)	(0.038)	(0.039)
QUALITY3	**0.223**	**0.223**	**0.224**	**0.212**	**0.211**	**0.199**
	(0.0222)	(0.0220)	(0.0234)	(0.037)	(0.037)	(0.039)
QUALITY4	**0.310**	**0.309**	**0.307**	**0.331**	**0.329**	**0.316**
	(0.027)	(0.026)	(0.028)	(0.040)	(0.041)	(0.042)

[1] Reference case.
Bold: significant at 0.05 level.
* significant at 0.10 level.

Continued on next page

Table 13.4 SAR Interval Regression: estimation results for different samples.

Variable	Sample1			Sample2		
	Inter.reg.	SAR	SAC	Inter.reg.	SAR	SAC
QUALITY5	**0.394**	**0.395**	**0.394**	**0.513**	**0.514**	**0.521**
	(0.039)	(0.040)	(0.041)	(0.046)	(0.047)	(0.049)
QUALITY1[1]						
FUELA	0.015	0.015	0.013	0.011	0.010	0.007
	(0.033)	(0.034)	(0.034)	(0.042)	(0.043)	(0.044)
FUELB	**-0.265**	**-0.268**	**-0.281**	**-0.895**	**-0.951**	**-0.770**
	(0.077)	(0.078)	(0.084)	(0.252)	(0.258)	(0.229)
FUELC	-1.578	-1.897*	**-5.083**	-0.235	-0.290	-0.483
	(881.810)	(1.343)	(0.484)	(0.296)	(0.303)	(0.387)
FUELH	-0.039	-0.040	-0.044*	-0.075*	**-0.074**	**-0.074**
	(0.031)	(0.032)	(0.032)	(0.041)	(0.042)	(0.042)
FUELE[1]						
WALLISO	**0.034**	**0.033**	**0.034**	**0.073**	**0.071**	**0.070**
	(0.015)	(0.015)	(0.016)	(0.018)	(0.019)	(0.019)
PARKB	**0.127**	**0.126**	**0.118**	**0.262**	**0.260**	**0.259**
	(0.017)	(0.017)	(0.018)	(0.019)	(0.019)	(0.019)
PARKC	**0.122**	**0.122**	**0.113**	**0.270**	**0.270**	**0.278**
	(0.025)	(0.025)	(0.025)	(0.032)	(0.033)	(0.034)
PARKA[1]						
GARDEN	-0.005	-0.007	-0.002	0.011	0.010	0.011
	(0.017)	(0.017)	(0.017)	(0.021)	(0.021)	(0.021)
LNPM10	**1.404**	0.774	0.882	4.735*	2.004	2.628
	(0.667)	(0.729)	(1.137)	(2.500)	(2.814)	(3.482)
LNPOTJOBS	-1.435	-0.623	-0.858	-0.701	0.087	-0.667
	(1.189)	(1.220)	(1.701)	(1.185)	(1.213)	(1.589)
LNSLOPE (AC)	0.011	0.002	0.052	-0.029	-0.022	-0.052
	(0.031)	(0.031)	(0.050)	(0.045)	(0.046)	(0.053)
LNPER_FOREST (AC)				**0.323**	0.140	0.179
				(0.163)	(0.185)	(0.229)
LNPER_ARABLE (AC)	-0.083	-0.034	-0.063			
	(0.088)	(0.088)	(0.120)			
LNREVMED (SS)	**0.283**	**0.240**	**0.241**	**0.468**	**0.434**	**0.454**
	(0.070)	(0.076)	(0.100)	(0.082)	(0.082)	(0.091)
scale(σ)	0.283	0.082	0.090	0.366	0.138	0.146
Nobs		2,969			2,565	

[1] Reference case.
Bold: significant at 0.05 level.
* significant at 0.10 level.

Concluded from previous page

Table 13.4 SAR Interval Regression: estimation results for different samples.

13.5.3 Two-stage estimation with locational fixed effects

While the SARIR model may mitigate the omitted variable bias by inserting a spatial lag term, it does not control for the endogeneity generated by the simultaneous causality between the dwelling rents and the observed locational characteristics, like the average income. Indeed, while we may expect dwelling rents to be high as a resultant of high average income in the area, we may as well consider areas with high dwelling rents as a sign of attractiveness for high income households because they are expected to host better schools or because they host socio-economic peers. Since reverse causality is more likely to concern location invariant attributes, the two-step estimation procedure by retrieving locational fixed effects addresses this endogeneity problem.

Table 13.8 in Appendix 13.C provides results for the first stage. They are as expected and in line with the previous estimations. Before discussing the effects of a choice of a specific delineation, few remarks are in order. The first one is that conversely to earlier results where the percentage of the basic spatial unit covered by agriculture had a negative impact on dwellings rents while potential accessibility had a positive elasticity, Table 13.5 indicates reverse effects for those two variables. The private dwelling market of the Brussels Metropolitan area values more dwellings close to agricultural areas and with a lower accessibility to the main employment centers. Clearly, inconveniences of urban life (congestion, pollution etc.) reflect in marginal prices.

Moreover, the SAC model is validated in this second-step estimation for the Urban Region and the ERUC samples since we obtain significant disturbance spatial dependence parameter. It provides more precise estimations (lower mean square error) and a better fit (Higher Pseudo $\bar{R}^2$). As before, choices of delineation impact statistical results. From the agglomeration to the Union samples, we obtain different estimates. This holds for OLS and spatial specifications. Therefore, controlling for substantive spatial dependence does not mitigate bias due to the delineation choices. This further highlights the necessity to delineate the study area consistently.

Regarding the impact of the choice of the basic spatial unit, the second-step estimation provides outcomes that are similar to the results obtained with the Interval Regression Model. Table 13.6 compares results when the median income by tax declaration is measured successively at the statistical sector and at the municipal levels. It confirms that the coefficient of the logarithm of the median income is higher when average income is measured at the municipal level.

13.6 Conclusion

Microsimulation tools require a massive amount of geocoded data, collected from several sources and often available at different spatial scales. Hence, choices have to be made about the relevant underlying basic spatial units (BSU), as well as the limits of the studied area. Those choices are generally suspected to influence or even bias econometric results. Moreover, spatial autocorrelation is also likely to have significant impacts on statistical findings.

The goal of this chapter was to address those issues by carrying out sensitivity analyses. Therefore, on the basis of the hedonic regression model, we have investigated the three aforementioned problems.

The delineation of the metropolitan area highly impacts the statistical estimations. We show that most of the coefficients highly and significantly vary with the definition of the study area. Hence, defining a city by functional

Variable	Agglomeration			Urban Region			ERUC			Union		
	OLS	**SAR**	**SAC**	**OLS**	**SAR**	**SAC**	**OLS**	**SAR**	**SAC**	**OLS**	**SAR**	**SAC**
Intercept	4.578*	0.740	0.764	**4.611**	**2.231**	0.610	**4.183**	**2.275**	**1.314**	**3.236**	**1.477**	0.928*
	(2.357)	(2.133)	(2.151)	(1.162)	(1.285)	(1.335)	(0.692)	(0.771)	(0.801)	(0.586)	(0.620)	(0.751)
LNREVMED (SS)	**0.711**	**0.523**	**0.514**	**0.501**	**0.440**	**0.431**	**0.461**	**0.416**	**0.392**	**0.460**	**0.404**	**0.388**
	(0.116)	(0.105)	(0.100)	(0.074)	(0.072)	(0.063)	(0.047)	(0.048)	(0.046)	(0.044)	(0.043)	(0.044)
LNPERAR (AC)	0.008	0.0016	0.001	**0.025**	**0.018**	0.011*	**0.020**	**0.014**	**0.010**	**0.024**	**0.018**	**0.015**
	(0.010)	(0.009)	(0.009)	(0.008)	(0.009)	(0.008)	(0.007)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)
LNPERFOR (AC)	0.002	-0.006	-0.006	0.009	0.004	0.003	0.004	0.002	0.001	0.005	0.003	0.003
	(0.010)	(0.009)	(0.008)	(0.007)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
LNSLOPE (AC)	0.026	0.049*	**0.048**	0.025	0.028*	0.024*	0.022*	**0.020**	**0.018**	**0.020**	0.015*	0.014*
	(0.037)	(0.033)	(0.030)	(0.022)	(0.021)	(0.017)	(0.012)	(0.012)	(0.010)	(0.010)	(0.010)	(0.009)
LNPOTJOBS	**-0.376**	**-0.236**	**-0.239**	**-0.240**	**-0.178**	**-0.139**	**-0.091**	**-0.086**	**-0.081**	-0.052	**-0.054**	**-0.053**
	(0.144)	(0.123)	(0.116)	(0.064)	(0.065)	(0.053)	(0.037)	(0.037)	(0.031)	(0.034)	(0.032)	(0.029)
LNPM10	-0.186	-0.024	-0.022	-0.238*	-0.120	-0.069	**-0.452**	**-0.290**	**-0.221**	**-0.316**	**-0.183**	**-0.151**
	(0.174)	(0.152)	(0.154)	(0.140)	(0.139)	(0.114)	(0.086)	(0.090)	(0.085)	(0.070)	(0.071)	(0.070)
ρ		**0.538**	**0.555**		**0.310**	**0.494**		**0.285**	**0.435**		**0.310**	**0.405**
		(0.110)	(0.134)		(0.091)	(0.124)		(0.057)	(0.090)		(0.051)	(0.096)
λ			-0.069			-0.375*			**-0.248**			-0.163
			(0.252)			(0.209)			(0.140)			(0.138)
scale(σ)	0.087	0.083	0.072	0.094	0.092	0.086	0.098	0.097	0.093	0.096	0.095	0.092
(Pseudo) R^2	0.763	0.788	0.835	0.653	0.668	0.711	0.394	0.405	0.455	0.362	0.373	0.423
(Pseudo) $\bar{R}^2$	0.741	0.768	0.820	0.640	0.655	0.700	0.387	0.398	0.449	0.356	0.367	0.417
Nobs		71			168			532			613	

Table 13.5 Two-stage estimation procedure with fixed effects: second step. Impact of the delineation of the study area.

Variable	(1)			(2)		
	OLS	**SAR**	**SAC**	**OLS**	**SAR**	**SAC**
Intercept	**4.183**	**2.267**	**1.416**	**2.280**	1.311*	1.316*
	(0.692)	(0.775)	(0.804)	(0.903)	(0.920)	(1.016)
LNREVMED (SS)	**0.461**	**0.415**	**0.395**			
	(0.047)	(0.047)	(0.047)			
LNREVMED (COM)				**0.614**	**0.508**	**0.502**
				(0.076)	(0.078)	(0.087)
LNPERAR (AC)	**0.020**	**0.015**	**0.011**	**0.022**	**0.019**	**0.019**
	(0.007)	(0.007)	(0.006)	(0.007)	(0.007)	(0.007)
LNPERFOR (AC)	0.004	0.002	0.001	0.001	0.001	0.001
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)
LNSLOPE (AC)	0.022*	0.020*	**0.018**	**0.031**	**0.027**	**0.027**
	(0.012)	(0.012)	(0.011)	(0.013)	(0.013)	(0.012)
LNPOTJOBS	**-0.091**	**-0.085**	**-0.082**	**-0.100**	**-0.084**	**-0.081**
	(0.037)	(0.037)	(0.032)	(0.040)	(0.039)	(0.040)
LNPM10	**-0.452**	**-0.290**	**-0.227**	**-0.309**	**-0.213**	**-0.212**
	(0.086)	(0.090)	(0.084)	(0.096)	(0.094)	(0.098)
ρ		**0.286**	**0.418**		**0.244**	**0.246**
		(0.056)	(0.093)		(0.061)	(0.127)
λ			-0.221*			-0.030
			(0.139)			(0.155)
scale(σ)	0.010	0.009	0.009	0.010	0.010	0.010
(Pseudo)R^2	0.394	0.405	0.452	0.364	0.368	0.391
(Pseudo)$\bar{R}^2$	0.387	0.398	0.446	0.357	0.361	0.384
Nobs	532					

(1): Estimation with **LNREVMED (SS)**, **(2)**: Estimation with **LNREVMED (COM)**.

Table 13.6 Two-stage estimation procedure with fixed effects, second step: impact of the choice of the BSU, Sample=ERUC

or morphological criteria, as each urban specialist or planner would do, will lead to different results to that defined by a transportation regional planner.

A second spatial aspect addressed by this chapter is the choice of the basic spatial units. The sensitivity of the coefficients to scale effect is empirically demonstrated on the example of Brussels. Our results are consistent with Gehlke & Biehl (1934) and the related literature.

Therefore, in order to minimize the biases, the delineation of the study area must be chosen in a way that is consistent with the phenomenon under investigation. Moreover, the finer the aggregation scale, the more precise the coefficients estimates. Indeed, bad choices in terms of the aggregation scale may lead to misspecification biases. In the ideal situation where all the statistical information is available at the individual level, biases inherent to the ecological fallacy would not exist.

We also accounted for the impact of substantive spatial dependence. As we obtained a statistically significant spatial dependence parameters from those estimations, our econometric results provide evidence of spatial dependence.

The estimation of this Spatial Model is likely to mitigate the omitted-variable bias which generally undermines traditional hedonic estimation.

Finally, we analyse the joint effects of nuisance and substantive spatial dependence through a two-step estimation procedure with locational fixed effects which has the clear advantage to alleviate the endogenous bias inherent to the hedonic regression model. This last estimation strategy shows up that while the explicit consideration of spatial effects in the econometric specification improves the fit of the model, it is still vulnerable to biases due to choices in terms of either the delineation of the study area or the basic spatial unit.

The current implementation of the Real Estate Price Model in UrbanSim does not allow the estimation of spatial econometrics models. Therefore, even more attention should be paid to issues related to nuisance spatial dependence and to spatial heterogeneity. The future implementations of UrbanSim should take that into consideration in order to avoid compounding all kinds of spatial biases.

Chapter appendices

13.A List of dwelling structural characteristics

Variable	Description
Type of dwelling dummies	
APP	dummy for appartments
OTHER	dummy for other kinds of dwelling
SING[1]	dummy for single family dwelling
Dwelling surface dummies	
SURFA	dwelling with a surface lower than 35 m^2
SURFB	dwelling with a surface between 35 and 54 m^2
SURFC	dwelling with a surface between 55 and 84 m^2
SURFD	dwelling with a surface between 85 and 104 m^2
SURFE	dwelling with a surface between 105 and 124 m^2
SURFF[1]	dwelling with a surface higher or equal to 125 m^2
Heating installation dummies	
HEATA	Individual central heating installation
HEATB	Central heating installation common to several dwellings in one building
HEATC	Central heating installation common to several dwellings in several buildings
HEATD[1]	Other heating installations

[1] Reference case.

Table 13.7 List of variables linked to physical attributes of the dwellings (*continued on next page*).

Variable	Description
Dummies related to the composite quality index	
QUALITY1	Insufficient quality
QUALITY2	Basic quality
QUALITY3	Good quality
QUALITY4	Good quality and spacious
QUALITY5[1]	Very good quality
Dummies related to the source of energy used for heating	
FUELA	Fuel
FUELB	Coal
FUELC	Wood
FUELD	Heat pump
FUELE	Electricity
FUELF	Natural gas
FUELG	Butane, propane
FUELH[1]	Other source of energy
Garage dummies	
PARKA	No garage available
PARKB	Garage for one car
PARKC[1]	Garage for more than one car
Other dummies	
LOFT	dummy for studio or loft
RECENTBUILT	dummy for dwelling built after 1981
RENOVATION	dummy for dwelling renovated after 1991
FURNISH	dummy for dwelling furnished
DGLAZING	double glazing dummy
WALLISO	wall isolation dummy
BATHROOM	bathroom dummy
TOILET	toilet dummy
GARDEN	garden dummy
LNROOMS	ln of the number of rooms

[1] Reference case.

Table 13.7 List of variables linked to physical attributes of the dwellings (*concluded from previous page*).

13.B Further details on the econometric strategies

Estimation of a Spatial Autoregressive Interval Regression Model

Specification of equation (13.1) describes a Spatial Autoregressive (SAR) model. The SAR model implicitly assumes that $W\tilde{y}$, the spatially weighted average of housing prices in a neighborhood, affects the price of each dwelling (indirect effects) in addition to the standard explanatory variables of housing and neighborhood characteristics (direct effects). The Spatial Error Model (SEM), characterized by the following specification:

$$\tilde{y} = X\beta + u \quad u = \lambda W u + \epsilon \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2 I_N\right),$$

and the SAR are quite similar mathematically. But, the logic underlying each model's structure is somewhat distinct. In a SEM, spatial autocorrelation is assumed to arise from omitted variables that follow a spatial pattern (Kim et al. 2003). The SEM is the most appropriate when there is no theoretical or apparent spatial interaction and the modeler is interested only in the correction of spatial autocorrelation (Anselin 2001). Since, structural spatial interaction is strongly expected in the dwelling market, in this chapter we will focus on the SAR model.[28]

As in Lesage & Pace (2009), we may express the likelihood for the SAR as

$$L\left(\tilde{y}, W|\rho, \beta, \sigma^2\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} |I_N - \rho W| \exp\left\{-\frac{1}{2\sigma^2}\left(\epsilon'\epsilon\right)\right\} \quad (13.2)$$

where

$$\epsilon = \left(I_N - \rho W\right)\tilde{y} - X\beta.$$

Using diffuse priors for $\left(\beta, \sigma^2, \rho\right)$ results in the following expression of the joint posterior density:

$$p\left(\beta, \sigma^2, \rho|\tilde{y}, X, W\right) \propto |I_N - \rho W|\, \sigma^{-(N/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left(\epsilon'\epsilon\right)\right\} \quad (13.3)$$

Estimates of this distribution should be sampled through a Gibbs sampler with the following 4 steps:

1. Drawing β from $p\left(\beta|\sigma^2_{(0)}, \rho_{(0)}, \tilde{y}_{(0)}\right) \propto \mathcal{N}\left(\tilde{\beta}, \sigma^2_{(0)}\left(X'X\right)^{-1}\right)$ with

$$\tilde{\beta} = \left(X'X\right)^{-1}\left(X'A\tilde{y}\right); \quad (13.4)$$
$$A = I_N - \rho W; \quad (13.5)$$

2. Drawing σ^2 from $p\left(\sigma^2|\beta_{(1)}, \rho_{(0)}, \tilde{y}_{(0)}\right) \propto \left(\sigma^2\right)^{-(N/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left(\epsilon'\epsilon\right)\right\}$;

3. Sample $p\left(\rho|\beta_{(1)}, \sigma^2_{(1)}, \tilde{y}_{(0)}\right) \propto |A| \exp\left\{-\frac{1}{2\sigma^2}\left(\epsilon'\epsilon\right)\right\}$ by inversion approach (Lesage & Pace 2009);

4. Drawing $\tilde{y}$ from the truncated multivariate normal distribution $TMVN\left(\mu, \Omega\right)$ with

$$\mu = \left(I_N - \rho W\right)^{-1} X\beta; \quad (13.6)$$
$$\Omega = \sigma^2\left[\left(I_N - \rho W\right)'\left(I_N - \rho W\right)\right]^{-1}. \quad (13.7)$$

[28] To have a broad picture on the kind of spatial autocorrelation at work in our model, we also estimate the General Spatial Model (SAC) which combines the SAR and the SEM model

$$\tilde{y} = \rho W\tilde{y} + X\beta + u \quad u = \lambda W u + \epsilon \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2 I_N\right).$$

The conditional distribution of ρ does not take a known form as in the case of the conditionals for the parameters β and σ. Therefore, sampling for the parameter ρ must proceed using an alternative approach, such as numerical integration or Metropolis-Hastings.[29]

Two-step estimation procedure with locational fixed effects

Equations (13.8) and (13.9) describe formally the first stage regression. This regression allow to obtain estimated values of dwelling rents.

$$\tilde{y}_i = z_i'\alpha + \mu_l + \epsilon_i \quad \text{or equivalently} \tag{13.8}$$

$$\tilde{y} = Z\alpha + Z_\mu\mu + \epsilon \tag{13.9}$$

with

$$Z = \begin{bmatrix} z_1' \\ z_2' \\ \vdots \\ z_N' \end{bmatrix}$$

and $\mu = (\mu_1, \mu_2, \ldots, \mu_S)'$ where Z is a $N \times k_s$ matrix of dwellings structural characteristics,[30] α is the corresponding vector of coefficients, μ is a vector of unobserved location fixed effects, and Z_μ is a $N \times S$ matrix of spatial entities dummies. After this first step, we may retrieve the fixed effects and obtain an estimate of the average log of dwelling rent within former townships LRENT, adjusted for dwellings characteristics.

Then in a second stage, we can regress LRENT on location characteristics

$$LRENT_j = LOC_j'\gamma + \eta_j \tag{13.10}$$

where LRENT is the average log of dwellings rents estimated from specification (13.9), LOC_j is a vector of location controls, γ is the corresponding vector of coefficient estimates and η is a random error term.

This two-step estimation strategy ensures that there is no endogeneity bias due to the correlation between unobserved location characteristics and observed neighborhood and environment quality attributes.

We may account for spatial dependence in this second stage by extending (13.10) into a spatial model

[29] The Gibbs algorithm of a SAC model will imply 5 steps as we have to draw λ, the disturbance spatial dependence parameter. Further details on the Gibbs Sampler implied by a SAC model may be found in Lesage & Pace (2009).

[30] With k_s the number of regressors derived from the structural characteristics plus 1 for the constant.

$$\text{LRENT} = \delta \text{M LRENT} + \text{LOC}\gamma + \eta \tag{13.11}$$

where δ is a spatial dependence parameter, M is a $S \times S$ standardized spatial weight matrices which tells whether two spatial entities are neighbors.[31]

13.C Two-stage estimation with fixed effects: first stage

Variable	Union		ERUC		Urban Region		Agglo.	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Intercept	**5.901**	0.109	**5.916**	0.110	**5.652**	0.173	**6.118**	0.060
APP	**-0.085**	0.004	**-0.087**	0.005	**-0.080**	0.006	**-0.071**	0.006
OTHER	0.012	0.022	0.002	0.024	-0.041	0.031	**-0.081**	0.037
SING[1]								
LOFT	0.006	0.011	0.018*	0.010	**0.056**	0.011	**0.053**	0.012
LNROOMS	**0.096**	0.004	**0.103**	0.005	**0.126**	0.006	**0.121**	0.007
SURFA	**-0.414**	0.009	**-0.434**	0.010	**-0.504**	0.012	**-0.528**	0.013
SURFB	**-0.366**	0.008	**-0.384**	0.008	**-0.451**	0.010	**-0.475**	0.012
SURFC	**-0.317**	0.007	**-0.331**	0.008	**-0.392**	0.009	**-0.414**	0.011
SURFD	**-0.231**	0.007	**-0.240**	0.008	**-0.287**	0.009	**-0.299**	0.011
SURFE	**-0.148**	0.008	**-0.153**	0.011	**-0.177**	0.010	**-0.179**	0.012
SURFF[1]								
RECENTBUIL	**0.175**	0.005	**0.175**	0.005	**0.195**	0.007	**0.199**	0.008
RENOVATION	**0.058**	0.005	**0.060**	0.005	**0.070**	0.006	**0.074**	0.007
FURNISH	**-0.191**	0.007	**-0.190**	0.007	**-0.188**	0.009	**-0.176**	0.009
HEATA	**0.181**	0.004	**0.175**	0.004	**0.171**	0.005	**0.164**	0.005
HEATB	**0.168**	0.005	**0.160**	0.005	**0.153**	0.006	**0.147**	0.006
HEATC	**0.151**	0.015	**0.143**	0.015	**0.135**	0.017	**0.149**	0.019
HEATD[1]								
FUELA	0.015	0.035	0.023	0.037	0.067	0.051	0.090*	0.054
FUELB	**-0.139**	0.037	**-0.133**	0.039	**-0.112**	0.054	-0.082	0.057
FUELC	-0.072*	0.040	-0.072*	0.043	-0.097	0.060	-0.040	0.067
FUELD	0.090	0.060	0.109*	0.063	0.142*	0.082	0.139	0.088
FUELE	**0.091**	0.035	**0.096**	0.038	**0.122**	0.052	**0.133**	0.054
FUELF	0.010	0.035	0.013	0.037	0.042	0.051	0.062	0.054
FUELG	-0.006	0.037	0.003	0.040	0.017	0.054	0.038	0.058
FUELH[1]								
DGLAZING	**0.028**	0.003	**0.025**	0.003	**0.021**	0.004	**0.019**	0.004
WALLISO	**0.045**	0.004	**0.043**	0.004	**0.037**	0.004	**0.033**	0.005
BATHROOM	**0.252**	0.007	**0.245**	0.007	**0.222**	0.008	**0.215**	0.008
TOILET	**0.040**	0.008	**0.042**	0.008	**0.041**	0.010	**0.037**	0.010
PARKA	**-0.169**	0.005	**-0.171**	0.006	**-0.190**	0.007	**-0.188**	0.008
PARKB	**-0.047**	0.005	**-0.044**	0.005	**-0.040**	0.007	**-0.026**	0.008
PARKC[1]								
GARDEN	**0.013**	0.004	**0.013**	0.004	**0.023**	0.005	**0.022**	0.005
scale(σ)	0.334		0.335		0.336		0.332	
Nobs	72,105		63,681		44,939		38,053	
Sample size	330,096		301,111		233,562		208,355	

[1] Reference case.
Bold: significant at 0.05 level.
* significant at 0.10 level.

Table 13.8 Interval regression with fixed effects: estimation results for different study areas.

[31] We may allow for spatial autocorrelation in the disturbance η in (13.11) by assuming that $\eta = \phi \mathcal{M} \eta + \xi$ with $\xi \sim \mathcal{N}\left(0, \sigma_\xi^2 \mathrm{I}_S\right)$.

Chapter 14

Incorporating equilibrium aspects in microsimulation models

André de Palma, Stef Proost, and Saskia van der Loo

How cities evolve is the result of complex interactions between many agents such as households, developers, firms, and transport agencies. Describing the behavior of all these agents simultaneously is not a simple task and is made even more complex due to the different speeds of adjustments characterizing the different urban changes. Construction of new transport infrastructure is typically a very slow process; large transport infrastructures can take a very long time to construct and once in place are more or less permanent elements. Buildings also have a long life-span, take several years to complete, and are costly to destroy. Population and employment, on the other hand, can change fast; firms are established or closed down, households formed or dissolved and all can move easily. Transport flows are the most flexible of all processes and can change in some cases almost instantaneously.

Here, we study the dynamics of housing and work decisions when a permanent-shock in labor-demand occurs in one of the two zones. The differences in speed of adjustment imply that decisions will be made on the basis of expectations rather than on basis of known market prices. Expectations may have different levels of rationality. We focus especially on the role of these expectations: on the future rental prices of the developers and the impact of government decisions on the transport infrastructure.

A few papers deal explicitly with this problem, but these only consider the housing market (interaction between developers and households). Anas & Arnott (1991, 1993) develop a model for one housing market in which they introduce heterogeneity in the housing stock and in consumer tastes. Moreover, they foresee different types of conversions of the housing stock by profit-maximizing developers. They propose a perfect foresight equilibrium concept

and an algorithm to compute it. Martinez & Hurtubia (2006) also propose a land use and housing model in which housing units are unique. Profit-maximizing developers supply housing units to the highest bidders. In contrast to Anas & Arnott (1991), they introduce myopic foresight for the developers and therefore explore another limiting case. Developers are unable to foresee prices correctly and the future prices they anticipate are modeled as a weighted average of present and past prices. Other land use and housing models are less explicit about the supply decisions and the equilibrium mechanism (see e.g. Wegener (2004), Iacono et al. (2008) for an overview on land use models and de Palma, Picard & Waddell (2007)). Still other, more theoretical, models only analyze the steady state and its comparative static properties (see e.g. Arnott & McMillan (2006), and Glaeser (2008)). Mercenier (2013) argues that the equilibrium on the housing price has to be addressed with the equilibrium on other markets. He points out the importance of the accounting rules between stocks and flows, as well as the contraints imposed by the inter-temporal maximization of the agents. His model is developed in the context of the Grand Paris study, initiated by the SGP (Société du Grand Paris).

We use a simple two zone model to study the dynamics of housing and work decisions when a permanent shock in labour demand occurs in one of the two zones. Using the simple analytic model, we were able to study the dynamics of the adjustment process by incorporating explicitly the information and expectation process of each of the economic agents. It will be the differences in the information and expectation of the different agents that will drive the adaptation process and be responsible for different speeds of adjustment. We focus especially on the role of these expectations on the future rental prices of the developers and the impact of government decisions on the transport infrastructure. This approach is relatively new as in the traditional approach the expectation process remains hidden or one uses extreme assumptions like the perfect foresight assumption or the myopic foresight hypothesis. In a sense, we aim here to bridge the gap between these extreme assumptions.

The chapter is organized as follows: first we introduce a simple two zone model which will be used to analyze the reaction of housing, labor, and transport markets to an exogenous productivity shock in one region. After a sketch of the different components of the model in Section 14.1, we study the different *laissez faire* equilibriums without transport investment (Sect. 14.2) for three different expectation mechanisms (perfect foresight, delayed reaction, and myopic expectation), and illustrate the equilibrium mechanisms numerically. Next, we introduce a government that has the option to change the transport infrastructure and study the impact of a reduction of the commuting costs on the housing and labor market assuming perfect foresight for all agents (Sect. 14.3). In Section 14.4 we illustrate the findings when expectations are no longer perfect. The final section concludes.

14.1 Model description

In this section we describe the different elements of the model. What is the spatial set-up and what is the role of individuals, developers, and transport agency.

We consider two regions, labeled A and B. There are N households consisting of a single worker which choose simultaneously their residence and their workplace. Workers are allowed to commute, so that job and residence location need not to be in the same region. In each region there are competitive firms employing n workers. Each household consumes one unit of housing which is provided by profit-maximizing developers. We use Figure 14.1 to introduce the basic notation.

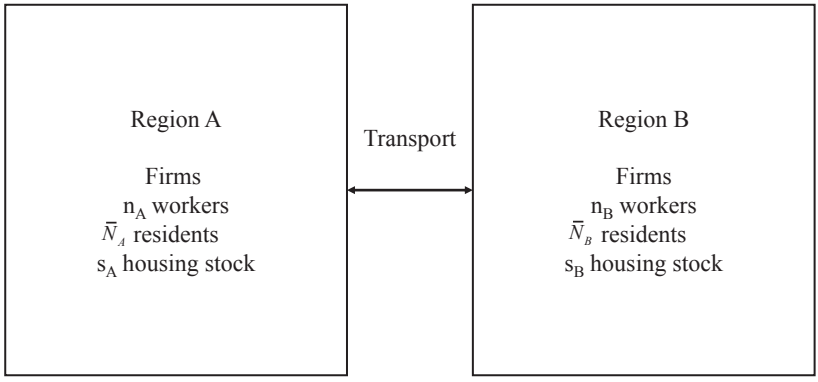


Figure 14.1 Elementary notations.

We distinguish different periods $t = 0, ...T$. At the beginning of each period, households choose their workplace and residence location, firms choose how much labor they need, and developers decide how many new houses will be built in each region.

14.1.1 Individuals

Let N_A (or N_B) denote the number of (homogenous) individuals living and working in A (or B). The number of commuters, i.e. individuals living in region I but working in region J, is labeled N_{IJ}. We assume that there is no migration, so that the number of individuals in the economy is fixed and equal to N. We further assume that all individuals work and that the number of individuals is large enough that they take wages and house prices as given. Individuals are assumed to jointly select a residential location and a work location so as to maximize utility. The indirect utility of an individual (U) depends on their net income which equals their wage (wt) minus housing costs (r) minus the commuting costs (tt), where the average commuting cost is a linear function

with slope μ of the number of commuters[32]:

$$U_{ij}(t) = U\left(w_j(t), r_i(t), tt_{ij}(t)\right) = w_j(t) - r_i(t) - tt_{ij}(t) \qquad i,j = A,B, \quad (14.1)$$

$$with tt_{ij}(t) = \mu N_{ij}(t). \quad (14.2)$$

As individuals can relocate at the beginning of every period, a spatial equilibrium for the individuals implies that the utility of every individual will be equal whatever their choice of residence and job. In addition, if there is commuting at equilibrium, the commuters have the same utility as those that do not commute:

$$U_{AA}(t) = U_{BB}(t) = U_{BA}(t) \quad (14.3)$$

$$w_A(t) - r_A(t) = w_B(t) - r_B(t) = w_A(t) - r_B(t) - \mu N_{BA}(t). \quad (14.4)$$

Workers will only be willing to commute from B to A if wages in region A are higher to compensate the commuting costs. The differences in rents between the two regions will compensate the wage differences between them.

14.1.2 Firms

On the production side of the economy, we assume that there are firms in each region producing a single homogenous good and use labor as the only input. Contrary to the individuals who can choose their residential location after each period, we assume the location of the firms as fixed. In each region all individual firms have the same quasi-linear production function. The aggregate marginal product function is decreasing in the number of workers n (this means that we neglect agglomeration effects which would imply an increase in marginal productivity with the number of workers):

$$MP_i\left(n_i(t)\right) = \alpha_i - \beta_i n_i(t), \qquad i = A,B. \quad (14.5)$$

Production is sold on the world market and its price is normalized to 1. The shareholders of the firm are assumed to be residents of the region where the firm is located, so that profits will directly benefit the region. Labour and product are homogeneous, and perfect competition on the product and labor market implies that the marginal product is equal to the wage (w):

$$MP_i\left(n_i(t)\right) = w_i(t) \;\Rightarrow\; w_i(t) = \alpha_i - \beta_i n_i(t), \qquad i = A,B. \quad (14.6)$$

We will refer to this condition as the *labor market equilibrium* condition.

[32] We only consider commuting from region B to A, therefore N_{BA} (the number of commuters) can be negative when commuting takes place in the other direction.

14.1.3 Developers

While households can reallocate every period and firms can adjust their wages without delay, we will assume that building a house does take time. We assume that a housing unit lasts for two periods and it takes one full period to construct a house. Houses built in period $t-1$ and available in period t will be referred to as the "new" houses in period t, the same housing units available in period $t+1$ will be referred to as the "old" houses (see Fig. 14.2).

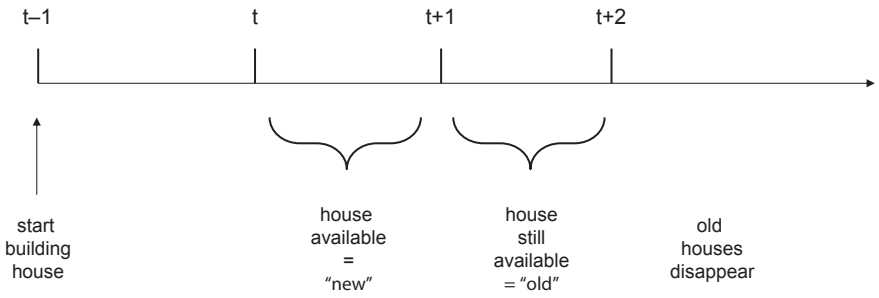


Figure 14.2 Evolution of housing stock over time.

We assume there is no depreciation or increased maintenance cost for old housing. We further assume that at the end of each period, individuals can move and relocate at a fixed moving cost M that in our case is taken to be equal to zero.

The housing stock $s(t)$ in each period t is the sum of the houses built $b_i(.)$ in period $t-1$ and $t-2$:

$$s_i(t) = b_i(t-1) + b_i(t-2). \tag{14.7}$$

The total construction cost (including the cost of land) of new houses $b(t)$ built in period t in zone i is given by:

$$TC_i(b_i(t)) = \gamma b_i(t) + \frac{\delta}{2}(b_i(t))^2 \quad \forall t,\ i = A, B. \tag{14.8}$$

Houses are rented to residents at a rental price $r_i(t),\ i = A, B$ at each period t. Rental prices are determined by demand and supply on the housing market and developers and households take rents as given (perfect competition assumption).

Developers maximize profits when deciding how many houses to build. The expected profits of a developer in region i in period t are given by:

$$\Pi_i(b(t)) = \rho\{r_i(t+1) + \rho r_i(t+2)\} b_i(t) - \left(\gamma + \frac{\delta}{2} b_i(t)\right) b_i(t), \tag{14.9}$$

where we use the discount factor $\rho = 1/(1+ii)$, where ii is the real interest rate for alternative investments on the international capital market. Their housing supply maximizes discounted profits in each period:

$$\frac{\partial \Pi_i(t)}{\partial b_i(t)} = 0 \Leftrightarrow \left\{\rho r_i(t+1) + \rho^2 r_i(t+2)\right\} = \gamma + \delta(b_i(t)). \tag{14.10}$$

Perfectly competitive developers take rental prices as given and developers build until the discounted profit equals the marginal cost. If discounted rents do not cover the marginal construction cost, building activity will decrease and can disappear. In Eq. (14.10) we can see the role played by future rental prices in the building decisions of the developers: the higher the expected rental prices the more houses will be built.

We now have all the elements to describe demand, supply, and equilibrium upon the housing market. The demand for houses equals the equilibrium number of inhabitants in that region. The supply of houses consists of the houses constructed in that region in the two previous periods:

$$N_A(t) = s_A(t) = b_A(t-1) + b_A(t-2). \tag{14.11}$$

Up to now we have nothing that prevents vacancies or a lack of housing units for the given population. A good treatment of vacancies requires reconversion options for older houses (as in Anas & Arnott (1991)). A shortage, on the other hand, requires the reconversion of housing space into smaller units. In this exploratory phase we avoid these complexities and introduce an artefact to deal with both problems. We put a minimum rental price in the market by assuming that at rent r_{MIN}there is an infinite demand for housing units for another purpose (commercial storage, etc.) And we use a maximum price r_{MAX} at which there is an infinite and immediate supply of temporary housing (containers, etc.)

To summarize, we give the four conditions that are valid in equilibruim:

1. spatial equilibrium for residents: $w_A(t) - r_A(t) = w_B(t) - r_B(t) = w_A(t) - r_B(t) - \mu N_{BA}(t)$
2. labor market equilibrium: $w_i(t) = \alpha_i - \beta_i n_i(t), \qquad i = A, B$
3. housing supply: $\rho r_i(t+1) + \rho^2 r_i(t+2) = \gamma + \delta(b_i(t)), \quad i = A, B$
4. demand for housing: $\bar{N}_i(t) = s_i(t) = b_i(t-1) + b_i(t-2), \quad i = A, B$

14.1.4 The transport agency

Since one of the purposes of this chapter is to understand the interaction between land developers and governments investing in transportation, we introduce a transport agency which is responsible for managing the transport infrastructure. We assume that there is only one such agency and that it can improve the existing transportation connection between the two regions. By

doing so, it reduces the commuting costs and may change residential location choices of the households, the rents, and development decisions. Improvements in the transport infrastructure are assumed to be slower than building houses, but once in place they will last longer: it takes two periods to complete the investments and they will last forever. This would be the case for rail and metro investments. The objective function of the agency is the sum of the welfare of the two regions, which is the sum of the total production (TP) minus the construction costs of housing (TC), the total commuting costs $\left(\mu\left(N_{BA}\right)^2\right)$ and the transport infrastructure costs (INV(μ)):

$$\begin{aligned} W(t) &= W_A(t) + W_B(t) \\ &= TP_A(t) + TP_B(t) - TC_A(t) \\ &\qquad - TC_B(t) - \mu\left(N_{BA}(t)\right)^2 - INV(\mu), \end{aligned} \tag{14.12}$$

where

$$TP_I(t) = \int_0^{n_I} MP_I(t)\, dn_I = \alpha_I n_I(t) - \frac{\beta_I}{2}\left(n_I(t)\right)^2, \tag{14.13}$$

$$TC_I(t) = \gamma b_I(t) + \frac{\delta}{2}(b_I(t))^2, \qquad I = A, B. \tag{14.14}$$

For the numerical simulation we will assume that the investment costs are inversely proportional to μ: higher investment levels correspond to lower levels of μ.

14.1.5 Shock and equilibrium concepts

Now that all the agents have been defined, we need to specify the exogenous productivity shock, the assumptions on the behavior of all agents, and the information available to each agent.

The shock is assumed to occur in region A in period $t = 1$ and is a permanent positive productivity shock, meaning that suddenly (e.g. due to exogenous investments in the zone) the marginal product function of region A will shift upwards from the start of period t=1 onwards:

$$MP_A(t) = \alpha_A^S - \beta_A n_A(t), \quad \alpha_A^S > \alpha_A \quad t \geqslant 1. \tag{14.15}$$

In order to characterize the equilibria, we need to specify the information and behavior of all agents that make decisions. For firms, we assume that they always hire workers until the point where the marginal product equals the wage cost. Individuals have to take two decisions: where to live and where to work. As they are fully mobile between the two zones, and as all individuals are identical, we need to satisfy the spatial equilibrium conditions: the location and work decisions will be in equilibrium when the utility of the individual cannot be improved by moving to another zone of residence or work.

As long as the transport agency is passive, the behavior of the developers will determine supply and equilibrium within the housing market and the economy as a whole. We can construct different types of equilibriums: the perfect foresight, the delayed reaction and the myopic equilibria. Which equilibrium occurs depends on the information and expectation rules of the developers. The equilibrium obtained when the shock is correctly anticipated – both in timing and magnitude – and developers anticipate the future rental prices correctly is the perfect foresight equilibrium. If the shock is not anticipated but the developers have correct expectations on the future rental prices we end up in the delayed reaction equilibrium. Finally, when the shock is not anticipated and building decisions are based on past and current rents we have the myopic equilibrium. In Table 14.1, we summarize the three possibilities.

Perfect foresight	**Delayed reaction**	**Myopic equilibrium**
Shock is fully anticipated	Shock is not anticipated but perfect foresight after the shock	Shock is not anticipated and decisions are based on weighted average of prices

Table 14.1 Different equilibriums.

In order to study the decisions on transport capacity we will, in addition, need to account for the expectations of the transport agency regarding the buiding activity of the developers. The agency can either anticipate that the developers will react to a productivity shock or not anticipate their reaction. Moreover, the agency can also anticipate (or fail to anticipate) that developers will adjust the housing stock to the change in transportation costs due to a transport investment.

14.1.6 Initial equilibrium

Initially the two regions are identical (same amenities, productivity, etc.). This means that, before the shock, exactly half of the total population lives in each region. As there is no difference in productivity between the two regions, wages will be identical in both (see Figure 14.3), and both regions employ half of the population. As a result there will be no commuting and rents will be equal:

$$\bar{N}_A^0 = \bar{N}_B^0 = n_A^0 = n_B^0 = \frac{N}{2} \quad \text{and} \quad N_{BA}^0 = 0, \tag{14.16}$$

$$w_A^0 = w_B^0 = w^0 \quad , r_A^0 = r^0 \equiv r^0 \quad \text{and} \quad s_A^0 = s_B^0 = \frac{N}{2}. \tag{14.17}$$

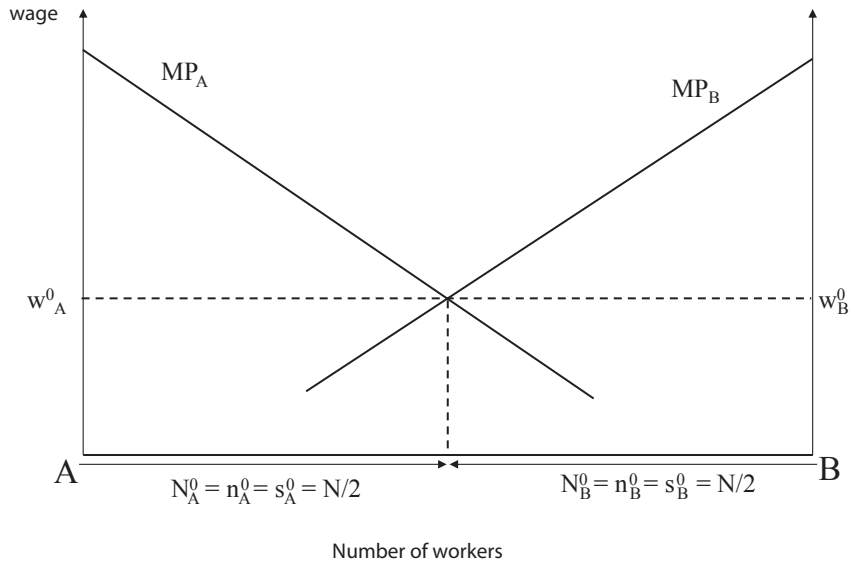


Figure 14.3 Equilibrium in the absence of shock and not accounting for housing and commuting costs.

In the absence of a shock there is a housing stock that is the result of a steady state building activity, namely: $s_i(t) = b_i(t-1) + b_i(t-2) = \frac{N}{2}, i = A, B$. As an example, consider that at the stationary state, the function $b(.)$ is constant, denoted by b^{st}, and equal to $N/4$. Then, $b_i(t-1) = b_i(t-2) = b^{st}, i = A, B$. It could however be that it follows a relation of the form

$$b_i(t-1) = \eta \quad \text{and} \quad b_i(t-2) = \frac{N}{2} - \eta, i = A, B.$$

The magnitude of η will have no implications for the rest of the analysis and we will therefore assume that in the stationary state there is a constant building activity and with b^{st} equal to $N/4$.

14.2 Equilibrium without government investments in transport

We first assume that the government is completely passive and developers know that no investments will be made in the transportation network. As already pointed out, the building decisions of the developers depend crucially on the future rental levels and how developers anticipate these levels. In the subsequent subsections we will assume different expectation rules, and analyze the resulting equilibria. First, we assume that developers have perfect knowledge of the timing and magnitude of the shock (perfect foresight). Next, we assume that they do not anticipate the shock but once it occurs they anticipate the

future rents correctly (delayed reaction). Finally, we assume that developers are myopic, meaning that their predictions of future rental levels are based on actual and past rental levels. Households are assumed to react passively and to move without costs every period in function of their spatial equilibrium condition.

14.2.1 Perfect foresight equilibrium

Steady state equilibrium

We first develop the steady state equilibrium for the periods after a shock. Next we discuss the adaptation process. In a steady state after the shock, the rental prices, the stock, and the population residing and working in each zone will remain constant over time. We proceed as follows. We first use the spatial equilibrium condition for the residents (Condition I in Section 14.1.3) and the equilibrium conditions for the labour market (Condition II) to derive the demand for housing in each of the zones as a function of the rents in the two zones. Next we confront demand for housing units with the supply of housing units by the developers (Condition II) to determine the equilibrium number of residents as well as the number of commuters, the wages, and the rental prices.

The demand for housing can be derived from the spatial equilibrium condition, where we have used the labour market equilibrium condition. We have, in the steady state (for $t >> 1$):

$$\begin{aligned}\alpha^{A} - \beta\left[N_{A}(t) + N_{BA}(t)\right] - r_{A}(t) &= \alpha - \beta\left[N_{B}(t)\right] - r_{B}(t)\\ &= \alpha^{A} - \beta\left[N_{A}(t) + N_{BA}(t)\right] - r_{B}(t) - \mu N_{BA}(t),\end{aligned}$$

where N_{I} is the number of people working and living in zone I. Inspecting the first and the third term implies that, in the presence of commuting, we always have that the two rental prices are linked via the commuting costs $(\Delta r\,(t) = r_{A}\,(t) - r_{B}\,(t))$,

$$\Delta r\,(t) = \mu N_{BA}(t). \tag{14.18}$$

The equality of the first and second term, together with the constraint on population $(N = N_{A}(t) + N_{B}(t) + N_{BA}(t))$ implies

$$N_{A}(t) = \frac{\left(\alpha^{A} - \alpha\right) - \Delta r\,(t) + \beta N - 2\beta N_{BA}\,(t)}{2\beta}. \tag{14.19}$$

Since we assume that there is only commuting from B to A, the number of people living and working in A (N_{A}) will be equal to the number of residents in A $\left(\bar{N}_{A}\right)$. Using the relationship between the two rental prices and the commuting costs gives us demand for housing units $\bar{N}_{A}(t)$ as a function of the difference in rental prices:

$$\bar{N}_{A}(t) = N_{A}(t) = \frac{1}{2}\left[N + \frac{\alpha^{A} - \alpha}{\beta} - \frac{(2\beta + \mu)}{\beta\mu}\Delta r\,(t)\right]. \tag{14.20}$$

The increase in the number of residents of A will be larger when the productivity shock is larger because this increases the marginal product so that more workers are needed in zone A. Whenever the commuter costs (μ) are higher, it is more attractive to move to A than to commute, so this also increases the population in A.[33] A higher rent in A compared to B will lead to a smaller increase of the population in zone A. Only the rent differences count and not the absolute rent because the number of individuals is fixed, and all individuals need exactly one housing unit. The number of residents working and living in B is:

$$N_B(t) = N - N_A(t) - N_{BA}(t) = \frac{1}{2}\left[N - \frac{\alpha^A - \alpha}{\beta} + \frac{1}{\beta}\Delta r(t)\right]. \quad (14.21)$$

For later use we also give the number of residents in A ($\bar{N}_A$) and B ($\bar{N}_B$):

$$\bar{N}_A(t) = N_A(t) = \frac{1}{2}\left[N + \frac{\alpha^A - \alpha^B}{\beta} - \left(\frac{2\beta + \mu}{\beta\mu}\right)\Delta r(t)\right], \quad (14.22)$$

$$\bar{N}_B(t) = N_B(t) + N_{BA}(t) = \frac{1}{2}\left[N - \frac{\alpha^A - \alpha}{\beta} + \frac{(2\beta + \mu)}{\beta\mu}\Delta r(t)\right]. \quad (14.23)$$

The difference between the two being:

$$\Delta\bar{N}(t) = \bar{N}_A(t) - \bar{N}_B(t) = \left[\frac{\alpha^A - \alpha}{\beta} - \frac{(2\beta + \mu)}{\beta\mu}\Delta r(t)\right]. \quad (14.24)$$

The next step is to use the supply of housing as a function of the rental prices. In a steady state the stock of houses and rental prices remain constant over time:

$$r_A(t) = r_A(t+1) \text{ and } b_A(t) = b_A(t+1) = \frac{\bar{N}_A(t)}{2}. \quad (14.25)$$

Combining the supply equations for zones A and B (condition III) we can derive an expression for the difference in rental prices in the steady state (denoted by superscript SS):

$$\Delta r^{SS}(t) = \frac{\delta}{\rho(1+\rho)}\frac{\bar{N}_A(t) - N_B(t) - N_{BA}(t)}{2} = \frac{\delta}{2\rho(1+\rho)}\left(2\bar{N}_A(t) - N\right). \quad (14.26)$$

Substituting the demand condition (14.20) in (14.26), we find the steady state population in region A:

$$\bar{N}_A^{SS}(t) = \frac{N}{2} + \frac{1}{1 + A/2}\frac{\left(\alpha^A - \alpha^B\right)}{2\beta}, \quad (14.27)$$

[33] $\frac{\partial N_A}{\partial \mu} = \frac{1}{\mu^2}(\Delta r(t)) > 0$

where

$$\frac{1}{1+A/2} \equiv \frac{1}{1+\frac{\delta(\mu+2\beta)}{2\beta\mu\rho(1+\rho)}} = \frac{1}{1+\frac{2\beta+\mu}{\beta}B} \quad \text{with} \quad B \equiv \frac{\delta}{2\mu\rho\,(1+\rho)}. \tag{14.28}$$

We see that, in equilibrium, the number of residents in A is increasing in the magnitude of the productivity shock $(\alpha^A - \alpha^B)$, is increasing in the commuting costs (μ) but decreasing in the slope of the building costs (δ). When the marginal cost of construction is constant, we see that the population of A is only determined by the productivity difference and that the discount factor and commuting costs play no role for the equilibrium. Substituting this back into (Eq. 14.26) and using the relationship between the rent difference and the number of commuters (Eq. 14.18) we get

$$N_{BA}^{SS} = \frac{2B}{1+A/2}\frac{\left(\alpha^A - \alpha^B\right)}{2\beta}. \tag{14.29}$$

An increase in the the productivity shock will lead to more commuting, since working in A is more attractive and, as could be expected, a decrease in commuting costs μ will increase the number of commuters.

The adaptation process

In the previous section we described the steady state equilibrium after the shock. However, we are also interested in the adaptation to the shock. Demand for housing and commuting reacts every period to the rental prices. The rental prices are, in the end, determined by construction activities so we need to study the behavior of developers. One of the important factors is knowing when the housing market starts to adapt to the (known) shock. Does the dynamic equilibrium mechanism simply consist of building more houses in period 0 in zone A and less houses in zone B so that the housing market is again at steady state from period 1 onwards?

The developers decide on new houses in function of the expected rents. This means that all rental prices are linked. Using the developers' equilibrium conditions (Eq. 14.10), we have the following relations:

$$N_A(t) - N_B(t) = \frac{\rho}{\delta}\left[\Delta r(t-1) + (1+\rho)\Delta r(t) + \rho\Delta r(t+1)\right]. \tag{14.30}$$

We see that all periods are interlinked via the construction decisions that depend on the rental prices in the two coming periods.

In the perfect foresight equilibrium, there will be no adaptation period; developers anticipate the shock correctly and will ensure that the steady state levels are satisfied when the shock takes place. The intuition goes as follows:

Taking period 0, the expectation of future rents implies that it pays for developers to increase the construction of houses in zone A in period 0 in order to satisfy the higher demand for houses in zone A in period 1 when the productivity shock increases the marginal product. Does this imply that the adaptation to the shock will consist for region A of the following oscillating construction activity:

1. higher building activity in period 0 so that in period 1, the new houses $b(0)$ + old houses $b(-1)$ equal the steady state stock s^{SS}?
2. lower building activity in period 1 as one only needs to replace old houses built in $t-1$
3. higher building activity in period 2 in order to replace the housing built in period 0
   etc.

This would mean that the stock is fully adapted in period 1 and that this adaptation only takes one period.

Does it make sense to smoothen the adaptation process by having a somewhat higher building activity already in period $t-1$? The answer is no for two reasons. First, the houses constructed in period $t-1$, will only serve one period after the shock at the beginning of period 1 while houses constructed in period 0 will serve in periods 1 and 2, two periods with higher demand in zone A. Second, a reason that developers would already increase construction in period -1 could be the construction cost function. The construction cost is increasing in $b(t)$ and this tends to make smoothing more interesting. But because the marginal housing construction cost function is linear, there is no net gain through spreading the extra construction over several periods.

14.2.2 Non anticipated shock but perfect adaptation (delayed reaction)

In this case one observes an unanticipated shock in period 1. Commuting can react immediately, but the change in the housing stock can only take effect in period 2. Future rental prices are perfectly anticipated. The new steady state is equal to the perfect foresight equilibrium. The only difference is that the adaptation is delayed by one period.

14.2.3 Myopic equilibrium

Analytical expressions

There are several versions of a myopic equilibrium. One commonly used assumption is that the rent in the next period is a weighted average of the rents in the past periods:

$$\tilde{r}_i(t+k) = \lambda r_i(t+k-1) + (1-\lambda) r_i(t+k-2). \tag{14.31}$$

This implies that developers now decide on construction on the basis of:

$$\left\{\rho \tilde{r}_i(t+1) + \rho^2 \tilde{r}_i(t+2)\right\} = \gamma + \delta(b_i(t)). \tag{14.32}$$

The optimal building strategy for the developer (Eq. 14.10) then becomes

$$b_i(t) = \frac{\rho(1+\rho)}{\delta}\left[\lambda r_i(t) + (1-\lambda) r_i(t-1)\right] - \frac{\gamma}{\delta}, \tag{14.33}$$

and the stock of housing in zone A at t is equal to

$$s_A(t) = \frac{\rho(1+\rho)}{\delta}\left[\lambda r_A(t-1) + r_A(t-2) + (1-\lambda) r_A(t-3)\right] - 2\frac{\gamma}{\delta}. \tag{14.34}$$

In the period of the shock there will be no adaptation yet, since $r_A(t<1) = r_B(t<1) \equiv r^0$ and

$$s_A(1) = s_B(1) = \frac{N}{2}, \tag{14.35}$$

housing stock will be equal in both regions. This means that there will be an oversupply in region B and an undersupply in A, and some people will commute from B to A. The rent difference between the two regions can be derived by equating $\Delta s(t=1)$ with $\Delta \bar{N}(t=1)$:

$$\Delta s(1) = \Delta \bar{N}(1) \quad \Rightarrow \quad \Delta r(1) = \frac{\mu}{2\beta+\mu}\left(\alpha^A - \alpha^B\right) > 0, \tag{14.36}$$

the rents in B will be lower to compensate the commuters and the higher wages in B (due to the shock). The number of commuters is then

$$N_{BA}(1) = \frac{\left(\alpha^A - \alpha^B\right)}{2\beta+\mu}. \tag{14.37}$$

It can be shown that the rental price difference will be greater than in the steady state, meaning that there are more commuters. The reason is that the myopic developers did not adjust the housing stock to accommodate the increase in demand in A. In the period just after the shock ($t=2$), the developers will react to the increased rents in region A by increasing construction in this region, indeed the difference in housing stock in the two regions at $t=2$ is

$$\Delta s(2) = \lambda \frac{\rho(1+\rho)}{\delta} \Delta r(1) > 0. \tag{14.38}$$

We can derive the rent differences of all subsequent periods after equating the housing stock to the number of residents $\Delta s(t) = \Delta \bar{N}(t)$:

$$\Delta r(t) = \Delta r(1) - \frac{1}{A}\left[\lambda \Delta r(t-1) + \Delta r(t-2) + (1-\lambda)\Delta r(t-3)\right] \tag{14.39}$$

where

$$\Delta r\,(t) = 0 \quad \text{for} \quad t < 1.$$

Analytically, it is not clear whether there is any convergence. It can, however, easily be checked, that whenever the system converges, it will converge to the steady state Δr^{SS}. In the next section we illustrate that for reasonable values of the parameters of the model the system does converge.

Numerical simulations

The values used for the numerical simulations are given in Table 14.2.

Variables	Value
total population N	100
construction parameter γ	1
construction parameter δ	1
discount factor ρ	0.9
marginal production slope β	1
old productivity parameter in A (α_A)	100
new productivity parameter in A (α_A^S)	150
productivity parameter in B (α_B)	100
congestion parameter μ	0.5

Table 14.2 Variables values used in the simulations.

In Tables 14.3 and 14.4 the results are shown for $\lambda = 0.5$ and 1 (where λ is the weight given to the rental price in the previous period by a myopic developer). The first column of each table represents the initial equilibrium; the shock takes place in period 1 (second column). The developers only react in this period and the stock of housing adapts from period 2 onwards. The last column is the steady state. (Note that we only give the first eight periods after the shock, there may be need for more to fully reach the steady state). As can be seen in Figure 14.4, where the rent differences for the two different lambdas are compared, the parameter λ does not seem crucial for the convergence. It seems that the value of the commuting costs plays a more important role. In Table 14.5 we give the simulation results for $\lambda = 0.5$ but this time $\mu = 0.1$ instead of 0.5 as previously, and in Figure 14.5 the rent differences for the two different values of μ are plotted in a single graph.

In conclusion, we found that the system does converge after more than eight periods. This has to be considered in relation to the lifetime of the building (2 periods) and the length of the rental period (1 period). If one period stands for 10 years, and the housing is fully renewed after 20 years, convergence after 80 years is very slow.

	SS1	Shock	Adaptation								SS2
N_A	50.00	50.00	54.28	62.09	63.57	59.91	58.70	59.90	60.74	60.39	60.15
N_B	50.00	30.00	29.15	27.58	27.29	28.02	28.26	28.02	27.85	27.92	27.97
N_{BA}	0.00	20.00	16.58	10.32	9.14	12.07	13.04	12.08	11.41	11.69	11.88
r_A	14.04	19.04	18.18	16.62	16.32	17.05	17.30	17.05	16.89	16.96	17.00
r_B	14.04	9.04	9.89	11.45	11.75	11.02	10.77	11.02	11.18	11.11	11.07
r_A - r_B	0.00	10.00	8.29	5.16	4.57	6.04	6.52	6.04	5.70	5.84	5.94
U_{AA}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
U_{BB}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
U_{BA}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
commuting costs	0.00	10.00	8.29	5.16	4.57	6.04	6.52	6.04	5.70	5.84	5.94
w_A	50.00	80.00	79.15	77.58	77.29	78.02	78.26	78.02	77.85	77.92	77.97
w_B	50.00	70.00	70.86	72.42	72.71	71.98	71.74	71.98	72.15	72.08	72.03
profit dev A	364	364	434	578	608	536	513	536	552	545	540
profit dev B	364	364	300	199	183	226	241	226	215	220	223
profit firms A	1250	2450	2510	2622	2644	2591	2573	2591	2603	2598	2594
profit firms B	1250	450	425	380	372	392	399	393	388	390	391
welfare A	2163	3663	3900	4304	4376	4196	4134	4195	4237	4220	4208
welfare B	2163	3563	3372	2903	2797	3050	3126	3050	2996	3019	3034

Table 14.3 Simulations for myopic developers with μ=0.5 and λ=0.5

	SS1	Shock	Adaptation								SS2
N_A	50.00	50.00	58.55	64.18	59.33	59.06	60.81	60.30	59.88	60.20	60.15
N_B	50.00	30.00	28.29	27.16	28.13	28.19	27.84	27.94	28.02	27.96	27.97
N_{BA}	0.00	20.00	13.16	8.66	12.54	12.75	11.35	11.76	12.10	11.84	11.88
r_A	14.04	19.04	17.33	16.20	17.17	17.22	16.87	16.97	17.06	17.00	17.00
r_B	14.04	9.04	10.75	11.87	10.90	10.85	11.20	11.10	11.01	11.07	11.07
r_A - r_B	0.00	10.00	6.58	4.33	6.27	6.38	5.68	5.88	6.05	5.92	5.94
U_{AA}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
U_{BB}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
U_{BA}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
commuting costs	0.00	10.00	6.58	4.33	6.27	6.38	5.68	5.88	6.05	5.92	5.94
w_A	50.00	80.00	78.29	77.16	78.13	78.19	77.84	77.94	78.02	77.96	77.97
w_B	50.00	70.00	71.71	72.84	71.87	71.81	72.16	72.06	71.98	72.04	72.03
profit dev A	364	364	510	620	525	520	553	543	535	541	540
profit dev B	364	364	243	176	233	236	215	221	226	222	223
profit firms A	1250	2450	2571	2652	2582	2579	2604	2596	2590	2595	2594
profit firms B	1250	450	400	369	396	397	387	390	393	391	391
welfare A	2163	3663	4126	4405	4166	4152	4241	4215	4194	4210	4208
welfare B	2163	3563	3135	2751	3087	3104	2991	3024	3052	3031	3034

Table 14.4 Simulations for myopic developers with μ=0.5 and λ=1.

	SS1	Shock	adaptation								SS2
N_A	50.00	50.00	51.02	53.01	53.87	53.63	53.49	53.48	53.50	53.50	53.50
N_B	50.00	26.19	26.14	26.05	26.01	26.02	26.02	26.02	26.02	26.02	26.02
N_{BA}	0.00	23.81	22.84	20.94	20.13	20.35	20.49	20.50	20.48	20.47	20.48
r_A	14.04	15.23	15.18	15.08	15.04	15.05	15.06	15.06	15.06	15.06	15.06
r_B	14.04	12.84	12.89	12.99	13.03	13.02	13.01	13.01	13.01	13.01	13.01
r_A - r_B		2.38	2.28	2.09	2.01	2.04	2.05	2.05	2.05	2.05	2.05
U_{AA}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
U_{BB}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
U_{BA}	35.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96	60.96
commuting costs	0.00	2.38	2.28	2.09	2.01	2.04	2.05	2.05	2.05	2.05	2.05
w_A	50.00	76.19	76.14	76.05	76.01	76.02	76.02	76.02	76.02	76.02	76.02
w_B	50.00	73.81	73.86	73.95	73.99	73.98	73.98	73.98	73.98	73.98	73.98
profit dev A	364	364	380	413	427	423	421	421	421	421	421
profit dev B	364	364	348	319	306	310	312	312	312	311	311
profit firms A	1250	2724	2731	2738	2737	2736	2736	2736	2736	2736	2736
profit firms B	1250	343	340	338	339	339	339	339	339	339	339
welfare A	2163	3472	3595	3707	3688	3680	3683	3683	3683	3683	3683
welfare B	2163	3920	3801	3683	3703	3711	3709	3709	3709	3709	3709

Table 14.5 Simulation results for myopic developers with μ=0.1 and λ=0.5.

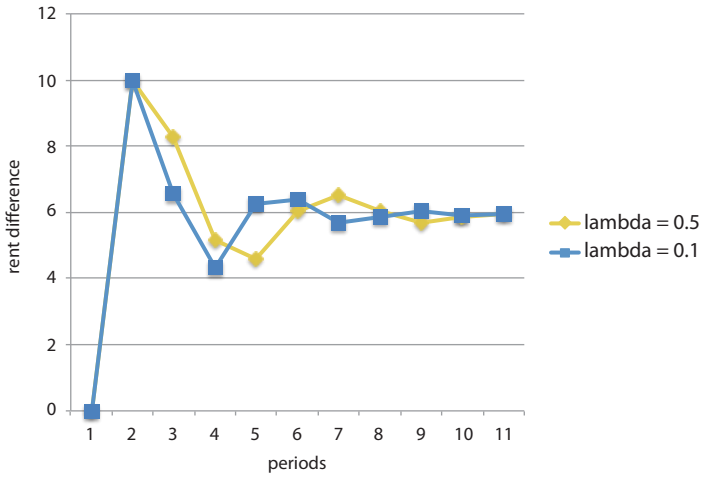


Figure 14.4 Differences in rents during the adaptation period for $\lambda=0.5$ and $\lambda=1$.

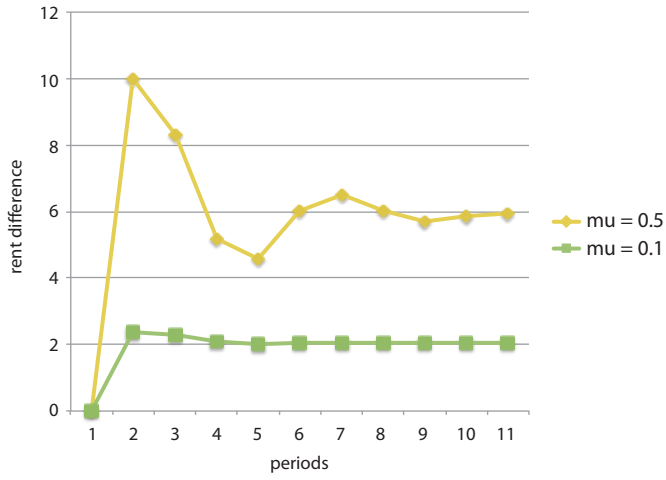


Figure 14.5 Differences in rents during the adaptation period for $\lambda = 0.5$ for two different commuting costs ($\mu=0.5$ and $\mu=0.1$).

14.3 Optimum with transport investment

We now introduce a government transportation agency that can invest in transport infrastructure. To focus on the effect of including the housing market, we first look into the case where the housing stock is constant and compare this to the case where developers are active players and react to a change in commuting costs. More specifically, we are interested in understanding how housing market developments affect the optimal transport infrastructure investment rule. To compute the optimal investment level we assume that the government

anticipates correctly the reaction of the developers. In the next section we will relax this assumption and analyze other equilibria.

14.3.1 Constant housing stock and rents

Assume that the government knows that, due to a productivity shock, there will be a high demand for commuting. To avoid too high commuting costs, it decides to improve the transportation network between the two regions. Once the shock occurs, the investments are made and the commuting costs decrease relative to the non-investment equilibrium.[34] In this section we will, moreover, assume that the stock of houses and the rental prices remain unchanged.

If rents cannot adjust, we cannot impose the spatial equilibrium conditions between residents of A and B (if we do, we get that wages need to be equal and there will be no commuting), we can, however, impose that the utility of a commuter and someone working in B are equal. This implies that

$$w_B = w_A - \mu N_{BA}.$$

Using $w_A = \alpha_A - \beta (N_A + N_{BA})$ and the fact that if stock do not adjust it remains at the initial level and: $\bar{N}_A = \bar{N}_B = N_A = N_B + N_{BA} = N/2$, we found that the number of commuters is:

$$N_{BA} = \frac{(\alpha^A - \alpha^B)}{2\beta + \mu}. \tag{14.40}$$

When commuting costs decrease, the number of commuters increases:

$$\frac{\partial}{\partial \mu} N_{BA} = -\frac{N_{BA}}{(2\beta + \mu)} < 0. \tag{14.41}$$

As the housing stock is constant, an increase in commuting implies an increase in the number of workers in region A which will lead to a decrease in wages in A (see Figure 14.6, where the wage in A decreases when going from point Y to V and X). This decrease implies that the utility of the residents of A will decrease. An increase in the wage in B, compared to the case where no investments are made, leads, on the other hand, to an increase in the utility of the residents of B (including the commuters).[35]

It is straightforward to derive the optimal investment rule. In the case that the housing stock remains constant we have

$$\frac{\partial \bar{N}_A}{\partial \mu} = \frac{\partial \bar{N}_B}{\partial \mu} = 0, \tag{14.42}$$

[34] So we assume that the government in charge of the investments has perfect knowledge of the timing and magnitude of the shock (perfectforesight).

[35] Note that there is still an increase in wages for both regions compared to the situation prior to the shock.

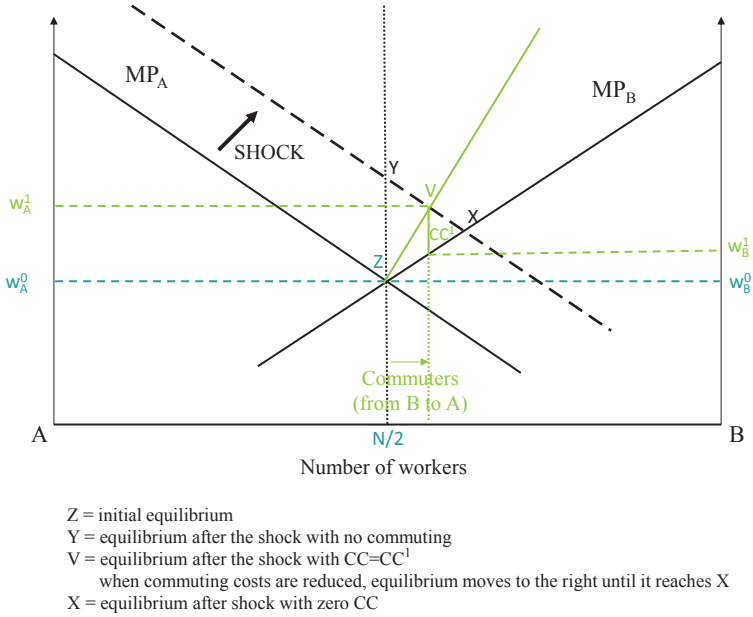


Figure 14.6 Equilibrium before and after a productivity shock in region A when rents and housing stock remain constant.

$$\frac{\partial TP_A}{\partial \mu} + \frac{\partial TP_B}{\partial \mu} = (\alpha_A - \alpha_B)\frac{\partial n_A}{\partial \mu} - \beta\,(n_A - n_B)\frac{\partial n_A}{\partial \mu}, \tag{14.43}$$

$$\frac{\partial TC_A}{\partial \mu} + \frac{\partial TC_B}{\partial \mu} = 0. \tag{14.44}$$

Using the expression for the welfare given in Eq. (14.12) and optimizing w.r.t. μ, it turns out that the government will invest up to the point where the marginal investment costs (MINV) equal:[36]

$$-\text{MINV}^A = \left(\frac{2\beta}{2\beta + \mu}\right)(N_{BA})^2 = \frac{2\beta\left(\alpha^A - \alpha^B\right)^2}{(2\beta + \mu)^3}. \tag{14.45}$$

The marginal benefits consist of two components, the marginal change in total production of the firms and the marginal change in total commuting costs. The marginal change in total commuting costs is squared in the number of commuters because the average cost is increasing in N_{BA}. The marginal change in total production is also a function of the square of the number of commuters. Since people cannot move, the change in number of workers is only determined by the change in the number of commuters. For high values μ (a large amount of congestion), there will be few commuters and a marginal change in the

[36] Note that since more investment in the capacity leads to a smaller μ, the marginal investment (MINV) is negative.

commuting costs will have less impact; the smaller the value of μ, the larger the impact. The marginal benefit is thus a decreasing function of μ. The larger the difference in productivity, the more people are willing to commute and the larger the benefits will be of an investment. The steeper the marginal product function of the firms (β larger), the less impact a change in commuting costs will have on wages and on the total production, and the smaller the benefits are of an investment. For later reference we will denote the resulting investment level as μ^A.

14.3.2 Constant housing stock but rents can adjust

When rents can adjust, the spatial equilibrium constraints for the individuals will hold, and rents adjust until all individuals have equal utility. Compared to the case where the rents do not adjust, the residents of A will lose some utility since they see their rent increase, the opposite being true for the residents of region B. Note that this is the situation occurring at the period of the shock when developers are myopic; they did not anticipate the shock, and thus housing stock is still at the old levels but rents can adjust.

14.3.3 Housing stock and rents can adjust

In the situation where housing and rents can adjust freely we can make use of the steady state equations (14.18). We can show that in the steady state:

$$\frac{\partial N_{BA}^{SS}}{\partial \mu} < 0, \qquad \frac{\partial N_{A}^{SS}}{\partial \mu} > 0, \qquad \frac{\partial N_{B}^{SS}}{\partial \mu} > 0 \text{ and } \frac{\partial \bar{N}_{B}^{SS}}{\partial \mu} < 0 \tag{14.46}$$

and

$$\left| \frac{\partial \bar{N}_{A}^{SS}}{\partial \mu} \right| < \left| \frac{\partial N_{BA}^{SS}}{\partial \mu} \right|. \tag{14.47}$$

A decrease in average commuting cost will make commuting more attractive and so, as one could expect, the number of commuters increases. This implies that region B becomes more attractive to live in and less people will choose to move to A. Region A will thus see its population decrease (in Figure 14.7 the population in A goes from N_A^1 to N_A^2). Although the total population in B increases, there will be a decrease in the number of workers in B (due to the increase in commuters) and thus wages will increase (see Figure 14.7 where wages in B go from w_B^1 to w_B^2). In region A we see the opposite: the population decreases but, due to the increase in the number of commuters, total workforce will increase and wages decrease.

Compared to the case where stock does not adjust, the number of commuters will be inferior since some of the population has actually moved to A.

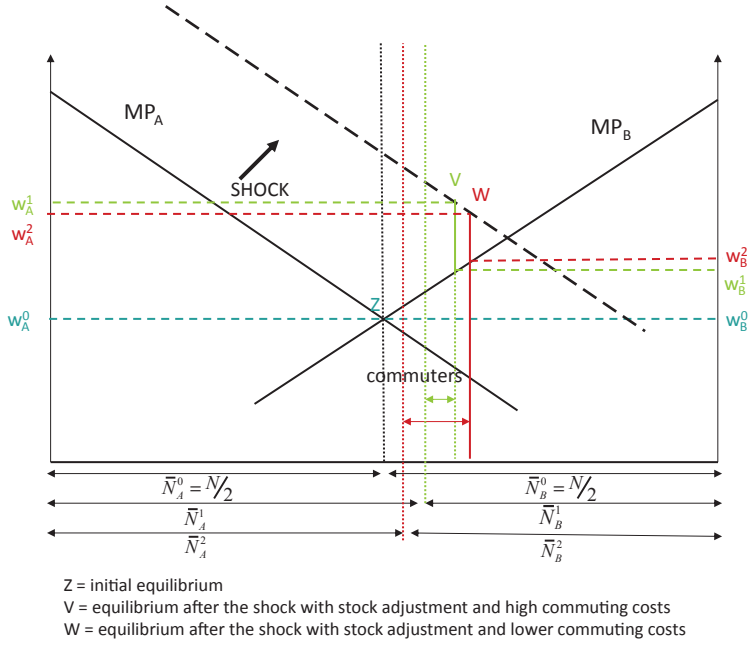


Figure 14.7 Equilibrium before and after a productivity shock in region A when rents and housing stock can adjust.

Next we look into the effects on rental prices. Using

$$s_A = \frac{\rho 2(1+\rho)}{\delta} r_A - \frac{2}{\delta}\gamma = N_A \text{ and } s_B = \frac{\rho 2(1+\rho)}{\delta} r_B - \frac{2}{\delta}\gamma = N - N_A, \quad (14.48)$$

we can derive

$$\frac{\partial r_A}{\partial \mu} = \frac{\delta}{2\rho(1+\rho)} \frac{\partial N_A^{SS}}{\partial \mu} = -\frac{\partial r_B}{\partial \mu}. \quad (14.49)$$

Rents in A decrease (to off-set the decrease in wages), while rents in B increase by just the same amount. Note that the total utility of all residents remains unchanged when investments are made since rents will adjust to counter the effect of the changes in wages and commuting costs.

The optimal investment rule will now look quite different. The number of commuters is in this case equal to N_{BA}^{SS}, deriving this w.r.t. μ yields:

$$\frac{\partial N_{BA}^{SS}}{\partial \mu} = -\frac{1+\mu}{\mu} \frac{1}{1+A/2} N_{BA}^{SS} < 0. \quad (14.50)$$

The derivatives of the total production and construction costs are now:

$$\frac{\partial TP}{\partial \mu} = \left[\left(\alpha^A - \alpha^B \right) - \beta \left(n_A - n_B \right) \right] \left(\frac{\partial \bar{N}_A}{\partial \mu} + \frac{\partial N_{BA}}{\partial \mu} \right), \quad (14.51)$$

and

$$\frac{\partial TC}{\partial \mu} = \frac{\delta}{4}\left(N_A - N_B\right)\frac{\partial N_A}{\partial \mu}. \tag{14.52}$$

Putting this together, we get as optimal investment rule:

$$-MINV^B = \Delta\alpha\left[\frac{\partial \bar{N}_A}{\partial \mu} + \frac{\partial N_{BA}}{\partial \mu}\right] - \beta\left(N_A + 2N_{BA} - N_B\right)\left[\frac{\partial N_A}{\partial \mu} + \frac{\partial N_{BA}}{\partial \mu}\right]$$
$$+\frac{\delta}{4}\left(N_A - N_B\right)\frac{\partial N_A}{\partial \mu} - \left(N_{BA}\right)^2 - 2\mu N_{BA}\frac{\partial N_{BA}}{\partial \mu}. \tag{14.53}$$

Again, we can express everything in terms of $\left(N_{BA}\right)^2$:

$$-MINV^B = \left[2B - 1 - \frac{\delta}{4\mu B}\right]\left(\frac{1}{1 + A/2}\right)\left(N_{BA}^{SS}\right)^2. \tag{14.54}$$

It can be shown that the right-hand side of Equation (14.54) is always smaller than the right-hand side of Equation (14.45). Since there will be less commuters when the housing stock can adjust, the marginal benefits of a reduction of the commuting costs will be smaller, the optimal investment level $\left(\mu^B\right)$ will therefore be smaller than in the case where the stock does not adjust. Governments who do not take the developers' reaction into account are thus likely to over-invest.

14.4 The developers, the transport agency, their expectations, and the resulting equilibrium

In the previous section we have studied the optimal transport investment when the transport agency anticipated correctly the reaction of developers to the productivity shock and transport investment. In this section we study the role of expectations of the transport agency and the developers for the housing and transport infrastructure decisions. We do this for the steady state.

Consider first the developers. They realize that the willingness to pay for additional housing in region A depends on the transport infrastructure in place. As the transport infrastructure has a longer life than housing, we know that, in the steady state, the housing stock will depend on the transport infrastructure in place. As long as we are only interested in the steady state we can therefore concentrate on the expectations and decisions of the transport agency. This will dictate the approach in this section.

Incorrect expectation by the transport agency can be based on an over- or under-estimate of the adjustment of the housing stock adjustment to a shock. As the transport agency has to make only one decision on the size of the infrastructure we need to specify what information is used by the transport agency. In the previous section we already studied two extreme cases for the transport infrastructure investment. The first was where the housing stock does

not adjust. The second had an optimal adjustment of the infrastructure stock. The transport agency could also anticipate (wrongly) that the developers react to the productivity shock but do not anticipate the (slower) adjustment of the transport infrastructure. In this case, the transport agency assumes that the developers do anticipate the shock and will build new houses in A up to the steady state level $s_A^{SS}(\mu)$, equal to Equation (14.27), but since the agency does not expect the developers to anticipate an investment in the transport capacity it will assume that μ is equal to the intial capacity μ^0. To find the level of investment chosen by the agency we maximize welfare for a given stock of houses equal to $\widehat{s}_A^{SS}$ which is assumed to be fixed (not depending on μ).

For a given stock s_A the government chooses μ that maximizes the welfare function. Since the housing stock is assumed to be constant we have again that

$$\frac{\partial \bar{N}_A}{\partial \mu} = \frac{\partial \bar{N}_B}{\partial \mu} = 0, \tag{14.55}$$

$$\frac{\partial TP_A}{\partial \mu} + \frac{\partial TP_B}{\partial \mu} = (\alpha_A - \alpha_B)\frac{\partial N_{BA}}{\partial \mu} - \beta(n_A - n_B)\frac{\partial N_{BA}}{\partial \mu}, \tag{14.56}$$

$$\frac{\partial TC_A}{\partial \mu} + \frac{\partial TC_B}{\partial \mu} = 0. \tag{14.57}$$

The number of commuters is expected to be equal to Eq. (14.29)

$$N_{BA} = \frac{\beta}{2\beta + \mu}\left[\frac{N}{2} + \frac{\alpha_A - \alpha_B}{2\beta} - \widehat{s}_A\right], \tag{14.58}$$

where the stock of housing in region A is expected to be:

$$\widehat{s}_A = \frac{N}{2} + \frac{2\beta\mu^0\rho(1+\rho)}{2\beta\mu^0\rho(1+\rho) + \delta(\mu^0 + 2\beta)}\frac{\alpha_A - \alpha_B}{2\beta}. \tag{14.59}$$

Substituting this in the first order condition, we get:

$$-MINV^C = \frac{4\beta + \mu}{2\beta + \mu}\left(N_{BA}^2\right)^2. \tag{14.60}$$

We can now sum up the three cases where the transport infrastructure is based on the following three expectation assumptions of the transport agency:

1. no housing adjustment to shock and transport infrastructure $\left(\mu^A\right)$
2. housing adjusts to the shock but not to the changes in transport infrastructure $\left(\mu^C\right)$
3. correct expectations of the developers reactions to the shock and the infrastructure extension $\left(\mu^B\right)$

where $\mu^A < \mu^B < \mu^C$. The last inequality is not easy to prove but it is possible to show that $N_{BA}^C < N_{BA}^B < N_{BA}^A$. The (expected) marginal benefit of a capacity extension will therefore be smaller in the case when the transport agency does not take into account that developers will react on an extra investment than when it does, and will thus invest less, resulting in a higher value of μ. We summarize the results in Table 14.6.

Description	Transport agency's expectation of the change in housing stock	Transport agency's expectation of the number of consumers	Resulting investment levels	Outcome
No housing	$\Delta s_A^A (= 0)$	N_{BA}^A	μ^A	Large investment
Optimal adjustment to shock and investment	Δs_A^B	N_{BA}^B	μ^B	Intermediate investment
Adjustment to housing stock but not to transport investment	Δs_A^C	N_{BA}^C	μ^C	Small investment

Table 14.6 Summary of the results according to the expectations of the transport agency.

14.5 Concluding comments

In this paper we analyzed the role of expectations in the housing market and the possible interplay with transport investments. We used a simple two region model, where one region is subject to an unanticipated productivity shock that increases the demand for labor in that region. The demand for labor can be met by additional commuting, by more housing, or by additional investment in transport. When no additional investment in transport is possible, the result depends on the expectations of the developers. Developers with myopic expectations also arrive at the optimal housing stock, but need more cycles to reach the optimum (see de Palma et al. 2011*a*, for a general overview in urban and transport economics).

When we add a transport agency that can invest to improve the transport infrastructure, there is a coordination issue. Transport investment tends to have longer construction periods and lifetimes than housing investments. In this case the steady state equilibrium will mainly depend on the specific expectations of the transport agency. Depending on these expectations, our model predicts over- or under-investments in transport infrastructure.

The model is a very simple two region model with only one mode of transport. This is to be seen as a first exploration of the interaction between transport investments and housing developments. Expectations here are common knowledge; however, this need not to be the case. Different types of extensions can be envisaged. The first type of extension is to integrate more complex behavior from the side of the agents. One could include the endogenous loca-

tion of firms. Now the location of the workplace is exogenous and is the initial shock. One could take other sources of shocks (say a land use decision by the planner) and let the firms decide themselves also on the location. A second type of extension is to give a richer set of instruments to the regional government include income taxes and commuting flows, decide on building permits, etc. The second type of extension is to make the analytical model more realistic by integrating the main mechanisms outlined in the paper into existing land use models. Finally, we have assumed here that the level of rationality and the information available by each agent is given. This, however may not be the case necessarily. Information can be costly to acquire. In this case the level of information of each agent is endogenous. In a similar manner, the importance of rationality can be also partially endogenised via market studies, consultancy, or more Research and Development for example.

14.A Chapter appendix

A summary of the notations used in the chapter can be found in Table 14.7.

Symbol	Description
N_I	residents of region I working in region I
N_{BA}	commuters: residents of region B working in region A
$\bar{N}_I$	residents of region I working in either region I or J
N	total population (fixed)
n_I	workers in region I
μ	slope of time cost function (inverse of capacity)
w_I	wage received in region I
α_I	intercept of marginal product function of firms in I
β_I	slope of marginal product function of firms in I
$s_I(t)$	stock of houses in region I at time t
$b_I(t)$	number of houses built in region I in period t
$r_I(t)$	rent of houses in region I at time t
ρ	discount factor
γ	intercept of marginal cost function of the developers in region I
δ	slope of marginal cost function of the developers in region I
$U_I(w_I, r_I, \mu)$	utility of residents of I
$MP_I(n_I)$	marginal product of firms in I
$TP_I(n_I)$	total product function of firms in I
$TC_I(b_I)$	total cost function of building houses in I
Π_I	profit of developers in I
W	welfare function
$INV(\mu)$	cost of investments in transportation

Table 14.7 Notations.

Chapter 15

Indicators of sustainable development for microsimulation models

Stef Proost, Saskia Van der Loo, Constantinos Antoniou, and Dimitrios Efthymiou

15.1 Introduction

Most cities want to be 'sustainable', and this leads to a huge variety of policies. There are sustainable land use policies that try to increase the city density and improve the access to public transit, shorter to medium term approaches that work via transport pricing (road pricing, parking), or investment policies (bypass, bus lanes, etc.) that aim at reducing the use of private cars in the city. Finally, there are policies that aim to influence the type of transport technologies that are used in cities (promotion of clean and electric vehicles, bike plans, etc.).

Several papers aim at comparing the effects of different policies. Most of them rely on a set of sustainability indicators, but these are often extremely broad and missing economic costs as one of the criteria. In this chapter, our first aim is, therefore, to give a more precise definition of sustainability. This will be an economic definition, but one that encloses environmental and social values. The second aim of the chapter is to translate sustainability at the city level. This is important for two reasons: some of the negative environmental effects are exported to the rest of the world, and citizens can migrate to other cities. The third aim of the chapter is to present how microsimulation models can assess how sustainable a policy is.

This chapter is organized as follows: Sections 15.2 and 15.3 deal with the concept of sustainability. In Section 15.2, we give a brief overview of how the subject has been treated in the literature. In Section 15.3, we give the economic approach to sustainability, which is the approach used subsequently.

Section 15.4 deals with the problems that arise when we want to operationalize the sustainability concept at the level of a city. The last two sections describe how to evaluate how sustainable a policy is using microsimulations models: first we give a theoretical description and in the last section, a more technical description of the implemented module is given.

15.2 What is meant by sustainability and what is a sustainable city?

Nowadays, every policy has to be sustainable, but the definition is fuzzy for most policy makers. There have been many overlapping and contradicting definitions over the past couple of decades (see Pezzey & Toman 2002, for a review of the literature). One of the most cited definition of sustainability is derived from the Brundlandt report (World Commission on Environment and Development 1987) that wants to guarantee good living conditions for the future generations in the world: 'Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs'. At the core of sustainability lies the concern for equity both within and between generations. Many definitions of sustainability used today are broader and include, besides economic and ecological issues, issues of social equity and justice. It proved, however, a difficult concept to precisely define and many interpretations have proliferated since then. One of the sources of confusion is the definition of what needs to be sustained. This confusion has led to the concepts of 'weak' and 'strong' sustainability. The distinction between the two is founded on the distinction between economic capital (manufactured capital, knowledge, natural resources that are used in the economic transformation process), ecological capital (renewable resource stock, natural land areas as well as ecological factors such as climatic conditions, etc.), natural capital (natural resource base, which consists of the ecological capital and non-renewable resources), and social capital (socio-cultural values, human capital and labor force, social institutions, social cohesion, human health, etc.) (Hediger 2000). Strong sustainability is defined with respect to ecological capital whereas weak sustainability is defined with respect to total capital, which is an aggregate of economic, natural and social capital: the objective of strong sustainability is to maintain the ecological capital at its initial level, while the objective of weak sustainability is to maintain the level of social welfare. Weak sustainability allows for trade-offs between consumption and environmental quality and integrates strong sustainability as a special case (by putting infinitely high value to the environment compared to consumption). For these reasons, weak sustainability seems to provide the most comprehensive approach to sustainable development.

Agreeing on the definition of sustainability is not the only challenge. Ultimately, policy makers want to be able to know which policies will lead to

a more sustainable economy or city. The multi-dimensional character of sustainability has given rise to two types of operationalization. The first is more typical for non-economists and policy makers and consists of developing a set of indicators that enclose the different dimensions of the sustainability objective. The second approach is more typical for economists and consists of trying to value each of the components and looking at the sum of the components. This is often presented as the distinction between Benefit Cost Analysis and multi-criteria analysis.

The first approach has been widely used and there is a buoyant literature to be found. How to operationalize the concept of sustainability will, however, also very much depend on the level at which this has to occur (global, country, city, etc.). The focus of this chapter lies at the city level and we therefore limit ourselves to two sources that deal explicitly with sustainable cities: a survey focusing on the world level (Shen et al. 2011) and an EU project PROPOLIS (Lautso et al. 2004, Spiekerman & Wegener 2004) concentrating on European cities.

Shen et al. (2011) compare the use of sustainability indicators for 9 cities (Melbourne, Hong–Kong, Iskandar, Barcelona, Mexico City, Taipei, Singapore, Chandigahr, Pune). Each of these cities have ambitious plans for the long term (10 to 30 years) with divergent visions as 'greenest city', but also 'economically vibrant and sustainable city'. Shen et al. (2011) are interested in comparing the sustainability definition used and not in comparing how sustainable the 9 cities are. The sustainability definitions are regrouped into 115 indicators that can be classified into 4 categories:

- Environmental: water, air, noise, biodiversity, etc. (not GHG)
- Economic: consumption and production patterns, economic growth, strengthening small enterprises, etc.
- Social: energy, water, education, transportation access, poverty, recreation, etc.
- Governance: participation, transparency, accountability, efficient governance (so also process matters)

Their survey reveals three characteristics of the sustainability indicator definitions in what can be called the 'non-economic' literature. First, not only purely environmental indicators are being used, but also social, economic and decision process aspects are taking into account. Second, there is no explicit trade-off between different dimensions (more environment for less economic growth, etc.). Third, some of the indicators are sectorial. 'Sustainable transport' is one of these criteria and it contains energy use, transport mode and transport time as separate elements that are not aggregated and therefore difficult to compare.

The PROPOLIS project (Lautso et al. 2004) aimed to research, develop, and test integrated land use and transport policies, tools, and comprehensive assessment methodologies in order to define sustainable long-term urban strategies and demonstrate their effects in European cities. It developed 3 sets of indicators: environmental indicators, social indicators, and economic indicators. Interestingly, they include the total net benefit from transport as an economic indicator. As a result, there is at least the option to compare environmental benefits and the costs of realizing them.

This first multi-criteria type of approach offers very disparate results and lacks an overall accounting methodology. This is the result of considering the changes in the environment as unmeasurable in economic terms.

The second approach to operationalize sustainability is closer to the Benefit Cost methodology, where an attempt is made to monetize most criteria and to put them in a common framework. This approach will be outlined in the next sections.

15.3 The economic approach to sustainability

As already mentioned, sustainability has both a short term (within a generation) and a long term (between generations). We will present how these two aspects have been dealt with by economists in the next two sections.

15.3.1 Intergenerational equity: the trade-off between man-made and natural capital

It took a while before economists understood and accepted the sustainability concept that deals with the trade-off between current and future stocks of environmental resources. Arrow et al. (2004) offer a good synthesis of the economists' view of sustainability. The ultimate objective is to offer quality of life and consumption options to the future generations. So, what matters is the preservation of enough productive capital for the future generations. This productive capital is then combined with labor to guarantee a sufficient quality of life. Productive capital should be understood in a very broad sense: it concerns traditional 'man-made capital' (knowledge, physical assets such as roads and buildings) as well as 'natural capital' (depletable resources and the quality of the environment). It is the degree of substitution between the two stocks of capital that will determine the relative scarcity of the environmental quality indicators in the future. Figure 15.1 shows how over time, a decrease of natural capital combined with an increase of man made capital can lead to a decrease or increase of production possibilities in the future. The solid line that passes through points A, C, and B is the evolution of the stocks of capital over time. The dotted lines give the production possibilities in an optimistic view of the world, where both types of capital can be substituted (each dotted line

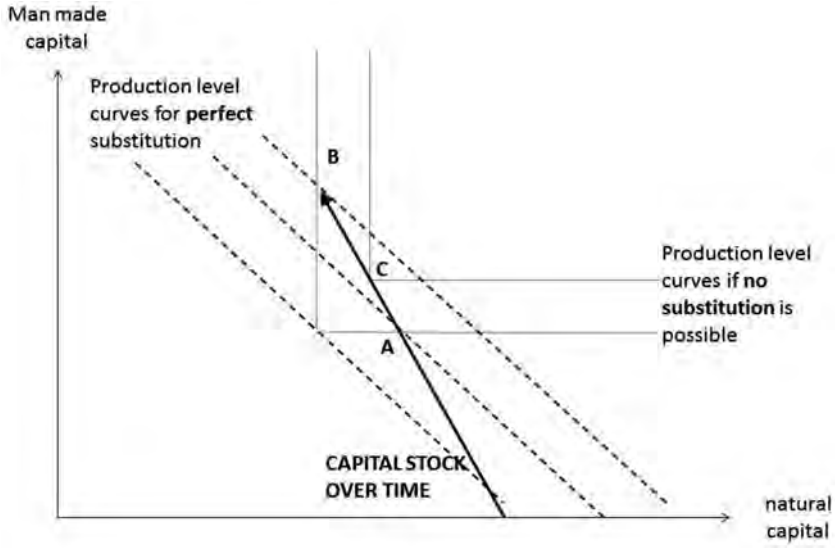


Figure 15.1 An economic definition of sustainability.

presents combinations of capital stocks that generate the same total production possibilities). The thin continuous lines give each a set of capital stocks that generate the same total production level when both types of capital need always to be used in the same proportion and are thus not substitutable.

Starting in point A and moving over time through points B and C, the future production levels are larger in C, according to the two views on substitutability. Moving from C to B is still improving if one counts on substitution possibilities (B is on a higher dotted curve than C), but this is no longer true if one believes that substitution between the two sources of capital is impossible: then point B implies a decrease of production possibilities.

Obviously, one needs to refine the analysis, and there could be a need to build in thresholds for both types of capital and to discuss the many types of natural capital in more detail. What is important for us is the idea of a stock of capital and the potential for substitutability among different types of capital.

15.3.2 Including Social Equity

Besides the economical dimension, sustainability also encompasses social aspects. Equity and a minimum access to basic amenities for the poor are considered as important components of 'sustainability' and are considered as separate indicators by some, but not all sources. Economists have a different view on this on two accounts. First, they prefer, in general, to integrate the equity issues into the overall costs and benefits of certain measures. Second, they prefer to respect the individual choices in the allocation of his income.

This means that the well-being of a poor individual is best measured by his disposable income rather than by the price he pays for a public transit ticket.

The equity dimension of a policy can be measured by first computing the costs and benefits of certain measures by income group and next to add them by giving a higher weight to the poorest. The degree of inequality aversion (Atkinson 1970) chosen by the policy maker will determine the relative weight of the different income classes. A simple weighting scheme could be to use weights inversely proportional to the income per household. One could also rely on more general non-welfaristic measures of equity that look into the capacity of households (education, etc.) to reach a given welfare (Fleurbaey 2010).

Other social indicators may be more difficult to measure and value. These are indicators that measure the quality and intensity of social interactions. What makes this more difficult is that there are strong externalities between individuals and network effects that are difficult to anticipate (see Schelling 1978).

15.4 Operationalization at the city level

To define sustainability at a city level we need to take the limited spatial area of a city into account. This will influence the definition in three aspects. First, there may be a need to deal differently with local stocks (green areas for daily use, historic buildings, water supply, air quality) or global stocks (stock of greenhouse gases, fish stocks, biodiversity). Second, one has to deal with the possible migration between cities. Finally, one should be aware that objective functions of a local government can be very different from that of a federal government; in other words, one needs to differentiate between positive and normative approach.

15.4.1 Local versus global stock

A city is, by definition, confined to a limited spatial area. This implies that the welfare of the individuals will be strongly influenced by the local stocks of capital that are more difficult to substitute (green areas for daily use, historic buildings, water supply, and air quality). The locals will be much less interested in global stocks of capital (stock of greenhouse gases, fish stocks, and biodiversity). These global stocks may be very important for the country, or world, as a whole, but they are the result of the actions of many cities. This first implies that the incentives for one city to take efficient action to protect the global stocks will be too small if they are not forced by a higher authority. Second, as the stock concerned is a global stock at country or world level, there is no need to put an absolute limit to the use of the resource in each city individually. It is sufficient to have a system of tradeable permits that limits the total use of the resource in the country or the world. Summarizing

for local stocks that are difficult to substitute, it will make sense to use absolute thresholds or take into account strongly increasing opportunity costs. For global stocks, it is better to make use of shadow prices that represent the national or world price of the tradeable permits and not to impose absolute thresholds.

15.4.2 Open versus closed city

The second issue we have to deal with is the possible migration between cities. In the US, the mobility of the population combined with the cheap supply of new housing is the main explaining factor for the difference in the city growth (Glaeser 2008). In Europe, the mobility of the population is much lower than in the US, but can still be important for long term policies. To see the importance of mobility for the type of sustainability policies put in place, consider an example with identical individuals and only two cities, X and Y. Let the utility of the population in a city depend on local factors x, y and on the population pop(X), pop (Y) so that the utility per capita equals U(x,pop(X)) or U(y,pop(Y)). The utility per capita of a city can be decreasing or increasing in the size of the population. Utility will be decreasing when there is increasing pressure of the population on local resources, but could be increasing when there are strong returns to scale in supply of public goods or when there are strong economies of agglomeration in production. If utility is downward sloping in population we can have a stable equilibrium. In Figure 15.2 we show how the initial allocation of the population implies that utility per capita is equal.

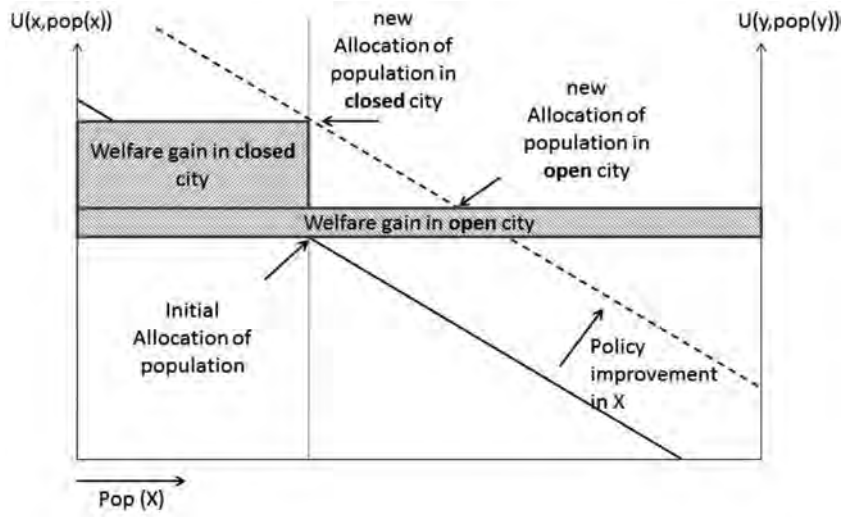


Figure 15.2 Differential effects of a local policy in a closed and open city.

Consider now a policy improvement in city X. This will result in an upward shift of the average utility curve for X. When the city does not allow any growth in number of inhabitants, this improvement will give rise to an important welfare gain for the inhabitants of X. It could be that the welfare gain is captured by the property owners rather than by the inhabitants. But when the city allows expansion, the welfare gains will be dissipated to the new entrants and to the remaining inhabitants of the other city that will see less pressure on its own resources. The **distinction closed/open** city is important because it determines the incentives of a city to implement a policy improvement: in an open city environment, a large part of the gains may dissipate to the rest of the country. The ultimate effect of a policy improvement depends therefore on the supply of extra housing in the city.

15.4.3 Positive versus Normative approach

When a city is only interested in the welfare of its inhabitants, there is also a difference in the objective function pursued by the city and a federal government. Cities neglect positive and negative externalities on the rest of the country. Also, federal countries neglect external effects to the rest of the world. This implies there is a need for clearly distinguishing the positive and normative approach. In the positive approach to policy making, one tries to understand what drives policy making. In the normative approach, one uses ethical criteria to state what 'should' best be done. The positive approach will be driven by the political institutions in place and in democratic regimes, this will be some weighted sum of the utility of the voters. The precise weights will depend on the functioning of the political system (Besley 2006), including the role of pressure groups (Dixit et al. 1997). The positive approach will also generate other policies when decisions are taken at city or at country level. Table 15.1 summarizes the different types of policy objectives. Most of the 'sustainability' literature is of a normative nature where most of the discussions focus on the trade–off between environmental values and other values. When it comes to policy making, we need to recognize that policy makers are not driven by sustainability as such but by what they think matters for their voters.

15.5 Evaluating sustainable urban policies with a microsimulation model

In this section we present a unifying and integrated evaluation framework for microsimulation models. To evaluate policies simulated by the microsimulation model, we couple the outcomes of the model with a social welfare function which will be detailed below. This social welfare function allows for trade-offs between different kinds of capitals, and equity can be taken into account by weighting income classes differently. To evaluate a policy package, we will

	Positive approach (what drives policy making)	**Normative approach (what should be done on the basis of ethical principles)**
City	Politically weighted sum of utility of current voters	Ethically weighted sum of welfare for current and future generations
Country of region	Politically weighted sum of utility of current voters	
World	Country objectives limited by international agreements	

Table 15.1 Different policy criteria to judge policies.

compare the social welfare outcome when the policy is implemented with the business as usual scenario. The social welfare function used is a (discounted) weighted sum of various components, representing the impacts of the policies on various groups, for example the residents of the study area, the commuters, the rest of the world, or the future generations. The welfare (or 'utility') of each group can be weighted differently, according to the objectives and criteria of the policy evaluator. Moreover, it includes the value of the local stock (of both natural and physical capital) left for the next generations at the end of the time horizon, the cost of implementation of the policy and the discounted sum of the revenues collected (these include the tax revenues as well as the toll and the revenues from public transport):

$$
\begin{aligned}
SWF_{city} = & \sum_{t=0}^{T} \frac{1}{(1+\delta)^t} \sum_{\text{inhabitants now}} \phi(U_i) U_{i,t}(I, H, tt) \\
& + \chi \sum_{t=0}^{T} \frac{1}{(1+\delta)^t} U_{commuters,t}(I, H, tt) \\
& + \xi \phi(U_w) U_w(GHG, \ldots) \\
& + \rho \text{ Local stock (T)} \\
& - INV \\
& + \sum_{t=0}^{T} \frac{1}{(1+\delta)^t} Rev(t)
\end{aligned}
\tag{15.1}
$$

where
$i = 1, \ldots, I$ are inhabitants
$t = 0 \ldots, T$ time horizons are considered
$\phi(U_i)$ = weight of individual determined normatively (inequality aversion)
U_i = utility of resident i

ξ = weight of commuters
$U_{commuters}$ = utility of commuters
ξ = value of effects on rest of the world
U_w = utility of rest of the world
ρ = weight of local stock
Local stock = remaining of capital stock and natural stock for future generations
INV = implementation cost of the policy (present value)
Rev = total revenues (taxes, tolls, tickets)

15.6 Overview of the indicator module

15.6.1 Data collection and preparation

Having defined the Social Welfare Function (SWF) to be used for the evaluation of the proposed policies, a tool for its computation using the results of the microsimulation has been developed. The first step of the implementation required a decomposition of the function into 'primary' and 'secondary' variables. 'Secondary' are the variables that compose the utility functions of the residents, commuters and the rest of the world, and 'primary' are the data required for the computation of these variables. In order to comply with the proposed methodology, we started from the most optimistic scenario of data availability. Tables 15.2 and 15.3 show the preliminary components of the proposed SWF.

Secondary variable	Primary variable	Units
Net Income (I)	Yearly wage before taxes	euros/year
	Property income	euros/year
	Local taxes on income	euros/year
	Global taxes on income	euros/year
	Local taxes on property values	euros/year
	Global taxes on property values	euros/year
Housing cost (H)	Yearly rent/Property cost/Return on property	euros/year
	Maintenance cost	euros/year
Travel cost (tc)	Total time spent traveling	hours/year
	Value of time	euro/hour
	Monetary cost	euros/km
	Total distance travelled	km/year
	Public transport ticket costs	euros/trip
	Public transport comfort level	euros/trip

Table 15.2 Utility of residents.

Secondary variable	Primary variable	Primary variable	Units
Global environment	Transport	Total distance traveled by mode and type of fuel	Vehkm/year
	Buildings	Floor space heated	euros/year
		Type of energy	euros/year
		Amount of electricity used	euros/year
		Electricity generation	euros/year

Table 15.3 Utility of rest of the world.

Being aware of the particularities of each case study due to the different spatial level of analysis and data availability, the case study groups were asked to identify which of the required data were available. Table 15.4 shows the available data from each case study. As a result, the 'translation' of the theoretical methodological proposition of the SWF to a technical, tangible indicator has been followed by several assumptions driven by the data availability and the particular characteristics of each case study.

The second decision that needed to be made, concerns the level of integration that the 'indicator module' should have with UrbanSim. After evaluating the options, namely to 1) develop the tool in OPUS GUI using the existing indicator generator, 2) create a single social welfare (SW) variable that could be accessible from UrbanSim's 'variable library', and 3) develop a tool that runs externally, the decision was made to develop a script in python that could run outside UrbanSim and use the microsimulation output data.

The 'indicator module' for the computation of the SW has therefore been coded in python using the UrbanSim libraries and can be executed externally and independently of the OPUS software (but against an existing data-base of previous runs). A single tool has been developed for both zone-level (Brussels and Paris) and parcel-level (Zürich) projects.

15.6.2 Zone-level project support

This version of the module supports zone-level projects (such as the Brussels and Paris cases in the SustainCity project). The Social Welfare Function components were first computed at the zone-level and then aggregated at city level.

- *Annual income*: The monthly income per household is multiplied by 12 and aggregated at the level of zone
- *Housing cost*: The housing cost is computed as a proxy of the transaction cost, which is the only price available in this case study. Following DiPasquale & Wheaton (1996), the annual rent is equal to the purchase

Secondary variable	Primary variable	Brussels	Paris	Zürich
Net Income (I)	Annual wage before taxes	✓	✓	✓
	Property income	✗	✗	✗
	Local taxes on income	✗	✗	✗
	Global taxes on income	✗	✓	✗
	Local taxes on property values	✗	✓	✗
	Global taxes on property values	✗	✗	✗
Housing cost (H)	Annual rent/Property cost/ Return on property	✓	✓	✓
	Maintenance cost	✗	✓	✗
Travel cost (tc)	Total time spent traveling	✓	✓	✓
	Value of time	✓	✓	✓
	Monetary cost	✓	✓	✓
	Total distance travelled	✓	✓	✓
	Public transport ticket costs	✓	✓	✓
	Public transport comfort level	✗	✗	✗
Global environment	Total distance traveled by mode and type of fuel	✗	✗	✗
	Floor space heated	✓	✓	✓
	Type of energy	✗	✓	✗
	Amount of electricity used	✗	✗	✗
	Electricity generation	✗	✗	✗

Table 15.4 Data availability by case study.

price multiplied by the interest rate, which is considered the housing cost in our case. The analyst must indicate the interest rate.

- *Total time spent traveling*: The travel times for trips between home and work are computed by the MATSim travel demand model, which has been integrated with UrbanSim, at an agent (individual) level. Depending on the percentage of the population that the modeler uses for the traffic simulation, the tool computes the total annual travel time between work and home for the entire population.
- *Utility of residents*: The utility of residents is then computed by summing the components per zone level.
- *Utility of commuters*: The utility of the commuters is assumed to be equal to the utility of the residents.

15.6.3 Parcel-level project support

This version of the tool supports parcel-level projects (such as the Zürich case study in the SustainCity project). The Social Welfare Function components

were first computed at the parcel-level and then aggregated at city level.

- *Annual income*: The monthly income per household is multiplied by 12 and aggregated at the level of parcel.
- *Housing cost*: The annual rent is considered as the annual housing cost.
- *Total time spent traveling*: This component is computed with the similar method used in the Brussels case study.
- *Utility of residents*: The utility of residents is then computed by summing the components per parcel level.
- *Utility of commuters*: The utility of the commuters is assumed to be equal to the utility of the residents.

15.6.4 Implementation aspects

The 'indicators module' requires data of two categories in order to compute the SW; the output of the microsimulation computed by UrbanSim and MATSim after a scenario analysis, and external data defined by the user, that are specific to the examined policy and case study. The user needs to specify the following:

- Basic parameters
  - Data paths
  - Level of the project (zone or parcel)
  - Name of the case study
- Policy–specific variables
  - Interest rate
  - Value of time
  - Percentage of agents simulated
  - Year of implementation of the policy
  - Year of evaluation

The individual components of the SW are computed at zone/parcel level and are then aggregated at the whole area of interest. The script generates an output file with the per year and total social welfare, but also the disaggregate values per zone/parcel in order to facilitate further/more disaggregate analysis. Figure 15.3 shows the flow diagram of the steps required to execute the script.

15.7 Conclusion

This chapter has proposed a methodology and a software module that allows the comparison of alternative urban developments. The methodology used is different than the one proposed in most urban simulation models. Traditionally, one relies on a set of indicators, one or more for each of the dimensions: the economic, the social, the equity, and the environmental and global sustainability dimension. We use the same basic variables, but use the social welfare

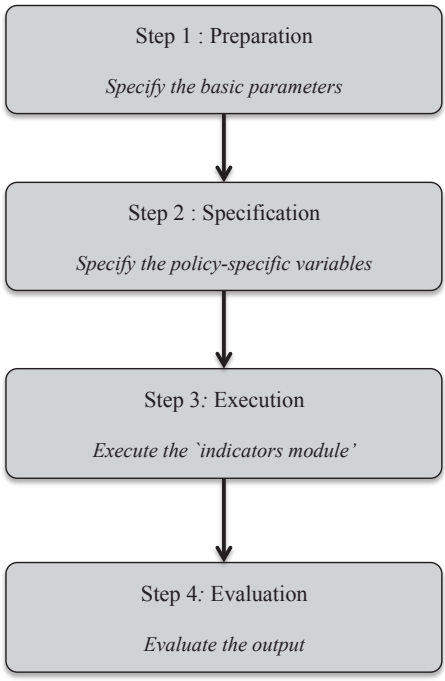


Figure 15.3 Flow diagram for scenario testing.

function to aggregate the different types of effects. This approach offers more guarantees for consistency than a pure indicator approach. Indicators are, of course, informative, but are usually insufficient in making reasoned trade-offs between objectives.

Two more issues were raised when our assessment methodology was put into operation. The first is the use of local and global stocks of man-made or natural capital. Variations in these stocks need to be measured and included in the assessment of alternatives, and local stocks (historic buildings, rivers, etc.) will carry a much higher weight than global stocks (CO_2 emissions, biodiversity). The second issue is the role of emigration and immigration into the urban area. A city that is sustainable will tend to attract more inhabitants, and this can create a dilemma as this may come down to simple shift of welfare between two areas. It will also create immigration conflicts as the initially better off urban area will tend to block the entry of newcomers that also want to benefit from the more sustainable city.

The methodology is operational and promises to generate new insights into urban sustainability strategies.

Part III

Integration of Transport and Land Use Models

Chapter 16

Agile modeling: adapting UrbanSim to the European context using the Open Platform for Urban Simulation

Paul Waddell, Liming Wang, Hana Ševčíková, and Alan Borning

16.1 Introduction

The SustainCity project was proposed to enable researchers dispersed across several universities throughout Europe to collaborate in the development and application of large-scale urban simulation models to three European cities. The project proposed to use a common software and modeling system, UrbanSim, in order to simulate alternative policy strategies for urban sustainability. The proposal emphasized the challenges of adapting UrbanSim to the European context, due to the many differences in historical, political, cultural, and economic circumstances as compared to North American cities for which UrbanSim has been more widely developed. The expectations among the research team at the beginning of the project were that such substantial adaptations would be needed, that an entirely new version of UrbanSim would be needed, and hence the name UrbanSimE was proposed to describe the European version. In fact, this expectation proved to be quite incorrect.

The role of this chapter in the current volume is to present in a coherent way the software platform that forms the basis of the applied work discussed in the applications of UrbanSim to Brussels, Paris, and Zürich elsewhere in this book. This chapter explores the Open Platform for Urban Simulation (OPUS) in which UrbanSim is implemented, and describes how this modular and adaptable software architecture enabled the adaptations needed for this project. The adaptions require no changes to be made to the core of the software and only very localized changes to user configurations, and a few

modest changes in user packages for their applications. This follows similar experiences in adapting UrbanSim to be used in metropolitan areas in North America, Europe, Asia, and Africa, and confirms that the platform is agile enough to be rapidly adapted to different circumstances.

OPUS is an open source software platform for *agile modeling* of dynamic urban and environmental systems. This chapter is intended to provide an introduction to OPUS for researchers interested in urban spatial and environmental modeling, and in the software tools with which to do so. OPUS emerged as a generalization of earlier software development efforts in support of the UrbanSim model system (Waddell 2002, Noth et al. 2003, Waddell et al. 2003). Its overarching goal is to create a robust, modular, and extensible open source system for developing and applying model components and integrated model systems that can represent complex urban and regional spatial dynamics, including real estate markets, transportation, emissions, resource use, and land cover change. The OPUS system is intended to provide a common platform to facilitate the growth and development of the integrated urban and environmental modeling community, just as the R project (Ihaka & Gentleman 1996) has supported the development of the statistical computing community, the SWARM project has facilitated the growth of the Agent Based Modeling community (Minar et al. 1996), the Biogeme software (Bierlaire 2003) has broadened access to discrete choice modeling, and the PySAL project (Anselin & Rey 2007) is beginning to support the area of spatial econometrics. In keeping with the spirit of open source software development, OPUS avoids re-inventing existing functionality, leveraging these related open source software libraries to maximize the productivity of model developers and users. A preliminary description of OPUS is provided in Waddell et al. (2005); this chapter is the first full publication describing the OPUS platform and its use.

Model systems for integrated assessment of human-natural systems (Alberti & Waddell 2000), and for land use-transportation-emissions modeling (Dowling et al. 2005), are emerging rapidly. These integrated model systems are designed to analyze the complex interactions among urban and environmental systems in order to better assess the social, economic, and environmental impacts of urbanization and to design and evaluate public policies to address them. The challenges of developing useful model systems in these domains are substantial (Waddell 2011), not the least of which is the need for a robust and widely available software platform that contains the necessary functionality to develop models, estimate their parameters, and integrate the models and the data they use into a dynamic simulation system that includes interactions among many urban and environmental processes. OPUS was developed to make it much easier to iteratively build and refine such systems.

We use the term *agile modeling* to describe a methodology that supports quick and fluid iteration among all the different phases of modeling, including data preparation, model creation and specification, scenario construction and

execution, result analysis, and interaction with policy stakeholders (Kriplean et al. 2010). A key element in supporting this methodology is a software platform that facilitates many of these varied tasks and their coordination. There are software platforms, both open source and proprietary, that partially meet these needs, but none focus on the kinds of spatial and microsimulation models we wish to support. The closest platform to meeting our needs would be R, which has many user-contributed packages for spatial models and discrete choice models. We also note the availability of Biogeme and PySAL (described earlier in this section as well). However, these platforms provide tools to estimate the parameters of individual models of different types, whereas we also want to combine multiple models, and efficiently manage simulations with the resulting model system. There are other platforms that have been developed for agile modeling within specific modeling frameworks, such as SWARM (Luna & Stefannson 2000) and numerous derivative platforms for Agent-based Models (ABM). There are several integrated model systems for the land use and transportation modeling domain, such as ITLUP (Putman 1983), TRANUS (de la Barra 1989), MEPLAN (Echenique et al. 1990). In the domain of demographic microsimulation, full integrated simulation frameworks include DYNAMOD (Abello et al. 2002), and SOCSIM (Hammel & Mason 1990), and at least one platform that can be described as an agile modeling platform for this domain: LifePaths (Gribble 2000). None of these existing platforms, however, covers the full scope of our objectives outlined below.

The design goals for the OPUS platform are listed below.

- Provide a comprehensive framework to support the development of integrated urban and environmental models, supporting the entire iterative life-cycle of model development and application for high-performance simulation using large datasets.
- Provide efficient methods for implementing microsimulation models using individual agents, such as households and persons, even for large regions.
- Make model development and testing as agile as possible, through modularity, inheritance, templates, and ease of use.
- Make model development more accessible by reducing the need for modelers to have extensive software development expertise, with model templates, an intuitive GUI, and a powerful domain-specific expression language.
- Make it easy for others to extend the system by offering a simple Application Programming Interface (API) for contributed packages.
- Keep the system open and widely accessible by building on platform-independent, open source libraries.

The remainder of this chapter is organized as follows. We start with describing the OPUS architecture in Section 16.2. This description includes a higher level discussion of OPUS' capabilities for using and manipulating data,

implementing models, combining models into a model system, defining variables and expressions, and creating indicators. In Section 16.3 we discuss OPUS packages that extend the basic OPUS core package, such as a graphical user interface and the UrbanSim package. Section 16.4 describes the adaptions of UrbanSim for the SustainCity applications. Section 16.5 summarizes the current status of the platform as well as plans for further development.

16.2 OPUS architecture

OPUS is implemented in the Python programming language. This high-level programming language has become widely popular in the modeling and GIS communities, due to its relatively rapid learning curve, development productivity, and the steadily increasing availability of specialized open-source libraries. OPUS leverages numerous Python libraries, but central to its computational efficiency is the numpy Python library, which provides fast multi-dimensional array operations. In addition to leveraging Python libraries, OPUS also contains interfaces to external packages, for example, in addition to several predefined R-based methods, users have access to all R functions via the Python package rpy2. Python is well-known as an excellent 'glue' language, making it relatively simple to connect software systems written in other languages, such as C, C++, or Fortran, with Python programs.

The core design for OPUS is object-oriented and highly modular, allowing users to easily compose models from small pre-defined modules and, using its inheritance feature, to implement their own modules by overwriting only the necessary part of an existing module without having to implement the whole model.

One of the main goals of OPUS is to support *agile modeling* (Sect. 16.1). From our experience, work flow in a dynamic simulation system includes iterating between analyzing available data, configuring models reflecting processes of the system, estimating model parameters and selecting a preferred specification; then running these models to simulate the evolution of the system over a period of time and analyzing the results, often including the creation of a variety of indicators and their visualization. OPUS provides tools supporting modelers in each of these phases, each of which will be discussed in the following sections. Figure 16.1 shows a typical work flow with OPUS. In practice, the process inevitably requires iteration among the steps and experimentation within each step.

Figure 16.2 shows a collection of main OPUS classes structured into groups that roughly correspond to such work flow: classes for developing and analyzing data, classes supporting developing and specifying models, and a miscellaneous group that supports the creation of an OPUS package and interactive explo-

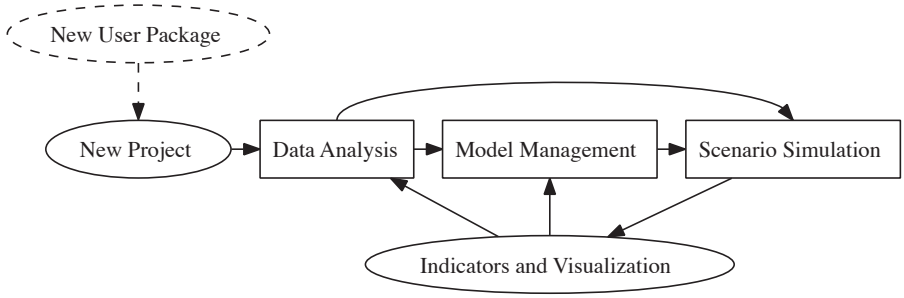


Figure 16.1 OPUS work flow. Model management includes creating, configuring and estimating models; Scenario simulation includes configuring and running a system of models.

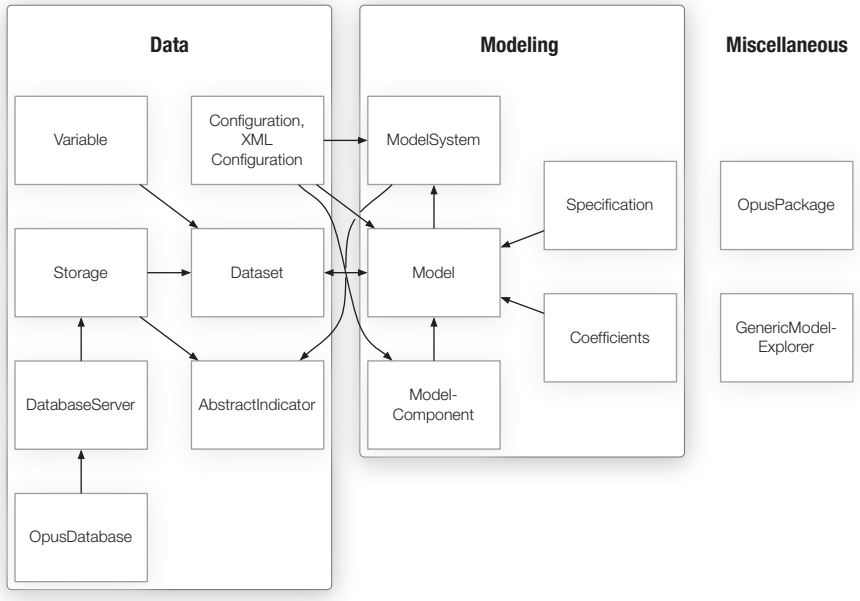


Figure 16.2 OPUS Architecture.

ration of models and data. Object-oriented and modular design allow modelers to experiment with and customize modeling system to fit their needs.

The basic unit of organization in OPUS is a *package*. An OPUS package is a standard Python package that conforms to a few additional requirements, mainly certain directories and files that define a standard API within OPUS. The most general functionality of OPUS is implemented in a package called

opus_core, which is the focus of this chapter. It provides general modeling functionality applicable to a variety of domains. Section 16.3 describes several extensions to *opus_core*, namely OPUS packages that are domain-specific.

16.2.1 Developing, analyzing, and organizing data

Data is both the beginning and end of modeling. It is a necessary pre requisite to begin building a model, and if a model produces new data, such as predictions, again, data analysis must be performed at the end of the modeling process in order to validate or use the results of the model.

Data is encapsulated in OPUS by a Python class called Dataset. A dataset is a collection of attributes for a particular type of entity or object, such as a set of counties, or a set of persons, or a set of buildings. Each member of this set has the same collections of attributes, such as age of persons or size of buildings. Conceptually, a dataset can be considered to be an $n \times m$ table where n is the number of entries and m is the number of attributes. Attributes can be read from a disk-based data store, which we refer to as *primary attributes*, or, computed on-the-fly with OPUS variables and expressions or models, which we call *computed attributes*.

When loaded, each dataset attribute is stored in memory as a numpy array, which allows highly efficient computations by performing any arithmetic operations on all elements of the array at once. This feature overcomes a well-known performance bottleneck in Python: iterating over an array or a list of elements.

OPUS Dataset class allows users to read from and write to different types of data stores, depending on their needs, while keeping a consistent internal representation of the data. Each type of data stores is implemented as a subclass of the Storage class. There are several pre-defined subclasses of Storage, while new ones can be easily incorporated into the system. Currently, users can choose from the following types of storage:

- ASCII, tab or comma delimited (CSV) files.
- SQL storage, which supports various relational database systems such as MySQL and Postgres.
- Dbase formatted files.
- ESRI formatted files.
- Numpy (Python) arrays on disk.
- In memory Python dictionaries.

Each dataset object contains a pointer to an object of the particular storage class, and any I/O operations on the dataset are delegated to this object. While a dataset object defaults to read and write from the same type of data storage, it can also use one storage object for reading data and another storage object

for writing. Thus, one can use the Dataset class for converting data between formats. Because the format conversion function is so commonly used, we provide tool scripts that wraps the data storage class to help users convert data between most of the data types.

The OPUS Dataset class also provides a rich set of functions for data manipulation and analysis, including adding and removing attributes, computing data descriptives (summary, correlation), joining, grouping, plotting (scatter plot, histogram, correlation image) or mapping (for raster data). In combination with variables and expressions, Dataset class provides the most commonly used functions in data development and analysis.

While individual Dataset objects can be instantiated from various sources, OPUS organizes multiple datasets used in a project into a *cache directory* structure in a binary format so that the system has fast access to all data without having to keep everything in memory. This also allows models to output datasets into a cache for use by other models. For each time increment in a simulation (e.g. a year), a new subdirectory is created in order to record results of each iteration. Thus, at the end of a simulation, the cache directory contains as many subdirectories as there were iterations, each of them containing datasets from that particular time point.

The architecture allows each model to use the cache location for collecting its own inputs. Moreover, model outputs can be again stored in the same location and used by other models as their inputs.

Each dataset is stored in a directory with the same name as the table name from which the data were imported, e.g. `households`. Each attribute of a dataset is a numpy binary file located in this directory, e.g. `income` and `parcel_id` in the example in a household dataset with income and parcel_id attribute.

Developing data for a project begins with creating a cache directory that contains all necessary datasets for the base year (a base year cache), OPUS provides several scripts for converting data from other formats into a cache type data. At the beginning of a simulation, the base year cache is copied into a given *run cache* directory as a sub-folder labeled with the base year, i.e. the start time of the simulation. The model system creates a new sub-folder for each time increment, or year, of the simulation. For example, after two time periods of running a model system that uses only the household dataset from the above example, starting in year 2005, the run cache will contain the following directories:

```
run_xx_time/
           2005/
                households/
                       household_id.li4
                       income.li4
                       parcel_id.li4
```

```
2006/
     households/
            household_id.li4
            income.li4
            parcel_id.li4
2007/
     households/
            household_id.li4
            income.li4
            parcel_id.li4
```

16.2.2 Implementing models

Models in OPUS are implemented in a highly modular way. Researchers should have the flexibility to experiment with different models by exchanging parts of a model using either ready-to-use model components or components developed by the modelers themselves. The collection of the ready-to-use model components can be easily extended as the need arises. OPUS also offers a functionality for managing a group of models combined into a model system that can be used for simulating over a period of time (discussed in more detail in Section 16.2.3).

There is only one requirement for a piece of Python code to be an OPUS model, namely that it should be defined as a class that inherits from the Model class. Additionally, if an OPUS model is to be combined in a model system, then it must have a method called 'run' which contains the functionality that is to be invoked in each simulated time period. Optionally, if the model is to be estimated for model parameters, the estimation functionality is implemented in a method called 'estimate'. There are no further requirements for an OPUS model.

To make the process of implementing models easier, OPUS offers a few commonly used models as well as various plug-in model components (Fig. 16.3). Table 16.1 gives a brief description of these models. The regression, choice, and hierarchical choice models all have a simulation component and an estimation component. Users can choose from an existing set of such components, or implement their own. In the following, we briefly describe the main functionality of the two choice models.

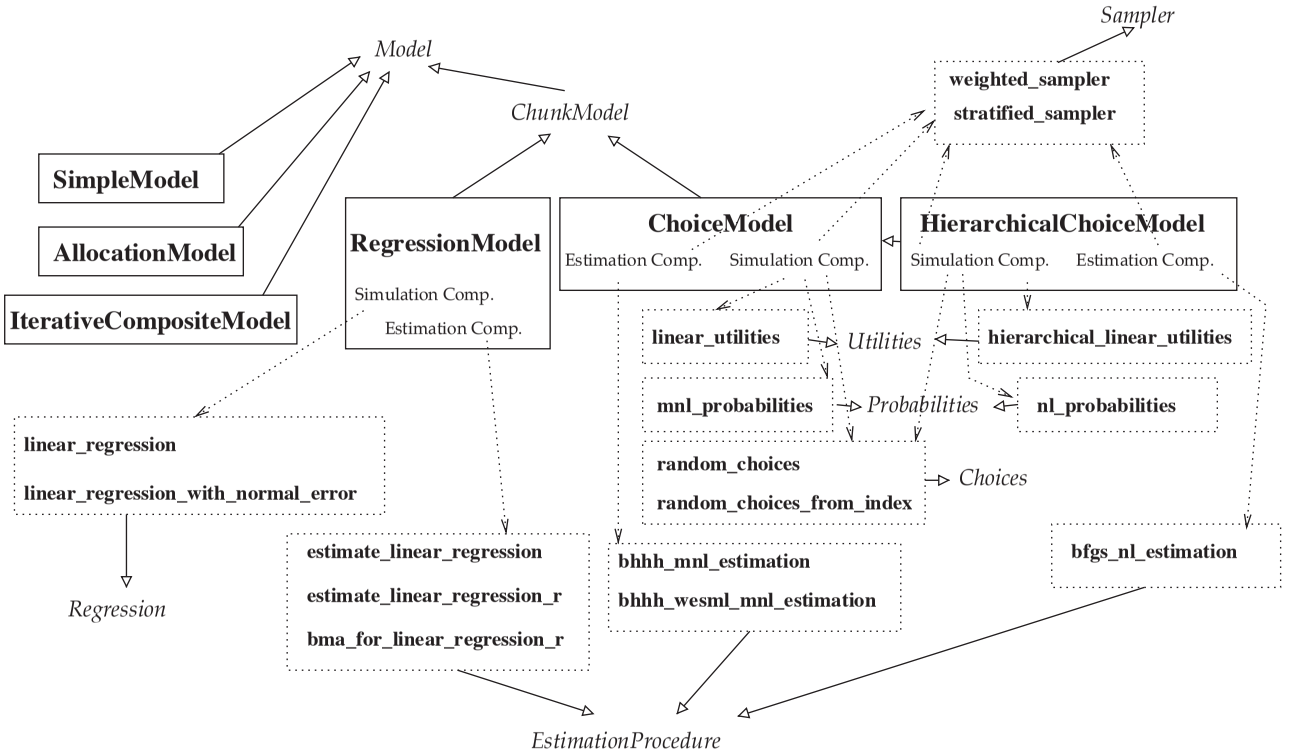


Figure 16.3 OPUS models and model components. The solid arrows show inheritance relationships; the dotted arrows show optional plug-in relationships. Abstract models are shown in italic font. Dotted boxes group components that have the same parent. mnl - multinomial logit; nl - nested logit; bma - Bayesian Model Averaging.

Model/Comp. name	Description	Example
Model	Generic class that is the superclass of all models. Includes some logging and timing features.	Any of the following models.
Chunk Model	Separates the observations into chunks (possibly in a random order) and runs the model by chunk.	Useful if computer memory is limited and the number of observations is too large for array operations as a whole.
Simple Model	Updates a dataset attribute with results from an expression.	An aging model: updates the age attribute of persons with age+1.
Allocation Model	Allocates an aggregated quantity according to weights while meeting capacity restrictions.	Proportionally distribute visitor population to zones based on the number of hotel rooms in each zone.
Iterative Composite Model	Iterates a set of given models until a given condition is reached; provides a capacity to configure two models that are inter-dependent and are to iterate until convergence.	Coupling a household location choice model with a housing price adjustment model; the two models iterate as an iterative composite model until markets clears.

Regression Model
 	
`estimate_linear_regression_r`
`bma_for_linear_regression_r`
	Framework for combining modules to estimate and simulate a multiple regression model.
	Uses R functions to estimate a linear model.
	Performs a variable selection using the R package BMA (Raftery et al. 1997).
		Real Estate Price model.

Choice Model	Framework for combining modules to estimate and simulate a discrete choice model (page 319).	Establishment location choice model.
Hierarchical Choice Model	Framework for combining modules to estimate and simulate a nested discrete choice model (page 319).	Household location choice model that nests household's staying or moving decision with location choices.

Samplers
 	
`weighted_sampler`
`stratified_sampler`
	Abstract class that groups modules used to draw samples from a dataset.
	Sample observations using weights.
	Sample observations using strata.
		As part of a choice model for sampling alternatives.

Table 16.1 Description of OPUS models and model components.

The choice models are used for constructing binary, multinomial, and nested logit models. The theoretical foundation for this approach is the Random Utility Model introduced by Daniel McFadden (McFadden 1974, 1981), which postulates that agents make choices to maximize their utility, and that their utility is derived from the underlying attributes of choice alternatives. Here, the simulation component consists of three parts: computing utilities, computing probabilities, and making choices. Each of those parts can be again either selected from predefined modules (such as `linear_utilities`, `mnl_probabilities`, `random_choices`), or user-defined modules can be easily plugged in. The `linear_utilities` component computes the systematic part of the utility under the assumption that it is given by a linear combination of the coefficients and predictors. The module `mnl_probabilities` computes probabilities (given utilities) using the multinomial logit formula, a widely used theory that assumes that the random part of the utility is independent identically Gumbel-distributed. Given probabilities, the module `random_choices` selects randomly a choice for each agent. To estimate the choice model, one can choose the module `bhhh_mnl_estimation` which implements the Berndt-Hall-Hall-Hausman maximum likelihood algorithm (BHHH) (Berndt et al. 1974). An extension to it is implemented in the `bhhh_wesml_mnl_estimation` module, namely weighted endogenous sampling maximum likelihood (Manski & Lerman 1977), which weights observations before passing them to BHHH, to take into account undersampled or oversampled observations.

The Hierarchical Choice Model is constructed in a similar way as the Choice Model, and is used for applications with hierarchical structure such as the nested logit models. In this case the utilities and probabilities can be computed using the modules `hierarchical_linear_utilities` and `nl_probabilities`, respectively. The model can be estimated using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimizations method for nested logit implemented in the `bfgs_nl_estimation` module.

All the modules described in this section and in Table 16.1 are heavily based on numpy's efficient matrix operations, and thus the computation is done for all observations (or chunks of observations) at once. Also note in Figure 16.3 that there is no strict line between models and model components. In fact, all root nodes in the figure, such as Model, Sampler, Regression, and Utilities, are implemented as children of the ModelComponent class (not shown in the figure).

As mentioned above, we define a model being an OPUS model if it is a child of the class Model and defines a method called 'run'. This method should implement a simulation of the model. Optionally, OPUS models can also have a method 'estimate' that implements an estimation of model parameters. This framework allows estimation and simulation to be performed on the same instance of a model.

If a model is to be estimated, its specification can be managed by the classes EquationSpecification and Coefficients, as for the regression and choice models. From a given specification object, the 'estimate' method estimates the parameters and creates a corresponding Coefficients object, which can be passed to the 'run' method. OPUS also supports saving estimation results for later use in a simulation session.

A model can have several sub-models, that is, the same piece of code (defined by the 'run' or 'estimate' method) is applied to different groups of dataset members separately. This allows for the same model to differentiate specifications by different groups of agents without having to take special care of the specific cases – the model applies it automatically. The EquationSpecification and Coefficients classes support this kind of grouping. In addition, in case of the choice models, the specification can distinguish between different equations or nests.

16.2.3 Configuring a model system and managing runs

In OPUS, one can combine a set of models into a model system that runs iteratively for a given time period. For example, a model system can be configured to run a set of models annually for a total of 30 years. Each model and the relationships between models can be configured using a Python dictionary object (supported by the class Configuration, as shown in Figure 16.2) or using an XML format (supported by the class XMLConfiguration). In the configuration, not only running sequence and dependencies between models can be specified, but outputs of models can be configured as inputs of another model that runs later in the running sequence.

Once a configuration is created, an estimation or a simulation of the model(s) included in the configuration can be launched via command line tool scripts or through the OPUS Graphic User Interface extension. In a model estimation process, the estimated model coefficients is displayed on the screen and, when becoming satisfactory, saved to the cache directory for use in simulation process; while in a simulation process, the model system writes any dataset with changed attribute values into the cache directory.

To help manage potentially large number of simulation runs in test and production stage, there is a tool for managing runs (start, kill, and restart a run) and archiving meta information of each run in a SQL database (including description, time stamp, host name, and information needed to restart the run). The run management system has been extended to manage runs across a cluster of different physical computers.

16.2.4 Variables and the expression language

Usually, the process of creating new models, calibrating, estimating and results evaluation involves considerable investigation and experimentation using

different model variables that describe attributes of actors, processes and geographies of the simulated environment. In many cases, the original variables must be transformed or combined to create new variables that are more suitable for the analysis.

As mentioned in Section 16.2.1, a variable (or computed attribute) is considered to be an additional attribute of a dataset, and thus is represented by a numpy array of the same size as number of records in the dataset. Its computation is invoked by the Dataset method `compute_variables`(*variable_name*, ...) where the syntax of *variable_name* depends on the definition of the variable. There are two ways to define variables in OPUS:

1. As a Python module.
2. Using a domain-specific expression language.

In the former case, the variable is implemented as a class in a Python module named variable_name located in a directory of the same name as the dataset name. It is a child of the class Variable and has a required method called 'compute' that contains the actual computation. The only requirements for this method is that it returns a numpy array of the same size as the dataset to which the variable belongs to. An optional method named 'dependencies' should return all (primary and/or computed) attributes that this variable depends on. The OPUS framework uses lazy evaluation for variable computation: the 'compute' method is only invoked if either the variable has not been computed before, or if the values of one or more of the dependent variables have been updated. The system maintains a version number for each variable to keep track of the bookkeeping for this. A variable defined this way will be referenced by a name composed from its actual name, its dataset name and the particular OPUS package. For example, if a user working in an OPUS package called 'my_package' implements a variable called 'my_var' for the dataset class 'my_dataset', the computation of this variable can be invoked by passing the expression as an argument to the compute_variables method (ds is an object of my_dataset class):

```
ds.compute_variables(['my_package.my_dataset.my_var'], ...)
```

OPUS includes a domain-specific programming language that allows users to compose new variables from existing ones with an expression and eliminates the need to code them as a Python module. An expression consists of the name of one or more existing variables, or a function or algorithmic operation applied to other expressions. The expression can be recursive, so that an expression can be composed from other expressions. Unitary functions available in numpy, such as `exp` and `sqrt`, , and numpy operators, including `+ - * / ** < > ==`, can be applied to OPUS expressions. Two important methods provide special operators beyond those in numpy for aggregating and disaggregating variable values over two or more datasets. The 'aggregate'

method associates information from one dataset to another for a many-to-one relationship, while the 'disaggregate' method does the same for a one-to-many relationship. Functions for aggregation include taking the sum, the mean, the min, or the max of the values being aggregated. Analogously, the 'disaggregate' method takes information from a coarse set of entities and allocates it to a finer set of entities. Aggregation and disaggregation can be done over multiple levels of geography or other agents in a single expression, for example aggregating the number of jobs from buildings, up to parcels, then up to zones, then to cities.

In order to work with variables that describe the interaction between two datasets for choice models, OPUS includes a subclass of Dataset called InteractionDataset. Attributes of this class are stored as two-dimensional arrays. A full range of operators and functions are available for these 2-D interaction arrays as well. For example, the expression

```
income_x_cost = ln(household.income) * zone.average_housing_cost
```

returns an $n \times m$ array, where n is the number of households and m is the number of zones. The $(i, j)^{th}$ element of the array is the log of the i^{th} household's income multiplied by the average housing cost for the j^{th} zone.

Expressions can be passed directly to the 'compute_variables' method or to model specifications. They can also be collected into an 'aliases.py' file placed in the same directory as other variables implemented as Python modules; then the user can refer to the variable simply by its alias which can be helpful in case of reusing definitions.

See Borning et al. (2008) and the *OPUS/UrbanSim Users Guide and Reference Manual* (2011) for more details on variables and the expression language.

Finally, the time-incremental structure of the cache directory described in Section 16.2.1 supports OPUS's framework for working with time series variables. Such variables are defined by simply adding the suffix '_lagx' to an existing variable name, where x is the number of time increments to move backwards when computing the specific variable. For example, say the computation of the 'my_var' from the beginning of this section is a part of a model system that simulates for five years, from 2010 to 2015. At the end of the simulation we can access values of 'my_var' from 2014 by invoking a computation of 'my_var_lag1' or from 2013 using 'my_var_lag2'. Similarly, we could use 'income_x_cost_lag1' to access values from the previous time period of the interaction expression above.

16.2.5 Indicator framework

One way to analyze data in the OPUS environment is via the indicator framework. It is designed to help analyze simulation results from running a model system as described in Section 16.2.3 and diagnose possible data and modeling anomalies. It operates on cache directories, i.e., data stored in the binary storage format.

The basic idea is to create a batch of indicator definitions (using variable names and the expression language), along with a pointer to the cache directory, the time point of interest, and the output format, one per each indicator. Available output formats are a tab or comma delimited ASCII file, map, chart, DBase table, Geotiff, or Arcgeotiff maps. The whole batch is processed and the results can be viewed through an HTML interface.

In addition to computing simple indicators, the framework also supports change indicators. These are quantities computed as a difference in indicator values between two time periods or between outcomes of different simulations.

16.3 Extending OPUS

Building on the *opus_core* package, OPUS can be extended in a modular way by adding other packages that import functionality from *opus_core* and adopt a standardized API for defining new OPUS packages. Interfaces to completely separate software systems, models, and tools can also be incorporated.

16.3.1 Creating an OPUS package

Users who wish to create new models, define new datasets or implement new variable definitions will need to create an OPUS package. The package is considered as a container for any new user-written Python code. Note that an OPUS package is not needed by users who simply wish to reuse existing models or construct model specifications using existing datasets, variables or by creating new expressions. In such a case, an XML configuration is sufficient to define the models and their specifications.

An OPUS package is a Python package conforming to OPUS conventions that extends and customizes existing OPUS functionality. By default, the package is placed in an OPUS workspace, alongside other OPUS packages such as *opus_core* and *urbansim* (see Section 16.3.2). The following Python code creates a skeleton of an OPUS package called 'atlantis' pre-populated with the required folder structure and files:

```
from opus_core.opus_package import create_package
create_package('c:/opusworkspace', 'atlantis')
```

16.3.2 UrbanSim

UrbanSim (Waddell 2002) is implemented as an OPUS package, called *urbansim*, bundling several land use models into one model system. These are models for predicting land prices, simulating real estate development, predicting increase and decrease in population of jobs and households, and simulating choices of moving jobs and households of new locations. The models are implemented as subclasses of the existing models in *opus_core*, as described in Section 16.2.2. For example, the LandPriceModel is a child of RegressionModel, and the EmploymentLocationChoiceModel and HouseholdLocationChoiceModel are both subclasses of the ChoiceModel class, supporting sampling agents using the *opus_core* sampling components. The *urbansim* package extends the collection of model components that can be plugged into models. For example, as an alternative to `random_choices` simulating agents making random choices (Figure 16.3), *urbansim* implements a component called `lottery_choices`, which takes into account capacity of locations and allows agents to re-decide if their choice is not available. This component is by default plugged into both the employment and household location choice models. (In general, there are ready-to-use configurations of all models containing useful default settings.)

The package pre-defines many useful variables for different geographies and different types of agents. These can be used as predictors for UrbanSim models or as indicators.

UrbanSim can be further customized to run at different geography levels: *urbansim_gridcell*, *urbansim_parcel*, and *urbansim_zone* are OPUS packages that extend and customize *urbansim* to run at the gridcell, parcel, and zone level respectively. Each package can override geographies defined in *urbansim* or type of agents being modeled, as well as the set of models being included in the model system.

16.3.3 Graphical User Interface (opus_gui)

A Graphical User Interface (GUI) for *opus_core* and *urbansim* is implemented as an OPUS package called *opus_gui*. It uses the Qt4 library and its Python wrapper pyQt, and exposes most of the underlying functions in a user-friendly interface. The GUI can be launched using the Python script `opus_gui/opus.py`.

The GUI is built around XML configurations and is organized to facilitate the work flow of data management, model management, scenario management, and results management. The main window of the GUI reflects this work flow, including four tabs in the left-hand panel of the main window labeled *Data*, *Models*, *Scenarios*, and *Results*, plus a *General* tab. Each of the four tabs provides a container for configuring and running a variety of tasks, organized

into the main functional areas involved in developing and using a simulation model.

- The *Data* tab organizes the processes related to moving data between the OPUS environment and, doing data processing both within OPUS, and also remotely in a database or GIS environment. OPUS can use Python to pass data and commands to a database system like Postgres or MS SQL Server, or to a GIS system like ArcGIS or PostGIS. Tasks can be organized in the Data Manager as scripts, and run as a batch, or alternatively, they may be run interactively.

- The *Models* tab organizes the work of developing, configuring, and estimating the parameters of models, and of combining models into a model system. Within this tab, new models can be configured from a template; arguments to individual models can be edited; specification of models whose coefficients need to be estimated from data can be entered and estimated.

- The *Scenarios* tab organizes the tasks related to configuring a scenario of input assumptions, and to interact with a run management system to actually run simulations on scenarios. Tools to monitor and interact with a running simulation are provided here.

- The *Results* tab provides the tools to explore results using indicators and visualization methods. It uses OPUS indicator framework to generate a variety of indicators within GUI for both diagnostic and evaluation purposes. It also provides functionality to visualize indicators as charts, maps, and tables, and to export results to other formats for use outside of OPUS.

The GUI can load an existing project file or create a new one, for example from a project template. A project file is an XML file containing configuration information. More details regarding the GUI can be found in (Kriplean et al. 2010) and in the *OPUS/UrbanSim Users Guide and Reference Manual* (2011).

16.3.4 Calibrating and assessing uncertainty

In large-scale microsimulation that involves stochastic, or random, components, it is clearly important to assess the uncertainty in model predictions. (It is important for other model types as well, but is often ignored by modelers.) A method for assessing uncertainty in urban simulations, called Bayesian Melding (Raftery et al. 1995, Poole & Raftery 2000), has been implemented in OPUS (Ševčíková et al. 2007, 2011). The idea is to draw model inputs (such as coefficients) from a prior distribution and run a Monte Carlo simulation with

these inputs. The simulation results at a specific time point are compared to observed data which yield weights for the different runs, a model bias, and variances. These quantities are propagated forward and can be applied to simulation results at the prediction time point.

In order to draw coefficient values from normal distribution given by the estimates and standard errors of the coefficients, the Coefficients class has the method 'sample_values_from_normal_distribution'. One can for example construct a model system where different draws from estimated coefficients are passed to models and run a set of simulations for these draws.

Having results from a set of simulation runs, each run corresponding to one cache directory, one can investigate the results using the class MultipleRuns. It is initialized by pointing to the set of cache directories. One can then use for example the methods 'compute_values_from_multiple_runs' or 'get_datasets_from_multiple_runs' to explore results across runs, the latter allowing the use of any Dataset methods.

A child class of MultipleRuns, called Bayesian Melding, supports passing observed data (in any storage type) and assessing uncertainty via a likelihood function. The likelihood function is implemented as a model component that is passed to the 'compute_weights' method, and thus it is easily exchangeable. By default, the likelihood is computed using a mixture of normal distributions. The likelihood for each run is proportional to its weight, which in principle evaluates that run with respect to the observed data. Thus, the highest weight implies that the run has the *best* inputs. This approach allows us to compare runs based not only on different coefficient values, but also runs that differ in other inputs, as well as in the set of models and their structure.

Furthermore, Bayesian Melding supports propagating the uncertainty assessed using observed data forward in time. This in turn allows the user to determine the theoretical posterior distribution of the outputs. Sampling from this posterior distribution is also possible, enabling the approximation of the posterior distribution of aggregated quantities.

16.4 Adaptations for SustainCity applications

16.4.1 Coupling of UrbanSim and Matsim

UrbanSim has been coupled to MATSim using a standard approach described above for creating packages in OPUS. It is integrated into the UrbanSim source code repository in the Subversion version control system, under opus/src/opus_matsim, in the same way that UrbanSim is stored as an OPUS package under opus/src/urbansim. Just as UrbanSim has been interfaced with many existing commercial travel modeling platforms such as Cube, TransCad, VISUM, and Emme, the same approach is used here to enable OPUS to run

UrbanSim on an annual basis, then on specific simulation years set in a user configuration and accessible via xml, OPUS writes out the data needed for the travel model, launches the travel model, and waits for the output from the travel model to be produced. Once the travel model is completed, OPUS loads this data into the UrbanSim cache, and restarts UrbanSim to run for the number of years specified in the configuration until the next travel model run is scheduled. Results from the travel model are generally in the form of zone to zone skims from assignment, but may take various forms, such as congested times on the travel network links, from which UrbanSim can compute accessibility measures.

A general diagram of this interface is shown in Figure 16.4. Full documentation on the interface is available from

`http://matsim.org/extensions/matsim4urbansim.`

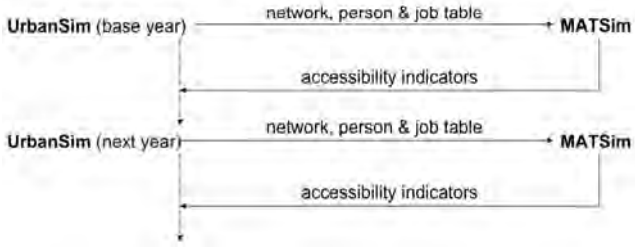


Figure 16.4 MATSim-UrbanSim Interface data flow.

16.4.2 Firmography models

Instead of modeling individual jobs on the employment market as it has commonly been done in UrbanSim projects, the Paris project models employment using establishment (firm) as the basic unit of analysis, as has also been done in applications of UrbanSim in San Francisco, Albuquerque, and Denver. A set of firmographic models predicts the birth, evolution (growth or shrinkage of employment), death, and location choice of establishments.

- Establishment disappearance model: a binary choice model predicting the probability of death of existing establishments.
- Establishment evolution model: a semi-log linear regression model predicting the growth or shrinkage of employment for existing establishments.
- Establishment transition model: a transition model sampling establishments to meet specified annual employment control totals by sector. (If

the sum of employment in existing establishments of a sector is less than the employment for the sector in control totals, the transition model first does a sample without replacement from establishments that have been flagged as disappeared by establishment disappearance model; if the sum of employment in establishments still falls short, the transition model then samples with replacement from existing establishments of the sector until the sum of employment meets the target. If the sum of employment in existing establishment of a sector is greater than the control total of the sector, the transition model samples existing establishments and marks them as disappeared until the sum of employment equals the control total of the sector.)

- Establishment location choice model: a multinomial logit model predicting the location (commune) choice of new establishments.

16.4.3 Calibration of establishment location choice model to meet targets

After estimating the establishment location choice model using the base year (1999) data, constants specific to zonal group and sector in the establishment location choice model specification was calibrated to match the observed employment targets at zonal group level of 2005. The manual calibration process is tedious and requires many iterations of simulation runs and estimation of new constant values. A calibration tool is being developed to automate the process.

16.4.4 External demographic model

An external demographic model was created in UrbanSim to load and process 30-year demographic simulation output from the INED demographic model. The model updates demographic characteristics of households and persons in UrbanSim with evolved demographic characteristics predicted by an external demographic model. In the case of the Paris and Zürich project, the external demographic model is developed by the Institut National d'Etudes Démographiques (INED). The INED model simulates demographic events including mortality, fertility, household union formation and dissolution, non-private residence, education, labor force participation, and migration. UrbanSim is able to detect demographic changes in households (e.g. with new born children), residence and jobs vacated (e.g. because of deceased households or retired workers), and so on, and handle them properly in quantity accounting (e.g. occupied residential units) and in models such as household relocation model and household location choice model. To handle the large volume of data from demographic output and to allow fast merging of demographic re-

sults and UrbanSim dataset, the external demographic model uses h5py for efficient data processing. The model also allows flexible configuration to accommodate different demographic characteristics used in the Paris and Zürich projects.

16.4.5 Residential real estate supply modeled at building and unit level

Since the Zürich application has detailed per unit data for residential buildings, it utilized this infromation and customized the data and configuration to model the residential real estate market at the detailed living unit level. Most of this customization does not require any change in code: the real estate price model, household location choice model, and residential building location choice model are all configured in XML project files from shipped UrbanSim models; they only needed to code their own real estate transition model that can handle both buildings and units simultaneously.

16.5 Status and further development

The OPUS platform is now relatively feature-complete for the initial objectives outlined at the beginning of this chapter. It has been used to re-implement the UrbanSim model system (Waddell 2002, Noth et al. 2003, Waddell et al. 2003, Waddell 2011), resulting in faster execution than the previous Java platform, and dramatically increased ease of model specification, estimation and simulation – benefits of the increased modularity of the platform, the integration of model estimation algorithms, and the incorporation of a high-level domain-specific expression language to make it much easier to create new variables without Python programming (Borning et al. 2008). Laboratory studies and interviews provided tangible evidence of these and other benefits to users from the development of a Graphical User Interface for OPUS and UrbanSim (Kriplean et al. 2010), as well as supporting the utility of an agile modeling methodology. The UrbanSim system, coupled with the OPUS platform, has been used by a rapidly growing list of researchers and planning agencies throughout the United States, Europe, Asia, and Africa. Applications using OPUS and UrbanSim have been published with empirical results for Detroit (Waddell et al. 2008), Paris (de Palma, Picard & Waddell 2007), Salt Lake City (Waddell, Ulfarsson, Franklin & Lobb 2007), San Francisco (Waddell et al. 2010), Seattle (Ševčíková et al. 2011), and several others, in addition to numerous unpublished applications.

The OPUS platform has also begun to be appropriated into environmental modeling domains as well. In the domain of landscape ecology, OPUS was used to develop and estimate a land cover change model for the Puget Sound region of Washington, and this in turn was used to assess impacts of

rapidly urbanizing environments on avian populations (Hepinstall et al. 2008). A second environmental application developed using the OPUS platform was a spatially-explicit model of water consumption, which also used predictions from UrbanSim as inputs to the water demand model (Polebitski et al. 2011).

Taken together, this considerable body of use, along with lab studies and interviews, provides substantial evidence of the utility of the OPUS platform. However, more remains to be done to make it even more productive for the urban and environmental geospatial modeling communities, and for researchers and planners more generally. Some specific areas that offer targets of opportunity for further development include the following:

- To make the Graphical User Interface more dynamic by using custom forms and wizards, while moving some of the clutter of declarative specifications embedded in XML out of the way (so that it is still available for those who wish to modify these settings, but does not confuse users who do not wish to do so).

- To develope a more robust mechanism for error trapping and handling of error messages for users. The heavy nesting of Python classes and calls from a Graphical User Interface can result in some very arcane Traceback messages that are daunting to many users.

- To experiment with alternative approaches in order to improve computational performance. Although performance is reasonably good for the current uses of the system, more interactive use (for example for interactively supporting deliberation and discussion of simulation alternatives) will require much faster execution. There are many opportunities here for the careful introduction of parallel processing methods.

- To create a well documented and tested Applications Programming Interface to allow programmers to extend the system and interface to OPUS platform through the API.

- To couple OPUS and UrbanSim to UrbanCanvas, a new 3D geometric analysis and visualization platform currently under development (see www.synthicity.com/products.html for details).

In closing, we reflect on 15 years of open source development with UrbanSim and OPUS, in which we have been motivated by optimism that the movement towards collaborative development efforts would gain momentum, and that platforms such as OPUS could contribute to making scientists, students, and practitioners more productive in collaboratively solving the urban and environmental problems that drive projects such as SustainCity. Whether this will actually occur is less clear. After one and a half decades of open source

development with UrbanSim, the degree of collaboration in the form of contribution to the code base and its improvement from outside the core team that developed it, has in reality been quite small. We will continue to explore avenues to increase collaboration within this domain, and to improve on what has become a robust platform to support agile modeling, as demonstrated by its adaptability to model European cities in the SustainCity project, and numerous adaptations to respond to varied challenges in integrated land use and transportation planning (Waddell 2011). Some of the advances proposed for model specification and estimation by Antoniou and Picard in Chapter 12, and the indicator framework proposed by Proost, Van der Loo, Antoniou and Efthymiou in Chapter 15, will inform further software innovation in the UrbanSim software platform. We anticipate that further efforts to streamline and simplify the software, better integration with visualization tools, and development of more robust support services will help further advance the use of integrated models to address important and challenging problems in urban planning. The roadmap for this effort is clear, and the efforts of projects such as SustainCity have provided valuable insights to further generalize UrbanSim and make it adaptable to diverse urban contexts.

Acknowledgements

This research has been funded in part by the National Science Foundation under Grants IIS-0705898 and IIS-0964412, by the SustainCity project, co–financed by the European Commission within the FP7 program, and by a number of planning agencies, including the Denver Regional Council of Governments, Puget Sound Regional Council, the Metropolitan Transportation Commission, and the Maricopa Associations of Governments. We also acknowledge the UrbanSim users in the SustainCity project who worked most closely with the software platform to adapt it to the case studies in Brussels, Paris, and Zürich, in alphabetical order, C. Antoniou, K. Axhausen, M. Bierlaire, D. Efthymiou, R. Hurtubia, K. Motamedi, K. Nagel, K. Müller, T. Nicolai, A. de Palma, N. Picard.

Chapter 17

Integration of agent-based transport and land use models

Thomas W. Nicolai, Kai Nagel

17.1 Introduction

The interaction between land use and transport plays an important role as mentioned in Chapter 2 of this handbook. Land use transport interaction (LUTI) models denote models that combine land use models and transport models with feedback mechanisms between them. Most urban simulation models such as UrbanSim, DELTA, CUFM, MUSSA, POLIS, or RURBAN are not modeling transport themselves; instead they rely on interaction with external transport models (Wegener 2004).

The present chapter focuses on the integration of the extensible, microscopic, agent-based urban land use model UrbanSim with the agent-based transport simulation model MATSim ("Multi-Agent Transport Simulation"; Balmer et al. 2005, Raney & Nagel 2006). To utilize the agent-based modeling technique, both models are directly linked at the person level. The goal is to provide improved access and accessibility indicators as feedback from the travel model in order to enhance UrbanSim sub-models reflecting the decisions of households, businesses, and developers.

A challenge for that software integration was that UrbanSim is written in Python and heavily using C/C++ libraries, while MATSim is written in Java. It is clear that many ways of coupling two pieces of software are technically feasible, but the question was which of these approaches might have a chance to remain robustly functional beyond the end of the project without additional maintenance. Thus, the first objective was to investigate how a *robust* coupling of the two software packages at the agent level could be achieved.

Travel models typically return zone-to-zone impedance matrices, containing, say, travel time or travel distance, to UrbanSim. A problem with zone-to-zone matrices is that they grow quadratically in the number of zones, which

for computational performance reasons puts a limit on the number of zones that can be used and thus on the spatial resolution. Thus, another objective was to investigate if other results of the travel model, in particular accessibility measures, could be computed and provided at the UrbanSim parcel level.

A final question was if the number of iterations, inherent in traffic assignment models and thus also in MATSim, could be reduced by re-using information from previous runs. This is called the "warm" and "hot" start capability.

This chapter first summarizes the simulation approach of MATSim (Sect. 17.2). Section 17.3 describes MATSim extensions that were used in the case studies. In Section 17.4 the integration approach is described, illustrating the feedback mechanism between both frameworks and describing the data requirements. The access and accessibility indicators provided by MATSim are explained in Section 17.5. Section 17.6 shows, as illustration, the accessibility consequences of a cordon toll for Brussels, obtained during work on the SustainCity Brussels case study. Computational issues are briefly discussed in Section 17.7. The chapter is concluded by a discussion (Sect. 17.8).

17.2 MATSim

MATSim (Balmer et al. 2005, Raney & Nagel 2006, Balmer et al. 2009) is a disaggregated, agent-based transport model that is designed to simulate several million travellers (agents) individually for large real-world transport scenarios. Advantages over traditional static assignment include time-dependent congestion, time-dependent mode choice, and the option to accelerate computations by running small samples. Compared to dynamic traffic assignment (DTA; Carey & Watling 2003), the main difference is that MATSim looks at full daily plans whereas DTA looks at trips. This gives MATSim more expressiveness with respect to individual replanning which goes beyond the trip, such as mode choice, moving the whole daily plan forward or backward in time, or activity location choice. In addition to this higher expressiveness, MATSim is often also considerably faster, at the expense of an often less realistic representation of traffic dynamics. MATSim has been applied to large scale scenarios in Zürich, Berlin, and many other cities (Balmer 2007).

17.2.1 MATSim process structure

The general MATSim process structure consists of the following parts (Fig. 17.1; Balmer et al. 2009):

Initial demand: MATSim requires the physical infrastructure, determined by the road network and facilities (i.e. activity locations like home, work, shopping or leisure) and the population including the demand of each individual person. The initial demand for each agent is usually generated

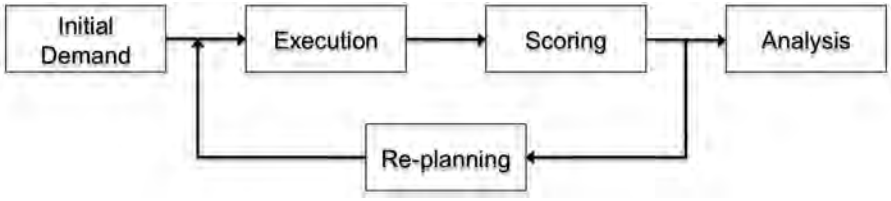


Figure 17.1 Process structure of MATSim.

based on microcensus and/or survey data. Initial mode choice is either based on data or heuristic assumptions; initial routes are usually the fastest paths on an empty network.

Iterative demand optimization: In an iterative demand optimization process the demand for each individual agent is improved. It takes into account physical constraints, e.g. the road network, and the interaction between the agents. The optimization process consists of an iteration cycle with three main steps, 'Execution', 'Scoring', and 'Re-planning', which are explained below in more detail.

Analysis: Finally, the simulation results such as the resulting population and demand and the traffic conditions on the network can be used for post-process analysis.

17.2.2 Agents and plans

Each person or traveler in MATSim is modeled as an individual agent. The demand of an agent is called plan (Balmer et al. 2009). A plan encodes the daily routine of an agent. It contains the agent's travel schedule including its intended activities and routing decisions between the activity locations (Balmer et al. 2005). Moreover, a plan captures (i) the order, type, location, duration as well as other time constraints for every activity and (ii) the selected mode, route, and expected departure and travel times of each leg (Balmer et al. 2005). A leg describes a part of a trip that uses exactly one transport mode (Balmer et al. 2005). An example plan is illustrated and explained in Figure 17.2.

17.2.3 Iterative demand optimization process

As stated before, the simulation takes the representation of the infrastructure and the population including their daily plans as input (Balmer et al. 2009). Balmer et al. (2005), Raney & Nagel (2006), Balmer et al. (2009) explain the simulation process. It consists of an iterative loop with three main steps, see Figure 17.1.

```
<person id="123456" employed="yes">
    <plan score="128.122" selected="yes">
        <act type="home" link="1000" x="100.0" y="100.0" end_time="07:00:00" />
        <leg mode="car" dep_time="07:00:00" trav_time="00:20:00">
            <route type="links"> 1000 1001 1002 1003 </route>
        </leg>
        <act type="work" link="1003" x="200.0" y="200.0" end_time="15:20:00" />
        <leg mode="car" dep_time="15:20:00" trav_time="00:20:00">
            <route type="links"> 1003 1004 1005 1000 </route>
        </leg>
        <act type="home" link="1000" x="100.0" y="100.0" />
    </plan>
</person>
```

Figure 17.2 This illustrates the demand or plan of a fictive MATSim agent. The agent with the id 123456 intends to leave home (located at link 1000) at 07:00 to go to work. The selected route consists of four links. The expected travel time by car takes 20 minutes. At 15:20, the agent intends to travel back home, which is expected to take again 20 minutes.

These three steps are summarized in the following:

Execution: The traffic flow simulation executes the selected plans of all agents simultaneously on the road network. At this stage agents are interacting with the physical environment and with other agents. The traffic flow simulation is implemented as a queue simulation, see Gawron (1998), Cetin et al. (2003) for more details. It has the advantage that it is computationally fast while still resolving each vehicle individually.

Scoring: All executed plans are scored by a utility function that determines the performance of each plan. The utility or scoring function is discussed in more detail below.

Re-planning: In this step, some agents obtain new plans (choice set extension), while all others choose between existing plans, typically based on a logit model. The choice set extension is implemented in a modular way so that different choice dimensions can be addressed. The modules used in conjunction with the Brussels case study (Ch. 19 of this book) are listed below.[37]

- **Time allocation module:** This module generates a new plan by taking an existing plan and randomly changing the planned ending times of activities (Balmer et al. 2005).
- **Router module:** This module generates a new plan by taking an existing plan and recomputing all routes. The router uses the

[37] The number of modules is not limited to those mentioned here. Every module is assigned a weight that determines the probability with which a module is applied to an agent. If one of these modules makes the choice set of an agent larger than some configurable limit, one of the plans, typically the one with the worst score, is removed.

updated generalized costs for each link from the last traffic flow iteration. The router is a time-dependent best path algorithm (Lefebvre & Balmer 2007), where best path is defined as the one with the least negative utility (Balmer et al. 2009).

- **Single trip mode switch module:** This module generates a new plan by taking an existing plan, randomly selecting a trip, and switching its mode to a randomly selected alternative mode. Other modules which take, say, subtours into account, are available (see Rieser et al. 2013).

The repetition of the iteration cycle coupled with the agent memory, i.e. the capability to remember more than one plan per agent, enables agents to improve their plans over several iterations (Balmer et al. 2005). The iteration cycle continues until a "relaxed" state of the system has been reached. MATSim currently does not use any quantitative measure of when this state is reached; usually the iteration cycle is repeated until the outcome is stable (Balmer et al. 2005). For simple situations (only two modes, only the car mode explicitly simulated) it is our experience that about 100 iterations are sufficient to obtain useful results (Nagel 2008, 2011, Röder et al. 2013). Under certain conditions, the result is a probabilistic integer version of the stochastic user equilibrium (SUE) (Nagel & Flötteröd 2012).

17.2.4 Evaluation of the performance of a plan with the scoring function

The utility of an executed plan is computed as (Charypar & Nagel 2005):

$$V_p = \sum_{i=1}^{n} \left(V_{perf,i} + V_{late,i} + V_{tr,i} \right) , \tag{17.1}$$

where V_p is the accumulated utility for a given plan p with n activities, $V_{perf,i}$ is the utility for performing activity i, $V_{late,i}$ is the disutility (negative utility) of being late at activity i, and $V_{tr,i}$ is a penalty for traveling from activity i to activity $i+1$. Plans are assumed to wrap around a 24-hour day, for that reason, the last activity is assumed to be the same as the first, and in consequence there are as many trips as there are activities.

The utility for performing an activity has a logarithmic form and is defined as:

$$V_{perf,i}(t_{perf,i}) = \beta_{perf} \cdot t_{*,i} \cdot \ln \left(\frac{t_{perf,i}}{t_{0,i}} \right) \tag{17.2}$$

Here t_{perf} is the actually performed duration of activity i, t_* is the "typical" duration of activity i, β_{perf} (positive) gives the marginal utility of any activity at its typical duration. $t_{0,i}$ has no effect as long as activity chains remain fixed

throughout the iterations, which will be assumed for all MATSim runs in this book. More details are provided by Rieser et al. (2013).

The disutility for being late is $V_{late,i}(t_{late,i}) = \beta_{late} \cdot t_{late,i}$, where β_{late} (normally negative) is the marginal utility for being late while $t_{late,i}$ gives the amount of time of being late at activity i.

The penalty for traveling is given by

$$V_{tr,i} = \beta_{tr,mode} \cdot t_i + \beta_{td} \cdot d_i + \beta_m \cdot m_i, \tag{17.3}$$

where t_i is the travel time, d_i the distance, m_i the change of the monetary position, and the different β_x are the usual pre-factors. In MATSim, m is negative when money is taken away from the traveler, so β_m is typically positive. β_{td} is typically negative. β_{tr} can be positive or negative.

The *effective* disutility for traveling is Eq. (17.3) plus the lost utility of time as a resource (opportunity cost of time). The marginal utility of time as a resource can be understood as the marginal loss of utility when the activity which follows the trip is shortened. This is computed as

$$\frac{\partial V_{perf}}{\partial t_{perf}} = \beta_{perf} \cdot t_* \cdot \frac{1}{t_{perf}} \approx \beta_{perf} \ , \tag{17.4}$$

where the approximation holds when the actual duration, t_{perf}, is close to the typical duration, t_*. That is, shortening an activity by Δt is penalized approximately by $-\beta_{perf} \cdot \Delta t$. If that time is spent travelling, Eq. (17.3) comes on top.

In consequence, the marginal utility of travel time savings, mUTTS, is, within the approximation of Eq. (17.4), equal to $\beta_{perf} - \beta_{tr,mode}$, and a monetary value of travel time savings is given by

$$VTTS = \frac{\beta_{perf} - \beta_{tr,mode}}{\beta_m} \ . \tag{17.5}$$

Hints for calibration based on those insights are given in the MATSim user guide (Rieser et al. 2013).

17.2.5 An illustration

Let us illustrate the MATSim iteration process. Initially, a synthetic traveler will have only one plan, say by car. This plan will be repeatedly executed in the traffic flow simulation.

Eventually, the synthetic traveler will be selected for re-planning, say for re-routing. The previous plan will be copied, all routes will be re-computed based on the previous iteration's travel times, the new plan will be made "selected", thus executed in the next traffic flow simulation, and scored based on its performance. The next couple of iterations, this synthetic traveler will

choose between these two plans, based on a logit model using the scores of these two plans.

Eventually, the synthetic traveler will again be selected for re-planning, say this time for single trip mode switch. The original plan will be copied, one trip will be randomly selected, and its mode will be switched to a randomly selected alternative mode (for example in this case, public transport or pt). That trip will be routed if necessary, the plan will be made "selected", thus executed in the next traffic flow simulation, and scored based on its performance. The next couple of iterations, this synthetic traveler will choose between these three plans, based on a logit model using the scores of these three plans.

In consequence, if the pt plan receives a much larger score than the two car plans, it will be selected with a much higher probability. Conversely, if the car plans have higher scores, then they will be selected with higher probability. Since for sufficiently many iterations, eventually all travelers will have all combinations of modes in their choice set, choice is entirely given by the calibration of the scoring function. This statement does even hold (approximately) when the choice set is limited in size, since only options with bad scores, which thus carry a low probability in terms of the logit model, are removed.

17.3 Special MATSim features used for the case studies

17.3.1 Public transit

MATSim provides several ways to simulate public transport (pt). The conceptually most straightforward, but also computationally and in terms of data procurement most demanding variant, is the direct microscopic simulation where every public transit vehicle is simulated individually according to its schedule and operating rules (Rieser & Nagel 2009, Rieser 2010). It includes information about public transit lines, their routes, the travel time between stops and the time of departure at the start of the route. Data in the GTFS format (General Transit Feed Specification www pages 2012) can be converted into MATSim transit schedules (MATSim extensions accessed 2013, under "GTFS2TransitSchedule").

A simple model, sometimes called "pseudo pt" (Grether et al. 2009, Rieser & Nagel 2009), estimates travel times by assuming that the travel distance between activity locations is composed of the beeline distance multiplied with a configurable beeline distance factor, often 1.3. The pt speed is configurable as well. The resulting travel times are determined by:

$$\mathrm{tt}_{\mathrm{pt}} = \frac{\mathrm{BeelineDistance} * \mathrm{BeelineDistanceFactor}}{\mathrm{PtSpeed}}. \qquad (17.6)$$

Alternatively, one can multiply the car free speed travel time with a configurable factor in order to obtain the public transit travel time. This approach

models public transport continuously and without capacity constraints and works without any knowledge about the public transport service in an area.

Within the project, an additional approach was implemented, called "matrix based pt", also without capacity constraints, but taking into account access and egress to pt stops and possibly other details of the pt system. Given an input table with pt stops and the associated coordinates, a matrix is generated including travel times and travel distances for any pair of stops. Travel times are determined based on Eq. (17.6). Travel distances are given by the numerator of Eq. (17.6), i.e. beeline distance between two stops multiplied by the beeline distance factor. The difference to the plain "pseudo pt" is that access and agress to the pt system are now modelled much more realistically. The stop-to-stop impedance matrix can also be taken from an existing VISUM (PTV AG 2009*b*,*a*) model or some other equivalent source, thereby taking connectivity characteristics of the pt system beyond its stop locations into account.

In both cases traveling by public transport is executed as so-called teleportation in MATSim, i.e. simulation of pt is not performed on the physical road network. The total travel time is composed of:

$$t_{pt,ik} = t_{wlk,gap,i} + t_{pt,matrix} + t_{wlk,gap,k} \ , \tag{17.7}$$

where $t_{wlk,gap,i}$ and $t_{wlk,gap,k}$ are the travel times on foot to overcome the gap between the activity locations i and k respectively to their nearest pt stop based on the beeline distance, and $t_{pt,matrix}$ is the travel time between the two stops, given by the matrix. The total travel distances are determined analogously:

$$d_{pt,ik} = d_{gap,i} + d_{pt,matrix} + d_{gap,k} \tag{17.8}$$

Here, $d_{gap,i}$ and $d_{gap,k}$ are the beeline distances between location i and k respectively to their nearest pt stop, and $d_{pt,matrix}$ is the travel distance between the two stops from the matrix. The pt travel disutility is computed on those combined travel times and travel distances; separate marginal disutilities for the walk and the pt parts are thus (unfortunately) currently not possible. For additional information, see the MATSim extensions (accessed 2013) under "matrix based pt router".

17.3.2 Road pricing

MATSim is capable of simulating different toll schemes such as distance or cordon toll (Nagel et al. 2008, Kickhöfer et al. 2011, Kickhöfer & Nagel 2013, forthcoming). Tolls can be limited to a part of the network. They can also be made time-dependent, i.e. the amount agents have to pay for the toll can vary depending on the time-of-day. For additional information, see the MATSim extensions (accessed 2013).

17.4 Integration approach to use MATSim as a travel model plugin to UrbanSim

In the past, some efforts towards integrating UrbanSim with external travel models like EMME (Babin et al. 1982) or VISUM (PTV AG 2009*a*,*b*) have been made. Both EMME and VISUM are traditional assignment models using origin-destination matrices (OD-matrices) as inputs (Ortúzar & Willumsen 2001). Thus, they do not make use of the disaggregated nature of UrbanSim. In this situation it seems quite natural to link UrbanSim with an agent-based travel model like MATSim directly at the agent level by directly feeding location and socio-economic characteristics of individual residents and firms from the land use model to the travel model and then having the travel model return updated accessibility measures back to the land use model.

By default, the feedback from external travel models to UrbanSim is a zone-to-zone impedance matrix including generalized costs of travel between any given pair of zones (Fig. 17.3). This is an $n \times n$ matrix, where n is the number of zones. UrbanSim uses this matrix as input for various models. Such matrices are growing quadratically with the number of zones and thus quickly become very large. This puts limits both on the spatial resolution that can be used with these zones, and on the number of attributes which can be returned from the travel model to UrbanSim. As one of the objectives of the present study, it therefore makes sense to search for alternative measures to feed back from a travel model.

from_zone_id	to_zone_id	am_single_vehicle_to_work_travel_time
601	601	1.2
602	601	10.5120915
603	601	10.84997029
604	601	47.72225589
605	601	54.58448668
606	601	33.78333333
607	601	60.57128651
608	601	10.78460873
609	601	88.72431786
610	601	35.32687035
611	601	10.71284562
612	601	11.51725003
613	601	13.42148847
614	601	48.76746873

Figure 17.3 Conventional zone-to-zone impedance matrix including travel times from any origin zone "from_zone_id" to any destination zone "to_zone_id".

17.4.1 MATSim4UrbanSim at a glance

The interaction between MATSim and UrbanSim (Nicolai et al. 2011) is a bi-directional relationship as depicted in Figure 17.4. It consists of three main steps:

1. When UrbanSim iterates in annual steps, it calls MATSim from time to time, at most once per annual step, and passes a path to the traffic network data together with the persons and jobs data tables as input. The tables include attributes such as the person id as well as the residence and job location of each individual person in UrbanSim.

2. Given the input tables from UrbanSim, MATSim generates the traffic assignment, as discussed in Section 17.2. Finally, it returns the resulting access and accessibility indicators, discussed below in Section 17.5. Then MATSim terminates.

3. Once MATSim has terminated, UrbanSim takes back control. It reads the indicators and updates the data store for the next UrbanSim iteration. UrbanSim models that make use of the updated traffic conditions are for example the Household and Employment Location Choice Models as well as the Real Estate Price Model, presented in Section 3.2. In fact, arbitrary UrbanSim models can make use of the travel model output, since the data is available as a regular UrbanSim table.

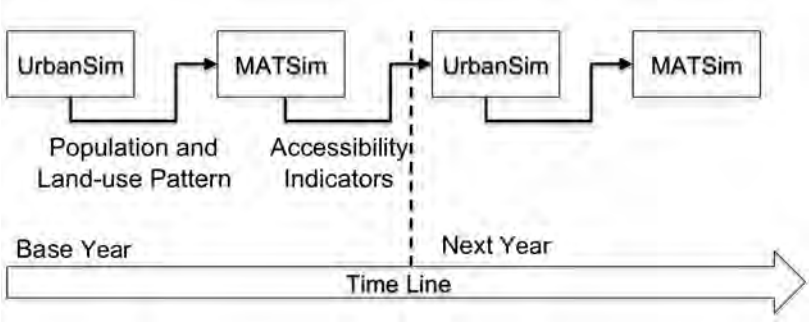


Figure 17.4 This shows the interaction sequence between UrbanSim and MATSim. UrbanSim calls MATSim in regular intervals and passes the current population and land-use pattern. MATSim computes the traffic based on the provided information and the resulting access and accessibility indicators. UrbanSim uses these indicators for the next year (iteration) as input for various models.

MATSim4UrbanSim simulates home-work-home commuting trips based on home and work locations of each individual person provided by the UrbanSim input tables. More complex, e.g. activity-based, demand patterns are possible with MATSim, but are not implemented in the present version of MATSim4UrbanSim.

17.4.2 Different options to couple UrbanSim with MATSim

Given that one of the objectives is to exploit the fact that both UrbanSim and MATSim are person-centric, it would be desirable to communicate between

both packages directly at the object level. This would mean to invoke and execute MATSim by UrbanSim like other UrbanSim routines, and to exchange data, such as current land use information in UrbanSim or computed access or accessibility indicators by MATSim, directly via the main memory.

As stated earlier, UrbanSim is implemented in Python, while MATSim uses Java. Both languages use different and incompatible byte code representations. Java programs are translated into Java byte code that is executed by the Java Virtual Machine (JVM). The reference implementation for Python is written in C, called CPython. In the same way as the JVM, CPython compiles Python source code into Python byte code. There is also a Java implementation for Python which would simplify the integration task. However, UrbanSim relies heavily on numerical and other libraries written in C or C++, so that was not an option.

Given the incompatible byte code representation, objects would need to be converted when passed from one package to the other. Arguably, the first package that comes to mind is the Java Native Interface (JNI; Liang 1999). JNI allows the calling of Java routines from C. Unfortunately, this does not yet resolve how to get the information from Python to C; such a layer would need to be separately implemented.

Other projects have taken up that challenge. The JPype project (JPype www pages accessed june 2013) allows Python full access to Java. This is achieved through interfacing the Python interpreter (CPython) and the JVM at the native level using JNI and PNI (Python Native Interface). The main problem is that JPype does not provide support for advanced data types, let alone objects. For example, to create an array, the number of dimensions and the number of elements in the array need to be declared at compile time. This is completely at odds with important concepts used both in UrbanSim and in MATSim, which decide on the size of data classes from reading the input files. Also, there is no support for making sure that the dimensions and sizes declared on the Python side are consistent with the dimensions and sizes used on the Java side. It is our belief that projects such as UrbanSim and MATSim have become possible for universities over the past years, because modern programming languages provide support for objects. Working without such support leads to code that neither the UrbanSim nor the MATSim team have the resources to maintain. In addition, the JPype project does not seem overly active, with the last blog entry from 2007, and the last two (minor) releases from 2007 and 2011, making it problematic to base an important part of the SustainCity project on JPype.

The only other project that we could find in that area was JEPP (JEPP www pages Accessed June 2013). That project, however, looks at calling Python from Java, which is exactly the wrong direction for what was needed here.

Thus, in the end it was decided to resort to the conventional technique of writing and reading files. UrbanSim already had methods to write and read comma-separated files, and for MATSim they were easy to write.

A challenge was how to install and configure MATSim in a convenient and robust way. This meant that preferably, at least to get started, one file containing all the MATSim material should be unzipped at the right place, and relevant entries should be made in the OPUS GUI (Graphical User Interface), presumably already known to the UrbanSim user. This, however, implied that a communcation channel from UrbanSim to MATSim would need to be established which could be automatically tested, so that eventual errors during future refactorings of UrbanSim or MATSim that would destroy that communication channel could be detected.

This was resolved passing an XML (W3C 2008) document from UrbanSim to MATSim, and using a so-called XML Schema Document (XSD; van der Vlist 2002) to standardize and certify the grammar of that document. For details, see Section 17.4.3.

Another challenge stems from the data type flexibility of UrbanSim, where virtually arbitrary attributes can be added to the tables, and related to each other through the configurable models. One consequence of this flexibility is that the attribute names are not standardized; for example, not even the keys (= attribute names) for coordinate entries are standardized in UrbanSim. In consequence, the passing of data between UrbanSim and MATSim has to rely on often-used names on the UrbanSim side (Nicolai 2012); projects which deviate from these naming conventions would need to implement their own data packing method on the UrbanSim side. This holds also for the direction from MATSim to UrbanSim: MATSim assumes the existence of tables under certain keys on the UrbanSim size; if they exist under a different key, the MATSim output will be added instead of replacing the data in the UrbanSim tables.

In short, the data exchange will work as long as the naming conventions from Nicolai (2012) are strictly followed. Any deviations from these naming conventions mean that changes in the computer codes need to be made.

Overall, not being able to couple the two packages at the object level is a bit disappointing, since this means that much of the agent-oriented expressiveness cannot overcome the barrier between the two packages. For example, it is not possible for a location-seeking agent on the UrbanSim side to query MATSim about detailed transportation options, and it is also not possible for a synthetic MATSim traveler stuck in congestion to, say, go shopping at a different location if landuse information was not passed to MATSim pre-emptively. It is to be hoped that future technological developments will overcome this barrier.

17.4.3 MATSim4Urbansim configuration

Both UrbanSim and MATSim have, in principle, their own configuration file. However, to run the frameworks together for different scenarios would have the consequence to set up both configuration files separately. This easily leads to inconsistent, error-prone and fragile configurations that are inconvenient to maintain by users. To address this issue, it was decided to provide one joint configuration for both frameworks, at least for simple situations. At this point a brief overview is provided; a comprehensive description is given by Nicolai & Nagel (2010).

The UrbanSim side was selected as the place to manage the joint configuration, since it embeds MATSim as a travel model plug-in. It is achieved by embedding necessary MATSim parameters into the travel model configuration section of the UrbanSim configuration, as shown in Figure 17.5. In this way, both simulation frameworks are conveniently configurable via the OPUS GUI, i.e. the graphical user interface that is available to configure the OPUS system and start UrbanSim runs. Necessary parameters, for instance, include input file locations, e.g. for the traffic network, the population sampling rate to accelerate computation times, the run description defining for which years to run MATSim, and more; in-depth information about this is provided in the user guide (MATSim extensions accessed 2013, see MATSim4UrbanSim).

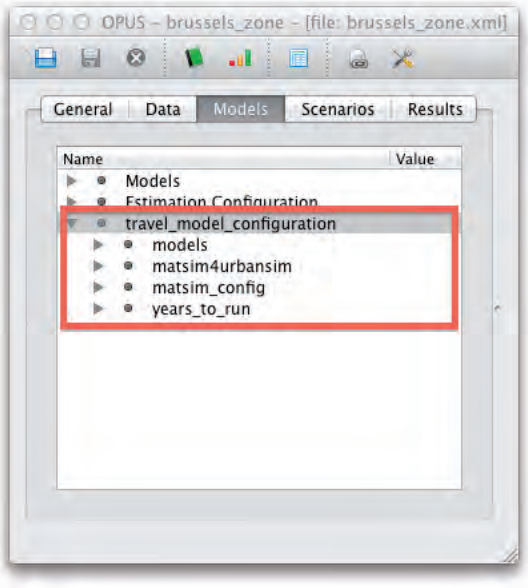


Figure 17.5 Relevant MATSim parameter are configurable via the OPUS GUI. They are embedded in the travel_model_configuration section (marked in red) as part of the UrbanSim configuration.

UrbanSim takes the parameter settings from the travel model configuration section and generates a separate configuration file in XML format to initialize MATSim. This is done each time the travel model is called by UrbanSim. To achieve a robust and reliable communication, a mutual consent between UrbanSim and MATSim regarding the structure and the parameter data types (e.g. string, int, or float) of the generated configuration file is crucial. Therefore, the generated MATSim XML configuration is specified by an XSD, which is an abstract collection of meta data about an XML document (van der Vlist 2002). The XSD is used to generate customized XML parsers in MATSim and UrbanSim that read, write, and validate XML documents; this technique is called XML data binding.

In addition, this approach allows a convenient adaption of the MATSim configuration to new requirements, e.g. when additional configuration parameters are required. This can be done by adjusting the XSD and regenerating the XML parsers on both sides.

17.4.4 Cold, warm and hot start

Here, three notations are introduced: cold, warm, and hot start. Cold start was in principle already described in Section 17.4.1 above. It means that MATSim generates the initial travel demand based on the provided UrbanSim population, see Figure 17.6.

The terms warm and hot start describe the capability of MATSim4UrbanSim to start a simulation from a pre-existing, relaxed plans file (e.g. see Section 17.2) and to recycle information, such as route decisions and departure time choices, from previous runs. In other words, MATSim "remembers" the travel schedule (daily plan) of each traveller. Consequently fewer iterations are made and thus less computing time is required to reach a relaxed state of the system.

In warm start, MATSim will always use the same relaxed plans file as shown (as "Inital Plans") in Figure 17.6. This plans file is generated from a preparatory run, which relaxes the travel model for the base year. In subsequent UrbanSim years, those synthetic travelers that neither change their residence nor their workplace will start with those relaxed plans; also see below. All other travelers will start from scratch.

When running UrbanSim over many years with a changing population, because of the relocations the initial plans file will become less and less correct, and the system will require more and more iterations to get MATSim back into a relaxed state. This issue is addressed by hot start. Here, MATSim stores an updated plans file after each run, including all changes of the current UrbanSim population. As a result the differences between the updated plans file and the UrbanSim population are kept small. As opposed to warm start, hot start uses the updated instead of the initial plans file after the first run, as shown in Figure 17.6.

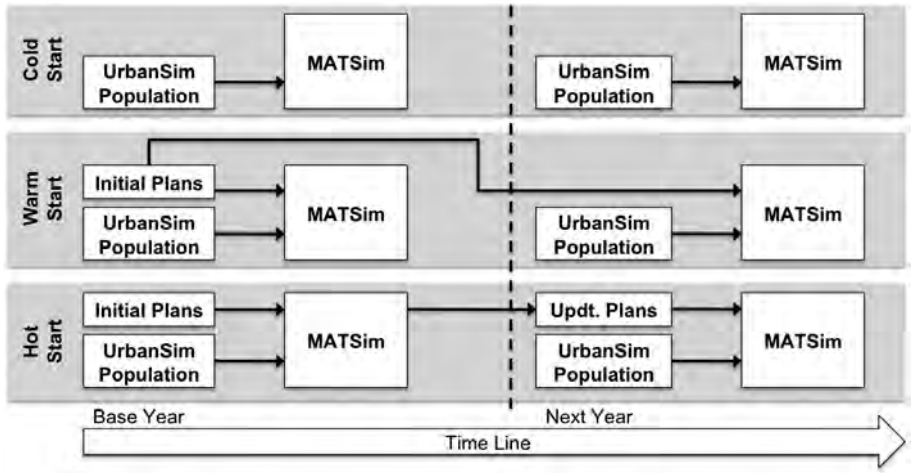


Figure 17.6 This illustrates the working of the (i) cold, (ii) warm, and (iii) hot start in MATSim. (i) Cold start MATSim generates the initial demand from the current UrbanSim population. With warm and hot start MATSim recycles information from previous runs; as a result less iterations are required to reach a relaxed state of the system. (ii) Warm start will always use the same plans file, which becomes less and less correct when running UrbanSim over a long time span. (iii) In hot start, opposed to warm start, after each run an updated plans file is stored, incorporates changes from the current UrbanSim population. In this case, MATSim starts from the updated plans file instead of the initial plans file.

Technically, when MATSim performs a warm or hot start it reads a plans file together with the current UrbanSim population. It keeps all persons from that plans file that have not changed. This means that

1. a person from the current UrbanSim population also exists in the plans file,
2. has the same employment status,
3. has the same home location.
4. and if applicable, has the same work location.

For all other persons, new initial plans are generated. Some computational results concerning warm and hot start are available from Nicolai (2013).

17.5 Access and accessibility indicators

This section looks at the access (= impedance) and accessibility indicators computed by MATSim as feedback for UrbanSim. UrbanSim uses the feedback to update its data base for the next year as stated in Section 17.4 above. The term "access" refers to a two-point-value such as travel time impedances between an origin and a destination; in contrast, "accessibility" is attached to one location and thus refers to an aggregated single-point-value. The MAT-

Sim feedback currently includes (i) so called zone-to-zone "skims", meaning a summary of key model predictions like travel times by mode or time of day (OPUS User Guide 2011), (ii) individual agent-based performances, as well as (iii) accessibilities to work places. The following provides a summary (see Nicolai 2013, for details) for details.

17.5.1 Zone-to-zone impedance matrix

The zone-to-zone impedance matrices contain travel times for several transport modes, travel distances, generalized travel costs and the number of vehicle trips for each pair of zones. Zones are assigned to the road network by determining the network node which is closest to the zone centroid. The centroid coordinates can be obtained from a variety of definitions, either (i) they are directly available in the UrbanSim model, or by (ii) averaging all parcel coordinates that belong to a zone. Typically, (i) applies to UrbanSim zone models and (ii) to UrbanSim parcel models.

Travel times are provided for congested- and free-speed car, public transport, bicycle, and walk. Congested- and free-speed car travel times are accordingly based on congested- or free-speed link travel times of the MATSim road network. By default, congested-speed car travel times are measured for the morning peak period, i.e. for trips that are assumed to start at 8am. Travel times on foot or by bicycle are based on the shortest path on the road network with a constant speed of 5 km/h (walk) and 15 km/h (bicycle). For public transport (pt) travel times are obtained by Eq. 17.7, which is composed of traveling on foot to the nearest pt stop from both the (i) origin, (ii) destination, and (iii) the pt travel times queried from the pt matrix.

Travel distances are given using the shortest path on the road network.[38] Generalized travel costs are given by Eq. 17.3, including congested-speed car travel times, distances and toll. The number of vehicle trips per OD pair are obtained from counting the actually simulated trips of each synthetic traveler. The numbers are then scaled to the full population sample.

17.5.2 Agent-based performance

This feedback contains the individual travel performance for each MATSim agent based on the selected mode, congested car or public transport (pt), including travel times and travel distances, for the selected plan, e.g. see agent plans in Section 17.2. These are given for both directions, commuting from home to work and back. For the time being, only information on the congested car mode and on the pt mode is returned, corresponding to the fact

[38] Note that a car will typically not take the shortest but rather the fastest/best path, and the bicycle/walk modes are not restricted to the road network. Thus, these distances should be seen as an approximation.

that the default version of MATSim4UrbanSim only uses these two modes. Additional modes could be integrated with relatively little effort.

17.5.3 Accessibility computation

A comprehensive description about high resolution accessibility computations in MATSim is given by Nicolai & Nagel (in press, 2012). At this point a brief overview built on these references is provided.

For the present implementation a utility-based measure is selected (e.g. Ben-Akiva & Lerman 1985, Train 2009). It reflects the (economic) benefits, computed as the expected maximum utility, that someone gains from access to spatially distributed opportunities (Geurs & Ritsema van Eck 2001, de Jong et al. 2007). This is also known as the logsum term; it is defined as

$$A_i := \ln \sum_k e^{V_{ik}} \ , \tag{17.9}$$

where k goes over all possible destinations. V_{ik} is the disutility of travel in order to get from location i to location k; in discrete choice theory interpretation, these are the so-called systematic parts of the utility.[39]

Origin and opportunity locations do not necessarily lie on the network. Thus, for the *network-oriented modes car, bicycle, and walk,* the calculation of V_{ik} includes the disutility of travel to overcome the gap between locations and the network; in the current implementation it is assumed that in these modes, opportunities can only be reached via the transport network. Overall, V_{ik} consists of the following components:

1. The walk disutility for reaching the transport network from a given origin location i. This gap is determined as the shortest distance to the network either given by (i) the distance to the nearest node or (ii) the orthogonal distance to the nearest link on the network.

   For the latter case, V_{ik} additionally includes the travel disutility to overcome the distance to the nearest node according to the given transport mode.

2. The travel disutility on the network towards k is given by

   $$V_{tr,i} = (\beta_{tr,mode} - \beta_{\mathrm{perf}}) \cdot t_i + \beta_{td} \cdot d_i + \beta_m \cdot m_i \ ; \tag{17.10}$$

   this is Eq. (17.3) *plus* the effect of using time as a resource. Using Eq. (17.5), the first term of the sum can also be expressed as $-\beta_m \cdot VTTS \cdot t_i$. The parameters are taken from the travel model.

[39] There is sometimes an additional scale parameter in that definition. Here it is assumed that this is absorbed into V_{ik}. In consequence, A_i is in the same units as V_{ik}; a conversion to monetary terms, for example, needs to be done separately.

3. The walk travel disutility for reaching opportunity k from the transport network. Here, the distance to the nearest node is used to determine the shortest distance to the network.

This results in

$$V_{ik,mode} := V_{wlk,gap,i} + V_{mode,tr,k} + V_{wlk,gap,k} , \quad (17.11)$$

where $V_{wlk,gap,i}$ is the walk disutility to overcome the gap between origin location i and the road network; $V_{mode,tr,k}$ is the travel disutility on the network traveling towards k (Eq. 17.10); and $V_{wlk,gap,k}$ is the walk disutility to overcome the gap between the road network and opportunity location k. The term $V_{wlk,gap,i}$ consists of

$$V_{wlk,gap,i} = \beta_{tt,wlk,gap,i} \cdot t_{wlk,gap,i} + \beta_{d,wlk,gap,i} \cdot d_{wlk,gap,i} , \quad (17.12)$$

where $t_{wlk,gap,i}$ is the walk time and $d_{wlk,gap,i}$ is the distance to overcome the gap between location i and the network. $\beta_{tt,wlk}$ and $\beta_{d,wlk}$ are marginal utilities that convert travel times and distances respectively into utils. $V_{wlk,gap,k}$ is defined in the same way.

For the *pt mode*, the accessibility computation is only implemented when the "matrix based pt" approach described in Section 17.3.1 is used. Then, the walk times and distances to the network are replaced by the walk times and distances to the nearest pt stop, and the travel times and distances on the network are replaced by the times and distances given by the pt matrix.

Accessibilities can be calculated for two spatial systems, (i) zones and (ii) parcels.

- For zones, zone centroids are used as measuring points. The centroid coordinates can be obtained from a variety of definitions, as explained in Section 17.5.1.

- For parcels, the study area is subdivided into square cells, where the resulting cell centroids serve as origins or measuring points for the accessibility calculation. The spatial resolution of this is configurable. Once the calculation for each cell is completed, MATSim interpolates the accessibility value from the grid-cell to each UrbanSim parcel. For more information see Thunig & Nicolai (2013).

  The intermediate result based on the grid cells can be saved as well.

The current version of MATSim4UrbanSim returns workplace accessibilities, i.e. the opportunities mentioned above are workplaces. Accessibilities to other types of opportunities could be added with relatively little effort.

17.6 Some accessibility results

As an illustration, intermediate results of the Brussels case study are shown. Figure 17.7 shows the morning workplace accessibility by car in the base case. One very clearly sees the influence of the network; since the code assumes that access to the network is by the walk mode, the distance to the transport network is the most decisive factor because of the low walk speed. This points to the necessity to make such studies with higher resolution networks – this was not an option here because of computational performance restrictions for the UrbanSim integration (see Sect. 17.7).

One of the scenarios of the Brussels UrbanSim case study was looking at the influence of a possible cordon toll of 5Eu, levied between 6am and 10am. The cordon was defined just outside the outer freeway ring; tolled links are shown in blue in Figure 17.8. Car users from outside thus need to pay toll when using the ring. MATSim was then run again with that toll enabled, and the resulting accessibility differences are shown in the same figure. One finds two major effects:

- The accessibility inside the cordon area is visibly improved. This is due to reduced congestion.

- The accessibility outside the cordon area is visibly reduced. This is because the toll is part of the travel disutility. In consequence, persons living outside the toll area now have a higher travel disutility to reach some of their possible workplaces, and thus their (car) accessibility to workplaces is reduced.

17.7 Computational issues

With the case study set-up for Brussels, one MATSim run with 100 iterations took about 40 minutes on a regular desktop computer. Figure 17.9 gives the breakdown between different modules, over the iterations. The accessibility computation itself is not contained in that breakdown, but takes about 2 minutes of computing time for a problem of this size (Nicolai & Nagel in press).

Compared to UrbanSim, which only needs a couple of minutes per simulated year, these 40 minutes per run were already a significant burden, and in consequence MATSim was only run for 3 different UrbanSim years. Keeping all other things constant, the computation time of a MATSim run scales roughly linearly in the number of links. Doubling or quadrupling the number of links, leading to MATSim computation times of 80 or 160 minutes, was not an option for the present study. Warm or hot start acceleration was not used since it had some isssues in conjunction with the matrix-based pt, and there was no time to resolve them.

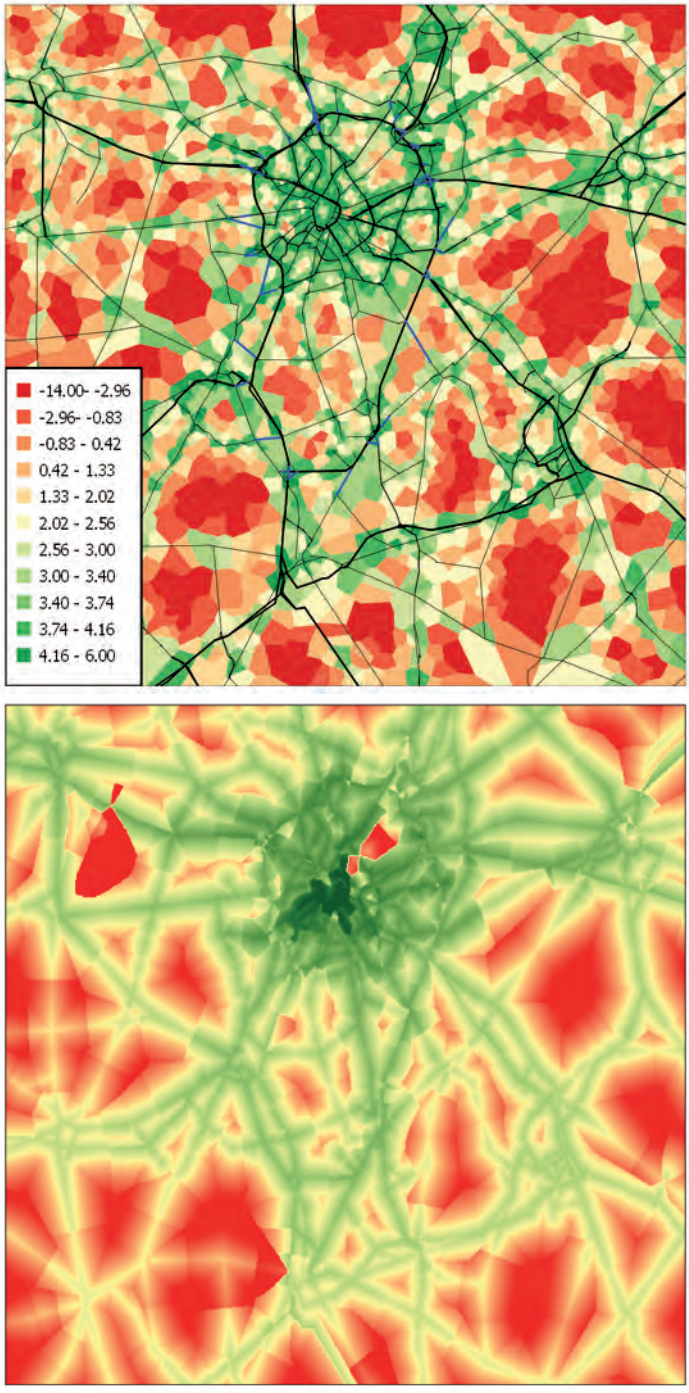


Figure 17.7 Brussels case study – intermediate results. Morning workplace accessibility by car in the base case. *Top:* Using zones. *Bottom:* Using high resolution (grid-based) accessibility.

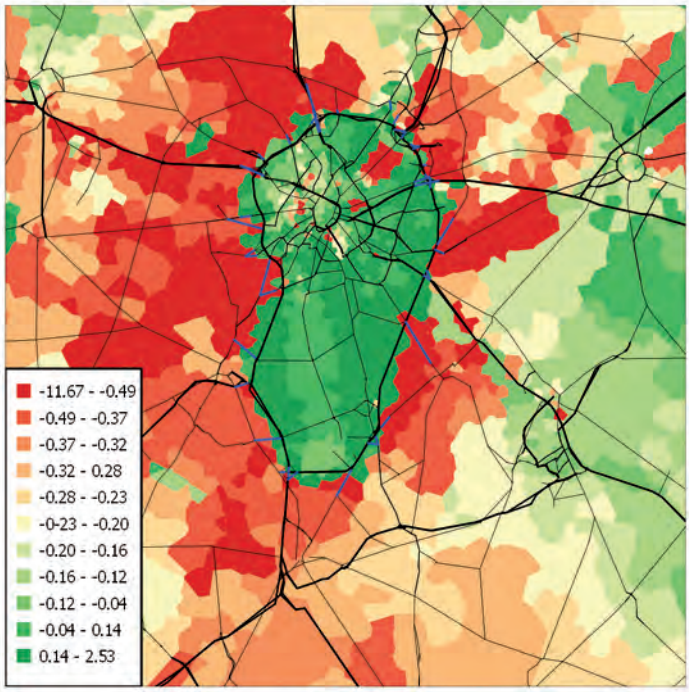


Figure 17.8 Changes in the morning workplace accessibility by car because of a morning cordon toll of 5 Eu in the Brussels area. The tolled links are marked in blue; the freeway ring is fully inside the cordon area.

17.8 Discussion and outlook for the SustainCity project

This chapter discusses a so-called agent-based model as travel model plugin, as opposed to a trip-based model which will be discussed in Chapter 18 of the book. The most important difference is that the agent-based approach keeps the simulated person intact throughout the day. This allows, for example, to consider the effect of a morning toll on the afternoon mode choice, or the effect of a morning delay on the afternoon schedule.

In principle, agent-oriented coupling should be very natural between MATSim and UrbanSim, because both are person-centric models. And indeed, one set of quantities that is fed back from MATSim to UrbanSim is person-centric.

Notwithstanding these statements, coupling Java-based MATSim with Python/C/C++-based UrbanSim, as was done here, remains problematic. The coupling between MATSim and UrbanSim was achieved using the traditional method of writing and reading regular files. This does, however, not allow to directly call one software from the other. In particular, it is impossible for "persons" in UrbanSim to query the MATSim model directly concerning, say, how easy or difficult it is to reach the workplace from a considered hous-

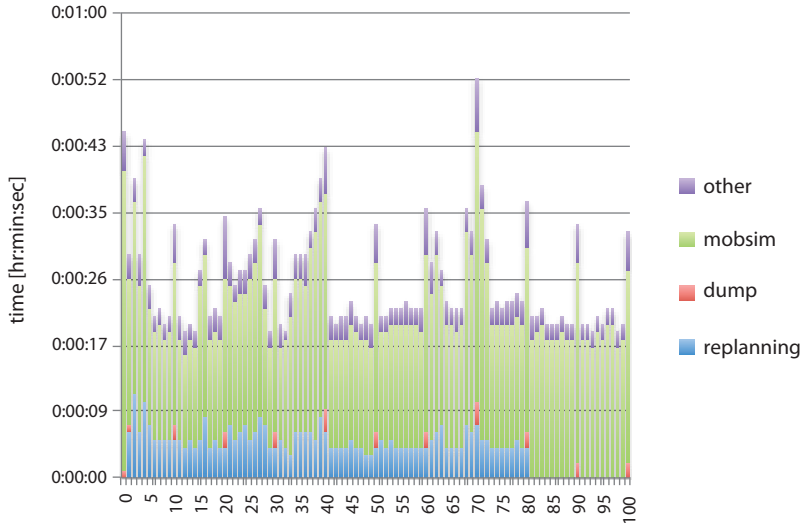


Figure 17.9 Computational performance of a typical MATSim run within the Brussels case study over the iterations. The machine had additional loads, which explains the fluctuations.

ing location. This reinforces the assessment that there are currently separate streams of software that will remain difficult to integrate fully until technology improves.

In contrast, the computation of high resolution accessibilities proved to be rather fruitful. As anticipated, it is quite feasible to compute accessibilities on a $100\text{m} \times 100\text{m}$ grid, while taking into account congestion. This provides better insights and thus also better modeling input to UrbanSim than conventional accessibility computations that aggregate over zones. It is intended to use this in the future for various scenarios and also for other activity types besides work. This should, for example, help with the identification of regions with poor access to certain services.

Chapter 18

Integration of dynamic transport models and agent-based land use models

André de Palma, Mohammad Saifuzzaman, and Kiarash Motamedi

18.1 Introduction

Accessibility plays a key role in location choice decision of urban actors as households, businesses, and urban developers. It is in a continuous evolution as a consequence of the evolution of spatio-temporal pattern of the congestion over the regional network. The experiences with dynamic integrated land-use-transportation models show that the equilibrium state of such a complex system is time path dependent and it is devoted to represent this dependence in the simulation results. Updating the congestion level and pattern over the simulation period is the objective of the interaction between the land use component and the transportation component of the model. We should consider that the agents not only adapt their choices in function of accessibility evolution but also they anticipate the future evolution (to read more about the transportation cost and city formation see de Palma et al. 2011*b*). The adaptation and anticipation behaviors are observed and studied in many local studies but they are not yet well implemented in the global simulations. This imposes the question of the frequency of the congestion update over the simulated period as the higher update frequency increases the simulation time and cost with a lower marginal benefit in the precision of the results.

This chapter describes the details of integrating an urban simulation model UrbanSim (Waddell, Ulfarsson, Franklin & Lobb 2007) and dynamic transport model METROPOLIS (de Palma et al. 1997) to increase the accuracy of the urban simulation model. UrbanSim is a land use model, which describes the behavior of residents and firms. These two type of agents are located in a

spatial setting (the city, or more precisely, the metropolitan area). Here, transportation is limited to the movement of residents, who can either commute or perform some other activities (leisure or personal affairs). Travel time depends on the size of the road, level of congestion, and time of day. These parameters are obtained from a transportation model.

The demand for travel is given by the origin-destination matrix (or O-D matrix). The elements of the O-D matrix are denoted by OD_{ij}, which corresponds to the number of individuals who wish to go from i to j. The demand for trip is given by the land use model. Usually, it is provided by a survey, and therefore extrapolation into the future, is very difficult, and somewhat ad hoc. Here, since the locations of residents and firms are endogenous, the travel patterns is also endogenous. It depends on the distribution of population and jobs, as well as on travel conditions. There is a fixed point problem here, since residential locations, for example, depend on travel conditions, which themselves depend on residential locations and locations of activities at destination. This fixed point problem is very hard to solve for this multi-agent game. Fortunately, these processes are not endogenous, but span over time. If an individual wishes to locate, he or she may safely consider that the level of congestion observed is exogenous. This is like a perfect equilibrium condition. The level of congestion (as the price is a general equilibrium model) is the outcome of a supply function (the capacity of the system) and a demand function (the O-D matrix). Of course, this equilibrium is never reached since the O-D matrix keeps on changing, and the capacity of the road does not remain constant. However, some dynamic is going on. This type of dynamic is hard to describe a priori. However, given the simplified assumptions, it is easier to describe. For example, it can be assumed that the time to adjust the level of congestion given an O-D matrix is rather short (on the order of magnitude of a few weeks or even a few months, but in any case, much less than a year). However, moving costs are high, so that the adjustment of couples to a new situation, such as loss of job, or a new kid, can span over a few years. This situation of slow and fast dynamics has been heavily described in physics. To simplify the analysis it is assumed that the fast process is always at the stationary process. This amounts to assume here that each year, the METROPOLIS model output is a stationary process, which is an input for UrbanSim, for one or a few years. In the next iteration of UrbanSim, the model produces a new distribution of firms and residents and creates a new or updated O-D matrix which is also an input to next METROPOLIS simulation, and the process goes on.

Making the interaction between transport and land use endogenous is at the heart of the research in urban economics. This will allow to study, for example, the short and long run impact of changes of infrastructure (long run corresponding to the situation where resident and firms relocated as a consequence of building new infrastructure). The short run corresponds in a sense where the model obeys some constraint (here this means that populations and firms are fixed). The questions raised by this discussion are whether or not

the long run effects of a policy (such as road pricing for example) are larger or smaller than the short run effect. Then development of our integrated land use and transportation model will allow to answer such a difficult, yet important, question.

The remaining of the chapter is organized as follows: Section 18.2 provides a brief description of UrbanSim and METROPOLIS, Section 18.3 describes the detail of the interface and how the interaction between the two software occurs and Section 18.4 draws the conclusion.

18.2 UrbanSim and METROPOLIS

18.2.1 UrbanSim

UrbanSim is an open source software-based urban simulation system designed by Paul Waddell (Waddell, Ulfarsson, Franklin & Lobb 2007) and developed with numerous collaborators to support planning and analysis of urban development, incorporating the interactions between land use, transportation, the economy, and the environment. UrbanSim can be seen to function in the following way over the course of a simulation year. Exogenous household and employment data are used as an input for the demographic and economic transition models. These models either remove jobs and households or create new jobs and households to be located later by the location models. Based on land-use and accessibility data, households and jobs are assigned locations. Based on the vacancy rate, the Development Project Transition Model creates a list of development projects to be placed. The location for these developments is then chosen by the Real Estate Development Model. Finally, the Land Price Model updates land-values that will be used in subsequent simulation years.

18.2.2 METROPOLIS

METROPOLIS model is based on simple economic principle developed initially by Vickrey (1969) and extended by Arnott et al. (1993). The software was developed in Geneva by de Palma et al. (1997). It describes joint departure time and route choice decisions of drivers. METROPOLIS proposes an interactive environment that simulates vehicle traffic flow and congestion in urban areas. The core of the system is a dynamic simulator that integrates commuters' departure time and route choice behavior over large networks. Drivers are assumed to minimize a generalized travel cost function that depends on travel time and schedule delay. METROPOLIS describes the joint departure time and route choice decisions of drivers. Each vehicle is described individually by the simulator. However, the modeling of congestion on the links is carried out at the aggregate or macroscopic level.

On the supply side a congestion function describes the travel delays of the links. METROPOLIS provides the possibility to introduce any relevant

function based on the traffic flow and capacity of the links. The usual forms of congestion function used in METROPOLIS are Bottleneck, Bureau of Public Roads (BPR 1964) or DAVIS (know also as INRETS function, see PTV 2012).

The demand is represented at microscopic level and each trip can be simulated. The user characteristics which are necessary for the modeling are Value of time (VOT), Schedule delay penalty, Distribution of desired time of travel, and mode choice parameters like VOT for public transport (PT) and PT penalty or fee. During simulation, each driver individually chooses his mode, departure time and route for travel. The mode choice could be modeled by either considering the average maximum expected utility offered by the car network in comparison with the other modes, or considering the utility of the available modes at the time of departure.

The choice of departure time for public transportation is not modeled, since the PT travel times are external inputs to METROPOLIS. The departure time choice model for car is a continuous logit model, either deterministic or stochastic. In the deterministic version the individual selects the departure time that minimizes the generalized cost function. The generalized cost function, $C(t)$, can be defined as travel time cost plus cost of arriving early or late:

$$C(t) = \alpha.tt(t_d) + \beta.\,(t^* - \Delta/2 - t_d - tt(t_d)) + \gamma.\,(t_d + tt(t_d) - t^* - \Delta/2)\,,$$

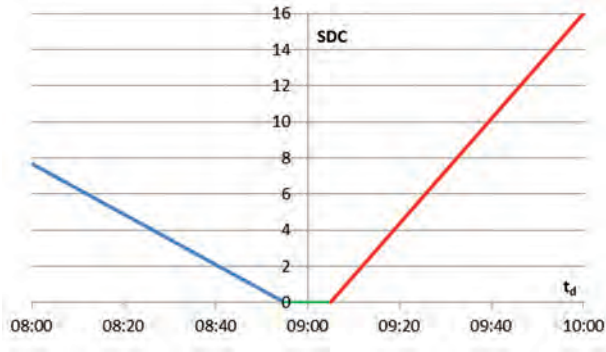


Figure 18.1 Schedule delay cost.

where, $tt(t_d)$ is travel time for a departure time t_d from origin, t^* is preferred arrival time, α is value of time, β is early arrival penalty, γ is late arrival penalty, and Δ is no penalty period. In the stochastic version of departure time choice, the probability of choosing a departure interval $(t, t + dt)$ is given by a continuous model logit:

$$P(t)dt = (1/A)\,.exp\left(-C(t)/\mu^T\right) dt,$$

where, the parameter $\mu > 0$ measures the heterogeneity of the departure time choice, and A is the accessibility which can be defined as

$$A = \mu^T \ln \int exp\left(-C(u)/\mu^T\right) du.$$

METROPOLIS uses a model of route choice based on point-to-point dynamic travel times. The user selects the dynamic shortest path from the origin node to the destination node. The decision is based on the real time situation of the immediate link and memorized information about the rest of the network up to the destination. It should also be noted that one day corresponds to one iteration in METROPOLIS. The software uses a learning process where users acquire knowledge about their travel and use this information to modify their trip for the next day. The learning process can be expressed by the following equation:

$$X^H(w+1) = f(X^H(w), X^S(w)),$$

where, $X^H(w)$ is the historical information acquired on day w and $X^S(w)$ represents the traffic conditions effectively simulated/incurred on day w. Several functional forms for $f(.)$ are available.

The output from METROPOLIS can be aggregate and disaggregate. Aggregate data includes the network measure of efficiency (MOE) like average travel time and average travel cost, free flow and delay cost, average speed, total collected revenues, consumer surplus, equity, congestion percentage, total mileage and so on. In the semi aggregate level we can have distribution of departure time, arrival time and so on. All the data of simulation are stored in the database. Therefore, various types of disaggregate data are possible to extract from the database based on the project needs. Mostly used disaggregate data are flow on some selected links collected in every 10 minutes (can also be stored in every second), travel time of some selected O-D pairs or road sections, travel time, cost, consumer surplus[40] and equity for each users and for different zones, etc. METROPOLIS also produces some environmental output like emissions of CO, CO_2, NO_X, and VOCs, and fuel consumption. The zone accessibility is calculated using either weighted average travel time or user surplus.

18.2.3 METROPOLIS application in Ile-de-France case

In this chapter we focus on the Ile-de-France case. METROPOLIS model has already been used for the traffic analysis of this region (de Palma & Lindsey 2006). The application is based on the network coded by two regional authorities: the Institute for Urban planning and Development of the Ile-de-France Region (IAU-IdF) and The Regional and Interdepartmental Direction of Public Works and Urban Development (DRIEA). Each of these two organizations has developed a separate transportation model for the region. The first one has been less detailed with 606 zones and the second one named "MODUS" is much more detailed with 1289 zones (for more details see DREIF 2008). These two models are based on conventional 4-step methodologies of travel demand

[40] logsum of the user's utility based on his mode, departure time and route choice that are modeled by nested Logit formulation.

analysis. The parameters of each step are estimated based on household travel survey (EGT)[41] at year 2001. The dynamic behavioral parameters are estimated based on a specialized survey named MADDIF (de Palma & Fontan 2001). They are presented in Table 18.1.

	α	β	γ	β/α	γ/α
Work to Paris and Near Suburbs	0.081	0.038	0.047	0.469	0.580
Work to Outer Ring	0.130	0.084	0.175	0.646	1.346
Other purposes	0.089	0.036	0.073	0.404	0.820

Table 18.1 Behavioral parameters, Source: de Palma & Fontan (2001).

In this phase of work we use the coded network of the MODUS. For the reasons of practical facility and comparability of the results, we plan to keep as far as possible the same procedure to make the OD demand as used in MODUS. We have used the **trip generation** matrices created by UrbanSim. The **trip distribution** and **mode choice** steps we have followed the exact same procedure as used for MODUS (DREIF 2008).

18.3 Integrated model: modeling approach

While UrbanSim represents a major improvement over urban models developed previously, the level of integration with the modeling of travel behavior is limited by the constraints of existing four-step travel models which imposes several significant constraints.

- The implementation of the four-step travel model use aggregate formulation. Therefore, the individual detail provided by micro-simulation in UrbanSim could not be applied in the travel model.
- User preference for mode choice, departure time choice and route choice is not reflected in this type of transport model, which are important parameters for detail modeling.
- Interaction between short-term and long-term choices made by households are not represented in this approach, leading to potential bias in evaluation of effects of transportation system changes.

UrbanSim requires a transport model so that the accessibility of an area (urban zone) can reflect the choice and preferences of the travelers who live/work/travel there. It provides the possibility of integration with a large variety of transportation models and interfaces have been developed for some of them like EMME (PSRC 2012) or VISUM (Kryvobokov et al. 2009). These models are

[41] The Enquêtes Globales de Transport (EGT) are household transport surveys in the Ile-de-France region. EGT surveys were conducted in 1976, 1983, 1991, 1997 and 2001. In 2001, some 10,500 households were interviewed, which means that some 25,000 persons described their travel behavior.

static and macroscopic. These models has been largely applied in transportation demand analysis practices and have well known characteristics but they suffer from the limitations mentioned above. They don't profit entirely from features offered by microsimulation framework that is the possibility to take into account the individual behaviors and preferences to provide economically relevant measures to be used in urban modelling or socio-economical analysis. We place our approach in the recent state of practice as discussed in TRB (2007) and Donnelly et al. (2010). Recently in the same research line the experiences are applied to the microscopic transportation model MATSim (Nicolai et al. 2011) and the activity-based transportation model like SF-CHAMP (Waddell et al. 2010). The microsimulation of traffic flow or activity based modeling of travel demand require much more detailed data and computational resources than current state of practice. The traffic flow microsimulation may not provide much more useful precision to our integrated model. Tour and activity-based models need adapted surveys that are not yet available in many cases including the Paris case study. In this study we are proposing an intermediate solution that provides a good equilibrium between data needs, computational resources and simulation time, and provision of realistic and theoretically consistent results. We use a dynamic trip based approach with microscopic demand representation. This approach takes into account the individual behaviors and preferences based on a Revealed Preference (RP)/Stated Preference (SP) survey with a limited sample size of 1000 respondents (MADDIF, see above). This survey completes usual travel survey data available at a regional level. METROPOLIS may serve this purpose. Its disaggregated nature describes the preferences of individual drivers and the congestion law finds the network equilibrium depending on variables specific to each link: the incoming flow and the average rate of occupancy. Future extension of this approach based on more detailed data (especially behavioral data) is possible. METROPOLIS deals with individual trips and it will be possible to introduce the individual values in place of distributions that are currently the only available data. Intermediate modules can be developed later to translate tour and activity data to trip data and introduce them in METROPOLIS.

18.4 Structure of the interface

METROPOLIS is coupled with UrbanSim to produce transport related inputs for the urban model. Obviously, we have to take into account the data availability constraints. For example, the dynamic behavioral parameters are estimated based on a limited size sample of about 1000 respondents. This does not permit us to estimate these parameters in function of the individual characteristics available in a microsimulation framework.

This section describes the integration process in detail along with the data transformations requirements. The METROPOLIS UrbanSim interface will be referred as MetroSim throughout this document. The interface is developed

using Python language and runs on every operating system available up to date. All data are stored in databases and managed by MySQL.

The interaction of the two simulation software can be simply described as following: UrbanSim provides the population and employment of different geographical units (in the case of Paris 1300 Communes over the region). The interface computes an Origin-Destination matrix. Then METROPOLIS calculates the accessibility of different zones based on the updated OD matrix, UrbanSim uses this new accessibility measures to simulate the evolution of population and employment. As the urban system evolves slower than transportation system, the interaction between METROPOLIS and UrbanSim is not performed each year, rather it is flexible for adjusting the scales. The developed interface handles all the data transfer and data modification, and runs the simulation automatically according to a pre-defined simulation plan.

The MetroSim interface performs all the data transfer and their modification. A Graphical User Interface (GUI) is used to control MetroSim. We have selected a file-based data transfer method. The data transfer plan is described in a simple diagram in Figure 18.2 showing the most important data transfers only.

The data transfer plan can be described as follows:

Initial input: At the beginning, the data about first OD demand and PT travel time is loaded in the MetroSim database. The first OD demand matrix represents the OD travel demand for the base year and can be obtained from survey data. The data file could be text or XML file or MySQL database.

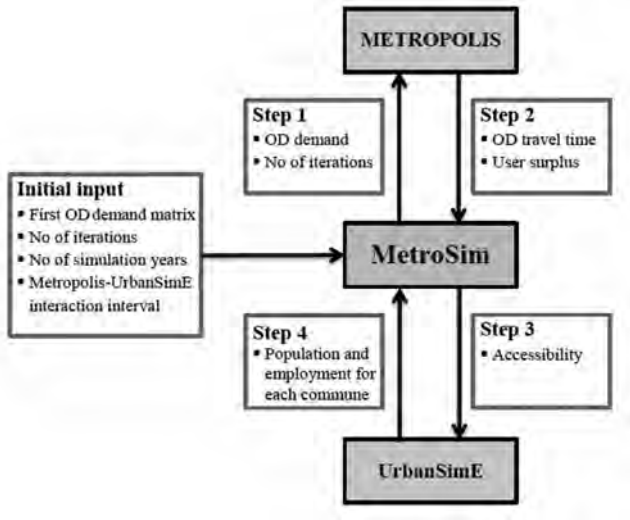


Figure 18.2 Data transfer plan.

Step 1): MetroSim supplies travel demand along with PT travel time and other simulation parameters to METROPOLIS and starts the simulation process.

Step 2): After the simulation, METROPOLIS provides the O-D travel time and user surplus for each user. As these data are stored in METROPOLIS MySQL database, MetroSim transfers the data to its own MySQL data store for further processing.

Step 3): MetroSim interface calculates accessibility based on the travel time and user surplus data generated by METROPOLIS and supplies that as an input to UrbanSim along with other simulation parameters. UrbanSim simulation runs for a fixed number of years which is equal to the UrbanSim-METROPOLIS interaction interval.

Step 4): After the simulation, UrbanSim supplies the updated population and employment for each commune to MetroSim interface. MetroSim stores these data to its MySQL database and calculates the new OD demand for METROPOLIS. Then Step 1 starts again and the process continues for a predefined number of iterations.

A flow chart showing the work flow diagram inside MetroSim interface is shown in Figure 18.3. The work flow diagram is explained briefly in Sections 18.1 to 18.3.

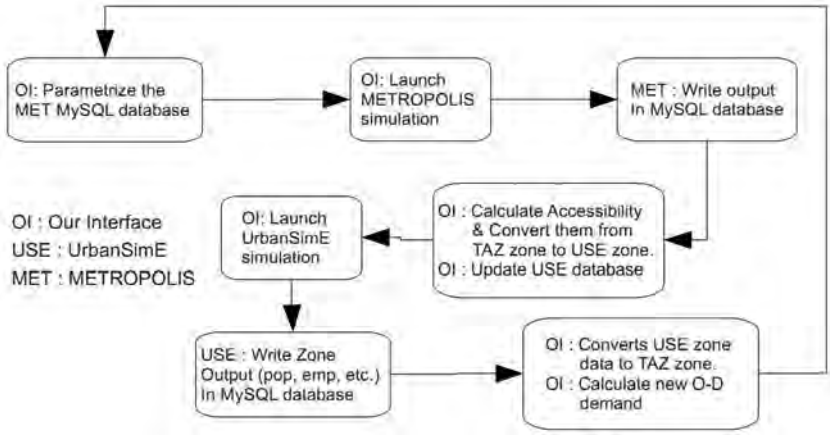


Figure 18.3 Work flow diagram inside the interface.

18.4.1 Parameterize the MySQL database and run simulation

This section describes how the METROPOLIS database is modified for a new simulation and how the simulation is launched from the interface. METROPOLIS has its own MySQL database and the structure of the database is known

to the interface (MetroSim) from the very beginning. The interface uploads new/modified OD matrices to METROPOLIS database. These new OD matrices could be externally supplied or created from the output of UrbanSim with a series of predefined computation processes by the interface. As the interface knows the structure of METROPOLIS database, it makes necessary changes in the database to link the new OD matrices with the existing tables required for simulation.

The run parameters are introduced in the METROPOLIS MySQL table named 'simulation'. Each line represents one simulation in this table. The interface makes a copy of previous simulation and modify it according to user needs. When METROPOLIS run is finished, the results are stored as tab delimited text files in the temp directory. The interface stores the tables in the MySQL database, assigns a new simulation id and updates the simulation table.

18.4.2 Processing METROPOLIS results for UrbanSim and running UrbanSim

This section describes how METROPOLIS results are processed to create the accessibility indicators. These accessibility indicators are used in UrbanSim simulation. There is two major steps in this calculation. First MetroSim computes accessibility of each zone, then it converts the data from METROPOLIS zone system to UrbanSim zone system. Indeed the zoning for transportation analysis is usually different than the zoning for UrbanSim simulation. Therefore, MetroSim is programmed to convert the data expressed for one zoning system to other, whenever the data transfer is necessary. The conversion process is described later in this section.

Accessibility calculation

METROPOLIS stores information for each user (e.g. departure time, arrival time, route used, mode used, etc.) from the simulation in one single User Vector (UV) file: a vector with all the data for the users. The accessibility is calculated based on this UV file. The indicators that are used for accessibility calculation are weighted average travel time based on Origin, weighted average travel time based on Destination, weighted average surplus based on Origin, and weighted average surplus based on Destination.

Conversion from METROPOLIS zone to UrbanSim Zone

METROPOLIS and UrbanSim use different zone systems. Therefore, conversion of zone is necessary for the interchange of data. METROPOLIS zones are named as Transport Analysis Zone (TAZ); UrbanSim zones are named USZ in this document. A simple rule is developed to convert TAZ data to USZ data.

Three conditions may arrive during conversion. First condition is when TAZ = USZ i.e. the zone location and area are same in both system. Second condition is when TAZ<USZ i.e. USZ contains more than one TAZ. The third condition is when TAZ>USZ i.e. TAZ contains more than one USZ. The situations are explained in Figure 18.4.

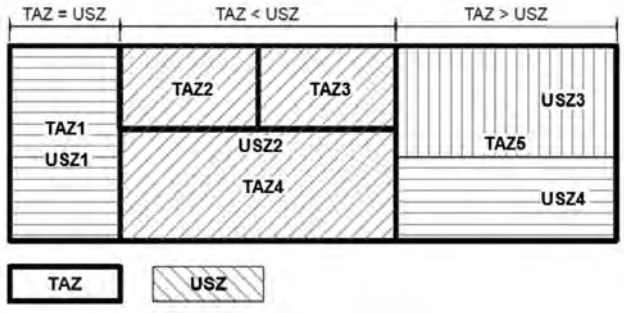


Figure 18.4 Possible combination of TAZ and USZ zones.

A conversion table is provided at the beginning of the simulation which remains unchanged during the simulation. The table contains respective zone ids and some factors to convert the data which are calculated based on some common indicators like population, employment, surface area, number of households etc. The conversion table is used for transforming the data from METROPOLIS to UrbanSim zone system.

The calculated accessibility works as the external input for UrbanSim simulation. The accessibility for each USZ is uploaded to the UrbanSim database and the simulation configuration is updated for next simulation. Then MetroSim starts a new simulation in UrbanSim. All these processes are automatically executed by MetroSim.

18.4.3 Calculation of O-D demand

UrbanSim does not produce O-D travel demand which is needed for simulation in METROPOLIS. However, it can produce trip generation matrices for each USZ. We use this information for our future calculation of O-D demand. We need to convert the trip generation matrices for METROPOLIS zone (TAZ). The procedure is almost the same as the conversion of TAZ to USZ as described before. In this case the conversion table is slightly different than before. The conversion factors are for USZ to TAZ zones.

The total transportation demand is computed for all modes and then the share of private cars is computed. As the travelers' behavior differs by their trip purpose and their captivity to the public transit, so the intermediate computations are performed separately for these classes of trips. These matrices

are produced by the interface by the following procedure described below (4.3.1 through 4.3.4).

Trip generation

At generation step, we categorize the trips in different trip purposes. In the Paris case, they are 8 as: Home to Work or professional affairs, Home to Commerces (shopping), Home to Entertainment, Work to Home, Commercial trip to Home, Entertainment to home, Work based but not home based and Home based but not work based. The trip emissions and attractions are computed as a function of population, employment or land-use statistics of each zone there is no limit for the choice of variables. The coefficients are estimated by simple linear regressions.

The independent variables are adapted to the available outputs of UrbanSim. We base our computations particularly on the number of employments rather than the floor spaces because we consider that in Paris case the first one is more reliable in the simulation results. The case of home to work trips is a special one. If the workplace choice model is implemented in UrbanSim model, the workplace of the workers will be known. In this case the active population should be divided in two categories: working people and unemployed people. The trip generation rate will be different for the two categories.

Trip distribution and mode choice

We use a nested Logit model for trip distribution and mode choice. The choices of each trip conditional to its origin are simulated based on a linear utility function for each mode and each destination. In the Paris case study, we have considered three modes: Car, Public Transport (PT), and Others. First of all we present the computation of the utilities.

Utility function

All the trips are also categorized in two groups based on their captivity to use PT or not. Therefore, we have a total of 48 (8 purposes $\times$ 3 modes $\times$ 2 captivity) matrices of O-D trips at this point of our calculation. The utility of each trip can be expressed as:

$$U_{ij}^{p.c.m} = \sum_{v} x_{v}^{p.c.m} X_{iv} + \sum_{w} y_{w}^{p.c.m} Y_{iw} + \sum_{u} z_{u}^{p.c.m} Z_{iju}^{m},$$

where, p is trip purpose; c is captive to PT or not; m is trip mode; i, j are origin and destination zone id; X_i is the matrix of characteristics for zones at origin (population density, average income, car ownership, etc.); Y_j is the matrix of characteristics for zones at destination (parking availability, employment density, etc.); Z_ij is the matrix of O-D distance or travel time that could be

differentiated by mode; x, y, z are the vectors of coefficients introduced by the operator and u, v, w represent the indexes of different variables considered in computation of utility.

Mode choice

The mode choice constitutes the lower level of the choice tree. It is conditional to a given origin and destination. The destination choice depends on the results of mode choice which we explain later. The mode choice can be expressed as:

$$N_{ij}^{p.c.m} = N_{ij}^{p.c.m^*} \left(\frac{e^{U_{ij}^{p.c.m}}}{e^{U_{ij}^{p.c.m^*}}} \right).$$

Where, N is number of trips, p, c, m stands for trip purpose, captivity and modes respectively, i, j denotes origin and destination respectively and U is the utility. Now, the utility for all mode together can be expressed as:

$$U_{ij}^{p.c.m^*} = \ln \left(\sum_{m=1,2,3} e^{U_{ij}^{p.c.m}} \right),$$

where, m^* stands for all modes.

Destination choice (trip distribution)

The trips are distributed according to the following formula:

$$N_{ij}^{p.c=1.m^*} = \left(E_i^p.CR_i^p\right) . \left(A_j^p.CR_j^p\right) .e^{\delta^{p.c}.U_{i,j}^{p.c.m^*}},$$

$$N_{ij}^{p.c=0.m^*} = \left(E_i^p.(1-CR_i^p)\right) . \left(A_j^p.(1-CR_j^p)\right) .e^{\delta^{p.c}.U_{i,j}^{p.c.m^*}},$$

where, CR_i^p and CR_j^p are captivity rates at origin and at destination respectively; δ is the vector of correction coefficients introduced by operator into the MetroSim database; E_i^p is trip generation at origin and A_j^p is trip attraction at destination.

The next step is to create O-D Matrices for METROPOLIS. The O-D matrices are sorted according to user type definition in METROPOLIS. In this study user types are defined according to certain number of origin and destination zone ids.

18.4.4 Software implementation

The interface (MetroSim) is developed in Python to provide the most possible compatibility with UrbanSim. It uses the NumPy library to improve the software efficiency in matrix computations. H5py library is used to provide a fast and semi-structured data storage facility that permits the rapid storage and

retrieval of large volumes of data. The interface software is designed to be able to run UrbanSim and METROPOLIS on two different servers. That permits the operator to use distant and specialized servers to run the two models. A graphical user interface facilitates the follow up and control of the overall simulation procedure and input of necessary parameters. All these parameters are stored in an XML file that can be merged with UrbanSim configuration XML file. Figure 18.5 shows one of the windows of the interface where the common parameters like the zoning and trip purposes are introduced.

18.5 Conclusions

In this chapter we have described a simple process to combine two software: UrbanSim and METROPOLIS, and run multiple simulations in them simultaneously. The process is automatic and an interface is built from where the simulations can be controlled. As the data structure is different for the two software, the interface automatically makes necessary data transformation. We have described the interaction process between UrbanSim and METROPOLIS and the data transformation steps.

The integrated model has been used in socioeconomic evaluation of the Grand Paris project. Grand Paris is a major public transportation and urban development project in the Paris area. About 160 km of automatic subway will be added to the existing subway network in majority around the Paris city and in near suburbs. 75 new subway stations will be constructed. In comparison with currently active 220 km subway lines and 303 stations, the project is a considerable evolution in the public transportation system of the area. The total projects investment is estimated at about 30 billion Euros over more than 20 years (up to 2035). The results will be presented in detail in the chapter dedicated to Paris case study. They show higher local congestion around revitalized urban areas that seems to be more realistic than the results of using MODUS transportation model run separately. In a parallel study integration of MATSim as transportation model has been carried out by other team of SustainCity project. They have chosen to integrate MATSim as an UrbanSim module. We refer the reader to the dedicated chapter for more information.

Acknowledgements

The authors wish to acknowledge the anonymous reviewers for their detailed and helpful comments. We also acknowledge the collaboration of Alexandre and Jeremy Grosman and the company Trace and Origin for software development.



Figure 18.5 MetroSim GUI and XML parameter file.

Part IV

Case studies

Chapter 19

Integrated land use and transport microsimulation for Brussels

Inês Cabrita, Sylvie Gayda, Ricardo Hurtubia, Dimitrios Efthymiou, Isabelle Thomas, Dominique Peeters, Jonathan Jones, Camille Cotteels, Kai Nagel, Thomas Nicolai, and Daniel Röder

19.1 Introduction

The Brussels case study consists in the development of an UrbanSim model covering Brussels and its surroundings. A MATSim travel model has also been developed for Brussels, to be plugged with the UrbanSim model, into an integrated land use and transport model. The objectives of this case study were to test the performance of UrbanSim as a land use microsimulation tool and to assess policy impacts for a few policy scenarios.

Unlike the two other case studies, no full-scale operational UrbanSim model of Brussels existed at the beginning of the SustainCity project. A prototype model had been built previously, in order to test the software abilities, but it had been developed at a very aggregate level and was unable to provide meaningful simulation results (Patterson & Bierlaire 2010). As the new model was planned to be far more ambitious and more disaggregated (both at the spatial level and at the level of the socio-economic segmentations), it has been developed on a completely new basis, requiring a completely new data collection process. New models have been developed, parameters have been estimated, validation checks have been performed, by comparison with observed data or with results from existing validated studies (notably for the transport side). Eventually, three scenarios have been simulated, a business-as-usual scenario and two policy scenarios. During this whole process, technical difficulties have been met and were overcome, in the limits of the scope of the project: mainly

difficulties in the data collection, in the validation process and in the simulation runs. Due to the high number of variables to be obtained at a very fine level of spatial analysis, some results are only partial and should be interpreted carefully. This is further detailed when discussing the results and the conclusions.

Specifically, to achieve the objectives of the case study, the following steps were carried out: (1) collection of the required data, estimation of the submodels making part of the UrbanSim model, and development of a comprehensive UrbanSim model; (2) use of MATSim as a travel model plug-in to UrbanSim; (3) definition and simulation of a set of policies with the integrated land use/transport model; (3) computation of indicators to assess the policies; (4) assessment of the performances of the simulation tool, notably on the following aspects: difficulty of the overall calibration, quality of the reconstitution of the base year situation, sensitivity of the model (direction and magnitude of the effects, elasticity), remaining limits and weaknesses.

This chapter is organized as follows: Section 19.2 contains a description of the city of Brussels and its current issues in the context of land use and transport; Section 19.3 provides a general description of the Brussels model (study area, base year, other horizons considered, structure of the model); Section 19.4 describes the datasets that were used and their limitations; Section 19.5 describes the hedonic-price and location-choice submodels estimated for the case study while Section 19.6 briefly explains the MATSim model for Brussels. Section 19.7 provides an analysis and validation of the business-as-usual (no intervention) scenario simulation results while Sections 19.8 and 19.9 describe the policy scenarios that were tested and the simulation results for each one of them.

Three partners were involved in the study: STRATEC, EPFL (École polytechnique fédérale de Lausanne), and UCL (Université Catholique de Louvain). The road traffic model, MATSim, and the mode choice model were developed by TU Berlin (Technische Universität Berlin).

19.2 The Brussels context

The area of study comprises a large region around Brussels, the capital and largest city of Belgium. This region contains well developed urban areas, as well as suburbs, semi-rural areas and several different types of land use. In Brussels city itself, an old industrial axis along a canal surrounded by poor neighborhoods with different ethnic communities makes its way through the whole city, cutting it in two parts. For some years now, this area is in a process of renewal, with important residential and office developments.

Another important feature is that Brussels, as many other European cities, undergoes a continuous out-migration of middle class families to the suburban

areas, yet for several decades, which causes urbanization of previously open spaces, commuting by car and traffic congestion. In reply to this, the Brussels-Capital regional authorities implement policies to improve the residential attractiveness of the region. On the other side, this out-migration process is compensated by a high growth of lower-income population.

Brussels also has to manage important administrative functions (including the European and other international institutions) and to elaborate strategies to integrate these functions in the city in a way both harmonious and efficient.

In the context of transport infrastructure, one of the main future projects is the Regional Express Railway Network (Réseau Express Régional or RER), to be fully implemented by 2025, which will significantly improve the accessibility between the center and the suburban areas, and is expected to have effects on household and economic activity location.

Another important point here is that, the three regions making up the Belgian federal state – the Brussels-Capital, the Walloon, and the Flemish Regions – have committed themselves to drastic reductions in their greenhouse gas emissions. This objective appears to be reachable only if it is supported by strong land use and transport policies.

In brief, some important issues that the city is currently facing are the following ones: (1) the housing supply to be developed in the Brussels-Capital Region to meet the needs of the demographic growth foreseen for the next decades; (2) the densification of office districts and namely of the European institutions district; (3) the possible implementation of an urban congestion pricing scheme; (4) the funding of the development of the public transport network and services; and (5) the pricing of the future RER (what would be the consequences of different pricing strategies in terms of modal shift from car to the RER and in terms of household and firm relocation?).

Consequently, from the very beginning of the SustainCity project, the policy scenarios which were envisaged to be simulated with UrbanSim were densification policies and urban road pricing.

19.3 General structure of the Brussels UrbanSim model

19.3.1 Study area

In the project, functional and transportation criteria were used to define the study area. It is hence much larger than the purely morphological and functional urban agglomeration.

The central part of the urban agglomeration corresponds to the Brussels-Capital Region (BCR). Then it spreads beyond the limits of the BCR, over a part of the Flemish Region and Walloon Region. The whole study area corresponds to a set of 151 municipalities (or "communes") around Brussels

(Fig. 19.1) and includes about 3 million inhabitants, which is much larger than the BCR (19 municipalities and about 1 million inhabitants).

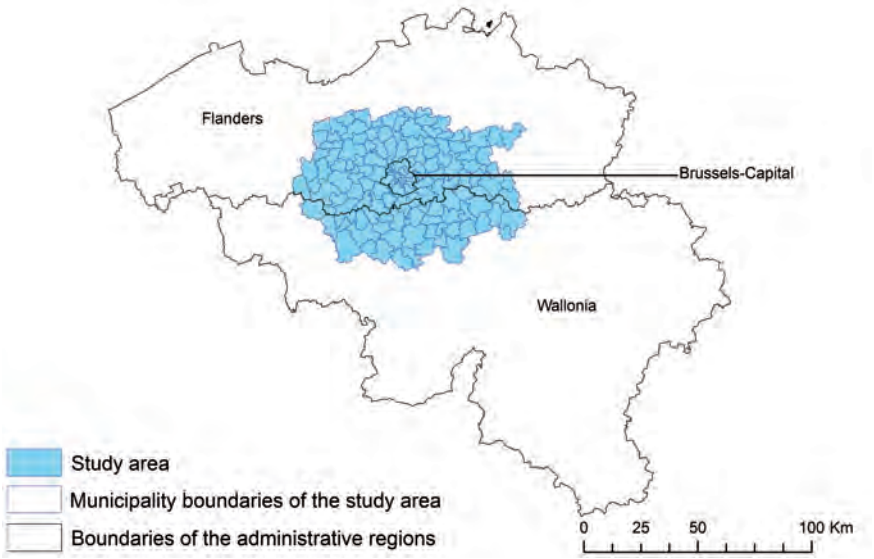


Figure 19.1 Commune boundaries of the study area of the Brussels case study (Source: IGN, Map: Stratec, 2011).

More precisely, the criteria used to delineate the study area are the following: (1) All municipalities that will be affected by the RER network. This "RER zone" is officially defined as "*all municipalities in about 30 kilometers from Brussels*" (Moniteur belge 2004), and represents 126 municipalities. (2) In addition, 21 municipalities, mainly in the northern and northwestern part of the study area, were added because they were included in the study area for the evaluation of the effects of the RER network by Stratec (Stratec 2003, 2005, Boon & Gayda 2008). (3) Finally, four municipalities (Landen, Linter, Hannut and Wasseige) located at the southeastern extremity were added since they are part of the Metropolitan Labor Area of Brussels defined by Van Hecke et al. (2009). We end up with 151 municipalities.

This study area is far larger than the delineations usually considered when studying Brussels (see Dujardin et al. 2007, Thomas et al. 2012 for reviews). Hence, the influence of the city center is "diluted" by two factors. First, several municipalities within the study area are in the commuting hinterland of other cities than Brussels. This is in particular the case in the northern part, due to the proximity of Antwerpen and Gent. Secondly, other municipalities, mainly in the southeastern part of the study area are more rural or periurban. Consequently, in terms of urban model, this study area does not correspond to an urban region, but rather to a polycentric area (this issue is discussed in details in Chapter 4). To mitigate that problem, it has been decided to use as

much as possible the delineations of urban areas proposed by Van Hecke et al. (2009) and macrozones for the evaluation of the results of the model (and the scenarios). In any case, the indicators computed for the policy evaluation were at least computed both for the whole study area and for the BCR, these latter results therefore having no border effect.

19.3.2 Zoning system

With regard to the spatial basic units, an UrbanSim model can be built at zone level, plot level or gridcell level. "Zone" here means any irregular polygon, like for example, the transport zones usually used in transport models or administrative units. For the Brussels model, considering the available data, it was decided to develop a zone-based model. The different levels of spatial disaggregation are the "municipalities" or "communes" that divide the study area in 151 units and the "zones" or "statistical sectors"[42] that divide it in 4945 units. Note that the terms "municipality" and "commune" are equivalent. In this chapter, the term "municipality" will be used.

19.3.3 Time horizons

The base year of the model is 2001 when the last National Population and Housing Census was carried out. The sub-models making up the Brussels model were therefore estimated on 2001 data. The year 2007 was used for validation purpose. The scenario simulations were carried out up to the horizon 2020.

19.3.4 Development of the model

The development of the model includes four main steps: (1) the definition of the overall structure of the model (which sub-models to include and which segmentation of population, employment, and buildings to adopt); (2) the building of the databases of agents and supply units constituting the structure of the model; (3) the estimation of the parameters of the sub-models; and (4) checks and validation of the model, as far as it is possible with the available data and resources. This last step consisted in checking the reconstitution by the model of the 2001 observed situation on various aspects (population, employment, and transport indicators), and comparing the 2007 modeled situation with the 2007 observed situation (population and employment).

19.3.5 Agents, supply units, and their segmentation

According to the basic principles of UrbanSim, the skeleton of the Brussels model consists of three databases containing the agents whose behavior is sim-

[42] The "statistical sectors" are official Belgian statistical spatial units, they are subdivisions of the municipalities.

Attribute	levels
Income level of the household (inc_h)	1 (0-744 Euros) 2 (745-1859 Euros) 2 (1860-3099 Euros) 4 (3100-4958 Euros) 5 (>4959 Euros)
Household size (hh_size_h)	1,2,3,4,5+
Number of children ($children_h$)	0,1,2+
Number of workers ($workers_h$)	0,1,2+
Number of cars ($cars_h$)	0,1,2,3+
Number of people with university degree ($univ_h$)	0,1,2+

Table 19.1 Household attributes.

ulated and the supply units: a household database, a job database, and a building database.

Households are characterized by their size, income level, number of children, number of workers and education level of its members. Table 19.1 describes the values for each attribute level.

Buildings were divided between residential and non-residential buildings. Residential buildings were classified in three types of single-family houses (fully detached, semi-detached, and attached) and one type of multi-family units (apartments). Each zone is assumed to have one representative building by type (with a variable number of residential units in each of them), generating a total of 19,780 possible location types (combinations of zones and building types). Ten types of non-residential buildings were considered, each being specific to an economic activity type. Table 19.2 summarizes the types of dwelling units and other buildings considered in the model.

19.4 Data collection

As seen above, one of the main characteristics of UrbanSim is the requirement of disaggregate data. The three main databases (households, jobs, and buildings) have to be provided for the base year (2001) and, besides that, total figures of households and jobs have to be provided for each simulation year (i.e. up to 2020 in the Brussels case). Additionally, a few other databases are required: data on the development constraints and on the past history regarding the development projects. A few general (non spatially disaggregated) parameters are needed, such as agent relocation annual rates and average building vacancy rates. Eventually, specific data for the MATSim road traffic model and the mode choice model are also required.

type id	description	category
1	isolated houses	residential
2	semi-attached houses	residential
3	attached houses	residential
4	apartment buildings	residential
5	agricultural	non-residential
6	quarrying, mining	non-residential
7	industry	non-residential
8	office (private sector, including bank)	non-residential
9	shops, retail	non-residential
10	hotels, bars, restaurant	non-residential
11	government and public service	non-residential
12	education	non-residential
13	health	non-residential
14	leisure activities	non-residential

Table 19.2 Building types.

According to the modeling dates, data was mainly collected for the years 2001 and 2007.

This section describes the main data sources that were used. In some cases, it also specifies some particular data processing that was performed.

19.4.1 Household data

In Belgium, the most exhaustive data source on population is the National Population and Housing Census, knowing that the last census was performed in 2001[43] (Vanneste et al. 2007, Verhetsel et al. 2009, Thomas et al. 2011). Beside the census, other rich sources exist: the National Population Register (yearly updated), the Crossroads Databank for Social Security, which combines itself various sources and provides an overall picture on the activity of the population, at individual level, and the National Labor Force Survey.

For the base year, a set of data from the 2001 census was obtained at an aggregate level from the Directorate-General for Statistics and Economic Information (Federal Public Service (FPS)-Economy).

The first idea was to request an extract of the census, with data at an individual level. However, the Belgian administration refused to provide these data because of privacy issues. Consequently, it was decided to build from the aggregated data a "synthetic population" of individual households, whose location is given at the zonal level, using a Markov Chain Monte Carlo simulation based approach (Farooq et al. 2013). The generation of synthetic populations

[43] Precisely the census is dated 1st of October 2001.

is a common practice in land use/transport modeling, especially when individual level data are not available due to strict privacy policies. In the US for example, it often happens that modelers do not have access to micro census records, and therefore must use synthetic population generation algorithms to create the data. For more details on the synthetic population, see Chapter 5 and 6.

Generally speaking, the main limitations of the 2001 census are that the rate of missing questionnaires (about 5%) is strongly related to the socio-economic level and that the survey is limited to persons legally registered in the National Population Register which leads to a certain lack of information, especially in big towns like Brussels.

Data on household income was available by statistical sector for the year 2001, from fiscal statistics set up by the FPS-Economy. This database provides average and median income at the level of the statistical sector for whole Belgium, on the basis of income tax return, as well as the interquartile difference, the interquartile coefficient and the interquartile asymmetry. It has to be noted that the number of income tax returns (or "fiscal households") is not always equal to the number of "social" households, in a statistical sector or a municipality. Indeed, there is one income tax return by fiscal household; and a fiscal household cannot include more than 2 persons having an income. Also, data are only available for statistical sectors that have more than 30 income tax returns.

For the validation year, population data were collected from the National Population Register, which provides the total population at the statistical sector level, classed by age group (of 5 years intervals), on the 1st of January 2008.

Data on activity of occupied active people, at the municipal level, were taken from the Crossroads Bank for the Social Security (Office National de l'Emploi, or ONEM). This database gathers data on employed, civil servants, and self-employed persons at the municipal level on the 31st of December 2007. All the information is registered at the home place of the persons.

19.4.2 Employment data

Data on employment for the base year and the validation year were collected from the same sources.

In general, data on employment are less accurate than data on population for various reasons: the level of spatial disaggregation is less detailed; often, a compilation of different databases (corresponding to different status: employees, self-employed workers, civil servants, etc.) has to be done in order to give the full picture, but with risks of double counting or missing data.

For the Brussels model, the main data sources on employment were the National Social Security Office for Employees and Civil Servants (Office National de Sécurité Sociale, or ONSS) and the National Institute for Social Security for self-employed persons (Institut National d'Assurances Sociales pour Travailleurs Indépendants, or INASTI). The databases were available from the FPS-Economy for each year until 2009.

The ONSS database includes all employee and civil servant jobs (i.e. all the jobs whose workers take part in the Belgian employee social security system) at the municipal level on 30th of June of each year. The employees are registered at their workplace, not at their place of residence.

The INASTI database describes the activity of self-employed persons at the municipal level (or district, province, region, national level) on the 31st of December of each year. The self-employees are registered at their company address. This leads to a difficulty: the address is the place where the company is registered, but in many cases this address corresponds to the home place of the self-employee and not to his actual workplace.

Further sources were exploited for estimating figures regarding the international jobs. The Brussels Institute of Statistics (Institut Bruxellois de la Statistique, or IBSA) provided data on jobs in international public institutions for the BCR (i.e. the European Commission, the Committee of the Regions, the European Council, the European Parliament, the NATO and the Western European Union) from 2001 to 2009. Estimation on international jobs in the BCR for 2001, 2005, and 2011 were also gathered from the Stratec study "Updating and adaptation of the Urban Transport Plan for the Brussels-Capital Region" (Stratec 2006).

Data from the national Labor Force Survey (LFS) was also used. The LFS is a socio-economic household survey conducted by the FPS-Economy, whose primary objective is to classify the population of working age (aged over 15) into three exhaustive and distinct population categories (employed, unemployed, and inactive). It provides, on each of these categories, descriptive and explanatory data. There is specific information concerning the home-based work status. The information is collected through face-to-face interviews and is valid for the 31st of December of each year, from 1999 to 2008. Households with only inactive members (aged over 64) may also be interviewed by telephone. The sample size is of 90,000 inhabitants aged 15 and over. Roughly, the representativity of the sample is sufficient for providing ratios, percentages, etc, by administrative Region, but not at a finer spatial level.

Employment data was only available at the municipal level. Distribution within zones and buildings was done through a Monte Carlo simulation, following the observed distribution of available non-residential surface by zone and building type. The number of activity sectors was assumed to be equal to the number of types of non-residential buildings, meaning that each type of

job can be located in a specific type of building (e.g. jobs in industrial sector can be located in industrial buildings).

19.4.3 Land use and real estate data

Informations on land-use and buildings are based on the Land Registry database. It is a cadastre of real estate goods, provided by the "Administration Générale de la Documentation Patrimoniale", at the plot and building levels. The database is federal and hence the same definitions of the variables are used in all three Belgian administrative Regions (since Brussels sprawls out of its administrative boundaries, these three regions are present in the study area). Its purpose is fiscal and juridical: it provides information for calculating the taxes to be paid by the owner of the plot/building(s). This ensures that all plots are taken into consideration (no sampling) but also implies some drawbacks.

For the entire country, 221 types of plots are defined (Table 19.3): 157 for built-up plots and 64 non-built. For instance, the most frequent types among the 2,047,675 plots in the study area are "houses" (42% of built-up plots) and "farmland" (44% of non-built plot). Data are geocoded allowing for crossing them in a Geographical Information System (GIS).

Field	Description	Note
CaPaKey	Identification code for the plot	
Nature	Type of plot	
IndiceCC	Classification index	Only for built-up plots
Type	Construction type (A: 2 side walls, B: 1 side wall), C: isolated building with 4 front walls)	Only for built-up plots
AnnéeCstr	Year the construction was finished	Only for built-up plots
AnnéeMod	Year of last renovation	Only for built-up plots

Table 19.3 Land Registry data for each plot.

2009 is the first year for which digitalized data were available. On this basis, the situation in 2001 was estimated by removing all buildings having a construction date posterior to 2001. The different plot types were manually re-aggregated into 14 categories to define the buildings represented in the model (see Table 19.2).

For each building type, the Land Registry database allows to compute the total surface occupied in t_0 by a building type i in the zone j. This surface is hereafter referred to as the *actual* land-use. Note that for residential buildings

the value of this actual land use is, in UrbanSim, the total number of residential units rather than the total surface. Hence, for such buildings, the number of plots is used, instead of the sum of their surface.

The database has to be interpreted with caution due to the following reasons:

- Types of land use were too numerous to be relevant as UrbanSim's buildings. Hence reclassification was absolutely required, but some discrepancies can be present in this reclassification.
- Types of land use are identified by a word (e.g. "houses" for houses plot) rather than an alpha-numerical id. Moreover, the word used changes according to the linguistic region where the plot is located, e.g. houses are registered as "maisons" in the French-speaking region and "huis" in the the Dutch-speaking region. This may also lead to errors in the reclassification process.
- Errors on the date of construction may also be observed, especially for renovation or extensions. This can be due to the fiscal nature of this database (fraud) but also to a lack of rigorous follow-up strategies of some local sections of the Administration. Hence, errors will be much smaller on recently built buildings than on old buildings.
- Single family houses plots generally encompass in a single plot both the footprint of the house and the garden. Nevertheless, looking closer at the data it was noticed that some houses plot follow the exact limit of the house, and that this plot is surrounded by a plot identified as "garden". Hence, it is not accurate to compute the built-up areas on the basis of the plot. Rather, an overlay of plot and buildings were made, and built-up surfaces were computed using the footprint of these buildings.
- Some features are not included in the Land Registry database, such as the transportation network.

UrbanSim also requires information on real estate prices. Data on housing real estate prices were available from the FPS-Economy at the municipal level. The database provides the average price of housings sales by type of dwelling (2, 3, or 4 front-wall houses and flats) on the 31st of December from year 1985 to 2008. It gives the number of sales and the total value of these transactions, on which the average selling price is computed.

19.4.4 Development constraints

Development constraints are not implemented explicitly in the current version of the model. Rather, the *actual* and *potential* area occupied by each building type in each basic spatial units are defined within the building table. This *potential* land use is defined as the total surface (footprint) that can be affected to a building type i in the zone j, whether the corresponding plots are vacant

or already built. Note that the *developable* surface, although not represented explicitly, is thus simply the difference between *potential* and *actual* land use.

The sources used were the Regional Master Plans of Wallonia, Brussels and Flanders. They define the type of buildings allowed on each plot: agricultural land, residential areas, industrial areas, natural areas, and so on (these data are geocoded and can be overlaid with the plot/building(s) from the Land Registry database). These databases suffer the following drawbacks:

- Since they are regional, the definitions used in one Plan do not converge completely with those of the two other Plans. Although main categories (e.g. agricultural areas, residential land use, industrial land use, etc.) are comparable, some discrepancies exist.
- The categories defined by the Regional Master Plans are not mono-functional. For instance, it is allowed to build small retail buildings in areas affected to residential developments.
- The areas defined by the Plans are subject to changes and/or revisions.

19.4.5 Development events history

Development events have been extracted from the Land Registry database, by selecting all plots having a construction date from 1991 to 2001 included.

19.4.6 Data for the transport model

A MATSim model was also developed for Brussels: this has been done in order to stay within an overall consistent microsimulation framework for both land use and transport. However the MATSim model has been developed with limited resources and has not been calibrated as far as it would have been desirable. As a reference for comparison, an existing model was used, the IRIS 2 transport model of the BCR (Stratec 2006), covering an area similar to the SustainCity study area. The IRIS 2 model was developed with the SATURN software for the road traffic model, which is a dynamic traffic model, and with the VISUM software for the public transport model.

The MATSim road traffic model requires a road network which was here built using the data from OpenStreetMap (2013). It consists of 10,861 nodes and 19,830 links.

It also requires a list of persons (workers) with information about their household (origin) and job (destination) location that was simulated following the synthetic households and observed origin-destination matrices for trips. The 2001 census provides information on home-to-work relationships by giving a matrix of persons distributed according to their residence place and their workplace at the municipal level.

The car kilometer costs for different years of simulation were estimated based on vehicle operating costs and fuel costs gathered from different sources:

(1) FPS-Economy for the official prices of fuel from 1988 to 2013; (2) Federal Planning Bureau (FPB)[44] for the predicted evolution of fuel prices until 2030; (3) MEET (1999) for the curves of fuel consumption depending on speed by vehicle type; (4) FPS-Finances for the vehicle operating costs; and (5) FEBIAC (Belgian federation of automative and cycle industry) for the distribution of cars according to the type of fuel in 2001.

For the mode choice model in MATSim, data on public transport travel times, public transport travel distances, public transport kilometer costs and car kilometer costs were collected. The public transport data were collected from time schedules.

In the IRIS 2 model, 7 macrozones were defined (Fig. 19.2). These macrozones are arranged concentrically around the city of Brussels: 1 corresponds to the historical city centre; 2 and 3 to the rest of the BCR; 4, 5, and 6 to the periurban area and 7 to the rest of Belgium. For the validation of the model, car travel times and mode shares were computed for the relationships between the 7 macrozones and were compared to the results of the IRIS 2 model.

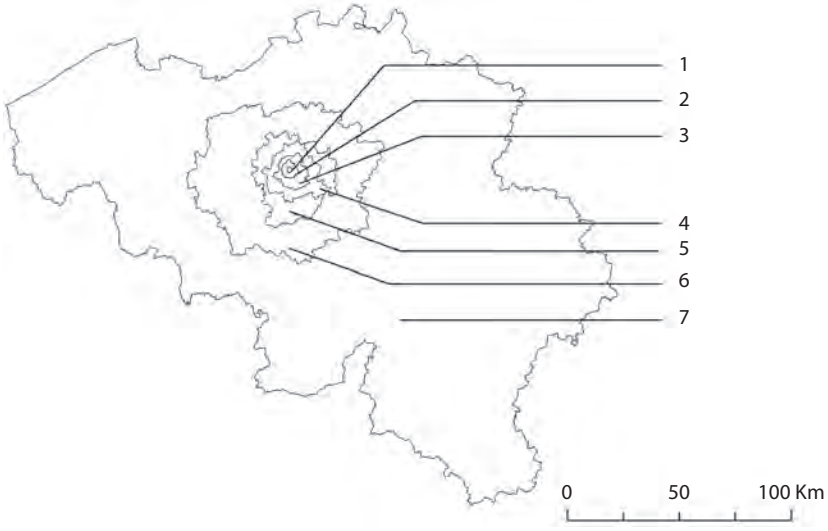


Figure 19.2 The 7 macrozones as defined by Stratec (2006): 1 corresponds to the most central zone and 7 to the rest of Belgium.

19.4.7 Other data

The macroeconomic data required by UrbanSim are the annual household control totals, the annual employment control totals, the annual household and

[44]The Federal Planning Agency is the Belgian public agency in charge of the official demographic and socioeconomic projections as well as the projections regarding socioeconomic and environmental policy issues.

job relocation rates and the target building vacancies.

The annual household control totals are the aggregate targets for households (total number of households for the study area) for the years of simulations (from 2001 to 2020). The data were estimated based on the 2001 population data census (FPS-Economy), on the 2007 National Population Register (FPS-Economy), on population projections (FPB), and on the evolution of the annual average household size between 1991 and 2008 per Region (FPS-Economy).

The annual employment control totals are the aggregate targets for employment (total number of jobs for the study area) by sector and for the years of simulations (from 2001 to 2020). The data were obtained from FPB. Data on activity of self-employed workers and employee/civil-servant jobs were available from 1995 to 2016 at the regional level and from 2017 to 2020 at the national level.

The main source for vacancy rates of dwellings was a 2005 publication from RBDH (Rassemblement Bruxellois pour le Droit à l'Habitat) which provides estimations based on water consumption for the BCR in 2003.

The main sources for the vacancy rates of buildings dedicated to economic activities were the following publications: the Retail Observatory (2005, 2007, and 2008), the Office Property Observatory (2007, 2008, 2009, and 2011) and the Production Activities Observatory from 1997 to 2011. These reports are produced by the BCR Administration and the Brussels Regional Development Agency (BRDA).

Estimates of the household annual relocation rates were found for Belgium in Bodeux (2002) and for the Walloon Brabant, all of whose municipalities belongs to the study area, in Zitouni et al. (2004) (Table 19.4).

Source	**Year**	**Study area**	**Annual households relocation rate**
Bodeux (2002)	1997-1998	Belgium	9.14%
Zitouni et al. (2004)	1998	Walloon Brabant	9.20%

Table 19.4 Estimates of the household annual relocations rates.

19.5 Sub-model estimation results

19.5.1 Real Estate Price Model

The real estate price model (REPM) used in the Brussels application of UrbanSim is a semi-log linear regression based on Ordinary Least Squares (Franklin & Waddell 2003, Waddell & Ulfarsson 2003*a*). It predicts the average value per dwelling unit, for every year of the simulation.

The results for the REPM are shown in Table 19.5. Two sub-models were estimated, one for the houses (detached, semi-detached, and attached), and one for the apartments. There were no available observations of non-residential real estate prices, and therefore no model was estimated for this case.

The price of houses and apartments is positively affected by the car accessibility of the zone and the percentage of green areas in the municipality. Sociodemographic characteristics that have a positive impact are the percentage of households with high income in the municipality, and the logarithm of the population density.

In order to reduce potential endogeneity issues, an instrument variable (de Palma & Picard 2005) was included in this model's specification. This is the municipal income tax, which is the part of the income tax levied by the municipalites, and has a negative effect.

The REPM presented in Table 19.5 is mostly based on location (neighborhood or municipality) attributes. The only building-specific attribute used is the residential surface, which is positive. Despite the fact that the literature shows that prices are largely explained by attributes of the buildings (Löchl & Axhausen 2010, Efthymiou & Antoniou 2013), the presented models are still able to capture land use effects that should be relevant for the modeling purposes.

19.5.2 Household Location Choice Model

The Household Location Choice Model (HLCM) estimation results are presented in Table 19.6. As the other location choice models, the HLCM predicts the probability that an agent chooses a zone (statistical sector) as a location.

All parameters are statistically significant and have the expected signs. Price has a negative effect in the utility for all households, no matter their income level. The presence of high income households attracts other households of high income but makes locations less affordable for low income households. Households with university degree holders prefer to be located in zones with high ratio of university degree holders. This is consistent with the expected social agglomeration and segregation effects usually observed in residential location.

Car accessibility increases the utility of car-owning households. Households with workers prefer to be located closer to the central business districts while households not owning any cars select locations close to the rail stations. Municipalities with high percentage of green areas are more attractive.

The spatial alternative-specific constant that accounts for unobserved attributes, indicates the attractiveness of the central locations. This constant is active when the location is inside the BCR.

Houses (n=14835)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
constant	-			11.5407	857.94
$\beta_{\text{car-acc}}$	Car accessibility	zone	%	0.0020	4.09
β_{green}	Green area score	commune	0 to 1	0.1349	10.81
$\beta_{\text{income-high}}$	Percentage of high income (>3) households	commune	%	0.0260	60.02
β_{tax}	Housing tax	commune	%	–0.0681	–47.75
$\beta_{\text{pop-den}}$	Logarithm of population density	commune	ln(pop/hectare)	0.0591	56.33
β_{sqm}	Surface	building	m^2	0.0005	8.751
R^2=0.59					
Apartments (n=4945)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
constant	-			11.2914	368.69
$\beta_{\text{car-acc}}$	Car accessibility	zone	%	0.0046	4.09
β_{green}	Green area score	commune	0 to 1	0.4128	14.24
$\beta_{\text{income-high}}$	Percentage of high income (>3) households	commune	%	0.0225	22.67
β_{tax}	Housing tax	commune	%	–0.0334	–10.13
$\beta_{\text{pop-den}}$	Logarithm of population density	zone	ln(pop/hectare)	0.0020	1.82
β_{sqm}	Surface	building	m^2	0.0002	1.89
R^2=0.31					

Table 19.5 Real Estate Price Model.

(n=48526)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Households with car * Car accessibility	household * zone	0 or 1 * (logsum)	0.0106	2.95
β_{educ}	Household with high education level * Ratio of university degree holders in zone	household * zone	0 or 1 * ratio	3.6401	27.97
β_{green}	Green area score	commune	0 to 1	0.1924	2.62
$\beta_{\text{income-low}}$	Households at income class 1 or 2 * Ratio of hh with high income over all hh	household * zone	0 or 1 * ratio	−2.8948	−13.12
$\beta_{\text{income-high}}$	Households at income class 4 or 5 * Ratio of hh with high income over all hh	household * zone	0 or 1 * ratio	4.8074	12.95
β_{workers}	Households with workers * Log distance from CBD	household * zone	0 or 1 * ln(meters)	−0.1453	−12.44
β_{rail}	Households without cars * Distance from rail station <1000m	household * zone	0 or 1 * meters	0.3681	12.82
β_{price}	Logarithm of transaction price	building	log(euros)	−1.0298	−29.09
ASC_{BCR}	Central Brussels area (central communes)	commune	0 or 1	0.8738	44.64
Log-likelihood=−95751					

Table 19.6 Household Location Choice Model.

19.5.3 Employment Location Choice Model

The Employment Location Choice Model (ELCM) is subdivided in eight sub-models, one for each type of economic activity. Table 19.7 shows the estimation results for each sub-model. Jobs in the agricultural and mining sectors are not considered for modeling purposes, because of their low share over all jobs.

Employment location choice of each sector is positively affected by the density of jobs of the same sector in the municipality. The logarithm of non-residential surface and the car accessibility in zone, have a positive impact, when significant. The results for the industry jobs location sub-model show that there is a negative effect of the density of jobs in the zone, which is only the case in this particular sub-model. Office jobs prefer to locate in zones with agglomeration economies and therefore favor density of jobs of the same type. Job density also has a positive effect in the utility for office jobs, probably because office jobs are service providers and prefer to locate near potential clients. Retail jobs also have benefits from the agglomeration economies and therefore the presence of jobs of the same type and of jobs in general have a positive effect in their location preferences. Finally, it is noted that health and leisure activities prefer to be located in municipalities with high percentage of high income households.

Industry (n=13943)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	-0.0627	-7.43
β_{sam}	Logarithm of non residential surface	building	ln(m^2)	1.2514	118.65
$\beta_{\text{ind-den}}$	Density of jobs in industry sector	commune	jobs/hectare	0.0782	27.47
Log-likelihood=-13634					
Office (n=14937)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{gov-den}}$	Density of jobs in public sector	commune	jobs/hectare	-0.0212	-6.36
$\beta_{\text{off-den}}$	Density of jobs in private sector (office)	commune	jobs/hectare	0.0152	4.93
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	0.6641	70.19
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	-0.0057	-12.08
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.5227	72.36
Log-likelihood=-22791					
Retail (n=3886)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Car accessibility	zone	logsum	0.0384	3.50
$\beta_{\text{ret-den}}$	Density of jobs in retail sector	commune	jobs/hectare	0.1643	4.43
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	0.0780	5.09
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	-0.0036	-2.19
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.8906	51.24
Log-likelihood=-6443					

Table 19.7 Employment Location Choice Model *(continued on next page)*.

Hotels/Bars/Restaurants (n=2013)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Car accessibility	zone	logsum	0.0427	3.21
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	0.3854	22.82
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	-0.0076	-7.10
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.3377	23.73
$\beta_{\text{hbr-den}}$	Density of jobs in hotels/bars/restaurants	commune	jobs/hectare	0.2018	10.44
Log-likelihood=-4923					
Government and public service (n=8471)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{off-den}}$	Density of jobs in private sector	commune	jobs/hectare	0.0125	6.69
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	0.7523	58.37
$\beta_{\text{pop-den}}$	Logarithm of population density	commune	ln(pop/hectare)	-0.0045	-7.69
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.5081	44.25
Log-likelihood=-10973					
Education (n=3775)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{edu-den}}$	Density of jobs in education sector	commune	jobs/hectare	0.2208	14.08
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	0.1824	11.37
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	-0.0075	-7.65
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.8405	46.04
Log-likelihood=-5995					
Health (n=5099)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{high-inc}}$	Percentage of households in high income scale (>3)	commune	%	0.0564	9.86
$\beta_{\text{hea-den}}$	Density of jobs in health sector	commune	jobs/hectare	0.1832	17.10
$\beta_{\text{job-den}}$	Logarithm of job density	zone	ln(jobs/hectare)	0.3708	32.00
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	-0.0129	-13.04
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.4908	41.29
Log-likelihood=-10493					
Leisure activities (n=1315)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{high-inc}}$	Percentage of households in high income scale (>3)	commune	%	0.0837	6.58
$\beta_{\text{lei-den}}$	Density of jobs in leisure sector	commune	jobs/hectare	0.2978	16.43
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	0.0133	11.17
β_{sqm}	Logarithm of non residential surface	building	ln(m^2)	0.6327	27.56
Log-likelihood=-2349					

Table 19.7 Employment Location Choice Model *(concludeded from the previous page)*.

19.5.4 Residential Development Project Location Choice Model

Estimation results for the Residential Development Location Choice Model (RDPLCM) are presented in Table 19.8. The models are estimated over data for real estate developments that took place in the ten year period previous to the base year and, therefore, are not representative of all existing supply in the city. All types of residential development tend to agglomerate and therefore have a positive parameter for the logarithm of the number of buildings of the same type. The dwelling categories "semi-detached" and "attached" were grouped for the purpose of this research, because of their similar characteristics.

Detached (n=59558)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
β_{price}	Logarithm of price of detached houses	zone	ln(euros)	1.5334	59.31
β_{units}	Logarithm of number of detached house units	zone	ln(sum)	1.6578	338.21
Log-likelihood=-160082					
Semi-detached and Attached (n=20119)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
β_{price}	Logarithm of price of semi-detached and attached houses	zone	ln(euros)	0.3013	7.06
β_{units}	Logarithm of number of semi-detached and attached house units	zone	ln(sum)	1.1172	164.97
$\beta_{pop\text{-}den}$	Population density	commune	pop/hectare	0.4097	37.48
Log-likelihood=-58729					
Apartments (n=5119)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
β_{price}	Logarithm price of apartments	zone	ln(euros)	0.1823	2.38
β_{units}	Logarithm of number of apartment units	zone	ln(sum)	0.1823	2.38
$\beta_{pop\text{-}den}$	Population density	commune	pop/hectare	1.0609	85.62
Log-likelihood=-12286					

Table 19.8 Residential Development Project Location Choice Model.

Residential buildings are developed preferentially in zones with a higher price of residential units. The population density of the municipality has a positive impact for semi-detached, attached, and apartments, but is insignificant for detached houses.

19.5.5 Non-Residential Development Project Location Choice Model

The Non-Residential Development Project Location Choice Model (NRDPLCM) models the location of the developed non-residential projects. Eight sub-models were estimated in UrbanSim, one for each of the building types: 1) industrial, 2) office, 3) shops, 4) hotels/bars/restaurants, 5) government and public service, 6) education, 7) health, and 8) leisure activities. Since there were not a significant number of development projects in the past, sub-models regarding quarrying and agricultural buildings were not estimated.

The estimation results are shown in Table 19.9. New non-residential supply tends to locate in places that already show agglomeration and with high concentration of other activities in general.

Development of building projects that host private services is negatively affected by the population density of the municipality and the logarithm of the total population in the zone, and positively by the logarithm of total number of jobs in zone. Another factor that affects the development of retail buildings is the number of jobs in the zone. The more jobs, the more preferable the zone is for the development of such infrastructure. The number of citizens in a zone is a positive determinant of the location.

Industry (n=2770)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Car accessibility	zone	logsum	-0.0659	-6.25
$\beta_{\text{ind-den}}$	Density of jobs in industrial sector	commune	jobs/hectare	-0.0705	-10.32
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.4251	33.19
Log-likelihood=-10949					
Office (private sector) (n=767)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Car accessibility	zone	logsum	0.0906	3.61
$\beta_{\text{off-den}}$	Density of jobs in private sector	commune	jobs/hectare	0.0269	3.32
$\beta_{\text{pop-den}}$	Population density	commune	pop/acre	-0.0318	-5.86
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	1.1741	34.43
$\beta_{\text{ln-pop-zone}}$	Logarithm of total number of population	zone	ln(sum)	-0.1460	-5.30
Log-likelihood=-1953					
Shops (n=1466)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.4451	21.84
$\beta_{\text{ln-pop-zone}}$	Logarithm of total number of population	zone	ln(sum)	0.3899	12.62
Log-likelihood=-5451					
Hotels, bar, restaurants (n=107)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Car accessibility	zone	logsum	0.2219	2.73
$\beta_{\text{hbr-den}}$	Density of jobs in hotels/bars/restaurants	commune	jobs/hectare	0.1600	1.56
$\beta_{\text{pop-den}}$	Population density	commune	pop/hectare	-0.0365	-2.89
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.7093	8.81
Log-likelihood=-359					
Government and public service (n=264)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.7184	15.57
$\beta_{\text{ln-pop-zone}}$	Logarithm of total number of population	zone	ln(sum)	0.1059	2.24
Log-likelihood=-932					
Education (n=140)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.3591	6.21
$\beta_{\text{ln-pop-zone}}$	Logarithm of total number of population	zone	ln(sum)	0.3539	5.77
Log-likelihood=-533					
Health (n=225)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.3523	5.45
$\beta_{\text{ln-pop-zone}}$	Logarithm of total number of population	zone	ln(sum)	0.5394	7.55
Log-likelihood=-840					
Leisure activities (n=970)					
Name	**Definition**	**Level**	**Specification**	**Coefficient**	**t-value**
$\beta_{\text{car-access}}$	Car accessibility	zone	logsum	-0.0240	-1.70
$\beta_{\text{lei-den}}$	Density of jobs in leisure sector	commune	jobs/hectare	3.2041	8.41
$\beta_{\text{ln-jobs-zone}}$	Logarithm of total number of jobs	zone	ln(sum)	0.2371	11.14
Log-likelihood=-3867					

Table 19.9 Non-Residential Development Project Location Choice Model.

19.5.6 Workplace Choice Model for Residents

The Workplace Choice Model for Residents (WCMR) assigns jobs to workers of the households. For its estimation, a table with information about the household and job of each person at the base year was created. The model con-

tains two variables: the car accessibility per zone, whose coefficient is positive, and the distance between the house and work location of the person, whose coefficient is negative.

19.6 The MATSim model for Brussels

The MATSim model is described in detail in Chapter 17. The main principles are summarized here and the model specifications for the Brussels case study are described.

MATSim is a microscopic ("agent-based") traffic simulation model. Every single part of the traffic system (e.g. a synthetic traveller) is simulated separately and is represented by a so-called agent. These agents typically have at least one full daily plan, each containing desired activities and desired transport modes. All agents compete for system-resources, e.g. the flow capacity of a street (Balmer et al. 2009).

When UrbanSim moves forward in time from year to year, it calls MATSim in regular intervals to generate the traffic assignment and update the accessibility of the zones and the travel indicators of the persons (travel time and chosen mode), that are used in the next iteration as input for the UrbanSim sub-models (see Chap. 17).

Besides the road traffic model, a mode choice model was implemented within MATSim. It includes a learning process: the travelers select transport options (modes and paths), try them out and compute a utility based on the performance of the options. Subsequently, they choose between the options that they have tested, based on a logit model. In this sense, the MATSim model is slightly different from the "classical" logit mode-choice models. In this model, agents are allowed to choose between car, public transport and walk.

As for the simulation of public transport, there are several variants of how the generalized cost between public transit stops can be calculated (see Chap. 17). For the Brussels case study, travel times and distances for each pair of public transport stops were provided as inputs. Agents that plan to use the public transport are forced to walk to the next stop. During the simulations of the integrated UrbanSim-MATSim model, there is no process that updates the public transport travel times. Moreover, the defined scenarios (see Sect. 19.8) are not scenarios of public transport policies, which means that the public transport supply is the same as in the reference scenario assuming business-as-usual.

Travel times for agents walking were calculated with the beeline distance and an average speed of 1 m/s. The mode walk was allowed in the simulations because the density of the street network in the model is very low, even in the center. Allowing the mode walk gives the agents a chance to reach nearby

activity locations by a short beeline walk rather than by a much longer trip on the network.

The relevant part of the MATSim mode choice utility function reads, in linearized form,

$$U_{\text{mode}} = \text{Const}_{\text{mode}} + \beta_{\text{money}} \cdot \text{money} + (\beta_{\text{ttime,mode}} - \beta_{\text{perf}}) \cdot \text{ttime}_{\text{mode}} \tag{19.1}$$

where

$\text{Const}_{\text{mode}}$ is the mode constant;

β_{money} is the marginal utility of money;

$\beta_{\text{ttime,mode}}$ is the marginal utility of spending time traveling by mode;

β_{perf} is the marginal opportunity cost of time.

The marginal utility of money and time were set in order to get a value of time equal to 9€/h (value for 2001, inspired from Stratec 2006). The alternative specific constant for the public transport, Const_{pt}, describes the inhibition threshold to use public transport. During the calibration process, a certain number of experiments, varying this parameter between −4 and −1, was done. −1 was found to deliver the closest results to the mode shares from the IRIS 2 model (Röder et al. 2013).

19.7 Validation of the business-as-usual scenario

Using the estimated models described in Section 19.5 a simulation with no modifications from the base-case scenario (business-as-usual) is performed. Comparing the situation simulated by the model with the observed situation provides an indication of the model strengths and weaknesses in predicting the real socio-economic (population and employment) and transport (mode shares and travel times) situation. The first task of the model is to reconstitute as far as possible the observed 2001 situation. The reconstitution of the year 2007 by the model was also analyzed.

The time interval for the simulation was set to 5 years for MATSim. A 5% sample of the agents that make trips to work was randomly drawn by MATSim. This sample size was used for all the simulations in order to speed up the computation and avoid memory limitation issues. 5%, and in fact even 1%, has been found sufficient in previous studies to obtain plausible congestion patterns (Nagel 2008, 2011), at least as long as no schedule-based public transit is included (Neumann et al. 2012). Clearly, larger sample sizes produce fewer fluctuations between runs and thus smaller statistical errors, e.g. for zone-to-zone travel times.

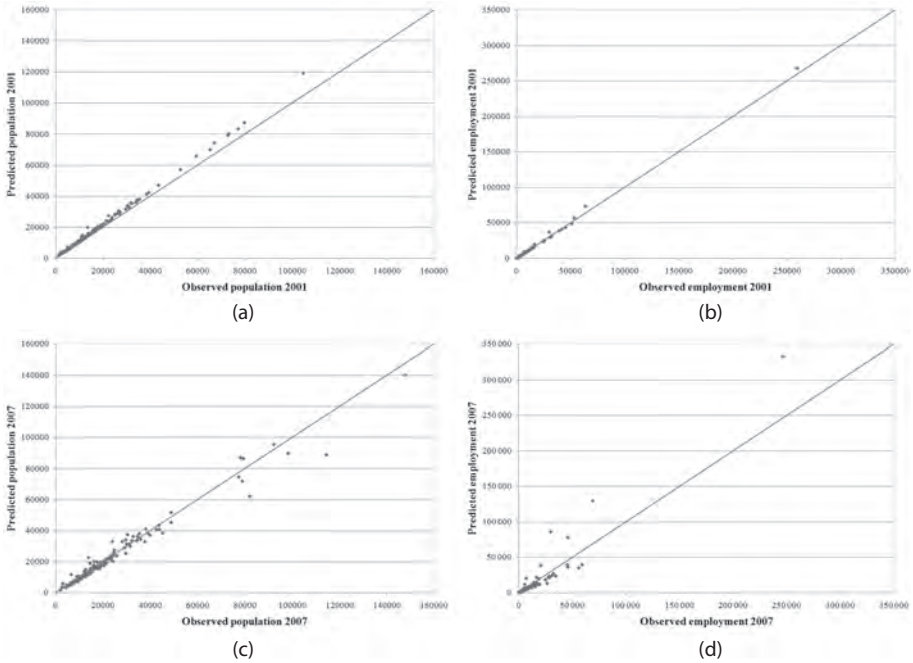


Figure 19.3 Comparison between the predicted and observed values: population 2001 (a), population 2007 (b), employment 2001 (c), and employment 2007 (d). The grey line is the 1:1 line.

The ability of a model to reproduce the observed situation can be assessed on the basis of a comparison between the predicted and observed values. Here the comparison was done at the municipal level.

Figure 19.3a and Figure 19.3b show the efficiency of the model in reconstituting with considerable accuracy the observed population ($R^2 = 0.996$) and employment ($R^2 = 0.997$) for the year 2001.

Figure 19.3c and Figure 19.3d show that the model also predicts with great accuracy the observed population ($R^2 = 0.972$) and employment ($R^2 = 0.921$) for the year 2007. As expected, the model predicts population and employment with less accuracy in 2007 than in 2001. The reconstitution of 2007 is interesting as a check but we should not expect the model to reconstitute perfectly the observed evolution 2001–2007, because the model considers only some variables of a "simplified" reality. In other words, the purpose of these models is to simulate and evaluate policies in a given reference framework, not to forecast the future.

The figures of the relative differences between observed and modeled values per municipality (Fig. 19.4) give a more nuanced picture. We conclude that the year 2001 is better predicted than the year 2007 and that population is better predicted than employment.

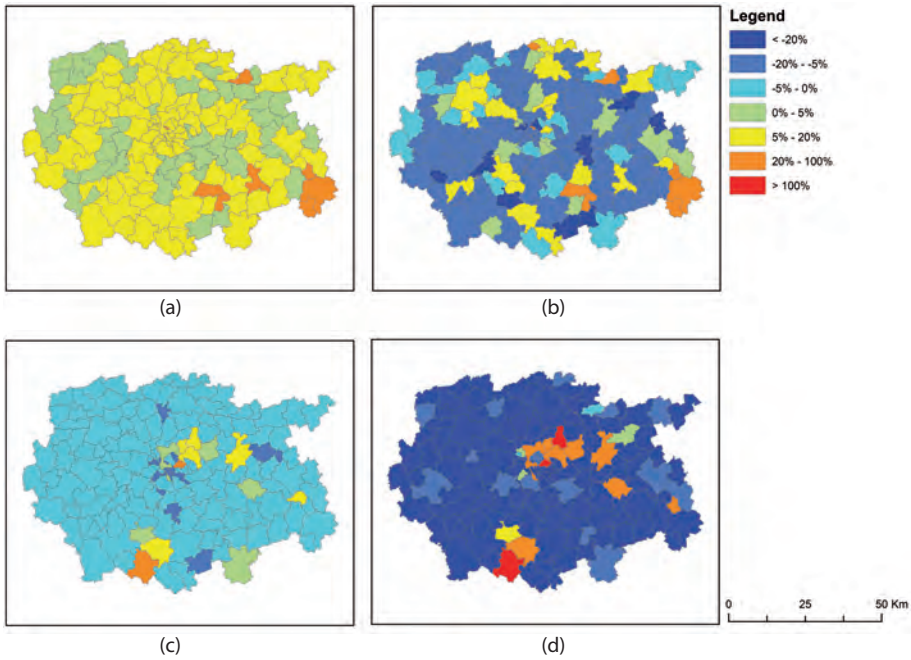


Figure 19.4 Relative differences between the predicted and observed values: population 2001 (a), population 2007 (b), employment 2001 (c), and employment 2007 (d).

As for the reconstitution of the transport situation, the 2001 car travel times and car shares predicted by MATSim per macrozone for the morning peak hours (6am–10am) were compared with the car travel times and car shares from the IRIS 2 model (Table 19.10).

The modeled average morning travel times by car were found to be higher compared to the values of the IRIS 2 model, which is due to the fact that the model was only partially calibrated. However tendencies are correct: trips heading towards the city center are significantly longer (in time) than trips heading towards the outside zones.

As shown in Table 19.10, the car shares predicted by MATSim are close to the IRIS 2 reference data. A clear tendency is noticeable: trips leaving the city center are mainly performed by car and trips heading towards the center are more often performed by another mode. Within a zone, the MATSim model clearly underestimates the car share, this results from the opportunity for the agents to walk (short trips are more often performed by foot) (Röder et al. 2013).

Note that a meaningful analysis for trips from and to macrozone 7 is not possible as the number of trips predicted by the model is very low. Indeed, this zone extends further than the study area of the Brussels case study. Other

Car shares from the IRIS 2 model							
	1	2	3	4	5	6	7
1	44%	69%	54%	75%	84%	48%	67%
2	43%	60%	66%	76%	73%	69%	82%
3	28%	67%	67%	78%	73%	72%	87%
4	24%	65%	78%	77%	82%	86%	95%
5	26%	67%	79%	89%	83%	83%	95%
6	22%	59%	78%	94%	91%	84%	85%
7	29%	59%	75%	96%	96%	86%	85%

Car shares from the MATSim model							
	1	2	3	4	5	6	7
1		64%	75%				
2	36%	44%	67%	85%	78%	81%	
3	34%	43%	61%	82%	86%	82%	
4	28%	47%	69%	75%	87%	86%	
5	20%	34%	62%	83%	81%	85%	
6	14%	23%	43%	67%	80%	85%	
7							

Differences							
	1	2	3	4	5	6	7
1		−5%	21%				
2	−7%	−16%	1%	9%	5%	12%	
3	6%	−24%	−6%	4%	13%	10%	
4	4%	−18%	−9%	−2%	5%	0%	
5	-6%	−33%	−17%	−6%	−2%	2%	
6	−8%	−36%	−35%	−27%	−11%	1%	
7							

Table 19.10 Comparison between the car shares from the IRIS 2 model and from the MATSim model per origin-destination relation at the macrozone level between 6am and 10am.

entries with small number of trips were also removed from the comparison. The reason why the model gives such low numbers is that MATSim simulations were run with a 5% sample of the population. Finally, it must also be reminded that the two models here compared are very different in their specifications and in their achieved level of calibration.

19.8 Definition of the scenarios

The following policies were simulated: the reference scenario assuming business-as-usual (BAU), the cordon pricing scenario and the densification scenario. The scenarios, which are further detailed below, were defined with a concern for realism, i.e. by taking into account policies that are known to be considered for

Brussels. Firstly, the BCR Administration is currently studying the possibility of introducing an urban road pricing in order to handle the congestion problem in the capital. Different scenarios have been explored, among which an area toll applied to the BCR (3€/day) and a kilometer pricing applied to the RER zone (0.07€/km). Secondly, demographic projections have shown that the population of Brussels is likely to increase. For over 10 years, the BCR Administration have been trying to fight against the urban sprawl because of its adverse effects on the environment (e.g. longer trips, increased dependence on car, increased congestion on the road network, increased fuel consumption, increased air pollution and noise pollution, consumption of agricultural and natural areas and pressure on ecosystems, etc.) but also for financial reasons. As a significant portion of the Region's revenues comes from the income tax (physical person), when the Region loses population, it also loses financial resources.

The scenarios were simulated from 2001 to 2020. The time interval for each simulation was set to 1 year for UrbanSim and 5 years for MATSim in order to update the accessibility values of zones and the travel related indicators of persons (i.e. chosen mode and travel time). Generally, land use changes occur at a slower rate than changes in choices related to commuting (i.e. mode choice and route choice) (Wegener 2007). However, the time intervals were imposed by technical constraints: the time interval of UrbanSim is set to 1 year by default; the time interval of MATSim could have been set to 1 year but the computations times would have been much higher. The results of each scenario are based on the average values from 4 simulation runs. The comparison of the runs revealed that the inter-run variation can be up to 1%.

19.8.1 Cordon pricing scenario

The cordon is located just outside the Ring road (orbital highway) which surrounds the BCR and some adjacent municipalities (Fig. 19.5). It was decided to include the Ring road inside the cordon as it is also affected by congestion. Also, on a more practical side, setting the limits of the cordon to the Ring road reduces the number of entries/exits and allows for easier controls. Every car entering the area inside the cordon during the morning peak period has to pay an additional cost of 5€. The cordon was here implemented at the end of the year 2001 in order to see the long-term effects on households and jobs relocation. Preliminary tests with the cordon implemented in 2015 had shown small effects on agent relocation.

19.8.2 Densification scenario

The objective of this scenario was to test the effects of household and job densification in zones defined as having a high accessibility by public transport. The definition of the highly accessible zones is different for the population densification and for the job densification.



Figure 19.5 Cordon pricing scenario zone located in the Brussels study area.

Population densification

The population densification scenario was implemented by increasing the housing supply in zones having a high accessibility by public transport. These zones highly accessible were located in the municipalities belonging to the "operational agglomeration" as defined by Van Hecke et al. (2009).[45] The "operational agglomeration" is formed by the 36 municipalities classified as the "center" and the "agglomeration" (Fig. 19.6). Note that this area is actually composed of 41 communes but 5 of them were excluded as they belong to the urban area of Leuven, Mechelen, and Antwerp.

The increase in housing supply was spread over time from 2013 to 2017 (i.e. 5 years). During this period, the number of residential units was increased each year by 2% in the BCR and by 4% in the other municipalities of the operational agglomeration compared to the existing residential units at the end of the previous year. [46]

[45] A municipality belongs to the operational agglomeration if more than 50% of its population lives in the morphological agglomeration. The definition of the morphological agglomeration is based on the inter-building distances (minimum 200 meters).

[46] Note that the 2% increase per year on the period 2013-2017 for the BCR (i.e. a total increase by about 50 000 units) is not far from what is currently discussed by politicians: indeed the Region will have to face an important demographic growth in the coming decade, which would require an increase of the housing supply by about 100 000 units by 2020.

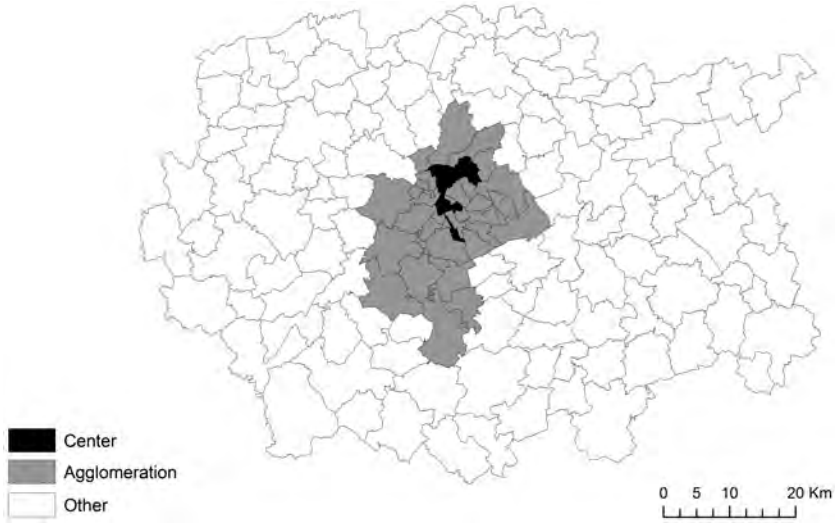


Figure 19.6 Target areas for population densification (center + agglomeration) in the densification scenario.

Job densification

It was decided to focus the job densification on the tertiary sector and to implement the scenario by increasing the available office floor space located in zones having a high accessibility by public transport. This policy was inspired from the "ABC policy", which was created in The Netherlands, about 25 years ago. The ABC policy aims to the best possible adequacy between *the mobility needs of enterprises and the accessibility profiles of locations*: it could be summarized as "The right business at the right place". In the general definition of this policy, the A locations are highly accessible by public transport (e.g. served by an inter-regional/inter-city railway station), the C locations have a car-oriented accessibility (e.g. near to a highway exit/entrance), the B locations lie somewhere in-between the A and C situations (medium-level accessibility both by public transport and by car). Eventually, the remaining locations classified as "R locations" offer a poor accessibility both by public transport and by car. As far as the enterprises are concerned, the A profile corresponds to a high concentration of employment (i.e. high number of jobs per floor space unit), a large number of visitors, and activity sectors not highly depending from the road (e.g. no road freight). The C profile, on the opposite side, is characterized by a low concentration of employment, a small number of visitors, and a high dependence on road accessibility. The B profile is characterized by a medium-level concentration of employment, a moderate number of visitors and a moderate dependency on road. As examples, the A profile includes the large private sector offices and public administrations, the C profile includes

the wholesale distribution centres, the B profile includes department stores and retail shops (Martens & Griethuysen 1999, Fontaine 2010).

In the Brussels case study, the A, B, C, R accessibility profiles of the statistical sectors were defined are listed below. Figure 19.7 shows the distribution of the zones in the study area.

- profile A: statistical sectors whose centroid is less than 1200 meters from an inter-regional/inter-city railway station;
- profile B: statistical sectors whose centroid is less than 800 meters from a RER station and within 2000 meters of a highway entrance/exit (excluding profile A);
- profile C: statistical sectors whose centroid is within 2000 meters of a highway entrance/exit (excluding profile A or B);
- profile R: all other statistical sectors.

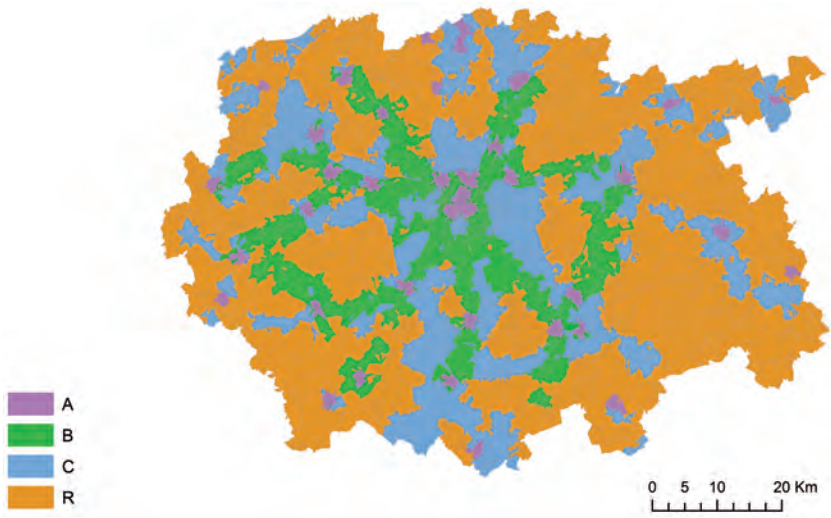


Figure 19.7 "ABC" typology of the statistical sectors of the study area.

The increase of office floor space capacity implemented by the densification scenario was set to 20% of the observed increase of occupied office floor space between 2001 and 2020 for the BAU scenario. The densification was split between statistical sectors classified as A and B, proportionally to the available floor space capacity of office buildings (i.e. developable office floor space). Densification was spread over time from 2013 to 2017, which means that 20% of the total increase in floor space of a particular building is built each year.

As mentioned above, scenarios were defined with a concern for realism. For this reason, densification was spread over time and the assumption for increasing housing supplies was stronger than the one for increasing office floor

spaces. Indeed, in the Brussels context, it is easier to increase housing supplies (e.g. by the conversion of single-family buildings into multiple units).

19.9 Simulation results

Each scenario was simulated from 2001 to 2020 and compared to the business-as-usual (BAU) scenario through different indicators (see Table 19.11). To assess the effect on land use, the number of jobs and households within the BCR (both scenarios), the area inside the cordon (cordon scenario only) and the target area of densification (densification scenario only) were computed. Note that this target area is the operational agglomeration for households and the zones belonging to profiles A and B for jobs. All these indicators measure the attractiveness of central places. An increase in their values is thus seen as desirable from a sustainability and/or social welfare point of view, since a concentration of jobs and households in these central places means a reduction of urban sprawl.

Transport indicators were also computed: (1) the average home-to-work travel times (car trips and all modes together), (2) the average home-to-work travel distance (car trips only), and (3) the average car share for home-to-work trips (all day or morning peak hours, in the BCR and the whole study area). Car share is a general measure of sustainability while travel times and distances are related to social welfare of the individuals. Hence, the desired effect of scenarios is here a decrease of the value of the indicators.

19.9.1 Cordon scenario

The main effect of the cordon scenario is noticeable on transport related indicators, i.e. on car shares and car travel times. Table 19.11 shows that the cordon pricing scenario leads to a decrease of the car share in the study area, compared to the BAU scenario. The decrease in the whole study area is modest (-1%) which can be explained by the fact that the cordon pricing only applies to cars entering the city center. This means that the room left on the network by the incoming commuters may be occupied by persons travelling inside or outside the cordon. In Table 19.11, we see that the car share increases slightly in the BCR (+0.4%). The analysis by macrozone confirms that the decrease in car share is more important for the incoming commuters: during the morning peak hours, the car share goes from 39.4% in the BAU scenario to 36.3% in the cordon scenario.

There is also a reduction in car travel time and car travel distance. The effects as regards to travel times (by all modes) result mainly from the combination of two effects going in opposite directions: the changes in the average trip distances and the modal shifts between on average faster modes (car) and on average slower modes (public transport and walk). For this reason, the

		Base year *in comparison*	BAU *Base year*	Cordon *BAU*	Densification *BAU*
Land Use Indicators					
Number of households in Brussels-Capital Region	number of households	411 047	459 214	458 718	498 123
Absolute variation			*48 167*	*-496*	*38 909*
Relative variation	*%*		*11.7%*	*-0.1%*	*8.5%*
Number of households inside the cordon	number of households	467 794	520 485	520 405	/
Absolute variation			*52 691*	*-80*	/
Relative variation	*%*		*11.3%*	*0.0%*	/
Number of households in the target area of densification	number of households	563 411	644 214	/	707 180
Absolute variation			*80 803*	/	*62 966*
Relative variation	*%*		*14.3%*	/	*9.8%*
Number of jobs in Brussels-Capital Region	number of jobs	678 251	871 142	873 919	865 326
Absolute variation			*192 891*	*2 778*	*-5 816*
Relative variation	*%*		*28.4%*	*0.3%*	*-0.7%*
Number of jobs inside the cordon	number of jobs	732 560	944 533	947 247	/
Absolute variation			*211 974*	*2 714*	/
Relative variation	*%*		*28.9%*	*0.3%*	/
Number of jobs in the target area of densification	number of jobs	905 542	1 138 070	/	1 133 418
Absolute variation			*232 528*	/	*-4 652*
Relative variation	*%*		*25.7%*	/	*-0.4%*

Table 19.11 Land use and transport indicators for the Brussels case study.

		Base year *in comparison*	BAU *Base year*	Cordon *BAU*	Densification *BAU*
Transport Indicators					
Average mode share of car, in the study area	%, all day	73.6%	67.5%	66.4%	67.6%
Absolute variation	*points*		*-0.1*	*-1.0*	*0.1*
Relative variation	*%*		*-8%*	*-1.5%*	*0.2%*
Average mode share of car, in Brussels-Capital Region	%, all day	60.2%	53.7%	54.0%	54.0%
Absolute variation	*points*		*-0.1*	*0.4*	*0.4*
Relative variation	*%*		*-11%*	*0.7%*	*0.7%*
Average mode share of car, in the study area	%, on the morning peak hours (6 am-10 am)	72.0%	65.2%	64.2%	65.4%
Absolute variation	*points*		*-0.1*	*-1.0*	*0.2*
Relative variation	*%*		*-9.5%*	*-1.5%*	*0.3%*
Average mode share of car, in Brussels-Capital Region	%, on the morning peak hours (6 am-10 am)	57.6%	50.9%	51.3%	51.1%
Absolute variation	*points*		*-0.1*	*0.4*	*0.2*
Relative variation	*%*		*-12%*	*0.8%*	*0.4%*
Average travel time (all modes), in the study area	minutes, all day	58.1	62.2	60.3	62.1
Absolute variation			*4.1*	*-1.9*	*-0.1*
Relative variation	*%*		*7.1%*	*-3.1%*	*-0.2%*
Average travel time for car trips, in the study area	minutes, all day	56.8	60.7	57.8	60.6
Absolute variation			*3.8*	*-2.8*	*0.0*
Relative variation	*%*		*7%*	*-4.7%*	*-0.1%*
Average home-work travel distance for car trips	kilometers, all day	14.2	13.4	13.2	13.4
Absolute variation			*-0.7*	*-0.2*	*0.0*
Relative variation	*%*		*-5%*	*-1.6%*	*-0.1%*

Table 19.11 Land use and transport indicators for the Brussels case study.

average travel time for car trips decreases more (-3 minutes) than the average travel time for all modes together (-2 minutes).

Finally, Figure 19.8 shows the changes in car accessibility between the BAU scenario and the cordon pricing scenario. Car accessibility is computed for each zone by the MATSim model. The calculation is based on the LogSum, which is a utility-based measure of accessibility reflecting the economic benefits as the maximum expected utility that someone gains from access to spatially distributed opportunities like work (Geurs & Ritsema van Eck 2001, de Jong et al. 2007) (for more details, see Chapter 17). The red and yellow zones experience a disadvantage due to roadpricing. The greenest areas experience an improvement.

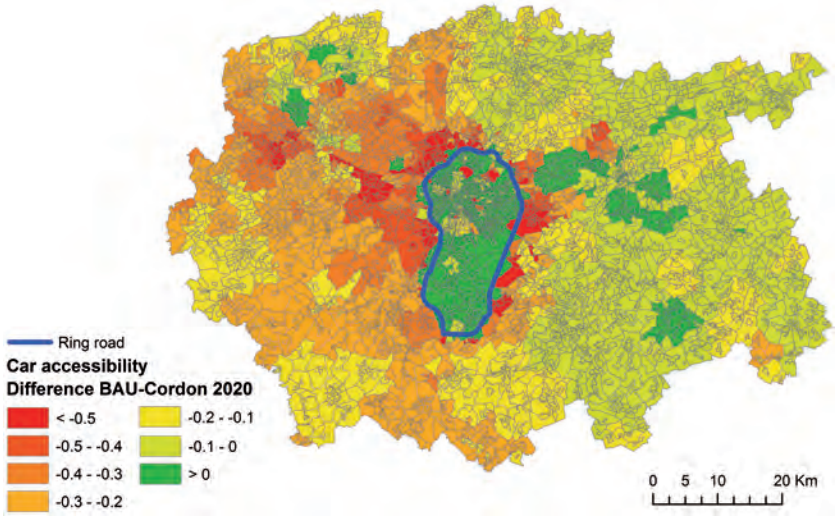


Figure 19.8 Absolute difference of car accessibility between BAU and cordon pricing scenarios.

The car accessibility may improve or deteriorate in the cordon scenario depending on the value of the savings of car travel time to the workplaces, compared to the toll. The savings in car travel time in turn depend on the initial level of congestion and the spatial distribution of the workplaces. This may lead to some local effects. [47]

Figure 19.8 shows a clear tendency: almost the whole area within the Ring benefits from the cordon. For areas outside the cordon, accessibility decreases significantly. The zones just outside the cordon experience the most important loss in car accessibility. This is due to the fact that the toll is included in the generalized cost of travel and that even when congestion is reduced, access

[47] Theoretically it also depends on the distribution of the values of time, but in the Brussels model, the value of time is assumed to be 9€2001/hour for all the travelling individuals.

to facilities behind the cordon now comes with a higher generalized cost than before. In other words, for these zones, the benefits from the reduction in car travel times and distances do not compensate the increase in monetary costs. Figure 19.8 also shows that a few zones outside the Ring benefit of an improvement in car accessibility. This may be due to local effects (e.g. local reduction of road congestion and therefore better accessibility to the workplaces located outside BCR). Similarly, a few zones inside the Ring undergo a decrease in car accessibility, which is due to the fact that locally the time savings do not compensate the cost due to the toll. Alternatively, the results may also be due to remaining weaknesses in the transport model.

Note that public transport travel times do not vary between the BAU and the cordon scenario. As we saw, the public transport travel times are defined beforehand and are not modeled by MATSim. However, public transport accessibility should change if the spatial distribution of households and jobs change.

Figure 19.9 and Figure 19.10 show that the cordon scenario has no significant effect on the location choice of households and jobs.

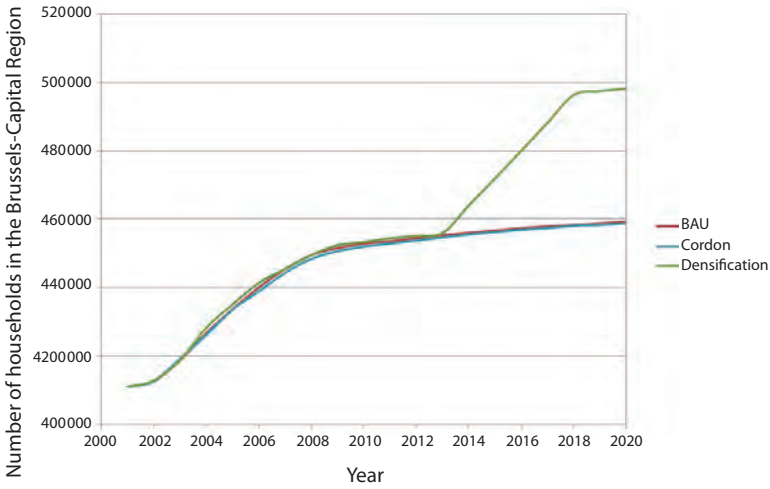


Figure 19.9 Evolution of the number of households in the Brussels-Capital Region for the three scenarios.

The land use indicators in Table 19.11 show that the relative variation of the number of households or jobs inside the BCR or inside the cordon between the cordon scenario and the BAU scenario, is less or equal to 0.3%, which is actually less than the inter-run variations.

An analysis of the estimation results of the household location choice model (Table 19.6) reveals that the car accessibility only influences the location choice of car-owning households. On the other hand, the transaction price seems to

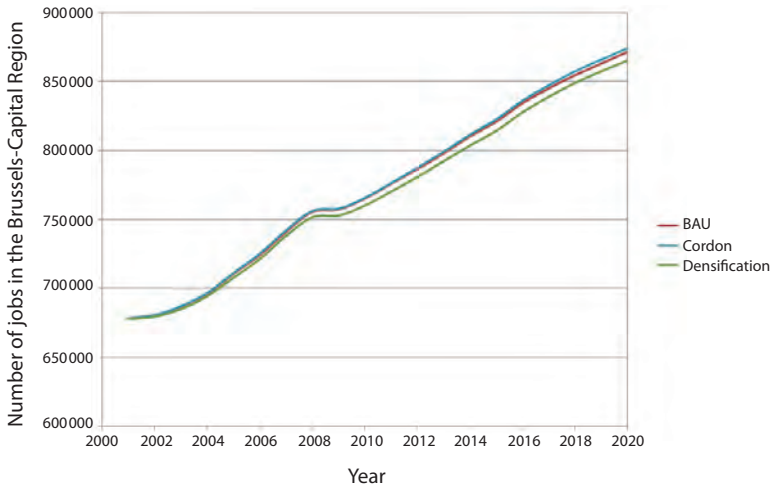


Figure 19.10 Evolution of the number of jobs in the Brussels-Capital Region for the three scenarios.

be a determining factor: the more expensive the real estate good is, the less attractive the location will be. Now looking at the specifications of the real estate price model (Table 19.5), we see that the car accessibility influences positively the real estate price: the more accessible the location is, the more expensive the real estate good will be. This means there is a kind of "feedback effect": households want to locate in accessible locations (and still prefer less expensive buildings) but at the same time the price of the buildings increases with the car accessibility.

As for the employment location choice model (Table 19.7), car accessibility only influences the location of the activity sectors "retail" and "hotels/bars/restaurants".

These results are counter-intuitive as a cordon pricing is generally expected to have a "centralizing effect". However, ex-post studies carried out after the implementation of a road pricing (Amstrong 1986, Larsen 1995) have shown that the effects on relocation are generally modest, especially when compared with the effects on traffic and on global land use changes (Eliasson & Lundberg 2003).

19.9.2 Densification

Table 19.11 shows that increasing housing supply in "highly accessible" locations leads to a relocation of households in BCR (+ 8.5%) and in the target zones of densification (+9.8%). In the BAU scenario, we see that there is already a trend of an increasing number of households in the BCR from 2001 to 2020. The densification scenario reinforces this trend. Figure 19.9 shows

that the number of households increases regularly between 2013 and 2017 and continues to increase after, but far more slowly. It suggests that all additional housing capacity implemented by the densification scenario are occupied, and that the increase observed in the number of households is then limited by the intensity of the densification process, rather than by any other factor.

Table 19.11 shows that increasing office floor space in "highly accessible" locations does not have a significant effect on the relocation of jobs in BCR and in the target zones of densification. Indeed, the difference between the BAU scenario and the densification scenario is of the same order of magnitude as the inter-run variation. A possible explanation is that the distance to rail stations and to highways, which defines the accessibility profiles, are not included in the employment location choice models (see Table 19.7). Even if these zones are seen as desirable locations for firms for sustainability, we do not know if they correspond to desirable locations for the firms themselves. Alternatively, the implemented policy is not enough incentive to induce a change in the choice location of jobs.

Table 19.11 shows that the car share tends to increase slightly in the study area between the BAU and the densification scenario. Generally, we would expect the densification to induce a modal shift from car to public transport as households relocate in denser areas better served by public transport. However, in this scenario, jobs do not relocate in the BCR. It is then possible that households relocating in the densified area still commute to peripheral areas, even if the average trip length (in time and distance) does not change compared to the BAU scenario.

19.10 Conclusions

This case study shows that, as long as data are available, an UrbanSim-MATSim land use/transport model can be calibrated for a large metropolitan area (3 millions inhabitants in the case of Brussels), within a reasonable time-schedule (12-18 months depending on the availability of the data). It also illustrates how land use/transport models can help in policy planning, by providing an estimation of the relocation effects, which in turn influence the transport indicators (i.e. mode shares, travel times, and accessibilities). Not only the final result is of interest to planners, but also the underlying mechanisms (changes in accessibilities, changes in the housing prices, reactions of the real estate developers, all elements that are represented in UrbanSim).

However, despite the significant achievements, the integrated UrbanSim-MATSim model for Brussels still has some weaknesses. Among others, elasticities (magnitude of the policy effects) have not been fully validated and in some cases are difficult to be interpreted.

The areas requiring improvements and further research are pointed hereafter. First, all the models (household location choice, job location choice, hedonic price model, etc.) have been estimated with observed data, in which correlations may exist (and probably exist). It is therefore difficult to estimate with accuracy the effect of the given explanatory variable x on the independent variable y. One solution would be to estimate the parameter values using "stated preference surveys", i.e. surveys where scenarios are presented to the respondents, with uncorrelated variables, so that the effect of each variable on their choices may be derived with much higher accuracy (Louviere et al. 2000). The collection of such data could provide an important improvement to the calibration of the UrbanSim sub-models, particularly the location choice models, where variables such as price, surface, number of rooms, etc., may be correlated in the reality.[48] More generally further research should be dedicated to correlation, autocorrelation and endogeneity issues in the calibration of these models (see Chap. 4).

Another field requiring further research is the validation of the models, e.g. by plausibility analysis, inter-city comparisons, inter-models comparisons, etc.

The calibration of the MATSim transport model for Brussels also requires improvement. Currently, the MATSim model is at a prototype state and was done with a small amount of time and budget. First, the model would benefit from a higher resolution network. More accurate public transport stops and travel times could be a possible improvement for the the mode choice model. A further problem seems to be the population-sampling. Experiments show a difference in spatial distribution of the resulting population. Probably, this may result from the drawing-procedure which is not stratified (e.g. by zone). As another point, the MATSim model only takes into account the home-to-work trips. The inclusion of other trip purposes (home-to-school, shopping, leisure) would be an important improvement. MATSim is, in general, designed to run with full daily activity chains. However, one would need to design and test a model that combines the home and work location information from UrbanSim with information about other activities from other sources. Finally, the current model system does not include transport demand elasticity to the generalized transport cost: whatever the travel time or travel cost, the demand does not change. Including this aspect in the model would make it possible to simulate demand induction/cancellation effects and, again, would make it more realistic.

Other more technical improvements also should be required, for UrbanSim and MATSim, such as a reduction of the computing time, to allow a better exploration of the stochastic variations (in microsimulation, each run provides

[48] To be complete, before being used in simulations, the parameters estimated by means of stated preference surveys have to be "scaled" to the actual behaviors (the scale factor being in fact a dispersion factor). Observed data are thus also necessary, together with the stated preference surveys.

a different result and a whole set of runs is necessary to provide an average result).

Note that another integrated land use/transport model has formerly been developed for Brussels but with a modeling framework based on the equilibrium theory (Stratec 2005). It has to be pointed out that the simulation results of the Brussels UrbanSim model cannot be easily compared with these previous results: because of the two different theoretical backgrounds (microsimulation versus equilibrium), the different underlying hypotheses, and because the UrbanSim model was calibrated in less detail, in a more limited time period.

As a general conclusion, we can say that for the first time an integrated land use/transportation microsimulation model was developed for Brussels. The Brussels case study tried to go beyond a theoretical simulation exercise in order to simulate actual policy simulations. In spite of its limitations, it provided policy simulation results which were analyzed and discussed. Most of them would nevertheless require further validation. This raises once again the crucial question of the need of validation data and validation processes.

Chapter 20

Application of UrbanSim in Paris (Ile-de-France)

André de Palma, Nathalie Picard, and Kiarash Motamedi

20.1 Scope and challenges

The integrated land use and transportation models have received increased attention in research and practice over the last decades. In particular, the researchers and the decision makers highlight the ability of these integrated models to examine the combined effects of land use and transport policies, to study the endogeneity of urban development and of travel patterns, as well as to analyze and quantify the effects of transport capacity expansion. The usefulness of integrated models goes beyond the examination of induced effects of transport demand. These integrated tools allow examining issues such as efficiency and fairness in markets for local public goods, as well as for local public bads. In addition, integrated models are ideal tools to study spatial patterns and the distributional effects of environmental externalities related to congestion, such as emissions from vehicles, local pollution, noise and accidents. In this chapter, we use the integrated land use and transport model system UrbanSim/METROPOLIS[49] to study the impact of a large transportation infrastructure project including the resulting inequalities in the spatial distribution of accessibility and environmental quality.

The main nine steps of the case study are described below. Each step has an interest in itself. However, here we will mainly focus on the results not of each module, but of the overall integrated system.

[49] The integration of two models is realized by an interface that manages the interaction between them and carries out the necessary data transformations. This interface is explained in chapter 18 of this handbook.

1. Data collection and analysis of data.
2. definition of the appropriate geographical scales (*Commune, Département, IRIS, Ilots, Ilots MOS*[50] are different geographical units used in France). Such spatial issues were also studied from a theoretical point of view in Chapter 4 - see also Chapter 12, which deals with econometric issues. The terminology will be explained in detail below in this chapter.
3. Integration of a demographic model.
4. Improvement and calibration of the real estate price model.
5. Implementation of the firmography model.
6. Improvement of the residential location choice model.
7. Calibration of the transportation model (METROPOLIS).
8. Definition of scenarios and evaluation of such scenarios.
9. Environmental and socio-economic policy analysis.

20.2 Introduction

Ile-de-France is the largest European application of UrbanSim with a total area of over 12 000 km^2 and more than 5 million households and jobs. It has also benefited from the richest database, including exhaustive census data. This case study has also benefited from the results of several theoretical developments (see Chapters 7, 8, 9, and Maurin et al. 2013). The results of this case study have been disseminated and used successfully in applied studies and socio-economic analyses, such as the Grand Paris project. We benefited from collaboration with the *Société du Grand Paris*, which helped us to define the scenarios to be considered in the context of the establishment of a network of rapid transit (Fig. 20.1), that will be created in the coming decades. The French Prime Minister has presented in 2013 the schedule for putting the different segments of the Grand Paris Express metro lines in operation. The first part of the project will be the southern part of the ring around Paris. Other segments will be completed consecutively to each other.

Several important questions should be answered. First, what should be the timing of the construction of different segments over the years to come? Second, what are the wide economic benefits? Third, how many jobs will be created (attracted) in the region in the coming years? Fourth, what will be the relocation decisions of different economic agents (residents and businesses)? This question is important because it will help the local authorities to plan to cope with the demand for housing (especially regarding collective dwellings) and for offices. And the fifth question: What will be the traffic demand, both in the existing and new public transport network and on roads?

[50] *Ilot* refers to the basic (homogeneous) census geographical unit which represents a block of buildings or land delimited by roads (streets). *IlotMOS* refers to the polygons with homogeneous land use cover represented in land use registry (*MOS*). *IRIS* refers to an aggregation

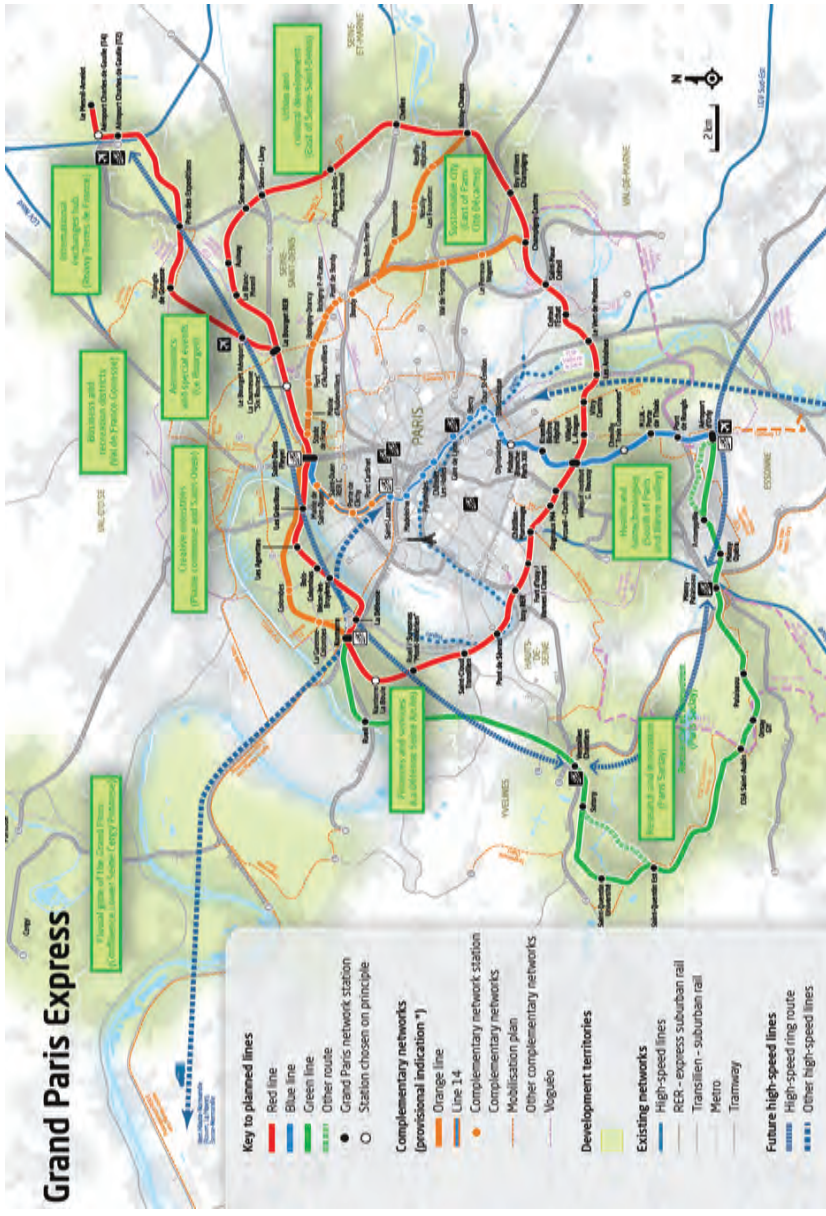


Figure 20.1 The Grand Paris Railway Network (Source: SGP 2012).

This chapter has 10 sections. The 8 remaining sections are presented as follows.

Section 20.3 presents the study area. Section 20.4 presents data collection. This is the first key task, since the quality of calibration and the quality of the prediction mainly depend on the quality of the data. Data were collected to estimate the four modules (residential location, firmography, real estate, and land use) and to calibrate the results produced by the integrated transport-land use model. It covers demography, population (exhaustive census data), firms and jobs (exhaustive Regional Employment Surveys), yearly real-estate prices in the largest 300 *communes*, land use, and many other marginal data sources. Moreover, transportation data (supply and demand) were collected. All the data needed to run UrbanSim were collected and carefully checked. The analysis of data is discussed in Section 20.5. The structure of the model and the description of the different sub-models are discussed in Section 20.6. The transportation model, METROPOLIS is presented in Section 20.7. The integrated land use and transportation model (detailed in Chapter 18) is summarized in Section 20.8. The major part of this chapter focuses on the results (Sect. 20.9). To the best of our knowledge, this is the first wide application of UrbanSim used to address major issues faced in a large metropolitan area. The description of policy scenario of the Grand Paris project as well as the results are presented in Section 20.9. Section 20.10 concludes.

20.3 Study area

The Paris area, namely *Ile-de-France*, embraces Paris and its suburbs. It is Europe's most populated region. The city of Paris has about 2 million inhabitants, on a regional total of 11 million. The total number of jobs is 5.1 million. With a surface of 12 000 km^2, *Ile-de-France* occupies 2.2% of the surface of France. It represents 19% of the population, 22% of the jobs and 29% of the GDP of the country. The GDP is at the level of EUR 552 billion (2007), which is the fifth-largest in the world, just before London, and after the urban areas of Tokyo, New-York, Los Angeles, and Chicago. *Ile-de-France* is Europe's biggest employment base with about 50% of them at managerial positions. The large fraction of senior professionals is due to the high density of company headquarters located in the Paris region. However, the Paris region economy is also diversified.

There are 3 administrative institutions in *Ile-de-France*: 1 '*Région*', 8 '*Départements*' (Counties) and 1300 '*Communes*' (Municipalities). In addition, we consider the 3 *départements* around Paris (*Hauts-de-Seine, Val-de-Marne*, and *Seine-Saint-Denis*) as close suburbs or the 'inner ring' and the 5 *départements* far away from Paris (*Yvelines, Essonne, Val-d'Oise*, and *Seine-et-Marne*)

of neighboring *Ilots* with about 2000 inhabitants or 1000 jobs.

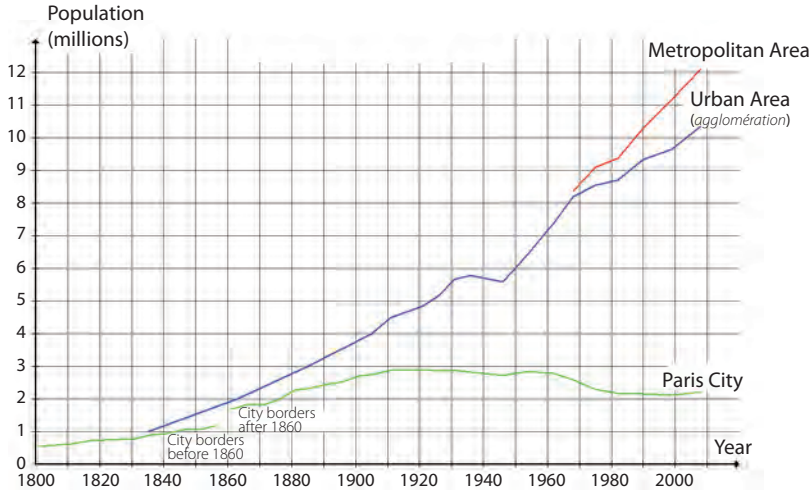


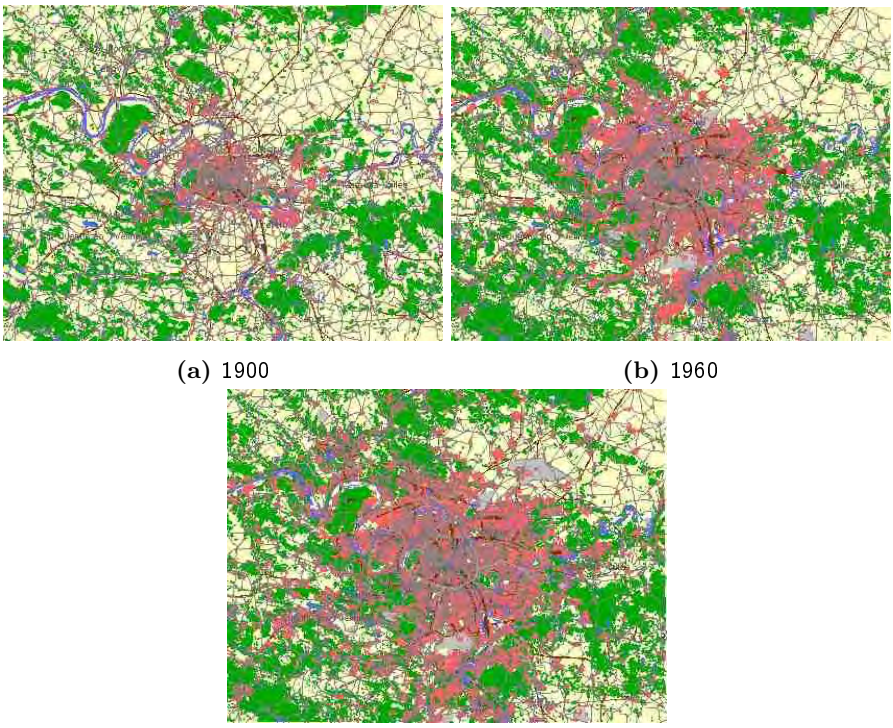
Figure 20.2 Population of Paris and its suburbs from 1801 to 2008 (Source: Wikipedia 2011).

as far suburbs or the 'outer ring'. Very large differences in population density can be observed between Paris and the outer periphery.

The genesis of Paris begins with *Ile de la Cité* a small island at the center of Paris city built in 56 BC with surface of 23ha. The population of Paris city reached its maximum at 2.9 million habitants in 1920. The borders of the current Paris city have been fixed in 1930s with a surface of 10.54 kha. Its population is about 2.26 million inhabitants in 2012. Figure 20.2 represents the evolution of the population of Paris city and its area. The population of Paris decreased during the 1960s and 1970s then remained steady up to 2000. Meanwhile, the population of the Paris region increased rapidly when it began to locate around the secondary poles and not directly in the influence zone of Paris. Figure 20.3 represents the evolution of the urbanization in Paris area. The urbanized area was concentrated in Paris and some nearby poles in 1900, occupied its main space in 1960 and then expanded its boundaries up to 2000.

Based on the English experience of new towns or *Villes nouvelles*[51] during the 1960s and 1980s, the French government created *Villes nouvelles* in the outer ring of Paris suburbs in order to multi-polarize the economy of the city (Combes et al. 2011). However, *Villes nouvelles* experiment showed the limits of the policy of developing population poles, since economic activities still

[51] *Ville Nouvelle* is a French administrative terminology corresponding to 4 cities that include Cergy-Pontoise, Marne-la-vallée, Sénart, Saint-Quentin-en-Yvelines, and Evry. They are connected to CBD by fast public transportation (RER) lines. They are aimed to be secondary centers in order to decentralize population and jobs. As a consequence, the Paris area can be seen as a polycentric city with one major center (historical Paris) and several sub-centers.



(a) 1900 **(b)** 1960

(c) 2000

Figure 20.3 Evolution of urbanization in Paris area (Source: *IAU-IdF*).

remain in a large measure concentrated in the center of the urban area of Paris Region, namely, Paris City and *Hauts-de-Seine*. As illustrated in Figure 20.4, the new towns have well absorbed the overflow of Paris population growth. Other parts of the region keep growing as before.

The land use is composed of built-up areas (30%), green areas (20%) and rural areas (50%). The public transportation network is made of 3 components:

- a main radial railway network, especially the RER lines (high speed train service between Paris and the suburbs);
- a subway network that provides comprehensive and punctual service in the city center;
- a bus network to complement the rail services.

The road network is organized into a hierarchy that is densely interconnected and often congested. The express network of the region is composed of 590 km of motorways and 250 km of expressways, with a total of 4 500 lane-km. Road traffic flows reach the highest levels known all over the country. But thanks to an efficient road network, and despite the traditional traffic jams (at rush-hour), traffic conditions are on the whole remarkably good for a

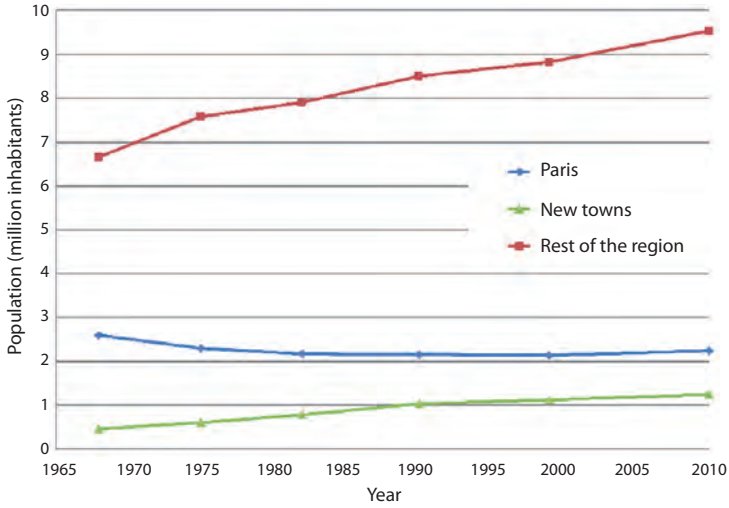


Figure 20.4 Population of Paris in comparison with new towns and rest of the region from 1968 to 2010 (Source: Census data, *INSEE*).

metropolis of this size, since the average duration of a car trip is 19 minutes, and no more than 25 minutes for commuting trips by car (DREIF 2005). The market shares for the home based work trips in 2001 were 50% private cars, 36% transit, and 14% bicycle/walk (and other modes). Over the last twenty years, the use of public transportation has decreased down to 6% in the region. This can be explained by structural effects, as increased mobility in areas with limited public transport tends to bring the average ridership lower. New investments are programmed to develop the transit network, which had not been significantly developed for more than 50 years.

20.4 Data

Access to good quality data is an important concern in every modeling work. This is particularly the case for microsimulation models that are avid in large volume of disaggregated data. Access to individual detailed data is costly in terms of their value and resources needed for their treatment. Moreover, it is also limited because of privacy protection legislation. In this case study, the partnership with public agencies allowed us to access rich data sources of particular interest to this study.

20.4.1 Data sources

Various data sources describing population, employment, land use, real-estate prices, and transportation system are used in this study. They are presented in the following sections.

Population data: Census

The French Census has been reformed since 2004. Before, a general census was carried out every 10 years. The new census becomes yearly, but only a sample of population is surveyed. We describe below the two Census types.

General Census of 1999 *(RGP99)*

The recent general census has been carried out in years 1982, 1990, and 1999. For year 1999, we have access to the exhaustive individuals database with many useful variables but not all of them. This database constitutes our main source of population data in this case study and in this case we did not need to create a synthetic population base. Based on this data, we estimate the residential location choice model and also the yearly residential relocation rates.

Census data are also used to provide the thematic data tables for 13 subjects:

- population (age, sex, matrimonial status, activity, etc.),
- activity of residents (activity rates by sex, unemployment, employment conditions, etc.),
- employment status (employee versus self-employed, part-time or full-time, etc.),
- commuting, education (students at their study location, diplomas, education level, etc.),
- migration (birth place, residence *commune* at the time of the previous census),
- citizenship and immigration (French or foreign nationals, immigrant or French at birth, etc.),
- households (sex, age, education level and activity of the household head, size, number of active members, etc.),
- dwellings (type, number of rooms, etc.),
- dwellings occupied as principal residence (date of move in, tenure, surface, etc.),
- buildings (date of construction, number of floors, number of housing units, etc.),
- information at the family level (couples, single parents, citizenship and socio-professional category of the family head, who may differ from household head in multi-family households, number and ages of children, etc.).

The probability to relocate is computed using the 5% sample of whole country base (which is the only one available for us with this variable) to be able to take into account households moving out of the region. The study of commuter behavior is also based on this 5% sample to observe simultaneously residence and workplace information. Other models are estimated using the

complete database. For the base year, the aggregate variables for different geographical units are computed using the census data table regarding either residents at residence place or jobs at workplace.

The yearly census from 2006

The results of yearly census are provided as processed tables (by the National Institute of Statistics and Economic Studies, *INSEE*) and also at an individual level (a sample including the variable "sampling rate" and a limited number of variables). We use that information for calibration and validation of models. As they are provided every year, they allow us to check the results against observed evolutions and trajectories.

Employment data

Regional Employment Survey *(Enquête Régionale sur l'Emploi, ERE)*

It constitutes our main data source about jobs and businesses (APUR 2001). It provides us with a set of fairly detailed indicators about activities and employees in both public and private establishments over the region. The survey has been carried out every 3–4 years up to 2001. Then it was replaced with another statistical database to which we could not have access. We had access to the *ERE* tables for 1997 and 2001. They have been matched based on the unique establishment ID number assigned by national statistical authority. The base provides the location of establishment at commune level and the number of employees as well as the activity sector in a 700 items list.

Land use and real estate data

MOS *(Mode d'Occupation du Sol)*

MOS provides at given points in time the surface of land covered by 83 land use categories (IAU-IdF 2012). Less detailed classifications are also used (based on 47, 21, or even 11 categories). **EvoluMOS** is a database which tracks the land use evolution in the *Ile-de-France* region. They are prepared and updated by the regional urbanism institute *IAU-Idf* in years 1982, 1987, 1990, 1999, 2003, and 2008 to monitor and analyze in great detail the evolution of land use on the whole area.

MOS and *EvoluMOS* together characterize the spatial and temporal evolution of land use in Ile-de-France. It is an essential source of data for the analysis of the choice made by urban developers concerning land use transformation. Currently, we use four generic types in order to keep the presentation manageable. These are single houses; collective housings; business; industry and service activities; and others. Our purpose is to capture the main trends characterizing land use changes.

Real estate prices: *Cote Callon*

Cote Callon provides the price and the rent values per unit floorspace for communes throughout France. These data span over a period of 10 to 12 years and distinguish old and new buildings and dwellings by number of rooms and standing category. It includes residential, industrial and commercial buildings, offices, and constructible and agricultural lands. Prices stored in the database are average ones per unit floor space at *commune* level. It should be noted that we don't take into account the diversity of prices within the same *commune*.

Information on local taxes

The information about the different taxes to be paid by firms and residents, residential owners and tenants are also available and used in the models.

Other data

Other data collected mainly on the *INSEE* website have been used to take account of inflation, vacancy rate by *département*, etc.

Transportation data

DRIEA (regional infrastructures authority) is the public authority that maintains a regional transportation model named *MODUS*. It is a standard 4-stage static model with public and private transportation modes and is considered as reference in different evaluation studies. The coded network and its planned evolution over coming years and simulation results were made available to us. We introduce directly the coded road network to METROPOLIS. That provides the travel time and accessibility measures by private car. The public transportation is not modeled in METROPOLIS and we use the MODUS results, which have a number of features:

- A user matrix for the morning peak.
- A loaded network for the morning peak (information on the number of passengers on each link of the public transportation network is provided).
- A travel time forthe morning peak for the different O-D pairs. Detailed information is also available such as in-vehicle time, access time to the public transportation network from home, walking time, waiting time, and walking time to destination.

To these data, we add the dynamic behavioral parameters needed for running the dynamic transportation model, METROPOLIS. These parameters describing the departure time choice of travelers are explained in Chapter 18. Their values are estimated based on a specialized survey named MADDIF (de Palma & Fontan 2001).

20.4.2 Data preparation

The raw data described in the previous section were usually not in a format which could be directly used in UrbanSim. The information contained in these data was also usually insufficiently detailed with respect to the need of the different modules of UrbanSim. Extrapolation was sometimes needed.

Typology of households (control totals)

UrbanSim needs the total number of households in the whole of region as input. It depends on the scenario and on the year. We have used 24 different types of households, which are described in Table 20.1. Households are distributed according to their size (1, 2, 3, 4, 5, and 6 or more persons), to their dwelling type (houses or flats) and to the ownership status (owner or tenant). Note that the number of households in these 24 types has changed considerably during the last decade (the typical household size is smaller, the fraction of owners is larger, etc.).

N. of persons	Dwelling Type	Tenure Status	1999	2006	2008
1	house	own	3.80%	3.89%	3.97%
2	house	own	7.84%	8.06%	8.09%
3	house	own	4.79%	4.54%	4.45%
4	house	own	5.49%	5.26%	5.22%
5	house	own	2.09%	2.08%	2.07%
6 and more	house	own	0.83%	0.77%	0.76%
1	flat	own	9.21%	9.43%	9.58%
2	flat	own	6.80%	7.00%	7.02%
3	flat	own	3.13%	2.97%	3.04%
4	flat	own	2.33%	2.24%	2.33%
5	flat	own	0.71%	0.70%	0.72%
6 and more	flat	own	0.27%	0.25%	0.25%
1	house	rent	0.71%	0.72%	0.72%
2	house	rent	0.80%	0.82%	0.81%
3	house	rent	0.67%	0.64%	0.63%
4	house	rent	0.65%	0.62%	0.60%
5	house	rent	0.29%	0.29%	0.29%
6 and more	house	rent	0.17%	0.16%	0.16%
1	flat	rent	20.90%	21.39%	21.35%
2	flat	rent	12.47%	12.83%	12.71%
3	flat	rent	7.11%	6.74%	6.73%
4	flat	rent	5.09%	4.88%	4.83%
5	flat	rent	2.28%	2.27%	2.29%
6 and more	flat	rent	1.56%	1.45%	1.39%

Table 20.1 Typology of households in different census years (Source: Census data, *INSEE*).

Typology of jobs

To take account of the distribution of jobs by activity sector, it was necessary to split the activities into 11 categories as follows:

- agriculture
- industry
- construction
- commercial activities and retail
- transportation
- financial activities
- real estate activities
- business services
- personal services
- education, health and social affairs
- administration.

In order to take into account the wide range of behaviors (in terms of birth, death, growth, and location) of firms as a function of their size, firms were classified in 2 wide categories of more or less than 10 employees.

Real estate prices

From data available on real estate prices in the *Cote Callon*, 5 types of real estate were defined: flats for rent or for sale, houses for rent or for sale and offices for rent. All prices are expressed in euro$\cdot$m^{-2}. The raw data obtained from *Cote Callon* are much more detailed and need to be grouped. The raw data differentiate the dwellings according to the age, comfort level, number of rooms, etc. The level of detail varies from one *commune* to another. For example, the partition according to the number of rooms is more detailed in the historical Paris than elsewhere in *Ile-de-France*.

To compute the average price value for the 5 types of real estate properties, we have evaluated for each *commune* the weighted average of the more detailed prices. The weights were inferred from Census data, which provide information for example on the number of rooms in the dwellings or on the construction age.

The prices of the *Cote Callon* were available for *communes* with more than 5000 inhabitants. To obtain the price for other *communes*, we have estimated the hedonic model of price on the sample, where they were observed. We then used such estimation to predict the price for all the *communes* in *Ile-de-France*.

Land use

For the land use transition module, the different types of land uses are classified into 4 general types: Single houses, Flats or Collective dwellings, Activities and others (see Table 20.2). These 4 categories capture the guidelines for the most relevant land use transition and to analyze and predict the urban development. Parallel to this aggregation process of the different types of use, we have studied to which extent each type of land use could accommodate different types of business or governmental activities. Some combinations have a very low probability to occur and on the contrary some types such as green areas and highly accessible places can only accommodate a limited number of activities. The institutional regulation, zoning, and fiscal constraints have also been taken into account.

20.5 Data analysis: spatial issues

20.5.1 Spatial issues

The area can be split in 3 rings, 8 *départements*, 44 *arrondissements*, 1 300 *communes*, 5 249 *IRIS*, and about 50 000 grid cells (500m side). Table 20.3 presents accessibility and environmental variables and indicates their average values in each *département*. The attributes are weighted by the number of households in each entity in order to compute these averages. Many statistics are only available at *commune* level. Table 20.3 indicates average value of these variables.

Table 20.4 shows the variance decomposition at the three relevant geographical levels: *Département*, *commune*, and grid cell. For each variable X (such as green area or accessibility, etc.), the model is:

$$X_{dcg} = \mu + \eta_d + \xi_c + \varepsilon_g, \tag{20.1}$$

where, according to the analysis of variance model, X_{dcg} denotes the value of X in grid cell g located in *commune* c and in district d, μ is the average level of X in *Ile-de-France*, η_d, ξ_c , and ε_g are independent random Gaussian variables with zero mean and variances σ_d^2, σ_c^2 , and σ_g^2, respectively. As the polygons are independent in all levels, there is no correlation and covariance and we have the total variance of X denoted by $\sigma^2 \equiv \sigma_d^2 + \sigma_c^2 + \sigma_g^2$, and the percentage of variance (reported in Table 20.4) at level i, $i = d, c, g$, is

$$\alpha_i = 100 * \sigma_i^2/\sigma^2,$$

with $\alpha_d + \alpha_c + \alpha_g = 100$.

Code	Label 47 types	4 types	Poten. to activ.	
			SR	MR
1	Forests	Others		
2	Clearings in forest	Others		
3	Poplar plantation	Others		
4	Plowed land	Activities		20%
5	Grassland dedicated to agriculture	Activities		20%
6	Orchards, tree nurseries	Activities		20%
7	Gardening, horticulture	Activities		20%
8	Intensive crops in greenhouse	Activities		20%
9	Water	Others		
10	Non-agricultural grass surfaces	Others		100%
11	Quarries, pits	Activities		
12	Landfills	Activities		100%
13	Rural vacant space	Others		100%
14	Parks or gardens	Activities		
15	Allotments	Houses		
16	House gardens	Houses		
17	Sport fields outdoor	Activities		
18	Large space for sport facilities	Activities		
19	Camping, caravaning	Activities		
20	Related to recreation park	Activities		
21	Vacant land	Others		100%
22	Individual housing	Houses		
23	Identical sets of individual housings	Houses		
24	Rural housing	Houses		
25	Dense single dwelling	Houses		
26	Collective housing (not fragmented)	Flats	20%	20%
27	Fragmented collective housing	Flats	10%	10%
28	Other housings	Flats	20%	20%
29	Water, sanitation and energy equipments	Activities	100%	100%
30	Dedicated spaces to activities	Activities	100%	100%
31	Logistic warehouses	Activities	100%	100%
32	Commercial spaces	Activities	100%	100%
33	Offices	Activities	100%	100%
34	Buildings or facilities for sport	Activities		
35	Educational facilities	Activities	100%	100%
36	Health facilities	Activities	100%	100%
37	Cemeteries	Activities		
38	Large convention and exhibition centers	Activities		
39	Cultural and leisure facilities	Activities		

Table 20.2 Land use categories and the 4 general types, short run and medium run potential to accommodate activities (Source: *IAU-IdF* and authors' appreciation) (*continued on next page*).

Code	Label 47 types	4 types	Poten. to activ.	
			SR	MR
40	Administration, official organization	Activities	100%	100%
41	Other facilities open to public	Activities	100%	100%
42	Railway transportation	Activities		
43	Roads	Activities		
44	Parking lots	Activities		
45	Bus stations, depots	Activities		
46	Airport facilities	Activities		
47	Construction sites	Activities		

Table 20.2 (*Concluded from previous page*).

Note that the term redevelopment area used in Table 20.3 is not a precise translation for the term '*Zones Urbaines Sensibles*'. These are areas with high concentrations of social and economic difficulty. They are targeted for government assistance or redevelopment investments. We will use the term Redevelopment Areas in this context.

We observe that for many of variables observable at an infra-communal level, a great part of variation is observed at that level finer than commune. That suggests to use an infra-communal geographical unit for modeling. But many important variables are not available at such a level. We denote also some communes with very low number of inhabitants and jobs (lower than 100) that are essentially situated at the far suburbs. We may suggest a multilevel (nested) geographical unit for modeling to take into account all available information. In this chapter we report the results of simulations done at commune level that was sufficient for the precision required for our policy analyses.

20.5.2 Households moving rate

Residential moving (relocation) strongly depends on the household types as can be seen in Table 20.5.

Globally the relocation rate decreases with the household size. It should be noted that the impact of household size also depends on tenure and type of housing. Tenants are generally more mobile than owners. Owners move more often when they initially own a flat. By contrast, tenants move more when they initially rent a house. These results are consistent with the fact that households either buy a house that they keep it several years (still, household competition has changed significantly) or rent a flat and move frequently in response to professional job change of the different members of the household.

Variable *département*	**75**	**92**	**94**	**93**	**78**	**91**	**95**	**77**
No. of Grid Cells	420	704	980	952	9 400	7 399	5 187	24 194
No. of *communes*	20	36	47	40	262	196	185	514
Distance and accessibility								
Subway and tramway stations	8.17	0.80	0.65	0.72	0.00	0.00	0.00	0.00
Subway and tramway stations around	71.93	7.14	5.77	6.49	0.00	0.00	0.00	0.00
Train stations	0.98	0.75	0.44	0.53	0.53	0.47	0.68	0.30
Train stations around	9.35	6.89	3.94	4.76	4.29	4.07	5.60	2.35
Accessibility to jobs (Public Transit)	50.93	49.29	48.02	48.25	45.20	44.30	45.97	41.86
Accessibility to jobs (Private Car)	54.45	53.88	53.42	53.48	51.18	51.15	51.82	48.60
Accessibility to jobs[52]	*-72.9*	*-72.3*	*-73.9*	*-74.0*	*-73.8*	*-93.3*	*-68.8*	*-89.9*
Average travel time (Public)	*28.08*	*31.54*	*38.45*	*38.54*	*47.84*	*52.04*	*46.24*	*48.86*
Average travel time (Private)	*16.19*	*16.08*	*16.31*	*16.90*	*24.13*	*36.58*	*20.31*	*35.59*
Accessibility to shops (Public)	33.24	32.25	31.98	32.06	30.28	30.25	30.94	28.66
Accessibility to shops (Private)	35.58	35.30	35.30	35.31	34.06	34.27	34.48	32.96
Private car travel time variability	*2.72*	*4.11*	*3.07*	*3.46*	*3.55*	*6.65*	*3.71*	*3.56*
Distance to arterial	2.03	1.28	0.74	0.67	3.27	1.62	2.71	2.15
Distance to highway	1.48	1.19	1.27	1.28	2.78	3.12	2.34	4.68
Distance to Châtelet	3.57	8.95	11.42	10.88	27.64	26.36	20.27	41.50
Environmental variable								
% surface in noisy areas (severe noise: >96dB)	0.00	0.00	4.44	0.03	0.00	2.32	1.36	0.76
% surface of parks and gardens	7.59	9.13	9.54	8.65	13.09	12.66	10.60	12.74
% surface covered by water	0.90	1.37	1.80	0.65	1.42	0.98	0.71	1.43
% surface covered by forests	0.27	1.54	1.20	1.01	7.15	7.28	3.97	8.02
% surface: Sport facilities	1.43	3.05	2.48	3.03	2.42	2.36	2.89	1.77
% surface: Public facilities	9.32	13.72	13.20	12.69	22.67	22.31	17.46	22.52
% surface: Open space	8.95	12.13	12.60	12.54	16.18	15.65	14.31	14.99
% surface: Redevelopment area	5.01	8.32	5.23	8.14	2.18	6.67	6.57	1.63

Table 20.3 Average accessibility and environmental variables, by *département* (over the grid cells and weighted by population). Variables in italics are computed at the *commune* level and not represented in Table 20.4 (Source: authors' computations from *IAU-IdF* GIS database and METROPOLIS computations).

Variable	*département*	*Commune*	Cell
Distance and accessibility			
Number of subway and tramway stations	60.5%	22.1%	17.4%
Number of subway and tramway stations around	65.2%	22.3%	12.4%
Number of train stations	10.0%	23.1%	66.9%
Number of train stations around	17.0%	33.4%	49.6%
Accessibility to jobs (Public Transit)	35.1%	64.9%	0.0%
Accessibility to jobs (Private Car)	39.1%	60.9%	0.0%
Accessibility to shops (Public Transit)	34.9%	65.1%	0.0%
Accessibility to shops (Private Car)	35.8%	64.2%	0.0%
Distance to nearest arterial	7.8%	84.1%	8.1%
Distance to nearest highway	7.8%	83.0%	9.2%
Distance to Châtelet (Paris center)	42.7%	56.9%	0.4%
Environment			
% surface in Noisy area (severe noise: >96dB)	1.8%	62.4%	35.9%
% surface of Parks and Gardens	5.2%	17.2%	77.6%
% surface of Water	7.4%	19.3%	73.4%
% surface of forests	2.5%	36.2%	61.3%
% surface in 'Redevelopment Areas'	3.9%	17.3%	78.8%
% surface of Public Facilities	0.5%	19.3%	80.3%
% surface of Open Space	1.9%	37.6%	60.5%
% surface of Sport Facilities	2.0%	11.3%	86.7%

Table 20.4 Variance decomposition of accessibility and environmental variables over different scales (Source: authors' computations from *IAU-IdF* GIS database and METROPOLIS computations).

	Household size					
Over 1 year (1998)	**1**	**2**	**3**	**4**	**5**	**6+**
house, own	4.94%	5.73%	7.06%	7.73%	7.50%	7.32%
flat, own	7.13%	6.35%	7.72%	6.60%	6.49%	7.68%
house, rent	22.20%	24.99%	25.51%	25.41%	23.21%	18.86%
flat, rent	19.18%	17.33%	15.08%	12.29%	11.65%	8.67%
all	13.98%	11.15%	11.60%	10.35%	10.19%	8.85%
Over 9 years (1990-1999)	**1**	**2**	**3**	**4**	**5**	**6+**
house, own	32.32%	35.71%	44.78%	57.93%	58.33%	52.90%
flat, own	41.46%	36.28%	49.73%	56.48%	58.08%	53.09%
house, rent	74.18%	76.31%	80.92%	85.55%	85.34%	78.86%
flat, rent	62.79%	59.85%	65.09%	67.22%	64.18%	54.83%

Table 20.5 Households moving rate, by household type (Source: Census data, *INSEE*).

20.5.3 Working zoning

In the context of Grand Paris, the results of model Relu-Tran (Anas & Liu 2007) based on about 50 zones have been compared with those of UrbanSim based on 1 300 zones. The definition of these zones depend on administrative boundaries on one hand and on constraints imposed by Grand Paris on the other hand. In the current form, the zones depend on the poles of economic development. They are 10 poles of economic development also referred to as *Contrats de Développement Territoriaux (CDT)*. Various difficulties arose during the process of defining the boundaries of the zones. Indeed several CDTs include part of several *départements* or *communes*. The final solution which was selected is a compromise between the logic behind the definition of CDT and the logic behind the zones used in the transportation model. Figure 20.5 presents the 50 zones selected which include 20 *arrondissements* within Paris and 10 zones which match approximately the *CDTs*.

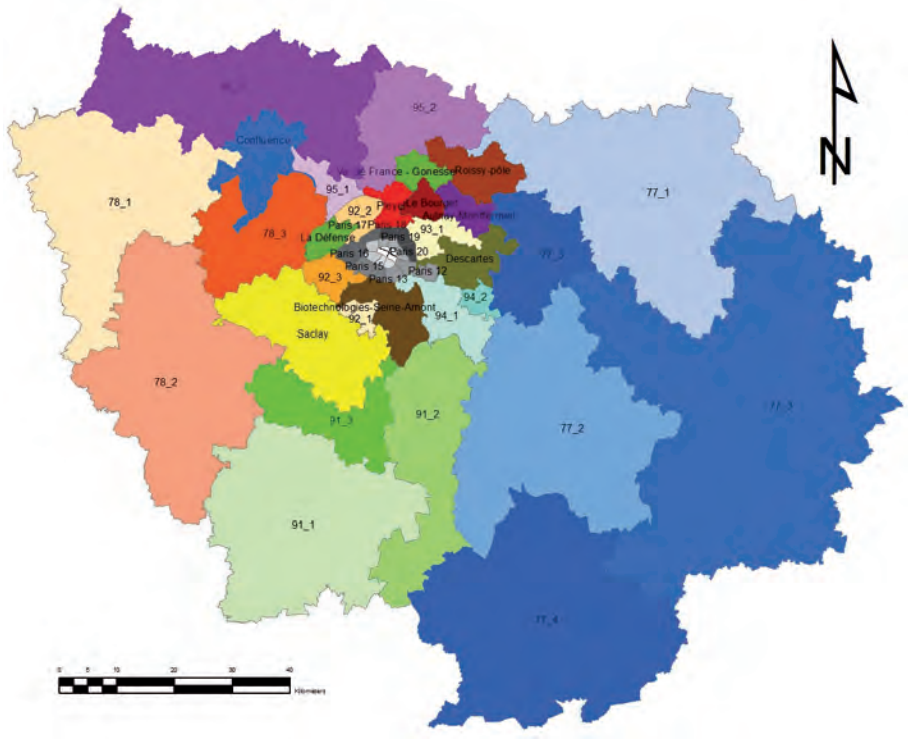


Figure 20.5 Aggregate zones defined in collaboration with *Société du Grand Paris*.

20.6 Model structure

UrbanSim is constituted by several models. The urban development model describes the endogenous land use change as an outcome of a dynamic process which is driven by the developers. The stakeholders (developers, public authorities, etc.) react themselves to satisfy the expected global demand of the households and of the firms (see Chapter 12 on this issue). A development project that consumes some surface of land and concerns a given number of dwellings or a given activity floor space. The project will be located in a parcel, consistent with regulation, which maximizes the profit of developer. For example, the probability that a given project of dwelling or activity floor space be located in a given parcel depends on the local prices as well as the demand for housing and commercial activities. This process provides a quite realistic description of the competitive use of land. Of course, competition remains imperfect and some political economy dimension should be included in the future work to account for this deviation from perfect competition.

The potential land use are defined exogenously. They depend on environmental regulation (e.g. preservation of forests) and on constraints imposed by the local urban plan *(Plan Local d'Urbanisme)*. Time lags are important. For example there is a time lag of one year between the decision to allocate land to some residence or commercial activity and the actual available supply. In other words the supply of housing and commercial unit is inelastic in the short run (1 year), but elastic in the medium run. It is assumed that, on the demand side, agents are price takers since they are small and therefore unable to have an impact on the price level (this is an assumption used in perfect competition studies). The price adjusts for one year to another as a function of supply and demand in each of the sub markets (e.g. flats, houses, purchase or the rent of dwellings and offices for rent). Household and firm behavior depends on various factors such as prices, accessibility, and local amenities. The main behavior modeled is the location choice.

Location choice modeling

The location choice of jobs and workers is determined by the location of the firm and of the households. Residential location is specific to each type of housing. There are several determinants of residential location. For example density, accessibility to job and shopping areas, local supply of dwellings, income of households of the vicinity, local amenities and the proximity to the city center. It is measured by the distance to Chatelet (the point of highest accessibility for public transportation) and to expressways. UrbanSim assumes inelastic supply of the dwelling in short run (over 1 year). The choice of residential location is a first come first served principle; households are sorted in a random way. At the beginning of the year the given number of locations are available at each location (*commune*). This determines the choice set of the first household

which wishes to find a residential location. The choice set of the next household in the list is the same as the previous one except that one vacant unit is now occupied and therefore unavailable. Housing prices are updated on a yearly basis but do not vary within a given year. Since the prices are fixed in the short run (one year of period) there may be rationing of some housing. This process is compounded by the fact that the markets are fragmented.

Employment location depends on the accessibility with respect to the worker, on agglomeration force, complementarity and substitutability between business activity sectors, on local population density, on the local distribution of land use, on the distance to the road infrastructures (streets, roads, and motorways), and on the distance to the center of the region.

20.6.1 Household models

Demographic model

The demographic module is described in Chapter 7. A microsimulation model was developed by *INED*[53] and allowed to create and update the yearly population database for the period 1999–2050.

Household Residential Location Choice Model

We have investigated what is the most appropriate spatial scale for studying preferences for locational amenities. This raises the question of the most relevant scale for estimating a household location choice model. We were asking whether residents have preferences for *communes* or for smaller more homogeneous geographical units. In order to answer this question, we estimated two location choice models: one at the *commune* level, and the other one at the grid cell level.

The *commune* j contains C_j dwellings, j=*1*, ..., *1300*. We assume that all the dwellings i located in *commune* j have the same observable attributes, and therefore the same expected utility $V^h_{i\in[1,C_j]} = V^h_j$ for household h, h=*1*, ..., *N*. The total number of dwellings in *Ile-de-France* is denoted by I. Typically the value of I is about five million dwellings over the area.

The probability that household h chooses a dwelling i located in *commune* j is given by the Multinomial Logit formula:

$$P^h_i = \frac{\exp\left(V^h_i\right)}{\sum_{i'=1}^{I} \exp\left(V^h_{i'}\right)}. \tag{20.2}$$

Since we have assumed that all the dwellings located in j have the same expected utility V^h_i, they therefore have the same probability of being selected

[53] *Institut national d'études démographiques*

$P^h_{j'}$, Equation (20.2) implies that the probability that household h selects *commune* j is

$$P^h_j = C_j P^h_i = \frac{C_j \exp\left(V^h_j\right)}{\sum_{j'=1}^{J}\left(\sum_{i' \in j'} \exp\left(V^h_{i'}\right)\right)} = \frac{\exp\left(V^h_j + \log\left(C_j\right)\right)}{\sum_{j'=1}^{J} \exp\left(V^h_{j'} + \log\left(C_{j'}\right)\right)}. \quad (20.3)$$

Under the IIA (Independence of Irrelevant Alternatives) assumption, one can obtain consistent estimates of the parameters in the expected utility function by selecting a random sample of alternatives, with uniform sampling of alternatives, provided the correcting term $\log(C_j)$ is added to the likelihood. However, more efficient estimates can be obtained with the technique of importance sampling of alternatives, that is if the probability that alternative j is included in the choice set is proportional to C_j, provided a second correcting term $-\log(C_j)$ is added to the likelihood. It can be shown that no correcting factor is necessary to obtain consistent estimates of the parameter when importance sampling is used. (See Ben-Akiva & Lerman (1985) and de Palma, Motamedi, Picard & Waddell (2005*a*) for details on this issue).

For the model at grid cell level, assuming that all the dwellings i located in the grid cell k, k =1,...,K, which is located in *commune* j, j =*1,...,J,* have the same observable attributes, household h, h =*1,...,N,* have the same expected utility $V^h_i = V^h_k$ for them. We assume that the expected utility is a linear combination of grid cell attributes X_k and *commune* attributes Z_j, in which the marginal utilities of grid cell and *commune* attributes can be household-specific. Expected utility is therefore of the form:

$$V^h_i = X_k \beta_h + Z_j \gamma_h, k \in j, \quad (20.4)$$

where β_h and γ_h denote the household-specific marginal utilities of grid cell and *commune* attributes, respectively.

If households choose among *communes* rather than grid cells, then the expected utility of a dwelling located in *commune* j only depends on Z_j and is of the form:

$$V^h_j = Z_j \delta_h = Z_j \gamma_h + E\left(X_k \beta_h | Z_j\right), \quad (20.5)$$

where δ_h mixes the marginal utilities of *commune* attributes and the marginal utilities of grid cells attributes, weighted by their distribution in the *commune*. The exact formula for $E\left(X_k \beta_h | Z_j\right)$ corresponds to the log-sum in a nested logit model (see Anderson et al. (1992) for details).

The number of dwellings in *commune* j and in grid cell k are denoted by C_j and G_k, respectively. Since all the dwellings located in k have the same expected utility, and the same probability of being selected, Equation (20.3)

implies that the probability that household h selects grid cell k is

$$P_k^h = G_k P_i^h = \frac{G_k \exp\left(V_k^h\right)}{\sum_{k'=1}^{K} \sum_{i' \in k'} \exp\left(V_{i'}^h\right)} = \frac{\exp\left(V_k^h + \log\left(G_k\right)\right)}{\sum_{k'=1}^{K} \exp\left(V_{k'}^h + \log\left(G_{k'}\right)\right)}. \tag{20.6}$$

Similarly, the probability that household h selects *commune* j is

$$P_j^h = C_j P_i^h = \frac{C_j \exp\left(V_j^h\right)}{\sum_{j'=1}^{J} \sum_{i' \in j'} \exp\left(V_{i'}^h\right)} = \frac{\exp\left(V_j^h + \log\left(C_j\right)\right)}{\sum_{k'=1}^{K} \exp\left(V_{j'}^h + \log\left(C_{j'}\right)\right)}. \tag{20.7}$$

We report in Table 20.6 the estimation results for the household location choice models at the level of 500m x 500m cells within the study area, and at the level of *communes*.

20.6.2 Employment models

In the Paris case study, we have estimated and implemented two versions of employment location models. The first one was a job location model and the second a firmographic model (Van Wissen 2000). Firmography provides a more realistic and comprehensive knowledge of the evolution of employment over the studied area and also about the establishments.

Background and definition

The term firmography, by analogy with demography, can be assigned to a four-fold model which includes establishment birth, its re/location, variation in the number of jobs within establishment, and its ultimate destruction. Consistently with available data, we propose some theoretical models which describe three stages of the establishments life cycle: disappearance of the establishments, evolution of the workforce within the stable units, and location of new establishments. For more discussion of the life cycle of firms (see Chapter 11).

To get an overview on the first stage of the establishment life cycle, the birth of the establishment, we compute the creation rates across eleven activity sectors for each county of *Ile-de-France*. All models are estimated separately for each activity sector. Given the volume of results, we refer the reader to Buczkowska et al. (2013) for more details. We will focus here on three key models. The first one describes establishment disappearance; the second one describes the evolution of the establishment workforce and the third one describes the location choice. Finally we briefly discuss the implementation of firmographic models in UrbanSim.

Model of establishment disappearance

We model the establishments death using a binary probit model. The death of the establishment is registered in two cases: when the establishment is closed,

Variable	*Commune* level		Gridcell level	
	Coeff	t-stat	Coeff	t-stat
Same district as before move	2.539	275.9	2.535	272.5
Paris	-0.254	-10.10	-0.168	-10.29
Log(Price of flat)	0.006	0.20	0.021	0.94
Log(Price of flat)* (Age-40)/10	-0.029	-1.74		
Log(Price of flat)* (Log(Income)- $\overline{\text{Log(Income)}}$)	0.409	8.20	0.248	5.84
Log(Price of flat)* (N. of hh members – 1)	-0.042	-2.40		
Log(Price of house)	0.087	3.70	0.139	7.41
Log(Price of house)* (Age-40)/10	-0.104	-8.78	-0.114	-20.90
Log(Price of house)* (Log(Income)– $\overline{\text{Log(Income)}}$)	0.187	5.78	0.149	5.17
Log(Price of house)* (N. of hh members – 1)	-0.109	-8.38	-0.130	-20.43
N. Subway stations around			-0.004	-2.37
N. Subway stations in the *commune* / cell	0.005	4.59	0.001	0.21
N. Railway stations around			0.013	3.85
N. Railway stations in the *commune* / cell	-0.009	-4.86	0.006	0.60
Average travel time from j, commuting (TC) [hr]	0.025	0.80	0.040	1.45
TC*(Dummy female) [hr]	-0.374	-8.28	-0.294	-6.56
Distance to highway [km]	-0.005	-3.10	-0.001	-0.74
Distance to arterial [km]	-0.008	-2.98	-0.015	-5.87
Distance to Châtelet (Paris center) [km]	0.002	2.94	-0.001	-1.06
% households with 1 member * 1 member in h	1.877	20.62	2.270	30.73
% hhs with 2 members* 2 members in h	1.336	4.20	1.773	10.03
% hhs with 3+ members* 3+ member in h	2.230	22.16	2.075	27.98
% hh with no working member * no working member in h	7.797	33.91	5.382	33.49
% hh with 1 working member * 1 working member in h	-0.931	-5.84	0.502	4.58
% hh with 2+ working member * 2+ working member in h	1.664	14.83	0.404	5.01
% hh with a young head	0.450	3.01	1.021	10.86
% hh with a young head * young head in h	4.777	28.34	3.243	29.59
% hh with a middle age head * middle aged head in h	-0.693	-4.16	-0.162	-1.51
% Rich hh * Rich h	3.280	28.36	2.876	36.58
% Medium Income hh * medium income h	1.621	10.11	1.923	17.00
% Poor hh * poor h	0.459	3.53	1.070	12.11
% hhs with a foreign head * foreign head in h	5.046	26.74	4.578	37.26
% hhs with a foreign head * French head in h	-1.995	-17.44	-1.305	-16.71
% of surf. in Redevelopment Area * Rich h	0.678	4.97	-0.228	-4.28
% of surf. in Redevelopment Area * Med. Inc. h	-0.083	-0.78	-0.193	-5.16
% of surf. in Redevelopment Area * Poor h	0.431	4.45	0.165	4.99
Log of the number of residential units	0.067	9.26	0.023	3.27
% of flats in total dwellings * Foreign head in h	1.368	15.42	0.676	11.37
% of flats in total dwellings * French head in h	0.485	7.28	0.161	3.76
% of flats in total dwellings * (N. of members - 1)	-0.085	-4.29	-0.056	-4.29
% of flats in total dwellings * young head in h	-0.131	-2.43	-0.037	-1.01
% of flats in total dwellings * old head in h	-0.766	-8.40	-0.188	-2.82

Table 20.6 Household location choice: estimation results at *commune* and grid cell levels (Source: authors' estimations) (*continued on next page*).

Variable	*Commune* level		Gridcell level	
	Coeff	t-stat	Coeff	t-stat
% of surface in Noisy area	-0.089	-1.56	-0.077	-1.52
% of surface covered by Forest	-0.117	-3.09	-0.297	-5.05
% of surf. covered by Forest * N. of Children	0.423	6.87	0.461	4.99
% of surface covered by Water	0.204	1.62	0.333	4.24
% of surface covered by Parks and Gardens	-0.428	-4.22	-0.152	-2.93
% of surf. covered by Parks * N. of children	-0.129	-0.71	0.615	6.67
% of surface covered by Sporting facilities	0.442	2.50	-0.371	-4.06
% of surf. covered by Sporting facilities * N of children	0.268	0.83	0.786	4.93

Table 20.6 (*Concluded from previous page*).

but also when it relocates. With the available data, we are not able to consistently track the establishments which move. For this reason, we do not model the relocation of the establishments. When the establishment moves, as its id[54] is automatically changed so that there is no possibility for us to identify the "same" establishment before and after the move. Then, we consider this event as a death and a birth of a new establishment. For an establishment present in 1997, the probability to disappear is:

$$\mathrm{P}\left(e \notin \mathrm{E}^{2001}\right) = \mathrm{P}\left(\mathrm{VL}_e > 0\right) = \mathrm{P}\left(\beta_s \mathrm{X}_c + \gamma_s \mathrm{Y}_e + \varepsilon_e > 0\right),$$

with

$$\mathrm{VL}_e = \beta_s \mathrm{X}_c + \gamma_s \mathrm{Y}_e + \varepsilon_e,$$

where

e - establishment present in the market in 1997
E^{2001} - set of establishments present in 2001
VL_e - value of leaving the market by the establishment e
ε_e - random part which measures the fragility of the establishment; $\varepsilon_e \sim \mathrm{N}(0,1)$
X_s - *commune* characteristics
β_s - vector of coefficients which stands by the *commune* variables
Y_e - establishment characteristics
γ_s - vector of coefficients which stands by the establishments characteristics
$\mathrm{P}\left(e \notin \mathrm{E}^{2001}\right)$ - probability that the establishment is not in the market any more in 2001.

We check whether the following variables have a significant effect on the disappearance of the establishment:

- the initial size of the establishment,
- accessibility to population with particular socio-economic characteristics,
- regional accessibility measures, such as number of metro stations, average time for public transport or private vehicles,

[54] The establishment has a unique id number assigned by the *INSEE*. It is named SIRET code.

- accessibility to labor (workforce from different socio-professional classes, workforce with different education levels),
- real estate prices.

Model of establishment evolution

To model the workforce evolution of the stable establishments, i.e. the establishments which are observed in the market for the two analyzed periods, we use regression model with the logarithmic value of the relative difference between establishments size in the final period and the initial year as a dependent variable. We first consider

$$\ln\left(\frac{\tau_e^{2001}}{\tau_e^{1997}}\right) = \text{const}_s + \sum_{t\in T}\beta_{s,t}I_t + \sum_{t\in T}\gamma_{s,t}I_t\left(\tau_e^{1997} - \tau_t^{\inf}\right) + \delta_s X_c + \varepsilon_e.$$

Alternatively, we consider another specification

$$\ln\left(\tau_e^{2001}\right) = \text{const}_s + \sum_{t\in T}\beta_{s,t}I_t + \sum_{t\in T}\gamma_{s,t}I_t\left(\tau_e^{1997} - \tau_t^{\inf}\right) + \delta_s X_c + \varepsilon_e.$$

The evolution model describes the growing or shrinking behavior of the establishment. Final workforce is a function of the initial workforce and the variables which represent the socioeconomic characteristics of the population and employees living in the *commune* surrounding of the establishment.

We use the following notation:

I_t - size class indicator
t- size class
τ_e^{2001} -workforce (size) of the establishment in 2001
τ_e^{1997} - workforce (size) of the establishment in 1997
$\tau_t^{\inf}$ - lower limit of each size class t
X_c - *commune* characteristics
δ_s - vector of coefficients which stand by the *commune* variables
ε_e - random perturbation.

To estimate the parameters in the evolution models, similar variables are used as in the disappearance models presented in the previous section.

Establishment Location Choice model

We use the multinomial logit model (McFadden 1974 and McFadden 1981) to study the location choice of the business units. The probability that a *commune* is included in the choice set is proportional to the total number of employees in this particular *commune*.

$P_i^e = \frac{\exp(V_i^e)}{\sum_{i'=1}^{I}\exp(V_{i'}^e)}$, and

$$P_j^e = C_j P_i^e = C_j \frac{\exp(V_i^e)}{\sum_{j'=1}^{J} \sum_{i' \in j'} \exp(V_{i'}^e)} = \frac{C_j \exp(V_i^e)}{\sum_{j'=1}^{J} C_{j'} \exp(V_{i'}^e)} = \frac{\exp(V_i^e + \ln C_j)}{\sum_{j'=1}^{J} \exp(V_{i'}^e + \ln C_{j'})}.$$

As the residential location choice model, we assume that all locations which belong to *commune* j have the same observable characteristics and lead to the same expected utility level. Thus all the establishments in the same *commune* j have the same probability P_j^e to be chosen by the establishment. These models address location behavior of units at the establishment level. We model only the location choice of new establishments. As was mentioned before, according to the way the employment database is built, there is no possibility to distinguish between the real disappearance of the establishment and its relocation. To avoid correlation between variables, de Bok & Sanders (2005) propose to construct a variety of *composed* accessibility that describe the distances to the physical infrastructure, the accessibility to labor, the accessibility to the customers and suppliers. They also suggest using a set of agglomeration variables to explore concentration or deconcentration patterns. They use the information on the urban environment and they treat an average rental level in a particular territory as a proxy for the real estate quality. All the models presented by de Bok and Sanders are computed by establishments size and by activity sector. Waddell & Ulfarsson (2003*b*) confirm the significance of the real estate characteristics, the neighborhood characteristics, local and regional accessibility measures (access to jobs, population and distance to highway, airport or central business districts) for the location models. Picard & Antoniou (2011) claim that commuting time is the most significant variable in explaining the location, much more important than variables which measure either accessibility or expected time commonly used in the location choice models.

In the location choice model, which we run, we take into consideration 10 out of the 11 activity sectors (Agriculture sector is excluded)[55] and we test the significance of a number of variables.

- the local competition in the employment market and the employment density in each activity sector,
- local and regional accessibility measures (variables indicating the distance to the national road, highways, travel time data, etc.),
- general characteristics of the region in which the establishment is being located,
- accessibility to population (financial situation of the households, the age of the head of the household, the presence of children in the household),
- real estate prices, treated as a proxy for the attractiveness of the area,
- land use data.

[55] The number of observed creations or moves of establishment in this sector was too low to be able to estimate a significant location choice model. That will be the case in the simulated period, too. This sector will be treated exogenously.

Firmographics implementation in UrbanSim

The three models described above have been implemented in UrbanSim by instantiation of the existing generic linear regression and multinomial logit models of OPUS core (see OPUS User Guide 2011). The estimated models represent the evolution over four years between 1997 and 2001. Also, the models are describing the establishment behavior as a function of the urban condition 4 years before. To obtain realistic results, at each year the model uses 4-year-old information and not the information of the last simulated year as it is the normal procedure of UrbanSim. The number of very large establishments is low and their behavior is different from other establishments. As a consequence, we have excluded the establishments with a workforce greater than 250 employees from our models. Any information about their evolution has been introduced as exogenous events. The control data that drive the simulation are given as the total number of jobs at each year differentiated by the activity sector. To guarantee the equality of the total number of jobs and the sum of the jobs in active establishments, we have added a 4th module that along with the employment transition model controls for the total number of jobs. If in a sector and for the simulated year, the total number of jobs exceeds the target one, the transition model will insure that the appropriate number of establishments disappear in order to meet the target value. If after the workforce evolution step, the number of jobs over the region is lower than the target control value, the establishment reappearance model will reactivate the appropriate number of previously disappeared establishments.

20.6.3 Real estate price models

The property prices are modeled using hedonic regression (Rosen 1974). We have performed simultaneous regressions for the price (rental price and purchase price) of offices, houses, and flats. The method of hedonic regression models the interaction between supply and demand. It allows to compute the implicit valuation of the characteristics of the dwellings as well as of the characteristics exogenous to this dwelling including several local amenities identified in the literature. Moreover, the price of the specific housing units varies over time as a response to fluctuation between supply and demand. The difference between short term vacancy rate (for given year) and long term vacancy rate plays also an important role in these evaluation. For example, when the short run vacancy rate is below the structural vacancy rate, price level is triggered upward, and vice versa for the other situation.

The prices are updated each year after the construction and the market transactions occurred last year. The prices at the end of a given year are used as the reference prices for the decisions occurring in the following year.

20.6.4 Urban development models

The urban development model simulates the construction of new housing projects which follow new infrastructure development or to transformation of existing projects. The necessary data to estimate the coefficients for urban development models, rely on data of *MOS*. We identify the changes between two consecutive dates in the land use. The model predicts the probability that a project with a given land use type be located at a given *commune*. A multinomial logit model is used to estimate that. This probability is then estimated and a Monte-Carlo simulation is used to select the specific location of new project. More precisely, each simulation year the model computes the difference between total spaces and occupied ones. If the vacancy rate is lower than some critical value, new projects will be created accordingly. The model will determine exactly where this project will be located. The urban development model has three modules.

- The transition module which determine the needs for construction of housing units and which computes the quantities to be built (number of dwellings and business units). This module also selects the projects that are required to be performed.
- The module which computes the geographical location of the project.
- The module which updates the stock of buildings which in particular takes into account new constructions.

During the simulation, the model performs sequentially the following steps:

- Computation of the real vacancy rate (number of housing units minus number of households);
- comparison of the above vacancy rate with the structural vacancy rate;
- decision to build;
- computation of the surface to be built;
- selection of the project of a given size required to meet specific needs;
- computation the choice set (the *communes* which can potentially accommodate the specific project on the basis of the nature of the land supply);
- location of the projects;
- update of the buildings capacity.

20.7 Transportation model

METROPOLIS is a fully dynamic transportation model that is especially adapted for large networks. It is a mesoscopic event based model and uses a multi-agent methodology with a disaggregated representation of travelers. The supply system relies on a macroscopic formulation that computes travel time as a function of the flow condition of the link. It has been developed since

the 90s and its main application on the region has been within the QUATUOR project (THEMA-TT-R 2002). It models the mode, departure time and route choices of users. The Logit formula is used for these models. The dynamic assignment procedure can be deterministic or stochastic.

The generalized cost function is:

$$C(t) = \alpha tt^{VP}(t) + \beta [t + tt(t) - t^*]^+ + \gamma [t^* - t - tt(t)]^+.$$

The generalized cost $C(.)$ includes the schedule delay cost terms in $(t^* - t_a)$, where t^* is the desired arrival time. Moreover t_a denotes arrival time and tt is the travel time. The operator can enter some distributions for α (VOT), β and γ (delay penalties), and t^*. These behavioral parameters are the only information that are necessary in addition to data required by the classical static traffic assignment models (such as network topology, links characteristics, and O-D matrices).

The mode choice is described by a binary logit model. The output of METROPOLIS that we mainly use in this study is the surplus for any traveler category and for any O-D pair. As the departure time choice is modeled by a continuous Logit model, the surplus is given by

$$S = \mu^T \ln \int exp\left(-C(t)/\mu^T\right) dt,$$

where $C(t)$ is the time dependent travel cost and where μ^T denotes the departure time choice heterogeneity parameter (see also de Palma et al. (1997) and de Palma & Marchal (2002)). For more information about METROPOLIS application and calibration in *Ile-de-France*, see SustainCity W.P. 6.7.

20.8 Integrated model of METROPOLIS and UrbanSim

20.8.1 Integration of METROPOLIS and UrbanSim

An interface is developed to manage the simulation process and data exchange between METROPOLIS and UrbanSim. The interface is described in Chapter 18. The interface computes and updates the O-D matrix based on the simulated population and employment data.

20.8.2 Runs and calibration of integrated model

The urban model consists of several modules which represent several urban subsystems. As we presented before, these modules are estimated using several types of data. Since the different sources do not necessary span on the same period, we had to find some time compromises. Each model has a margin of error which is more or less important and depends among others on the quality of available data. The difference between simulated values and observed ones is greater when the simulated period is longer and when the calibration

period is shorter. Moreover the simulation represents the observed phenomenon on the basis of the explanatory variables. However, the future is not always simple extrapolation of the past and we have decided to restrict the number of explanatory variables in order to improve the robustness of our prediction. This partially explains the difference between the simulation results and the observation for periods beyond the estimation period.

We use a 2-step procedure to calibrate UrbanSim. In the first step we use estimated coefficients to calibrate the simulation results (as a matter of fact the estimations are based on average values of the coefficients and some sensitivity analysis is required to analyze the results when value of coefficient can vary within the confidence interval of the estimation). The idea of calibration is to determine the value of the coefficient within the confidence interval in order to reproduce in the best possible way the observed data. The second step is concerned with the period beyond the estimation period. In this case, we adjust various constants specific to the geographical units and to the different groups of agents in order to get simulation results coherent with actual data.

With *Ile-de-France*, the first step corresponds to the period 1990–1999 and the second step corresponds to the period 1999–2006. We mainly fit the residential location model as well as the establishment location choice model. A constant was introduced for each *CDT* (10 zones), for each type of Household (24 types), and for each business activity (10 sectors). We have obtained a mean relative weighted error equal to 0.07% for the 10 *CDT* and the R^2 is larger than 99% for the employment (simulated versus observed). These are very encouraging results. We believe that the quality of the calibration is good, and then it can be used to study different scenarios for the Grand Paris. This will be discussed below.

20.9 Policy scenarios and results: The Grand Paris project

Several policy scenarios have been considered to be simulated by the integrated model. Among them, we describe the Grand Paris project which has several particularities that make it the most interesting to present in this chapter.

20.9.1 The Grand Paris project

In the long run, the Grand Paris project aims at transforming Paris area in its role as a large world metropolis. The realization of the Grand Paris project will rely on large economic centers (named *CDTs* for *Contrats de Développement Territorial*) around the historical center, connected through a new fast public transportation network (named *RTPGP* for *Réseau de Transport Public du Grand Paris*), and governed by an institutional structure, the *Société du Grand Paris*.

The radial structure of the current transportation network in the Paris area favors trips inside and towards Paris, but it is ill-adapted for transversal trips from one suburb to another suburb. Currently, most of the trips between two cities located in the suburbs must transit through Paris, which induces an inefficient congestion in the transportation network.

The project analyzed here is made of 3 automatic Metro lines representing a total of nearly 160 km, and 95 stops in 84 stations, for a budget of nearly 30 billion euro. These lines complete and improve the structure and efficiency of the existing network and go through 117 *communes*. The objective is to avoid useless trips through Paris, attract a large enough fraction of suburb-to-suburb trips, and thus release the pressure currently affecting the central Paris network.

Two main phases are considered for the realization of the Grand Paris Project, in 2025 and in 2035. The main characteristics of the Grand Paris Metro lines are presented in Table 20.7.

Line	Peak hour trains interval	Line Capacity	Commercial speed
Blue	85 s	38,000	45 $km.hr^{-1}$
Red	120 s	32,000	60 $km.hr^{-1}$
Green	150 s	6,000	65 $km.hr^{-1}$

Table 20.7 Main characteristics of the Grand Paris automatic Metro lines (Source: SGP 2011).

20.9.2 Expected impact on urban development and envisaged scenarios

The Grand Paris Project increases by more than 50 percent the total length of Metro lines, and by 20 percent the number of Metro stops in Paris area. The transport supply improvement induced by this project will dramatically improve accessibility across the region, and will considerably improve regional attractiveness and competitiveness. Measuring the aggregate effect of the project at the regional level would be out of the scope of this study, and it would require different tools adapted for macroeconomic and interregional analysis. Such macroeconomic studies were financed by the *Société du Grand Paris*, and their results are used as inputs here.

Three scenarios were established concerning population and employment:

- *Reference scenario.* This scenario corresponds to the regional evolution without the Grand Paris Project. It relies on the projections used in the draft of *SDRIF*[56] prepared by the *Institut d'Aménagement et*

[56] *Schema Directeur de la Région Ile-de-France* is the regional master plan for *Ile-de-France* (SDRIF 2008).

d'Urbanisme de la région Île-de-France (IAU-IdF). This scenario predicts 1.348 million inhabitants more and 0.683 million jobs more by 2030.

- *Low scenario* about regional attractiveness. This scenario corresponds to the lower limit of the projections of the effect of the the Grand Paris Project on regional attractiveness. This scenario predicts 1.406 million inhabitants more and 0.8 million jobs more by 2030.

- *High scenario* about regional attractiveness. This scenario corresponds to the upper limit of the projections of the effect of the the Grand Paris Project on regional attractiveness. This scenario predicts 1.5 million inhabitants more and 1 million jobs more by 2030.

Since the phases of the Grand Paris project are scheduled in 2025 and 2035, we translated the above values into projections for 2025 and 2035, and we used a piecewise-linear function for intermediate years. Table 20.8 presents rounded and exact values for the three scenarios.

Scenario	Values	Population Increase			Employment Increase		
		2025	2030	2035	2025	2030	2035
Ref.	Round.	1 100 000	1 300 000	1 600 000	500 000	700 000	800 000
	Exact	1 078 966	1 348 706	1 618 446	546 825	683 535	820 245
Low	Round.	1 100 000	1 400 000	1 700 000	600 000	800 000	1 000 000
	Exact	1 125 070	1 406 322	1 687 573	639 996	799 996	959 995
High	Round.	1 200 000	1 500 000	1 800 000	800 000	1 000 000	1 200 000
	Exact	1 199 997	1 500 006	1 800 014	800 004	1 000 006	1 200 008

Table 20.8 Input data for the 3 scenarios (Source: *SGP, DRIEA*).

The main objective of the study was to evaluate the effects of the Grand Paris Project until year 2035. Additional projections were run until 2050 assuming constant aggregate population and employment, in order to evaluate the long run evolution of the distribution of population and employment within the region after the network and total population and employment remained steady. Such very-long-run projections are less precise because the duration of the projection is very large.

In addition, the Grand Paris Project should improve the local economic development of some fractions of the region, such as the *départements* Seine-Saint-Denis or the *Plateau de Saclay*. The local development in those subregions is currently restricted by the lack of an efficient transportation network.

20.9.3 Data preparation for the 3 scenarios

Yearly control totals for the 3 scenarios, over the period 1999–2050, were required to run UrbanSim. To perform this exercise, we used and extrapolated

the input data provided by the *Société du Grand Paris* for the 3 scenarios in 2005, 2025, 2030, 2035, and 2050, as well as census data and other data provided by *INSEE*. The yearly control totals and intermediary variables used to compute them are listed below.

- Total population and number of households, distributed over the 24 types chosen and total employment, by activity sector;
- Transportation supply and service levels.
  - Matrices of flow, distances and travel times, by mode, at each period, under each scenario.
  - Graphs of road and public transport networks, metro and rail stations and stops, at each period, under each scenario.

Total population

We extrapolated the total population from the input data described in Table 20.9. This input data comes from Census and *INSEE* for the past, and comes through *SGP* for the projections.

Year	Population			Source
	Reference	Low	High	
1999	10 952 011	10 952 011	10 952 011	Census
2005	11 433 302	11 433 302	11 433 302	*INSEE*
2006	11 532 409	11 532 409	11 532 409	Census
2008	11 659 620	11 659 620	11 659 620	Census
2025	12 512 268	12 558 372	12 633 299	*IAU, SGP*
2030	12 782 008	12 839 624	12 933 308	*IAU, SGP*
2035	13 051 748	13 120 875	13 233 316	*IAU, SGP*
2050	13 051 748	13 120 875	13 233 316	*IAU, SGP*

Table 20.9 Input data used to extrapolate total population (Source: Census, *INSEE, IAU-IdF, SGP*).

For the yearly extrapolation of total population, we assumed a constant yearly growth over each period between two lines in Table 20.9. This yearly growth rate is specific to each scenario after 2008. Each growth rate was computed to fit the figures of Table 20.9, and is reported in Table 20.10. The resulting population control totals are represented on Figure 20.6.

The decrease in growth rate between the two periods 1999–2006 and 2006–2008, both concerning census data, probably result more from changes in census data collection method (from exhaustive census each 10 year to partial yearly census) than from real changes in population growth rates. Such inconsistencies in our input data do not affect our simulations.

Period	Population growth rate		
	Reference	Low	High
1999 – 2005	0.72%	0.72%	0.72%
2005 – 2006	0.87%	0.87%	0.87%
2006 – 2008	0.55%	0.55%	0.55%
2008 – 2025	0.45%	0.47%	0.51%
2025 – 2030	0.43%	0.44%	0.47%
2030 – 2035	0.42%	0.43%	0.46%

Table 20.10 Population growth rate, by period and scenario (Source: authors' computations).

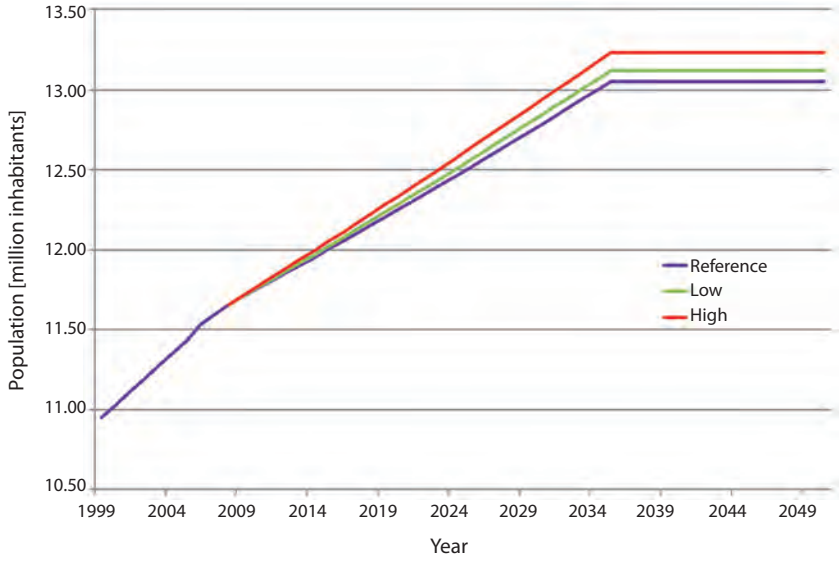


Figure 20.6 Control totals for population, 1999–2050 (Source: Census, authors' computations).

Number and size of households

We translated the yearly population control totals into number of households, by size (and type). In order to do this, we took into account and extrapolated to the period 2008–2035 the decreasing trend in household size observed over the period 1999–2008. The evolution of the distribution of households by size is assumed the same for the 3 scenarios. The resulting average household size is represented on Figure 20.7. The yearly number of households is obtained as the ratio between yearly population and yearly average size. The evolution of the number of households by type is specific to each scenario, like the evolution of total population.

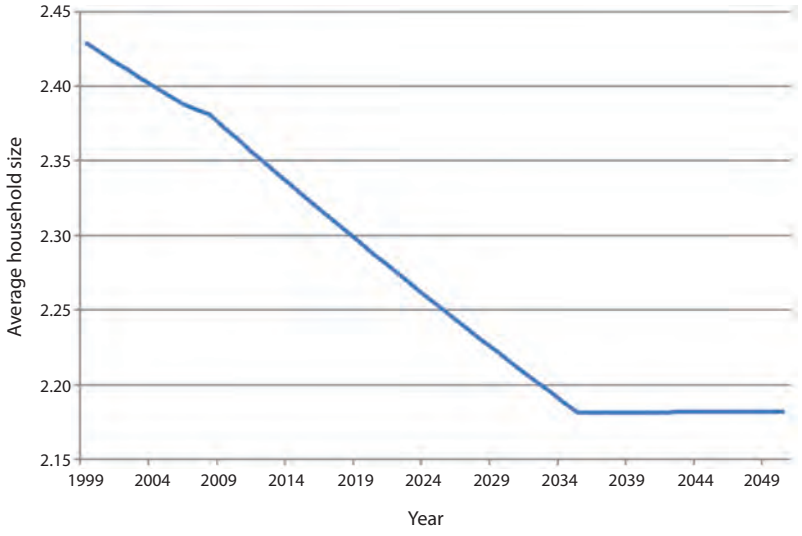


Figure 20.7 Average household size over the period 1999–2050 (Source: Census, authors' computations).

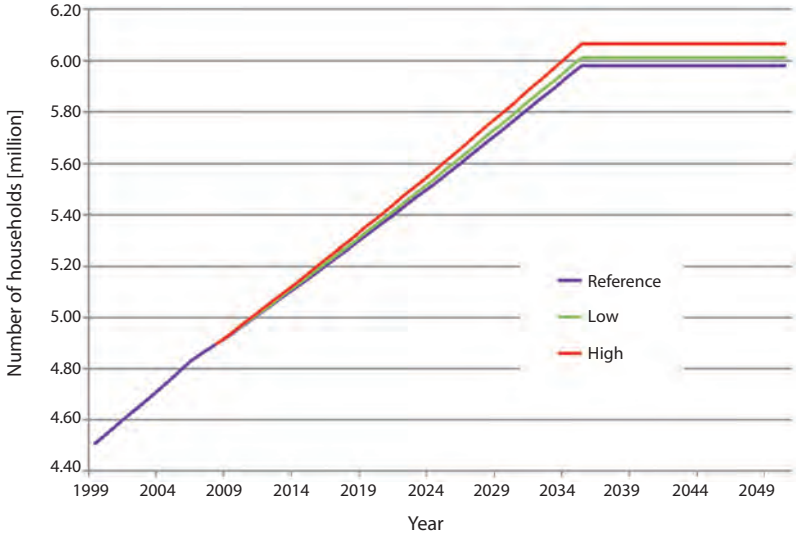


Figure 20.8 Total number of households for period 1999–2050 (Source: Census, authors' computations).

We then computed the yearly distribution of households of by size, extrapolating the trend observed in 1999–2006. This distribution is represented in Table 20.11 for the key years, and in Figure 20.9 for the yearly evolution.

Finally, households are distributed by dwelling type and tenure status, conditional on their size. This defines 24 types (6 size categories * 2 dwelling

types * 2 tenure status categories). To do this, we assumed that the distribution of household dwelling/tenure types conditional on household size is constant over time, and equal to the distribution observed in 2008.

Year	Distribution of household size						Average size	
	1	2	3	4	5	6+	6+	Total
1999	0.346	0.279	0.157	0.136	0.054	0.028	8.539	2.429
2006	0.354	0.287	0.149	0.130	0.053	0.026	8.601	2.388
2025	0.381	0.304	0.133	0.114	0.049	0.019	8.614	2.251
2035	0.397	0.311	0.124	0.106	0.046	0.015	8.644	2.181
2050	0.397	0.311	0.124	0.106	0.046	0.015	8.644	2.181

Table 20.11 Distribution of household size and average size, at key periods (Source: Census, authors' computations).

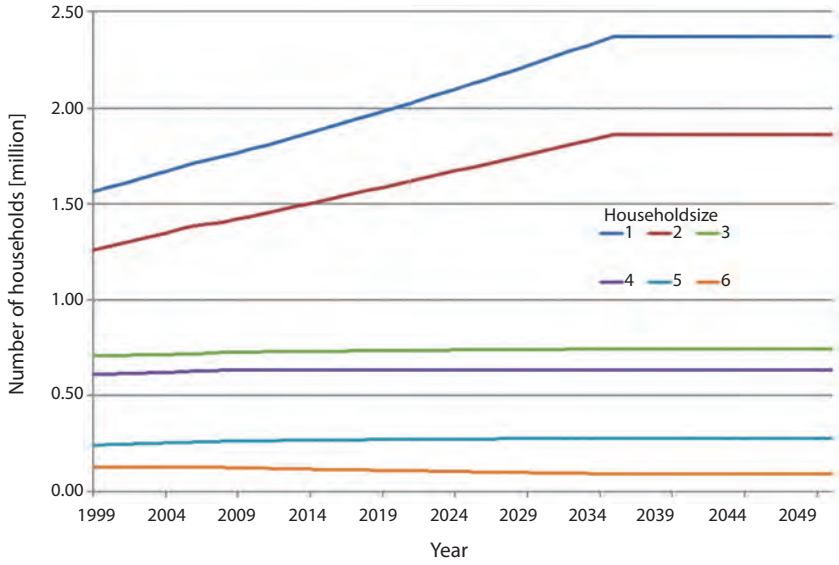


Figure 20.9 Number of households by size category for period 1999–2050 (Source: Census, authors' computations).

Total employment

A methodology similar to the one used for population was applied to build employment control totals:

- First step: Build control totals for total employment over the period 1999–2050;
- Second step: Distribute the yearly control totals for total employment over the 11 activity sectors considered.

The extrapolation of total employment control totals relies either on input values obtained through *SGP* (for projections), or from census data (for past data). Table 20.12 sums up the information available at key dates and the assumptions used to extrapolate this information to yearly control totals, and in Figure 20.10, represent the yearly evolution of total employment control totals over the period 1999–2050.

Year	Number of jobs			Source
	Reference	Low	High	
1999	5 041 995	5 041 995	5 041 995	Census
...	Constant growth rate			
2005	5 360 447	5 360 447	5 360 447	Census
...	Constant growth rate			
2025	5 907 272	6 000 443	6 160 451	*SGP*, input data
...	Linear evolution			
2035	6 180 692	6 320 442	6 560 455	*SGP*, input data
...	Constant			
2050	6 180 692	6 320 442	6 560 455	*SGP*, input data

Table 20.12 Input values and assumptions for extrapolating total employment (Source: Census, *DRIEA*, *SGP*).

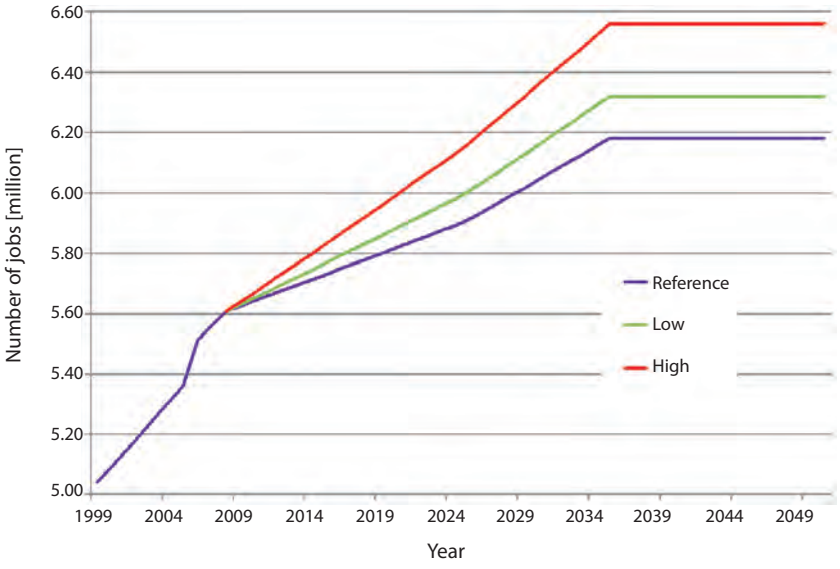


Figure 20.10 Total number of jobs for period 1999–2050 (Source: Census, authors' computations).

Although we are aware that the distribution of employment across activity sectors evolves across time, we do not have access to relevant data to measure this evolution in the calibration period 1999–2008, and extrapolate it in the projection period. This is partly due to a significant change in the typology of jobs used to provide employment data, which makes is difficult to follow sectoral distribution of employment over time. We therefore have to rely on simplifying hypotheses, and assumed that the sectoral distribution of employment aggregated at the regional level is constant over time.

Transportation supply

The improvement of the public transportation network will directly reduce public transportation travel times on the lines concerned. This will have different indirect effects: some individuals will change their mode from car to public transit as a response to improvements in public transportation infrastructure, which will reduce congestion on the roads; other individuals will possibly change their mode from public transit to car as a response to congestion reduction (this second effect should be small compared to the first one); some individuals will change residential and/or job location as a response to accessibility changes resulting directly from the infrastructure and indirectly from changes in mode choice. This will induce a significant change in the O-D matrix of flows and, consequently, in the O-D matrix of travel times. METROPOLIS computes the equilibrium resulting from such interactions, and produces new O-D matrices of travel times by private car.

O-D matrices of travel times and flows for public transit are computed by *MODUS* at key dates, for each scenario, as indicated in Table 20.13, which also presents the assumptions used to extrapolate these matrices between the key dates. These matrices were obtained under assumptions about household and job location which may a priori not be consistent with our own results. However, the differences in the distribution of household and job locations between the data provided by *DRIEA* and the output of METROPOLIS are not very large, so such potential inconsistencies are limited.

In combination with the population and employment scenarios defined in Section 20.9.2, we consider two different transport scenarios: the *Reference* scenario does not benefit from the Grand Paris Project infrastructure, whereas the *Grand Paris* (or *GP*) scenario does. In the results section, the *Reference* scenario will correspond to the reference both for population, employment and transportation; the *Ref.-GP* scenario will correspond to the *Reference* scenario for population and employment, but to the *Grand Paris* scenario for transportation, whereas only the *Grand Paris* transportation scenario will be considered in the (population and employment) *Low* and *High* scenarios.

For the simulations with METROPOLIS, the population and employment data produced by UrbanSim and introduced in METROPOLIS via interface.

Year	Ref: Public	GP: Public
1999	*MODUS* 2004	*MODUS* 2004
2003		
2004		
2005	*MODUS* 2005	*MODUS* 2005
...		
2020	*IAU-SDRIF* 2025	GPX-REF 2025
...		Evolution at a constant rate
2025		GPA 2025
...	Evolution at a constant rate	
2035	*IAU-SDRIF*2035	GPX 2035
...		
2050		

Table 20.13 Data sources at key dates and assumptions for extrapolating O-D matrices.

The road network characteristics are updated according to the put in operation planning of different projects.

O-D matrices are at the *MODUS zones* level (1289 zones). In the dense part of the Paris area, *MODUS zones* mainly correspond to IRIS zones or groups of several IRIS zones included in the same *communes*, whereas *MODUS zones* mainly correspond to *communes* or groups of *communes* in the less dense part of the Paris area. We built a correspondence between the 1289 *MODUS zones* and the 1300 *communes* of the Paris area. We used this correspondence to build O-D matrices of flows and travel times at the *commune* level. Matrices of flows, travel times, and distances correspond to weighted average values of the same variables, averaged over all the itineraries between the relevant zones. The weights correspond to the share of the itineraries in the relevant O-D pair.

20.9.4 Results

Evolution of the population

The aggregate population levels were given in the description of the scenarios in Section 20.9.2. Table 20.14 presents the population increase between 2005 and 2035 in the different scenarios, in Paris, in each CDT, and in the rest of the region. Figure 20.11 represents the evolution across time of the fraction of the cumulated increase in population (compared to 2005) which goes to the historical Paris, to the secondary centers (the *CDTs*, which are mainly located in the close suburbs and in some relatively dense parts of the far away suburbs), and to the core of the region, which is made of Paris plus CDTs.

Horizon	2035				Input data,
Scenario	*Reference*	*Ref.-GP*	*Low*	*High*	*High*
Zone					
Region	1 618 446	1 618 446	1 687 573	1 800 014	1 800 003
Paris +CDTs	**623 846**	**861 361**	**954 312**	**1 009 113**	**1 186 986**
Rest of the region	994 600	757 085	733 261	790 901	613 017
Paris	**93 229**	**89 555**	**89 586**	**90 661**	**141 042**
Aulnay-Montfermeil	23 780	37 929	47 548	58 101	96 084
Biotechnologies S.-A.	98 645	154 814	175 079	184 985	206 145
Confluence	65 128	52 653	51 075	54 058	49 525
Descartes	49 337	67 795	77 875	86 034	104 132
La Défense	56 339	71 423	78 514	83 531	98 487
Le Bourget	18 778	71 964	87 449	89 014	102 000
Pleyel	72 249	112 897	123 706	128 930	102 666
Roissy-pôle	28 965	29 409	30 560	32 291	58 033
Saclay	100 040	156 522	175 761	182 764	223 313
Val de France - Gonesse	17 356	16 400	17 159	18 744	5 559

Table 20.14 The evolution of population between 2005 and 2035 (Source: simulation results).

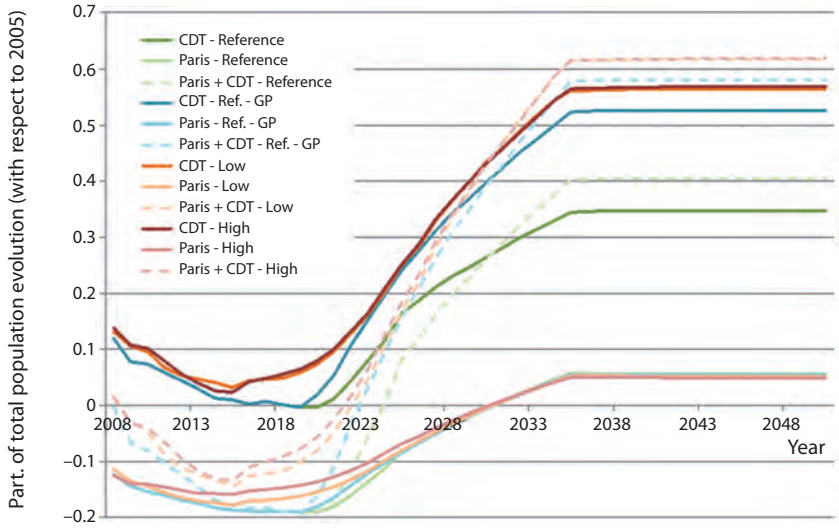


Figure 20.11 Evolution of the population for period 1999 to 2050 (Source: simulation results).

The negative figures for Paris until 2030 means that, whatever the scenario, population inside Paris will decrease, whereas the regional population will increase. Such a decrease was already observed in 2008 according to census data. In 2008, the decrease in the share of Paris in regional population growth was exactly compensated by an increase in the CDTs, so that the population in the core of Paris area is the same in 2008 as it was in 2005. The population decrease in Paris results from the conjunction of two phenomena: On the one side, strong capacity constraints inside Paris making it very difficult to build new dwellings; on the other side, the decreasing trend in household size implies a decrease in population when the number of households remains constant. Our simulation results suggest that the share of Paris in total population will remain stable until 2020, and will then increase until 2035, with small differences between scenarios.

Our results also suggest a decrease in the fraction of the additional population which goes to CDTs from 2008 to 2015 in all the scenarios. This reflects and extrapolates the current trend of the population to move from the core of the city to the far away suburbs. This is because the CDTs currently suffer both from high real estate prices and relatively poor accessibility. From 2015 to about 2020, we observe a very small increase in the share of CDTs in the *High* and *Low* scenarios, and a stabilization of this share in the in the *Reference* scenario, either with or without the *Grand Paris* infrastructure. In the *High* and *Low* scenarios, population growth in the CDTs will anticipate the Grand Paris transportation infrastructure and start locating in the places which will become more accessible a few years before accessibility actually improves. This can be explained by urban policies aimed at favoring housing construction, which will start several years before the transportation infrastructure actually works. This phenomenon starts later (only in 2019, which is only 1 or 2 years anticipation) in the Reference scenario for population and employment, with the Grand Paris infrastructure, because urban policies are less needed and will therefore be less active in this scenario.

In the long run, the CDTs attract 56% of the increase in population in the Low and High scenarios, and 52% in the Reference scenario with Grand Paris infrastructure, whereas this fraction is only 35% in the Reference scenario. The improvement of accessibility and urban policies favoring the construction of dwellings in the CDTs attract the majority of the population in the CDTs.

This fraction goes up to 61% for the core of the region in the Low and High scenarios, 58% in the Reference scenario with Grand Paris infrastructure, and 40% in the Reference scenario.

In the intermediary period 2020–2035, the rhythm of population growth in the CDT relatively to the rest of the region significantly varies across scenarios. In the Reference scenario with Grand Paris infrastructure, the relative growth started later than in the Low and High scenarios (2019 instead of 2015) because

total population growth is less important, so there is less need to anticipate the construction of the infrastructure. Relative population growth rate in the Reference scenario with Grand Paris infrastructure then rapidly catches up the gap compared to the Low and High scenarios between 2019 and 2023. The infrastructure is anticipated later when total population is smaller, but the difference disappears just before the first phase of the project ends. The relative evolution of population growth is then the same in the three scenarios with Grand Paris infrastructure until 2026, which means that it does not depend on the size of the aggregate population in the region during this period. Finally, the relative share of the population in the CDTs grows faster in the period between the two phases of the construction of the infrastructure (2026–2035) in the Low and High scenarios than in the Reference scenario with Grand Paris infrastructure. This may indicate that the second phase of the project is needed to attract the additional aggregate population assumed to come to the region in the Low and High scenarios compared to the Reference scenario. Without the second phase of the project, this additional population would induce too much congestion in the CDTs if it attracted the same share of the growth in Low and High scenarios than in the Reference scenario. Stated differently, the first phase of the project is sufficient to attract population in the CDTs in the reference population scenario, whereas the second phase is needed in the Low and High scenarios.

Evolution of employment

The aggregate employment levels were given in the description of the scenarios in Section 20.9.2. Table 20.15 presents the employment increase between 2005 and 2035 in the different scenarios, in Paris, in each CDT, and in the rest of the region. Figure 20.12 represents the evolution across time of the fraction of the cumulated increase in employment (compared to 2005) which goes to the historical center, to the secondary centers (the *CDTs*), and to the core of the region.

The evolution of employment in Paris is subject to capacity constraints similar to those applying to the population, but the trends in the needs are totally different. Whereas each individual tends to require more and more dwelling capacity (because of the decreasing trend in household size), each job requires less floor space because of the decreasing trend in surface per worker. This decreasing trend allows the number of jobs to increase inside Paris from the beginning of the simulation period although the floor space devoted to employment does not increase before year 2020, except for some projects already planned, and introduced exogenously in the system. Capacity constraints and the decreasing trend in the surface per worker are the main forces driving the evolution of the number of jobs, which leaves no room for economic considerations.

Horizon	2035				Input data,
Scenario	*Reference*	*Ref.-GP*	*Low*	*High*	*High*
Zone					
Region	820 245	820 245	959 995	1 200 008	1 199 999
Paris +CDT	**495 723**	**650 979**	**776 876**	**927 578**	**1 010 657**
Rest of the region	324 522	169 266	183 119	272 430	189 342
Paris	**112 138**	**91 678**	**90 575**	**94 430**	**74 339**
Aulnay-Montfermeil	2 381	14 552	19 021	21 260	23 019
Biotechnologies S.-A.	75 856	69 868	90 885	128 976	144 836
Confluence	28 296	26 446	27 116	28 931	17 885
Descartes	39 719	60 782	69 696	79 001	84 246
La Défense	96 097	128 586	139 625	143 918	146 794
Le Bourget	15 092	33 629	42 416	49 640	60 000
Pleyel	55 980	75 183	85 043	93 163	115 489
Roissy-pôle	29 014	53 369	84 130	126 819	170 841
Saclay	35 187	94 820	121 294	144 125	159 019
Val de France - Gonesse	5 963	2 064	7 075	17 315	14 189

Table 20.15 Evolution of the number of jobs between 2005 and 2035 (Source: simulation results).

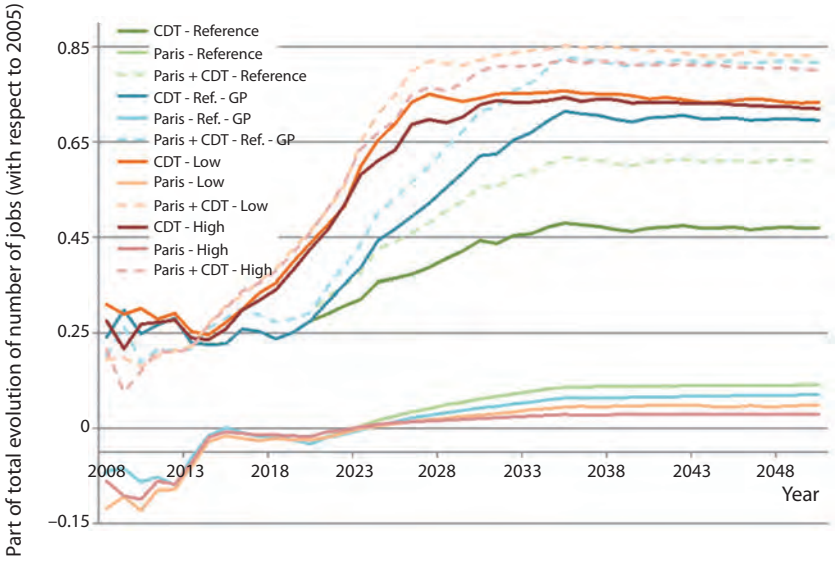


Figure 20.12 Evolution of the number of jobs for period 1999–2050 (Source: simulation results).

Strong capacity constraints in Paris imply that the jobs which locate inside Paris are those that can do so, not all that want to do so, which are more numerous in all the scenarios with the Grand Paris infrastructure. When comparing High, Low, and Reference scenarios with Grand Paris project, the share of Paris in total employment decreases when the size of total regional employment increases, but the number of jobs located inside Paris in these three scenarios does not differ significantly. The release in capacity constraints inside Paris allowed by the decrease in floor space needed per job explains that the share of the CDTs slightly decreases between 2035 and 2050 to the benefit of Paris, the share of the core remaining constant over this period.

The capacity constraints mentioned above were short run constraints, i.e. without changes in land use type. In the Reference scenario without the Grand Paris project, the demand for jobs to locate inside Paris is significantly larger than in the Reference scenario with the Grand Paris project. This is because, when the infrastructure is built but the aggregate regional employment is the one of the Reference scenario, improvements of accessibility will attract many jobs to the CDTs and decrease the demand for jobs to locate inside Paris. This excess demand for jobs inside Paris in the Reference scenario will be so large that it will move upwards the short run capacity constraints: The economic pressure will be so strong in the Reference scenario that it will force some changes in land use type and construction of additional office buildings. As a consequence, the number of jobs located in Paris, and the relative share of Paris in the increase of regional employment are larger in the Reference scenario without Grand Paris infrastructure than with this infrastructure.

Compared to the population, the final share of employment which goes to the CDTs and to the core of Paris area, and the pace at which these shares evolve across time slightly differ between Low and High scenarios. The total number of additional people working in CDTs or in the core are larger in the High scenario than in the Low scenario, but the share of new jobs going to CDTs or to the core is smaller (and its evolution is slower between 2022 and 2030) in the High scenario than in the Low scenario. This suggests that more jobs could be attracted by the CDTs and by the core of the region if an even more ambitious infrastructure was built in phase 2. Stated differently, the improvement in accessibility is not enough to attract the full potential of jobs which could be attracted by CDTs and the core of Paris area in the High scenario.

Nevertheless, more than 70% of the additional jobs will be attracted by the CDTs (more than 80% by the core of Paris area) if the Grand Paris infrastructure is built, whereas CDTs would attract only 47% (and the core 60%) of the new jobs in the Reference scenario without Grand Paris infrastructure.

Evolution of inequalities within the region

Based on UrbanSim simulation results, we analyzed the evolution of of inequalities, expressed in terms of income per capita and in demographic terms. Since the household income is not observed directly in census data, we predicted it, using the French Family Budget Survey (Cérani & Camus 2004). Income is predicted as a function of household characteristics such as size (by age group), number of working members, age, activity, diploma, and nationality of household head. In order to take into account economies of scale, the household income is divided by the square root of the number of household members, which is an equivalence scale commonly used. Without any equivalence scale, Paris would appear poor compared to the rest of the region just because households are smaller in Paris than in the rest of the region, and more generally, the apparent ranking of *commune* or *départements* by income level would be inconsistent. This problem is solved by our equivalence scale, and the ranking of *commune* or *départements* by income level is totally consistent.

An equivalence scale is intended to reflect the way the needs vary with household size. With this equivalence scale, the economic well-being of a household with one member and 2 000 euro per month is the same as the economic well-being of a household with 4 members and 4 000 euro per month. The distribution of income per capita is assumed constant over the period, at the region level. As a consequence, the income considered here has to be interpreted as a relative income (relative to the average regional level). This normalization has no consequences on our simulation results. We simply assume that the distribution of the population over the region would be the same if the income of each household increased by 10%. This assumption does not seem really restrictive.

Income per capita is divided in 3 classes (rich, middle income, and poor), built so that each of them represents about a third of the sample (actually, there are 30.7% rich households, 33.1% poor households and 36.2% middle income households in the region). There is a strong tendency to the increase of intra-regional inequalities over time, whatever the scenario.

Paris tends to become poorer as the richest households will tend to move to the suburbs, whereas the poorest ones will tend to stay inside Paris. As a result, in all the scenarios, there will be more than 41% poor households and less than 24% rich households inside Paris in 2035, whereas there were only 36.7% poor households and as much as 31.5% rich households inside Paris in 1999. The change in the income distribution is less severe in the other zones, although it systematically goes in the direction of an increase of inequalities over time. Outside Paris, the richest zones tend to become even richer, and the poorest zones tend to become poorer.

Variable	Percentage Rich hhs			Percentage Poor hhs			Poor-to-Rich ratio		
Year	2035			2035			2035		
Scenario	Ref.	Low	High	Ref.	Low	High	Ref.	Low	High
Other *communes*	34.7	34.7	34.5	28.7	28.6	28.6	0.83	0.82	0.83
Paris	23.7	23.8	23.7	41.5	41.4	41.4	1.75	1.74	1.75
Aulnay-Montfermeil	19.4	19.0	19.5	40.6	41.1	41.4	2.09	2.17	2.13
Biotechnologies S.-A.	28.6	30.5	30.6	33.1	31.8	31.7	1.16	1.04	1.04
Confluence	36.5	35.9	36.8	27.1	27.2	26.6	0.74	0.76	0.72
Descartes	33.0	33.2	34.2	27.8	27.6	27.3	0.84	0.83	0.80
La Défense	37.0	37.6	37.5	26.2	25.9	26.0	0.71	0.69	0.69
Le Bourget	12.2	12.4	14.4	55.6	54.5	53.4	4.56	4.40	3.72
Pleyel	9.6	9.8	9.9	60.3	60.0	60.1	6.31	6.12	6.05
Roissy-pôle	26.9	28.8	29.5	32.6	30.7	31.1	1.21	1.07	1.05
Saclay	42.7	44.1	43.5	22.0	21.3	21.6	0.52	0.48	0.50
Val de France - Gonesse	15.9	15.6	16.0	45.5	45.9	45.9	2.86	2.95	2.87
Region	30.7	30.7	30.7	33.1	33.1	33.1	1.08	1.08	1.08

Table 20.16 Distribution of income and Poor-to-Rich ratio (Source: simulation results).

We computed the Poor-to-Rich ratio as an indicator of inequalities, and we represent this ratio in Table 20.16 for each zone in each scenario.

Overall, the Grand Paris project marginally reduces intra-regional inequalities in income compared to the Reference scenario. Income distribution and differences in poverty are not significantly affected by it, although there are large differences between zones in all the scenarios. The poorest zones such as Pleyel or Le Bourget remain the poorest ones in the region, but the fraction of poor households is marginally reduced in those CDTs.

20.10 Conclusions

The Paris case study in SustainCity has benefited from several years of research in UrbanSim application in *Ile-de-France*. The data have been gathered over several years and they were updated and enriched with new sources within SustainCity framework. As urban modeling and in particular microsimulation urban models need a huge volume of data that often are not completely available, the model structure and modeling strategy is driven by available data. On the other hand, more data with a better quality become available and impose us to update the models with new possibilities.

The most important task, once data have been collected is the calibration task. The UrbanSim modules have been calibrated separately against observed data in particular the household and individual population data and employment data by activity sector. To present the results and calibrate the model, we have used an aggregated zoning with 50 zones over the region. These zones are designed to be compatible with urban policy zones and also transportation analysis zones. The compatibility with the urban policy zones allowed us to introduce some calibration constants that represent the effects of policies that are not explicitly modeled. The UrbanSim and integrated model results have been calibrated against the census data for the period 2006–2008 for population and employment. We used UrbanSim in order to evaluate the local effect of the Grand Paris project on the location of jobs and population within the region, and on real estate prices. UrbanSim is indeed very efficient in taking into account the heterogeneity of population and firms' responses to large infrastructure projects, and in reproducing the dynamics of the evolution induced by such large projects. More generally, microsimulation models like UrbanSim are well suited for evaluating the intra-regional micro-economic effects of large infrastructure projects, whereas general equilibrium models are well suited for evaluating the inter-regional macro-economic effects. These two types of models were successfully combined for evaluating the effects of the Grand Paris project. The results of calibration procedure presented in Section 20.8, show that the fit with collected data is very good. The forecasts we have made (and discussed in Section 20.9) were presented to the *Société du Grand Paris*, who is currently using this integrated model. Online reports provide the results which are often publicly challenged during animated debates. This is, after all, the ultimate scope of large scale models: to challenge intuition and occasionally, present new results and intuition.

In SustainCity, there are two other case studies: Brussels and Zürich. These three case studies are very different since they focus on different cities and are based on different sub-models (residential location models are different for example), market clearing in Paris and Zürich are also different, and the like.

Of course, a model is never finished. One of the reasons, is that any perfect model is likely to become poor after a few years because the behaviors change, and because the technology changes. Stability of behavior is challenged by various economists who advocate the study of endogenous preferences.

In fact, the pure cost-benefit analysis of a project such as the Grand Paris project, must be on a minimum of 50 years, or more. We may encounter qualitative changes, which are very difficult to forecast. For example, how long the 2008 crisis will last in France? Another implicit assumption of stationarity made in our approach is related to the technology. For example, the electric car that can change mobility, but we cannot ignore other forms of or-

ganization of mobility. We list those that seem most significant: Car-sharing, collective taxis, transport exploiting synergies between individual and collective transport, modular toll, deregulation of taxis, teleworking, teleeducation, etc. Obviously, these management methods will play a major role in mobility and in particular in residential location, in location of business, in fixing the price of land, and the use of space. Researchers still have serious work ahead of them.

Acknowledgments

Part of this research also benefitted from the financial support of Labex MME-DII (ANR11-LBX-0023-01), Ademe and DRI (Direction de la Recherche et de l'Innovation du ministère chargé du développement durable) as part of the French national research programme in land transportation PREDIT, ANR project ELITISME and *Société de Grand Paris*.

Chapter 21

Land use and transport microsimulation in the canton of Zürich using UrbanSim

Patrick M. Schirmer, Christof Zöllig Renner, Kirill Müller, and Kay W. Axhausen

21.1 Study area and setup

21.1.1 The Zürich case study

The case study described here is part of our ongoing work on the usability and extension of agent-based land use simulations in the European context. The IVT has set up a case study of UrbanSim for the Canton of Zürich.

The Zürich case study is running the latest version of the land use simulation UrbanSim in interaction with the transport simulation MATSim, the latter being developed at IVT in cooperation with TU Berlin. While the other case studies of the overall SustainCity project implemented the zonal version of UrbanSim, the availability of data on a very disaggregated level allowed the implementation of the parcel based version of UrbanSim.

The case study took advantage of a previous implementation of UrbanSim in the project *Zukunft urbaner Kulturlandschaften* (Loechl et al. 2007). Although that project used the gridcell version of UrbanSim, various findings could be reused here, e.g. probability rates and control totals. The decision to create a new implementation reflected the limitations found in the earlier project using the gridcell based version, as well as the different extent of the study area.

After having calibrated the simulation, three scenarios were tested and evaluated for their spatial effects: road-pricing (Sect. 21.4.2), densification of center areas (Sect. 21.4.3), and the combination of both (Sect. 21.4.4).

21.1.2 Study area and simulation period

The European research program ESPON 2006 defines the metropolitan region of Zürich as "one of 16 European motors having a leading role in economy, innovation, transportation, controlling and decision taking" (Metropolitanraum 2013). This metropolitan region contains 238 municipalities across eight cantons, having a total of 1.7m inhabitants and 820 000 jobs, with the City of Zürich as economic center (Metropolitanraum 2013).

The canton of Zürich (Fig. 21.1) has its administrative border inside this metropolitan region and covers a large part it, including the two major cities Zürich (375 000 inhabitants in 2011) and Winterthur (107 000 inhabitants in 2012). In recent years the population in the canton has grown from 1.2m (2000) to 1.4m (2010) inhabitants and is expected to grow further to 1.6m inhabitants in 2030 (Cantonal Statistical Office of Zürich 2013). The number of jobs can be estimated to increase from 725 000 to 975 000 by extrapolating sectoral trends between 1996 and 2001. This makes the Agglomeration of Zürich one of the fastest growing regions, not only in the Swiss context but also on the European scale.

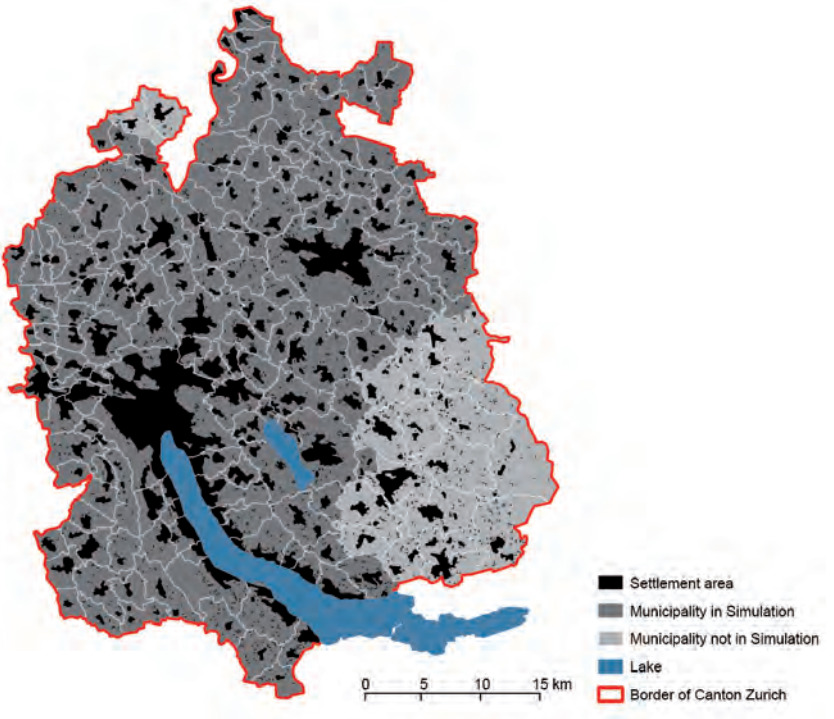


Figure 21.1 Urban areas in the canton of Zürich (©2011 Vector200 Swisstopo (JA100120)).

These developments result in vast changes of the urban structure and the dynamics taking place in the urban landscape. Due to the complexity and interactions taking place, the outcome of such developments are difficult if not impossible to foresee unaided. The use of an agent-based land use transport interaction model can help to understand these interactions. Through modeling the behavior of the individuals and their reaction to the urban space not only development trends can be simulated, but also alternative scenarios can be evaluated for their effects.

Due to the heterogeneity in data structures among the neighboring cantons we had to limit the study area to the administrative border of the canton of Zürich. Even within the canton data was not available on an homogeneous level at the time of this study. This is especially true for the cadastral information needed for the parcel based version of UrbanSim (see Sect. 21.1.3). The sparsely populated or uninhabited areas in the south-east of the canton could not be included into the case study for this reason, but are also of reduced influence to the studied area.

The different datasets had to be processed and combined to create a base year representation for the simulation. This process, which is described in Section 21.2, showed further inconsistencies in the data.

The period of simulation is 2000 to 2030, while 2010 is used for validation. Besides the current administrative permissions for construction, two scenarios have been tested that represent selective measures in form of road pricing and a change of zoning definitions for densification of center areas (Sect. 21.4.2 to 21.4.4).

21.1.3 Primary spatial unit

Each UrbanSim model defines a primary spatial unit, i.e., the only location information available within UrbanSim for the agents (persons, households, jobs, etc.). UrbanSim supports zones, grid cells and parcels as primary spatial units. The Zürich case study simulates at the parcel level, the most disaggregated level of spatial resolution available within UrbanSim. Both the Brussels and the Paris case study simulate at the zone level, see Chapters 19 and 20 for more detail.

An obvious advantage of a parcel-level model is the enhanced spatial resolution leading to a more precise placement of the agents. Additionally, as ownership and regulation of construction and density is defined at the parcel level, simulating development becomes more straightforward with a parcel-level model. The zone model combines several parcels into one zone, so that ownership and development cannot be simulated easily. Grid cells have uniform size which can be very different from the size of a parcel; in addition to the problems of the zonal model, some buildings cannot be mapped uniquely to a single grid cell.

On the other hand, detailed information at parcel level is more difficult, and sometimes impossible, to obtain. For the Zürich case study, for each high-resolution data set, access had to be negotiated individually, and a dedicated non-disclosure agreement had to be signed. In addition, detailed data comes with a greater potential for inconsistencies. This has substantially increased the effort required for collecting and preparing the data (cf.Sect. 21.2, see also Chap. 4 for more detail on spatial issues).

21.1.4 UrbanSim

The UrbanSim model system simulates each year in a given range (2000 to 2030 for the Zürich case study). For each year to be simulated, the same set of models is run in a sequence, updating the simulation data. Interaction between the models is defined only by the sequence in which they are run and the resulting changes in the simulation data. Each year is simulated only once, and in each year each model runs only once. No equilibrium of any kind is computed. The only notable exception is the transportation model (cf. Sect. 21.1.5); however, the equilibrium is computed in isolation from the other models in UrbanSim. For further details on the simulation technique please refer to the publications of the developers (Waddell et al. 2003).

The initial state of the simulation data is the result of the data preparation (cf. Sect. 21.2.2), and the final state of each year's simulation is saved for further analysis. Furthermore, scenarios can be defined by predefining part of the simulation data for a specific year in addition to the base year. This is used for the densification scenario presented in Section 21.4.3: the regulations on the floor area ratio are changed in the year 2015, which is specified by providing altered definitions for the parcel dataset from this year onwards.

The project followed the development of UrbanSim by continuously using and integrating the latest source code and not a stable release (currently version 4.3). Several enhancements to the common code base, mostly concerning the graphical user interface (GUI) and usability, have been contributed to the UrbanSim code base by the Zürich-based team.

21.1.5 Transport model

Land use models and transportation models operate on very different time scales: years or decades vs. hours or days. The former simulate a development over a long period of time, while the latter usually only capture a momentary state of the transportation system and not its development. In an integrated land use and transportation simulation, the transportation model usually constitutes a part of, and is controlled by, the land use model.

The UrbanSim model framework can optionally invoke an external transportation model to update accessibility and travel time information. For the

Zürich case study, MATSim (Balmer 2007, Meister et al. 2009, Balmer et al. 2008, MATSim 2013) has been integrated in the model. MATSim is an agent-based transportation microsimulation framework that simulates the movement of individuals in the transportation network. This allows a one-to-one mapping between the agents and facilities in both models, and a direct computation of travel time within the transportation model. See also Chapter 16 for a more detailed description of MATSim.

Both models operate in different computing environments (Python vs. Java) and require differently formatted input data. The technical part of the integration has been provided by TUB (cf. Chap. 16). Most MATSim parameters can be specified directly within the UrbanSim GUI, these parameters are seamlessly translated into a MATSim configuration. In addition, the integration layer offers full control by supporting an additional MATSim configuration for each simulation year. This allows to specify scenarios such as the cordon toll (cf. Sect. 21.4.2).

MATSim estimates a stochastic user equilibrium for the utilisation of the transportation infrastructure. It proceeds in iterations. At the beginning of each iteration, each agent has a schedule for its activities and trips during that day. All plans are executed in simulated real time to obtain a utility-based score for each agent. After that, a fraction of the agents are allowed to replan their schedule or to choose one of their previously used schedules before the next iteration starts. The process is presented in detail in Chapter 17.

21.2 Data processing and data model

21.2.1 Data model for simulation

The Zürich case study uses the parcel version of UrbanSim, while the preceding project Zukunft urbaner Kulturlandschaften (ZUK, 2007) used the gridcell based version for the same study area. These versions allow the use of the same modelling methodologies for simulation, but differ significantly in their data model, which has an impact on the kind of variables available for integration.

The gridcell version can be assumed to allow easier set up in the Swiss context, as most census data are available for a hectare grid. Its disadvantage is that buildings and zoning definitions often cut across multiple gridcells or change within gridcells and thus demand unsatisfying workarounds in the data preparation (Loechl et al. 2007). The zonal version of UrbanSim avoids this limitation and allows the use of irregular shapes as spatial representation. However, a zonal representation has limitations when different zoning definitions stand in conflict to each other, e.g. when traffic analysis zones do not match with construction zones. For further aspects on this issues see Chapter 13.

The parcel version of UrbanSim goes a step further and disaggregates each zone into single parcels, assuming that these are subject to regulations and

ownerships. Agents (jobs and persons) are located in buildings, which are situated on parcels. The parcels can be aggregated to different entities as zones, cities or counties. In our study we additionally disaggregate the buildings into individual residential units to allow the integration of dwelling characteristics into the household location choice and real estate models (see Sect. 21.3.6 and 21.3.5). Figure 21.2 gives an overview of the underlying data structures.

The drawback of such as disaggregated model is that the set up of the simulation data can create substantial additional work, while aggregated data is often available on a homogeneous level from the official statistic departments. Here we will explain the approach we took to put this into practice. A more detailed description of this process has been reported in (Schirmer et al. 2011).

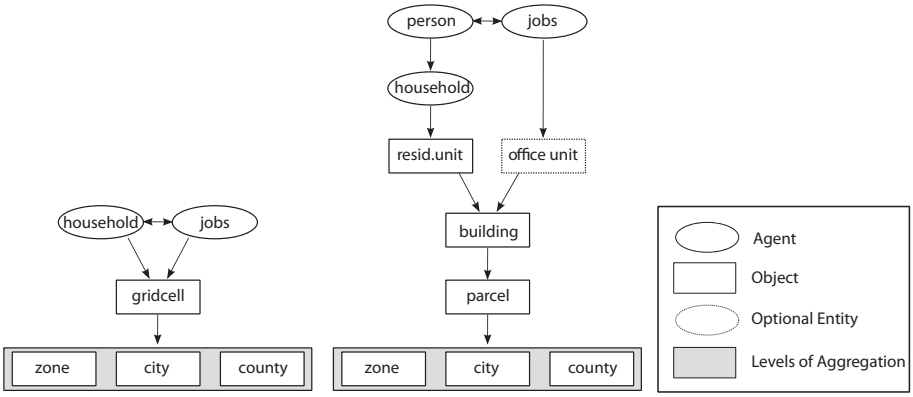
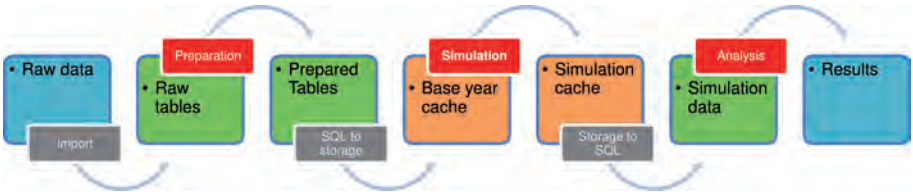


Figure 21.2 Structure of the datasets in UrbanSim. A simplified gridcell version (left), and the parcel version as used in the Zürich case study (right).

21.2.2 Data processing infrastructure

The data is delivered as one or more files, usually with accompanying metadata and documentation. Each dataset is stored unchanged in the project's Subversion server. From there, a collection of scripts automatically converts and prepares the data so that it fits the requirements of the UrbanSim model. This follows the principle of reproducible research, i.e. at any stage of the project, all required data can be recreated from scratch. Deficiencies in data preparation process can be fixed at the appropriate location instead of having to apply error-prone manual patching. In addition, consistent test data for a small part of the study area can be created easily; this test data is used for rapid prototyping and interactive development of models which would take prohibitively long when using the complete simulation data. Figure 21.3 outlines the creation of the UrbanSim base year data from the different data sources.

The original data sources are provided as ESRI shapefiles, CSV files, Excel spreadsheets, or INTERLIS files, among others. During the first stage of



The colors in this process diagram indicate the kind of process (gray: copying, red: processing) and the storage location (orange: UrbanSim, green: PostgreSQL, blue: file system).

Figure 21.3 Structure of the data processing.

the data processing, all datasets are copied into one schema of a PostgreSQL database. The purpose of this copy procedure is to create a faithful copy of the raw data in a common format. No other processing happens at this stage, except for adding an identically named municipality identifier to all imported tables. This first import stage requires just under one day, the bottleneck is the import of one large INTERLIS data set. Fortunately, this stage needs to be run only rarely, as the results of this stage are treated as read-only by the subsequent preparation stage.

In the subsequent processing stage, custom scripts written in SQL, R, and bash, access the imported data, convert it to the format required by UrbanSim and store the converted data to the same Postgres database mentioned above. The data preparation includes simple techniques such as vertical editing (recoding of attributes, aggregation), horizontal editing (filtering, imputation), as well as more sophisticated ones such as deterministic joins on categorical or spatial attributes, record linkage, and stochastic disaggregation and distribution (cf. Sect. 21.2.3). The overall performance is controlled by assertions – built-in conditions that are checked during execution of the scripts and alert in case of potential inconsistencies. For all generated datasets, the percentage of records in the original data that had to be dropped during the preparation is logged (cf. Fig. 21.4).

Finally, the prepared data is faithfully converted to the UrbanSim cache format. Conversely, the results of simulation runs are imported back to the Postgres database and analysed using another set of scripts.

21.2.3 Data flow

Most of the processing is carried out municipality by municipality, in parallel. Exceptions include the allocation of persons to jobs (see below), this happens in one step after all municipalities have been processed.

Table 21.1 shows the principal data sources and the year of representation. Abbreviations will be used in the description of the data processing. As already

Abbrev	Description	Year	Entities
AV	Cadastral information	2010	Parcels, buildings
GWR	Register of buildings and dwellings	2010	Buildings
GVZ	Building insurance data	2000	Buildings
VZ	Population census	2000	Persons, households
MZ	Transportation microcensus	2005	Persons, households
BZ	Enterprise census	2001	Jobs

Table 21.1 Principal data sources used for preparing the base year data.

described in our previous work (Schirmer et al. 2011) the AV and GWR datasets representing the year 2000 have such a low quality that we decided to use the 2010 data and remove all new constructions.

Municipalities, zones, and parcels After initializing the target database schema, spatial entities are imported first, ordered by increasing level of detail. For shapes of municipalities and parcels (entities *city* and *parcel* in Figure 21.2), data from the AV has been used. The cantonal transportation model provides the zoning (*zone* entity) as an intermediate aggregation level for analyzing the data. There is only a dummy entry in the *county* dataset.

In addition, the parcel data has been enriched with information computed in previous studies (Loechl et al. 2007). The parcel data covers the entire study area without gap, and those parcels that correspond to the road network are identified in advance. Unfortunately, the original datasets are not perfectly consistent, and a nearest-neighbor match is required for establishing the spatial hierarchy.

Buildings and dwellings In the data model a building is an entity with one or more dwellings.[57] These entities constitute the most difficult part of the data preparation. Three data sources are used to compose the simulation data for buildings: the AV provides the shape of the buildings, the GWR contains information on dwellings, and the GVZ defines the improvement value of a building. In a preprocessing step, the number of rooms is imputed for those dwellings where this information is missing. The non-residential area of a building is not available in the data and was estimated from the floor area, the number of stories and the (known) residential area. Rent prices are added to a part of the building data based on observations of real estate advertisements from the Comparis website, which itself aggregates advertisements from different sources. This forms the basis data for the real estate price model (Section 21.3.5).

[57] A dwelling is a general term like living unit. It comprises as well single family homes. An apartment is a part of a building and thus more specific.

A spatial match establishes the relationship between AV buildings and parcels. The GVZ dataset is matched to the AV dataset by insurance ID, and the GWR dataset is matched partially by federal building ID and partially by a greedy nearest neighbor match with a uniqueness constraint. Those buildings that were constructed after the simulation's base year are separated and used as a pool for newly constructed buildings (Sect. 21.3.4).

Jobs The data on workplaces is derived from the Swiss enterprise census (Swiss Federal Statistical Office (BfS) 2008). Unfortunately, this dataset suffers from the headquarters reporting problem – for large businesses, only the headquarters are listed with the total number of employees (up to several thousands), instead of listing all branches and their location. This leads to a severe misfit between the enterprise and the building data, which was corrected by a greedy nearest neighbor match with capacity constraints. The space occupied by a job was assumed to be a constant between 35 and 80 square metres, depending on a nine-level classification of the employment sectors based on the NOGA code (Swiss Federal Statistical Office (BfS) 2011). These rates have been defined based on architectural expert knowledge. The use of a building depends on the enterprises that have been assigned to that building. The sectors, along with abbreviations used throughout this chapter, are shown in Table 21.2.

Sector	Abbr	Sector	Abbr	Sector	Abbr
Manufacturing	Mfg	Retail	Rtl	Service	Srv
Construction	Cns	Hotels, restaurants	HR	Health	Hlt
Trade	Trd	Transport	Trn	Other	—

Table 21.2 Employment sectors.

There is a mismatch between the total available space and the total space allocated to jobs based on the space per job rates mentioned above, leading to a severe underutilisation of the buildings. Options to handle this include increasing the rates per job (leading to a larger area of redistribution of the jobs) or uniformly downscaling the area available in the buildings (affecting density and hence land use development). As a compromise, only the space per job for the employment sector "others" is adapted to fit the available space in the buildings, and jobs from this sector are not added or relocated during the simulation.

Households and persons Households and persons were derived from the full population census (Swiss Federal Statistical Office (BFS) 2000*a*). In the data available for the project, location was precise to the hectare, so a greedy nearest neighbour match with capacity constraints has been employed to map

households to apartments. Income and car ownership are imputed based on ad-hoc models derived from the transportation microcensus (Swiss Federal Statistical Office (BfS) 2005).

Allocation of persons to jobs For the workplace facility of a person, the population census lists the municipality, but lacks more precise location information. Assignment of a workplace to each person is carried out by randomly matching the jobs of a municipality to persons with workplace in that municipality. Commuting to or from other cantons is not supported; for census persons with a workplace outside the Canton of Zürich, a workplace in the canton is chosen at random from the remaining workplaces during the first simulated year.

Control totals and constraints This data is mostly static, apart from the employment control totals which are given at the canton level and have to be rescaled if only a part of the study area is prepared for testing purposes.

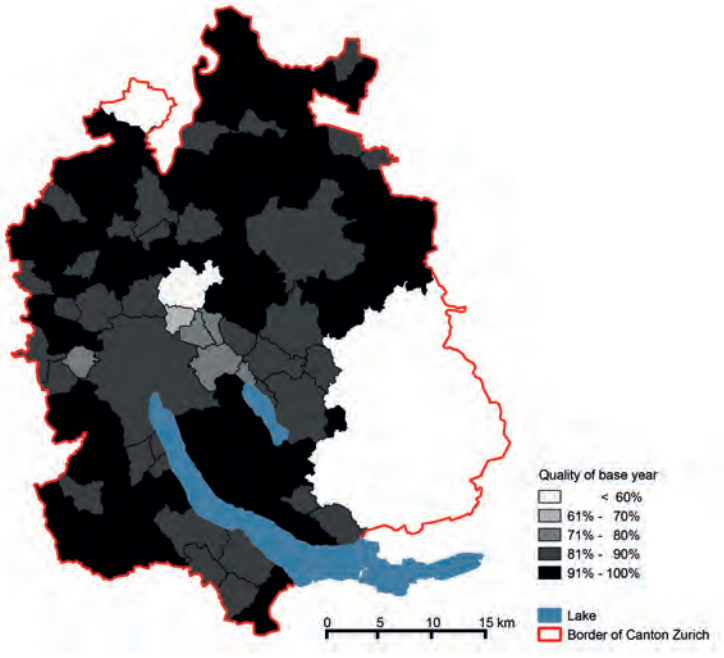
21.2.4 Quality of base year

Figure 21.4 shows the quality of the base year data used for the simulation, that is the matching of different datasets to each other. The datasets group information of various data, e.g. to locate agents in the buildings we check the available units with the number of households. This data processing can result in unsolved matches, so that households have to be distributed in the first year. For the quality check we compare the agents that form the initial baseyear of UrbanSim in each municipality, with their origin data, i.e. the agents of persons with the populations census. Afterwards, we group these individual quality maps to an overall quality (Fig. 21.4a). While matching of most data could be performed on a level above 80%, we find a low quality of the created jobs, especially in the airport region (Fig. 21.4b). These areas colored white are strongly lacking of building space for jobs, which could be put down to a problem in the enterprise census (see Sect. 21.2.3).

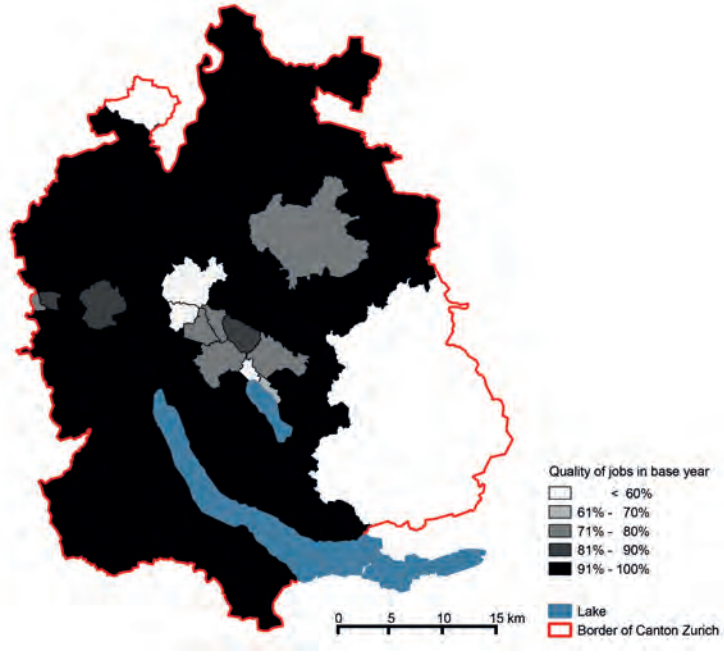
21.3 Modeling

21.3.1 Modeling framework

The modeling framework OPUS (Open Platform for Urban Simulation) used includes different types of models for modeling the behavior of agents, objects and the economic development (Waddell et al. 2003). While continuous variables such as rent prices are modeled through regression analysis; behavior, especially location choice, is modelled using the discrete choice modeling framework, initially developed and introduced to residential location choice by



(a) Overall



(b) Jobs

Figure 21.4 Amount of matched entities in dataset after data processing (©2010 Amt für Raumordnung und Vermessung (GV2113)).

McFadden (1978). Numerous extensions have been developed on this method, that allow for correlations data or modeling of latent classes (Train 2009).

Regression models

Hedonic regression models assume the observed value (e.g. price) is the result of the different aspects of the object, i.e. the weighted sum of the objects' characteristics (Rosen 1974). These characteristics can be individual to the object (e.g. size and age of an apartment), the location (e.g. distance to centre) or represent the overall economic market (e.g. vacancy and demand of single-room-apartments). We used this method to model the rent prices (see Section 21.3.5) and the income of a household. The regression output is kept as an array in the simulation. To make it available for following year of simulation and for analyses outside UrbanSim, it is written to the data tables of agents and objects with a post-processing model. As Löchl has shown in the preceding work (Löchl 2010) hedonic price models are influenced by their spatial distribution and it is necessary to use geographically weighted regression (GWR) or spatial autocorrelation models (SAR), which account for the spatial autocorrelation and spatial heteroscedasticity in the observations (Fotheringham & Rogerson 2009). These and other spatial issues are also discussed in Chapter 13. Due to unavailability of these methods in the OPUS framework, these have not been considered in the work presented here and a linear hedonic model has been used instead. However, to reduce the impact of spatial autocorrelation, several spatial variables have been included in the model.

Location choice models

The relocation of agents and objects is represented through several models that rely on each other. The transition model uses control totals that define the total number of agents per year as an exogenous input to create new agents and objects representing immigration and birth. Meanwhile relocation models represent the dynamics inside the study area. They select agents that will relocate within a simulation year using given probabilities. These exogenous defined relocation rates are based on the agents characteristics such as age and income.

After the creation or selection for relocation, the location choice is performed through a discrete choice model. A set of alternatives is chosen using random sampling or weighted sampling methods out of all available alternatives.

The simplest location choice model are multinomial logit models (MNL) which assume that the error term is independently and identically Gumbel distributed (i.i.d.). Common extensions of this model, are nested logit models (NL) and mixed logit models (MMNL) but only the earlier ones (MNL, NL) have recently been included in the OPUS framework. Nested logit models do

not enforce the assumption on independence of irrelevant alternatives (IIA) and allow for correlation among the error components. A common example is the decision to move and the choice of a new location. As the relocation and location choice is defined in separate models, we do not need to represent these nests and the simpler and faster MNL configurations has been used in our simulation instead.

The estimation of regression and location choice models can be performed on the baseyear data within UrbanSim or outside of the simulation environment. The estimation outside of UrbanSim has the advantage that it allows the easy integration of survey data, i.e. of observations instead of the given distribution. While the integration of survey data into the simulation environment is possible, it requires substantial work when variables rely on various datasets or temporal changes. To give an example, the integration of our household location choice survey (Belart 2011) would demand the creation of households, persons, jobs, apartments, and buildings that match to the survey data, are linked to each other and that match the neighborhood characteristics in the simulation data. Dynamic variables additionally require the temporal representation through creation of this data in several base years.

As this might not be feasible for all survey data, some estimation has been performed outside of UrbanSim with an import of the estimation results. A comparison of the base year and the real world data is recommended in that case, to varify their comparability.

21.3.2 Demography model

The Zürich case study uses the microsimulation demography model that was developed by National Institute of Demographic Studies (INED) in France and that is described in detail in Chapter 7. This simulates the evolution of our case study population from 2000 to 2030. The output is a cross-section of the population for each year of simulation. The records represent individual households and persons which are read in by UrbanSim. We added two simple models within UrbanSim for updating the attributes *car ownership* and *income* because these are not treated within the demography model (see also Sect. 21.3.9).

The demography model requires a base year population, a population of potentially immigrating persons (from abroad or from another region in the same country) and case study specific transition probabilities.

The base year populations are generated within the data preparation process described in Section 21.2.2. The original data is the population census 2000 from the Federal Statistical Office of Switzerland. Some of the required attributes are only available in limited accuracy, e.g. identification of family relations (mother, father, partner) is only possible if the family members still live in the same household.

The demography model has 24 dimensions that would allow the fitting of up to 1 660 696 parameters to tailor the population dynamics to a study area. Transition probabilities can be varied along various dimensions (e.g. simulation time and gender) which leads to this huge parameter set. To facilitate the calibration, probabilities for the majority of the dimensions were kept constant. Dimensions where varying parameters could be found are highlighted using boldface in Table 21.3. The fertility and the emigration module uses the default probabilities, which are Paris specific values, since no suitable values fitting to the dimensions required in the demography module could be found.

Parameter (probability distribution)	**Dimensions**
Education	
Level of education	**Sex**, **Mother's level of education**, **Father's level of education**, **Number of education levels**
Age at the end of studies	**Child's age**, **Possible ages at the end of studies**
Age at the end of studies assuming a given level of education	**Education level**, Simulation time, Sex, **Possible ages at the end of studies**
Passing to a higher school level	**School levels**, **Possible ages to enter next school level**
Labor	
Entering the labor market	Union status, Enrollment status, Level of education, Possible ages at the end of studies, Possible ages for entering the labor market, Dummy for having at least one child, **Sex**
Exiting the labor market	Union status, Enrollment status, Level of education, Possible ages at the end of studies, Possible ages for entering the labor market, Dummy for having at least one child, **Sex**
Population size	
Fertility	(Default values from Paris model)
Mortality	Level of education, Type of residence, Marital status, **Simulation time**, **Sex**, **Possible ages**
Number of immigrating households	**Origin (within country/abroad)**, **Simulation time**
Emigration	(Default values from Paris model)
Unions	
Union dissolution	Level of education, **Level of union duration**, Type of union, Possible ages for entering into a union, 1-year age partition up to 10
Union formation	Number of children, Not-in-union marital status, **Possible ages for entering into a union**, Dummy for living with parents
Leaving the parents' household	Union status, Working status, Enrollment status, **Possible ages**, **Sex**

Table 21.3 Parameters of demography model. Dimensions where varying parameters could be found are highlighted using boldface.

As the only registered partnerships are marriages, these were considered as unions in the demography model. Consequently the marriage module, which has no other consequences than adding marriages, is deactivated. Also the module determining whether a household lives in a communal establishment is not activated, since the location choice of households is modeled and simulated within UrbanSim (see Sect. 21.3.6). Further the model on complex households is inactive due to missing representation of these in the base year data.

The simulation is calibrated against the overall population size of the cantonal statistics.

21.3.3 Transport model

As noted in Section 21.1.5, the agent-based transportation microsimulation framework MATSim (Balmer 2007) has been integrated in the UrbanSim model. The network, shown in Figure 21.5, is derived from the cantonal travel model and contains only major roads. For simplicity, only car and public transport



Figure 21.5 Network used for the transport model (©2007 Office for Transport of the Canton Zürich).

modes are simulated, the latter by teleportation with a constant speed. Each agent uses the same mode for all trips.

Conceptually, the agent-based nature of MATSim would allow using the travel times computed by the transportation model as simulated agent variables and as input to other models. However, only a 10% sample of the persons simulated in UrbanSim are actually used in the MATSim model, and the travel times would be available only for a fraction of the population, so that these are not used in our land use simulation. Nevertheless, workplace accessibility values (Geurs & van Wee 2004) can be computed and used by the other UrbanSim models. The computation of the accessibility internally uses a grid cell partitioning which is then mapped back to the parcel level (cf. Chap. 17).

The transportation model is executed every 5 years, this is sufficient given the relatively low dynamics of job relocation between consecutive years. In addition, the transportation model is executed once prior to model estimation, as part of the data preparation, in order to compute the initial accessibility values. Executing the transportation model more frequently or with a larger sampling fraction (or even with all agents) would have resulted in prohibitive run times for the entire simulation. Options to reuse computation results from previous runs ("warm start", cf. Sect. 17.4.4) are available but have not been employed.

21.3.4 Land development models

The purpose of land development models is to predict the usage of a piece of land. This can be done with spatial interaction modeling (Lowry 1964, Putman 1974) using gravity models or economic modeling representing real estate markets. In literature we find two approaches: The bid-rent theory pioneered by Alonso (1964) and the bid-choice theory developed by Martínez (1992) as extension to the earlier.

These bid models assume market equilibrium, which means that prices adjust to supply and demand. Recent work moves towards quasi-equilibrium models (Hurtubia et al. 2012) for the sake of which Farooq & Miller (2012) adds principles from game theory. UrbanSim (Waddell 2002) is based on a second approach using hedonic price modeling (Rosen 1974) for price formation and allowing for a quasi dynamic market. Recent works also show the possibility to enforce equilibrium (Wang & Waddell 2013).

While heterogeneous actors are modeled on the demand side, most real estate supply models include a representative developer. Usually a distinction is made according to building types (Dong & Gliebe 2011, Haider & Miller 2004). Recent work extends these and models heterogeneous actors on the supply side of land and real estate markets (Martínez & Donoso 2010) (cf. Chap. 10).

Parceling of land is also part of land development but rarely modeled within land use models. A recent exception is the work of Zhou & Kockelman (2008) where supply of parcels for single family housing is modeled by subdividing big parcels following an observed distribution.

In this case study land development is modeled through two models. The building transition model (BTM) determines the number of buildings to build according to the user defined vacancy rate. These buildings are then passed to the Building Location Choice Models (BLCM) that evaluates the utility of different locations (parcels).

The building transition model calculates the vacancy for 10 sub-markets. We discriminate a residential market and nine non-residential sectors according to employment sectors (cf. Table 21.2). If the simulated vacancy falls below a predefined level, buildings are added from a pool of development templates until the target vacancy is met. The pool comprises all buildings built between 2000 and 2010 according to the GWR.

The target vacancy is set to 0.66% for residential and to 4.0% for non-residential. While this vacancy rate represents the value reported in the Swiss statistics, it is very low in comparison to other European cities. As Thalmann (2012) shows, it includes all dwellings, that are available on the rent- or sales market at a certain date, while our simulation includes all empty dwellings in the simulation, i.e. also those that might be a second home or currently under renovation. We can thus expect the vacancy to be too low, without having detailed information on this bias.

The location of the sampled buildings is determined by four sub-models depending on the type of building to be located. We differentiate buildings for single-family housing, buildings for multi-family housing, mixed use buildings and non-residential buildings. The locations to be chosen are individual parcels. The eligibility of a parcel is determined by its plan type and its remaining capacity of allowed floor area. The model can add floor area to a parcel which is already built but it will not replace existing buildings.

For each building to be located, 30 alternatives are sampled and evaluated with a linear utility function. The choice set size is a tradeoff between accuracy and computational efficiency, the default setting in UrbanSim has been used here. Table 21.4 describes the variables included in the respective models, their effects are shown in Table 21.5. The models are estimated in UrbanSim using the observations on buildings built between 2001 and 2010 in the study area. However, these models are still in an early stage and need further improvement.

The negative sign for car accessibility for residential buildings can be interpreted as preference for remote locations being more quiet and rural. These variables should be controlled for, most importantly because we must expect counter intuitive results when simulating policy scenarios. It is intuitive that public transport is more important in respect of multi-family housing, mixed

Variable Name	Description	Unit
Car accessibility	Accessibility of jobs by car according to formula (17.9) in Chapter 17. Travel times are calculated with MATSim.	[-]
PT accessibility	Accessibility of jobs by public transport according to formula (17.9) in Chapter 17. Travel times are calculated with the cantonal transport model (Vrtic et al. 2005).	[-]
Distance to school	Euclidean distance to next school facility (all education levels).	[m]
Fit of building to parcel	Step function of the difference between permitted floor-space on the considered parcel and floor-space of the development project. If the difference is larger than zero, the natural logarithm of the difference is taken. If the difference is smaller than zero, five times the difference is calculated.	[m^2]
New neighbouring buildings	Number of buildings with build year bigger 1995 within 150 m.	[-]
Price per permitted floor space	Price per permitted square meter floor-space.	[CHF/m^2]
Slope	Mean slope of parcel in percent.	[%]

Table 21.4 Description of variables included in building location choice models.

	Single family housing		Multi family housing		Mixed use		Non-residential	
Parameter Name	Effect	Sign.	Effect	Sign.	Effect	Sign.	Effect	Sign.
Car accessibility	-0.9987	***	-0.6133	***	0.4245	insig.	0.0397	insig.
PT accessibility	-0.0384	***	0.0318	insig.	0.1782	***	0.0657	***
Distance to school	0.0004	***						
Price per permitted floor space	-0.00002	***	0.0002	***	-0.0003	***	-0.0005	***
Slope	0.0278	***	-0.0040	insig.			-0.0996	***
Fit of building to parcel	0.00005	insig.	0.0008	***	0.0005	***	0.0002	***
New neighbouring buildings	0.0436	***	0.0426	***	0.0294	***	0.0190	***
Adj. likelihood ratio index	0.15		0.34		0.31		0.12	
Significance level	0.5%	***	1.0%	**				
Number of observations	9102		4856		337		1400	
Total number of observations	15695							

Table 21.5 Variables and their effects in building location choice models.

use and non-residential buildings since we can expect lower car ownership at such places.

Distance to school is only significant for single family housing and is positive. This is surprising since the majority of developers or potential buyers can be expected to be families which should prefer proximity to schools. Maybe we are again capturing preference for living at the fringe of settlements.

Positive signs for prices per permitted floor space are unexpected in the first place. Given that prices are averages within municipalities the positive signs could reflect the attractiveness of the municipalities capitalised in land value.

The preference for parcels with steeper slope indicates that hillside locations are preferred for single family housing. For non-residential buildings we see an opposite effect which can be explained by higher construction cost at steeper locations.

The fit of a buildings' floor area to the allowed floor area on a parcel is positive for all building types. This fits the expectation that development capacity is exploited as much as possible.

New neighboring buildings around a parcel increase the probability of development, which shows that development has a tendency to cluster or represent large size developments of certain developers. This highlights the need to model individual developers in the ongoing work.

With the current formulation we expect that vacancy is kept slightly above the assumed level. The location of the buildings has constraints through regulations on the parcel. Residential buildings can only be located in residential and mixed zones, non-residential buildings can only be located in non-residential and mixed zones where capacities for new floor area are available. Within these restrictions the location of new buildings largely depends on the accessibility but with different signs for residential and non-residential buildings. The model would benefit of the use of accessibility of residents rather than workplaces for residential buildings, which has not been realized in this first implementation (Sect. 21.3.3).

The current model is limited by not being sensitive to price which is a major drawback. The mechanism of maintaining vacancies for the whole study area is a crude representation of the market and deserves improvement. One possibility is to calculate vacancy per municipality rather than the whole canton. Building location choice would then be constrained to the municipality where the building is generated which makes the model more demand driven.

The current model only features new constructions, so densification of built space is represented by adding floor area to the already existing stock. This model would benefit from adding further options such as renewal, deconstruction or replacement. The model can also be improved by accounting for correlations among alternatives, e.g. considering spatial effects.

21.3.5 Hedonic real estate prices

The real estate price model simulates the rent price of individual residential units based on their characteristics and their location. A detailed review of literature on hedonic rent price and a model development for Zürich can be found in Löchl & Axhausen (2010). This work shows the benefit of modeling the rent price through integration of the living unit and its specific characteristics as an additional dataset.

The interaction of rent market and owners market is not respected in the current simulation and will be subject to ongoing work. Further work could also extend this model to include commercial rents and selling prices. Haase (2011) shows that the commercial rent prices vary substantially depending on the contract types. As these are not available for the study area the integration of his model was not possible. However, Bodenmann (2011) shows that the land price has a low impact on the location choice of firms, meaning that this shortcoming should not be a major problem.

Our research suggests that a large part of reported rent price is explained through the characteristics of the dwelling and not through its location. An alternative approach that has been tested was to model the rent price independent of total size as a rent price per squaremeter, but these models showed a low explanatory power, so that we favored a dedicated model for residential units. Not all variables reported by Löchl & Axhausen (2010) are available in the simulation environment. One of these are the taxes, that appear to have a high impact. No model is representing the tax changes over the simulation period is available, so that these could not be implemented. Future work should envision to include these as user defined scenarios instead. The model estimation was done inside UrbanSim using the asking prices of real estate advertisements (see Sect. 21.2.3).

The comparison of the models with Löchl & Axhausen (2010) shows the expected sign of all variables except foreigner density, which described a different content (Table 21.7). While Löchl & Axhausen (2010) defines only those persons as foreigners who visually differ from the typical western European, our estimation defines all Non-Swiss persons as foreigners. That these number differ can also be observed in the descriptive statistics of the choicesets (Table 21.6). Most variables in the base year data represent real world data used by Löchl & Axhausen (2010) fairly good. Only the density of jobs has been included through means of a different spatial extend. We use the absolute sum of jobs in 1km radius, while Löchl & Axhausen (2010) calculated these per gridcell of 100m x 100m.

In Table 21.7 we can observe that non-Swiss persons are mainly high-educated persons that can afford to locate in the attractive districts of the study area. The positive sign of that variable is thus representing the high attractivity of certain districts. We observe that exclusivity in form of lake

	UrbanSim mean	Löchl (2010) mean	Diff (%)
Car accessibility	9.97		
PT accessibility	11.34	10.69	6.06
Built in 1921 to 1930	0.02	0.03	-37.30
Built in 1981 to 1990	0.14	0.17	-17.95
Built after 1991	0.18	0.20	-11.89
Built before 1921	0.09	0.11	-10.90
Distance to station	0.90	0.91	-1.1
Proximity to highway (< 100 m)	0.02	0.02	9.14
Is a single family house	0.02	0.02	12.52
Jobs in hotels and gastronomy	448.51	4.54	(9781.89)
View of lake (ha)	1559.37	1465.57	6.40
Population density (ln)	4.47	4.26	5.12
Size in m^2 (ln)	4.38	4.43	-1.12
Slope of terrain	-3.28	-3.27	0.10
Sunshine index (evening)	-0.78	-0.64	20.51
Foreigners within 300 m	184.35	59.80	(208.29)

Table 21.6 Comparison of the real estate price model estimated in UrbanSim with the one estimated by Löchl (2010).

	UrbanSim		Loechl et al. (2007)	
	Effect	Sign.	Effect	Sign.
Constant	+	**	+	**
Car accessibility	+	**	n.a.	n.a.
PT accessibility	+	**	+	**
Built in 1921 to 1930	+	**	+	**
Built in 1981 to 1990	+		+	**
Built after 1991	+	**	+	**
Built before 1921	+	**	+	**
Distance to station	-	**	-	**
Proximity to highway (< 100 m)	-	**	-	**
Is a single family house	+	**	+	**
Jobs in hotels and gastronomy	+	**	+	**
View of lake (ha)	+	**	+	**
Population density (ln)	-	**	-	**
Size in m^2 (ln)	+	**	+	**
Slope of terrain	+	**	+	**
Sunshine index (evening)	+	**	+	**
Foreigners within 300 m	+	**	(-)	(**)
Adj. Likelihood ratio index:	0.78		0.85	
Number of observations:	6497		8592	

Table 21.7 Comparison of the model estimated in UrbanSim with the estimation outside of UrbanSim.

view and a lot of sunshine, raises the rent price. Also centrality measured through the density of jobs and the accessibility augments the value of a living unit, while the proximity to noise reduces this one. Finally we find a preference for certain building periods, especially for very old or very new buildings.

The real estate price model estimates the rent price of all living units in the simulation process, so that yearly changes and adoptions of the rent price are respected. The estimated price is written out to the dataset through means of a post-processing model (Sect. 21.3.9).

21.3.6 Household location choice

The household location choice model represents the residential choice of a relocating household. For simplification only the rental market is modeled. We can expect that renting households are less bound to their residential unit and relocating in shorter time intervals, so we assume a higher dynamic of household relocation and location choice in our simulation than in real world.

A detailed review of residential location choice on a disaggregated level can be found in our previous work (Schirmer et al. 2012, 2013). Some of these model the decision of relocation and the location choice in form of a Nested Logit Model (Kim et al. 2005, Zondag & Pieters 2005, Lee & Waddell 2010). This decision generally relies on changes of the household characteristics, e.g. the birth and death of household members or the formation and split of relationships. In our model these changes are modeled through the external demographic model (see Sect. 21.3.2), while decisions on moving are modeled through probabilities in the household relocation model based on the work of Beige (2008). Other studies model the workplace choice and residential location choice through means of nested logit models or joint models (Freedman & Kern 1997, Waddell, Bhat, Eluru, Wang & Pendyala 2007), which can not be represented in our case study as the workplace choice and residential choice form different models running in sequential order. We thus included the households location choice as a multinomial logit model representing only the location choice after having decided to move.

The accessibilities used in the model come from the transport model MATSim (see Sect. 21.3.3 for details). Meanwhile the distance to the center and the distance to workplace are euclidean distances. Using the travel times of MATSim would be of benefit, but demands running the transport model every year for the whole population, which seemed not feasible in terms of computation time. The same applies for the mode choice. Instead we differentiate between car owners and non car owners, which are defined through the car availability model (see Sect. 21.3.9).

The price of an apartment and the income are updated every year through the income regression model (Sect. 21.3.9) and the real estate price model (Sect. 21.3.5).

The models were estimated using a 2010 survey on 1000 recent movers (Belart 2011). Details on the model developments have been reported in Schirmer et al. (2013). This study showed the importance of modeling the residential location through means of the residential unit and individual attributes of price and size for that unit. Using the dwelling register allowed to model the residential location on this disaggregated level and to include individual characteristics of the residential unit. The rent price of the dwelling is simulated using the real-estate-price model to react to changes during the simulated period (see Sect. 21.3.5).

To avoid the integration of the survey in the base year (see Sect. 21.3.1), the residential location choice was estimated in *biogeme* (Bierlaire 2003) outside of *UrbanSim* using the survey data. The results of this estimation are presented in Table 21.10. Most variables have the expected sign as reported in Schirmer et al. (2012), but we do find different perceptions for some measures of location, i.e. for car accessibilities and densities of service and retail. Our models suggest that car owners dislike accessibility, that retail and service density is generally disliked and that households tend to search for distance to schools. The negative perception of car accessibility is in coherence with previous studies on the canton of Zürich (Belart 2011) and supposedly reflects the small extend of the study area, as most municipalities have access to the city of Zürich as economic center within less than 30 minutes by car. Living in a low accessible location does therefore rather represent self-selection effects of car owning households that seek the quality of rural regions, that still have a good access and good connections. Meanwhile we expect the negative influence of service and retail to be caused by the general tendency to dislike urban centers (see Schirmer et al. (2013) for details).

Rather surprising is the significant positive impact of distance to school, that keeps its sign and significance even for families. We expect this reflects the high number of schools and thus rather catches negative effect of the proximity to schools, e.g. being noisy areas with screaming children in the courtyards. A differentiation of different levels of distance would eventually be advantageous, but has not been studies and still needs to be evaluated in further research.

For control of the variables used in the simulation this model was also estimated inside UrbanSim on the base year, which represents the distribution instead of recent choices. Table 21.9 shows the comparison of the choice-sets of these models. Most variables are comparable with only the density of service and the retail being higher in the base year data. One possible explanation for this observation is the data-processing of jobs, which leads to a different distribution of jobs (Sect. 21.2.3). Another explanation is the location of buildings in the survey data. These represent recent movers that tend to move into new buildings (variables *New building?* and *Building age*). As these are often located at the fringe of cities, they tend to have less mixed use activity in the

close neighborhood. We thus can expect that our variables do not include any misspecification even if manually specified.

Variable	Definition	Expected sign[a]
Household		
Distance to workplace	mean Euclidean distance to workplace (dist * X^{η})	-
η for distance to workplace		
Distance to previous location	mean Euclidean distance to previous location (dist * X^{η})	-
η for distance to previous location		
Accessibility		
Car accessibility (car)	accessibility of jobs by car * household has a car	+
PT accessibility (no car)	accessibility of jobs by public transport * household has no car	+
Built environment		
Building density	density of buildings in 300m	+
Building density (families)	density of buildings in 300m * is a family (non-single household with children)	-
Building density (young HH)	density of buildings in 300m * is a young household (age <40)	+
Open space	open space within 2km radius	+
Distance to lake	distance to lake (size of lake <1 km^2)	-
Close to network?	is within buffer on arterials (50m) and railways (100m) (dummy)	-
Points of interest		
Distance to school	distance to school	-
Retail density	density of jobs in retail (NOGA-Code G47) in 300m	+
Service density	density of jobs on service (NOGA-Codes K to N) in 300m	+
Distance to station (no car)	distance to station * household has no car	-
Distance to highway access (car)	distance to highway-access * household has a car	-
Distance to CBD Zürich (young HH)	distance to CBD Zürich * is a young household (age <40)	-
Distance to CBD Winterthur (young HH)	distance to CBD Winterthur * is a young household (age <40)	-
Socioeconomic structure		
Population density	density of population in 300m	-
Population density (young HH)	density of population in 300m * is a young household	+
Population density (single HH)	density of population in 300m * is a single household	+
Households of same age	share of households with same age (3 groups) in 300m	+
Households of same size	share of households with same size (7 groups) in 1km	+
Households with children (has children)	share of households with children in 300m * household has children	+
Households of same nationality	share of households with same nationality (Swiss/non-Swiss) in 300m	+

[a] Expectations based on Schirmer et al. (2012).

Table 21.8 Explanation of variables and their expected influence (*Continued on next page*).

Variable	Definition	Expected sign[a]
Residential unit		
Room size	m^2 per room	+
Rooms per person (single HH)	rooms per person * is not a single household	+
Rooms per person (non-single HH)	rooms per person * is a single household	-
Rent-income ratio	ratio of rent to income	-
Building age	logarithm of age of building	-
New building?	building is a newly constructed building (<10 years)	+

[a] Expectations based on Schirmer et al. (2012).

Table 21.8 Explanation of variables and their expected influence (*concluded from the previous page*).

	UrbanSim		Survey		Difference
Variable	mean	stdev	mean	stdev	mean
Car accessibility (car)	9.89	0.49	9.07	0.61	9.03
PT accessibility (no car)	11.97	0.79	11.16	1.09	7.33
Close to network?	0.02	0.14	0.08	0.27	-76.38
Distance to highway access (car)	2147.81	1499.42	2333.57	1696.85	-7.96
Distance to school	420.50	289.42	446.98	357.36	-5.93
Distance to station (no car)	800.86	463.51	723.72	450.86	10.70
Distance to CBD Winterthur (young HH)	10299.55	7368.73	12104.99	7834.76	-14.91
Retail density	83.82	190.46	16.46	56.91	409.24
Service density	255.84	628.82	61.10	206.60	318.75
Households of age 0–39	0.38	0.09	0.37	0.11	1.30
Households of age 40–64	0.37	0.08	0.46	0.10	-20.91
Households of age 65 and above	0.21	0.07	0.20	0.09	7.07
Building age	3.46	0.94	3.03	1.43	14.25
New building?	0.10	0.31	0.24	0.42	-55.96
Rent-income ratio	0.27	0.19	0.24	0.18	9.65
Rooms per person	1.93	1.12	1.90	0.83	1.90
Room size	26.48	8.19	28.32	9.13	-6.52

Table 21.9 Comparison of the choicesets on the baseyear data and the survey.

21.3.7 Employment location choice

The employment location choice model is simulating the distribution of jobs through means of a multinomial logit model. Similar to the household location choice it distributes jobs that have been chosen by the relocation model. This relocation model uses probabilities for different sectors. We differentiate the following types of jobs using the NOGA codes of the enterprise census: manufacturing, construction, trade, retail, hotels and restaurants, transport, services, health.

Variable	estimate		β * median(choice)
Household			
Distance to previous location	-7.070	**	
Distance to workplace	-3.220	*	
η for distance to previous location	0.163	**	
η for distance to workplace	0.203	**	
(interaction distance to previous location)			*-26.892*
(interaction distance to workplace)			*-16.627*
Accessibility			
Car accessibility (car)	-0.302	**	-2.758
PT accessibility (no car)	0.541	**	5.877
Built environment			
Close to network?	-0.304	*	0.075
Points of interest			
Distance to CBD Zürich (young HH)	0.067	**	0.760
Distance to highway access (car)[a]	-0.092	**	-0.159
Distance to station (no car)[a]	-0.230		-0.171
Retail density	-0.003	**	0.000
Distance to school[a]	0.298	**	0.107
Service density	-0.001	**	0.000
Socioeconomic structure			
Households of same age	0.684	**	0.271
Residential unit			
Building age	0.360	**	1.248
New building?	0.578	**	0.578
Rent-income ratio	-3.400	**	-16.609
Rooms per person	-0.677	**	-1.124
Room size	0.000	**	0.000
Initial LL	-2679.736		
Final LL	-1261.079		
R^2	0.529		
Adj. R^2	0.522		
** 5% significance			
* 10% significance			

[a] Variable has been scaled from m to km

Table 21.10 Estimation results of household location choice model as implemented.

With an average of approximately 0.04 the relocation probability is significantly lower than for households (0.3). Bodenmann (2011) could show that the relocation probability is becoming lower over the lifetime of a firm. As we model the relocation of single jobs instead of firms and as the relocation model does not include information of the current firm age, a constant relocation probability has been included in our current setup. A possible extension for the ongoing work will be the grouping of jobs into firms and the integration of a demography model for firms respecting the lifetime and sector of a firm.

A transition model is replacing the demography model used in the household location choice and creates or deletes jobs based on control totals per year

and sector. As only three enterprise census are available for the calibration period, a rate-based growth has been assumed to define the control totals.

In preceding works, Bodenmann (2011) has reported on the relocation of firms (on the scale of municipalities) in the region of St Gallen and Appenzell. His models suggest that the main effect on the utility for a location are financial aspects, while the location is of lower importance. Also the rent price or land-value is found to be of low importance in his models. We would expect a similar behavior for the jobs in our study area, although our jobs are not grouped into enterprises. However, a comparative model using the base year data has shown that this thesis does not hold for the study area. On the one side not all the variables were available in our simulation data, on the other side the location characteristics of these study areas are too different, i.e. Appenzell and St Gallen being rather rural regions of low connectivity against Zürich being an economic center of high access and high density in terms of population and jobs. As a result the models did not converge or showed a very low model fit.

Due to this constraint, a new model has been set up that takes advantage of the disaggregated level of information given in the base year. This model has been estimated on the base year inside UrbanSim, which represents the distribution of the enterprise census (see also Sect. 21.2.3). The results of this estimation are given in Table 21.11 and show consistent and expected effects in terms of the sign for most variables. This model suggests that accessibility, job accessibility and the share of jobs in the same sector, tend to be attractive to a relocating job, while the zonal income and population density is rather disliked. This implies a tendency for functional separation across the urban landscape. As one would expect the service and health sector, which depends less on highway access search for proximity to urban centers while all other sectors depend on the road network and tend to search the proximity to highway and distance to the center of Zürich.

As mentioned in Section 21.2.2 the enterprise census required the redistribution of jobs in headquarters as the available building space did not fit to the amount of jobs reported. This results in a high number of jobs being redistributed in the first year, which has an effect on the whole simulation as jobs tend to search areas with a high share of jobs in the same sector. Besides, this also has some effect on the estimation results as some of the jobs used in the relocation have been relocated.

In conclusion, it is recommended that the ongoing work will estimate the employment location choice model using data of recent movers and extend the model to a firm location choice model grouping jobs into firms, similar to households grouping persons. This would allow for representation of different job types (leading position, collaborator, etc.) within the firm and link these to different incomes. Also the workplace choice model would benefit from this information.

	Mfg	Cns	Trd	Rtl	HR	Trn	Srv	Hlt
Average zonal income	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**
Accessibilty car	+1.36**	+1.24**	+1.65**	+2.42**	+1.67**	+1.86**	+1.85**	+0.63**
Accessibilty PT	+0.04**	+0.04**	+0.11**	+0.06**	+0.01	+0.00	+0.05**	+0.01
Distance to highway access	-0.00**	-0.00**	-0.00**	-0.00 *	+0.00	-0.00**	+0.00**	+0.00**
Distance to station	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**
Distance to Zürich CBD	+0.00**	+0.00	+0.00**	+0.00**	+0.00	+0.00**	-0.00	-0.00**
Household density (km^2)	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**	-0.00**
Job density (km^2)	+0.00**	+0.00**	+0.00**	+0.00**	+0.00**	+0.00**	+0.00**	+0.00**
Share of same jobs (zone)	+5.82**	+5.26**	+8.63**	+8.19**	+7.37**	+6.13**	+3.85**	+4.85**
Adj. likelihood ratio index:	0.17	0.11	0.23	0.18	0.13	0.26	0.21	0.17
Number of observations:	15714	9187	11895	10143	7038	14390	33170	12382

Table 21.11 Employment location choice models for the sectors defined in the Zürich case study.

21.3.8 Workplace choice

The workplace choice model represents a persons' choice of job when changing the job. Every work-seeking person is included and can match a job that is not linked to a person yet. The work status of a person is modeled through the external demography which includes a labor market model. Besides the workplace choice respects persons that have lost their job as this one got removed in the employment transition, and persons that are newly created in the demography model, i.e. immigrants and newborn.

As reported by Inoa et al. (2013), we can assume interdependencies of workplace choice with household characteristics, person characteristics and the residential location, i.e. one can expect that workplace choice strongly depends on the person's education, work experience, age, and partnership status, as well as the residential location. Inoa et al. (2013) thus model the job type, workplace locatino and residential location in a three-level nested logit model. However, as the job choice and residential location choice are modeled independently from each other through two different models in our study, this interaction has not been simulated and will be subject to the ongoing work.

Modeling the job search in that detail demands a precise knowledge of the job's characteristics. As reported in Section 21.2.3 the link of persons and jobs is mainly imputed as only given through the location on scale of the municipality. We thus estimated this model outside of UrbanSim using the population census. As no other characteristics than the sector are given for a jobs, we can only differentiate sectors and the information of the neighborhood. We thus decided to model the job search only in reference to the residential location including only the Euclidean distance in form of an exponential function.

Variable	estimate	t-value
$\beta_{\text{dist}} * X^{\eta_{\text{dist}}}$		
distance to job (β_{dist})	4.9	27.97
distance to job (η_{dist})	-0.106	-9.75
Adj. likeli ratio index	0.18	
Number of observations	5291	

Table 21.12 Estimation result of workplace choice model.

21.3.9 Additional models

Besides these models that represent the land use changes, we implemented additional models to keep interaction between the three agent-based simulations consistant. Changes that are generated in the demographic model can have an impact on the job-options and the car-availability, these are described in the following in detail.

Car availability The car ownership model is a binary choice model that will estimate the probability of a household to have a car or not. It would be of advantage to include the mode choice of the transport model as variable in this model, to avoid persons never using the car of keeping this one, but as MATSim runs only on a 10% sample of the population this was not feasible. The availability of a car is mainly of influence in the household location choice model (Sect. 21.3.6).

This model has two utility functions for the option of having a car or not: not having a car is explained by the households level of education, while the choice of having a car is explained through the distance to CBD, its income and the number of persons in the household. It has been estimated on the base year, which represents the information given in the population census. A simple model is used to post-process the outcome of this model and update the dataset.

Income regression The demography models does not update the income of a household. As the households income an important variable in the location choice of households, a model has been developed to update the income of a household during the simulation. This regression model estimated the income based on the households' level of education, availability of cars in the household and the number of persons in the household including children. The model has been estimated on the base year which represents the census information. A limitation at current time is that it does not include the work status of household workers.

Update workstatus The update to work status is a simple model, that deletes the links between jobs and persons when the work status of a person has been set to unemployed by the demography model. As a result the person is defined as work-seeking and will be respected in the workplace choice model.

21.3.10 Interaction of models

As mentioned in Section 21.1.4, UrbanSim is a non-equilibrium model and the sequential order of the models has an impact on the outcome of the simulation. To integrate changes that occur from the labor market or the new household composition, modeled in the external demography, the post-processing of income, car availability and the update of the job links is needed. The real estate model (REPM) evaluates the prices of all buildings, including the new construction based on their (new) location, so that it has to run directly after the building location choice. Running the workplace choice after the employment transition makes sense, as this one links the new created jobs to household. We decided to run the household location as the last model, so that it can take advantage of new buildings and new jobs being already linked to person.

The order of running the models is shown in Table 21.13. Figure 21.6 additionally gives an overview of the interactions currently implemented in the main models. As can be seen all models include the accessibility provided by MATSim Section 21.3.3. This allows to implement scenarios as roadpricing, which is described in Section 21.4.2.

#	Model	Abbr	Section	
1	Demography		21.3.2	Demography model
2	Income regression		21.3.9	Additional models
3	Postprocess income regression		21.3.9	Additional models
4	Car availabilty		21.3.9	Additional models
5	Unset job id		21.3.9	Additional models
6	Postprocess car availabilty		21.3.9	Additional models
7	Building transition	BTM	21.3.4	Land development models
8	Building location choice	BLCM	21.3.4	Land development models
9	Real estate price	REPM	21.3.5	Hedonic real estate prices
10	Postprocess real estate price		21.3.9	Additional models
11	Employment transition	ETM	21.3.7	Employment location choice
12	Employment relocation	ERM	21.3.7	Employment location choice
13	Employment location choice	ELCM	21.3.7	Employment location choice
14	Workplace choice	WCM	21.3.8	Workplace choice
15	Household relocation model	HRM	21.3.6	Household location choice
16	Household location choice	HLCM	21.3.6	Household location choice
17	MATSim		21.3.3	Transport model

Table 21.13 Order, abbreviations, and location of the description for all models.

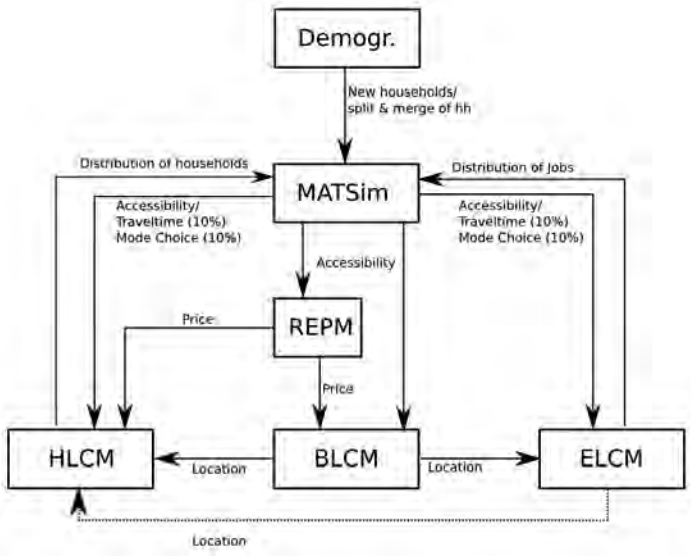


Figure 21.6 Interaction of the main models in the case study Zürich.

The real estate price model evaluates the rental price, which is only used in the household location choice. The building location choice evaluated the buildable space according to the zoning definitions and creates the space for jobs and households. It thus has a strong impact on their location options in case of low vacancies. Finally, the distribution of of jobs is of influence for the household location, as households tend to search for living space close to their workplace.

The use of same variables in several models causes correlations between the models that might cause unexpected simulation results. An example of such a variable is the accessibility. Price as evaluated by the real estate price model includes accessibility, which increases the rent price, on the other side price and accessibility are included in the household location choice, with price having strong negative effect. Depending of the influence of accessibility in both models, a lower accessibility as expected by the road pricing scenario might create unexpected results, e.g. that road pricing attracts persons having a car.

Future work should envision to include travel times and travel costs, which are given through the transport model. Several models currently use Euclidean distances instead, which does not change for a location (see Sect. 21.3.4 to 21.3.6). In this study this was not included as MATSim only ran on a 10% sample of the population so that travel times were not available for each agent.

21.4 Scenarios: simulation and results

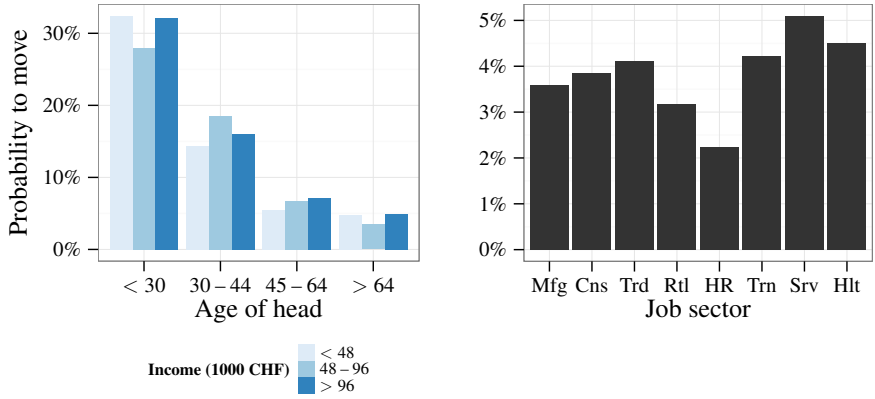
In the following paragraphs we briefly describe the assumptions common to the baseline and all scenarios.

We assume the same evolution of the population. The demography model is run once and the same output is used for all four scenarios.

The number of jobs is defined by the modeler and included in the simulation as control totals. We distinguish nine sectors on which we define these control totals. We assume growing number of jobs for retail, hotels and restaurants, service, health and other, while the number of jobs in manufacturing, construction, trade and transport are assumed to be shrinking according to the current developments. Growth and shrinking rates are derived from the Business Censuses 2001 and 2008.

The relocation rates for households and jobs are given exogenously. They vary according to age of household head, and household income (Fig. 21.7a). In case of jobs different relocation rates apply for the distinguished sectors (Fig. 21.7b, see also Table 21.2).The high relocation rates lead to a high dynamic concerning spatial distributions.

The building square foot needed for a job in one of the distinguished sectors is specified as well. The values assumed range from 35 to 80 square meters.



(a) Relocation rates for households **(b)** Relocation rates for jobs

Figure 21.7 Relocation rates as used in preceding project Zukunft urbaner Kulturlandschaften (Löchl et al. 2007).

These space requirements are kept equal over the whole simulation area and neither vary in time.

We distinguish 145 plan types. A plan type is a zoning class with associated development constraints. The constraints refer to specific land uses. Here the constraints are given as minimum and maximum of allowed floor area ratios (FAR). An example is shown in Table 21.14. The plan types are assigned to individual parcels. These legal geographies remain constant during the whole simulation as well as zones and municipalities. They are also the spatial units for aggregation.

Plan type id	Constraint id	Constraint type	Constraint use	Max	Min
1105	11051	FAR	Residential	1	0.9
1105	11052	FAR	Residential and commercial	0.1	0
1105	11053	FAR	Center	0.1	0
1105	11054	FAR	Governmental	0	0
1105	11055	FAR	Industry and commercial	0	0

Table 21.14 Development constraints of plan type 1105 (*More than 90% residential use with five levels.*).

The configuration of the transportation model (network, utility parameters, etc.) also remains unchanged during the entire simulation. In addition, no traffic from, into or through the study area is simulated, and for the study area only the work trips are considered.

21.4.1 Validation of baseline

The baseline is compared to validation data from the Cantonal Statistical Office. Validation is done in respect of the main entities in the simulation: households and persons, jobs, living units, and buildings. As it is a dynamic simulation we can look at time series and cross-sections in specific years. In the following we will show selected time series and focus on the cross-section in year 2010 for the baseline validation. Thus we have a validation period of ten years. Further we can look at locations since the simulation is spatially explicit. We choose municipalities as geographical aggregation level because statistics are collected at that level and political stakeholders are associated with them.

The validation of the total sum of agents and objects is mainly of relevance for the population and building stock, as the are the outcome of the demography model (Sect.. Section 21.3.2) and the building transition model (Sect. 21.3.4), while the total number of jobs is given to the simulation as control total.

The simulation contains 6.1% more living units than in the validation statistic after ten years of simulation (2010). A closer look at the development of the living units stock shows, while overestimated production of living units leads to a stock that is too big after three years of simulation (Fig. 21.8). The reason might the low assumption for the vacancy rate as discussed in Section 21.3.4.

The control of the demographic model using the control totals shows that the simulation contains to few persons (−1.5%) after 10 years, but the longitudinal validation shows that the demography model produces fairly accurate yearly population totals with no more than 0.017 relative error (Fig. 21.8).

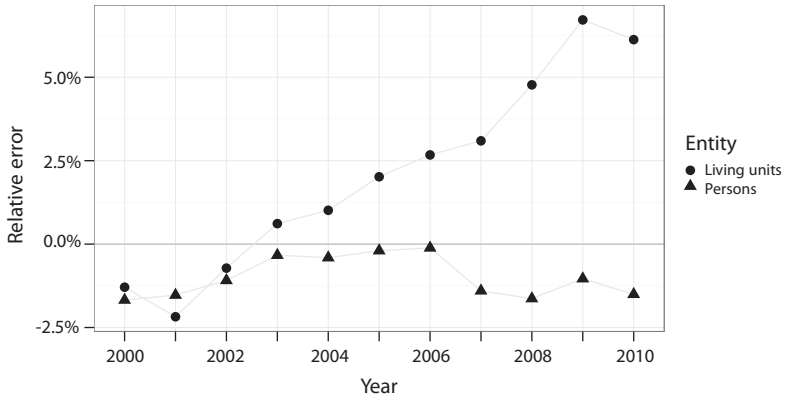


Figure 21.8 Deviation of overall number of living units and persons for study area to validation.

However, looking at the spatial distribution of the simulated entities might be more relevant here since most models are about location choice.

Figure 21.9 shows the prediction error for various entities. The analysis is carried out separately for municipalities with at least or less than 3000 inhabitants, respectively, denoted by *large* or *small* for brevity. Large are generally better predicted. It can be seen that the deviation of building counts in not larger than 50% for all but one municipality. In most municipalities, more living units are predicted than actually built in the validation period due overall overestimation stated earlier. The low spread of the density curves, which is wider for the living units than for the buildings, shows that spatial distribution of these entities is better reproduced compared to jobs and persons. Four outliers above 50% for living units indicate that built space dedicated for residential use is located less accurately. This has to be kept in mind when looking at the households and persons since the location of living units determines the location options for households. A reason for the lower quality in predicting location of living units is supposedly more available capacity in residential zones for construction.

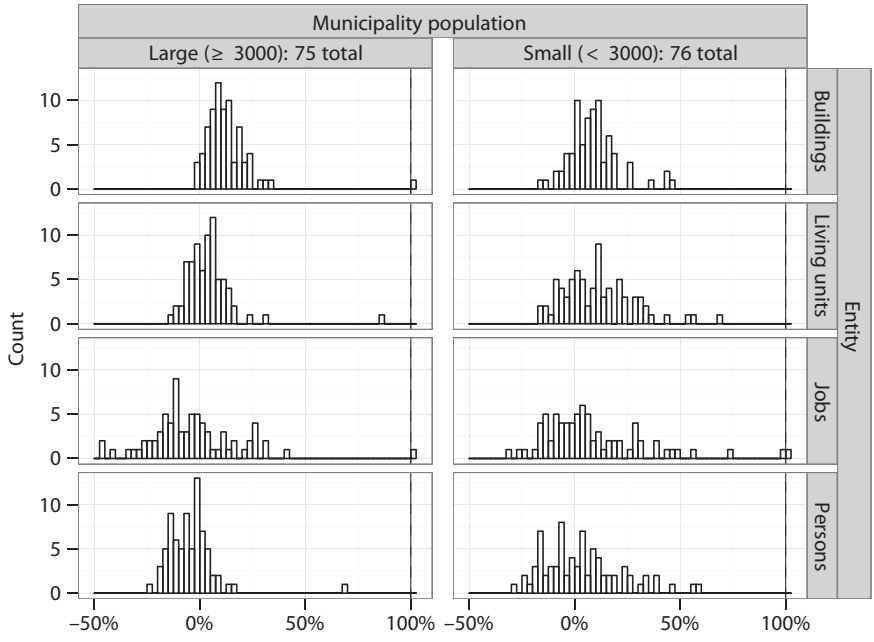


Figure 21.9 Histograms of per-municipality deviation for various entities. The horizontal axis is limited to 100%, all observations above this threshold are shown right of it. The censoring threshold is denoted by a vertical dashed line.

For jobs, the majority of the municipalities are simulated with less than 30% difference to the observed development. The number of persons in the large municipalities is predicted more accurately than the number of jobs. As the base year data demands more routines of distribution in the data processing (Sect. 21.2.3) this is not surprising. There are only three outliers with more than 40% deviation.

The presentation of the results in maps showing the percental difference, reveals some details on the spatial distribution of the simulation quality. Municipalities in red are overestimated whereas municipalities in blue are underestimated.

In ?? we see that the number of buildings is uniformly overestimated across the study area. The major overestimations are located in a few rural municipalities which possibly have high capacity for new built space. Comparing ?? and ?? we find that the spatial distribution of persons strongly depends on the one of living units. The low vacancy which is assumed (see Sect. 21.3.4) leaves reduced flexibility to households' location choice. Spatial distribution of jobs is less good than the one of persons. ?? shows a scattered image over the study area.

As these outcomes are from a stochastic simulation, the scenario effects are put in relation to the baseline model's inherent variability. We compute the coefficient of variation over multiple simulations. The set *Baselines* includes four baseline simulations whereas the set *Scenarios* includes the baseline and the three scenarios described later-on. Figure 21.12 shows that the coefficient of variation for households and jobs is higher after introducing new policies in 2015. This shows that the effects of the policies are more than model's inherent variability. The plots also show that simulation outcome spreads throughout the simulation period.

21.4.2 Scenario 1: road pricing

In the road pricing scenario a cordon toll for the city of Zürich is assumed. The toll is set to 6 Swiss Francs (5 Euros) and is applied to all travelers crossing the borders of the municipality of Zürich on a link towards the city. It is applied from 6 o'clock in the morning until midnight. The cordon pricing is introduced in simulation year 2015, through the use of a different configuration file for MATSim which specifies the tolled links, the monetary amount, and the time of day.

One way of validation is to look at the dynamics of the main indicators. Figure 21.13 shows three indicators for the persons crossing the cordon (left-hand side) or living within the cordon (right-hand side). The upper left plot of Figure 21.13 shows the mode share for traffic crossing the cordon. This mode share shift towards the use of public transport when cordon pricing is applied.

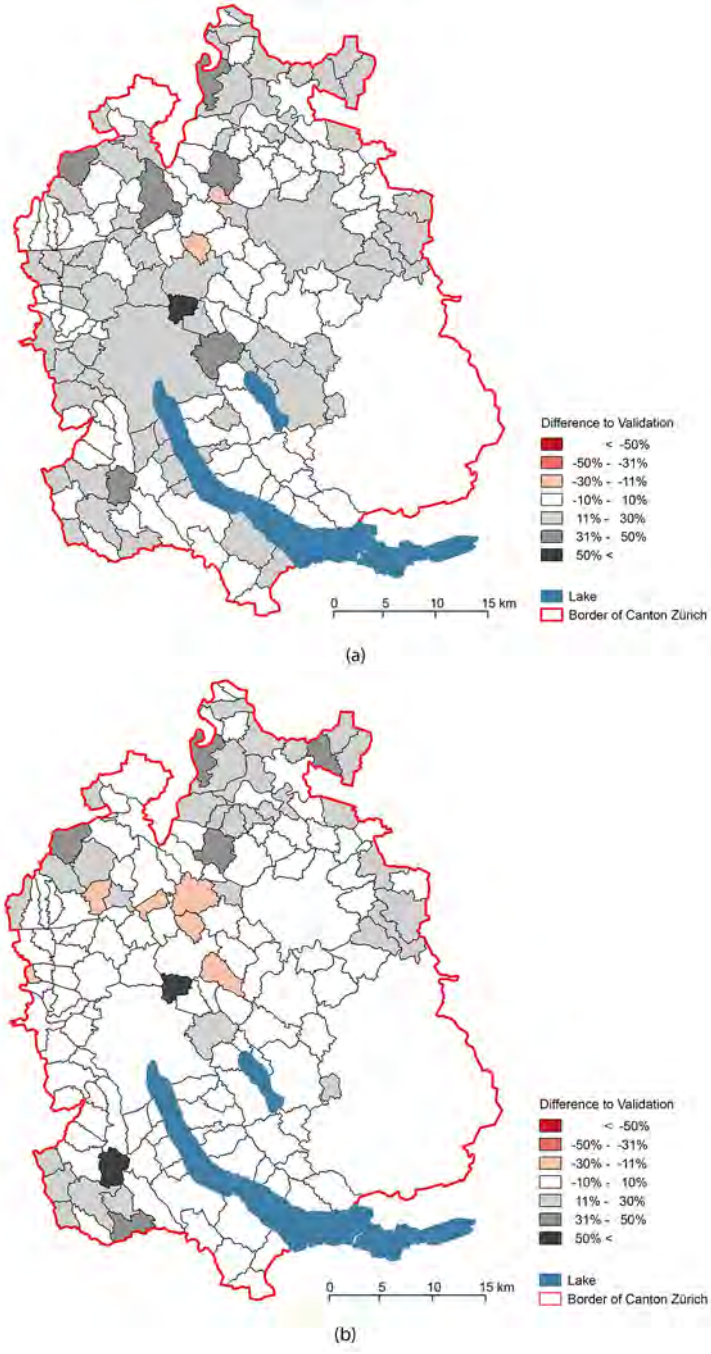


Figure 21.10 Relative deviations of buildings (a) and living units (b) per municipality after eight years of simulation (©2010 Amt für Raumordnung und Vermessung (GV2154) and ©2011 Swisstopo (JA100120)).

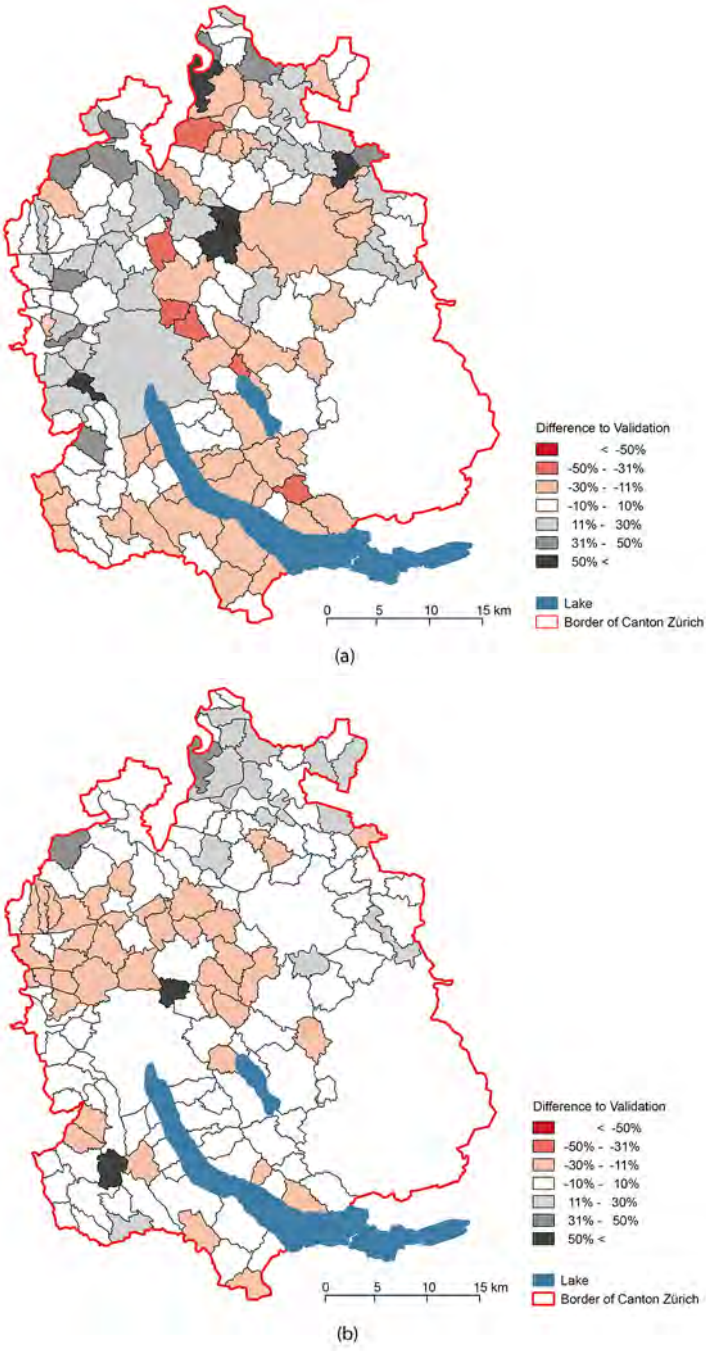


Figure 21.11 Relative deviations of jobs (a) and persons (b) per municipality after eight years of simulation (©2010 Amt für Raumordnung und Vermessung (GV2154) and ©2011 Swisstopo (JA100120)).

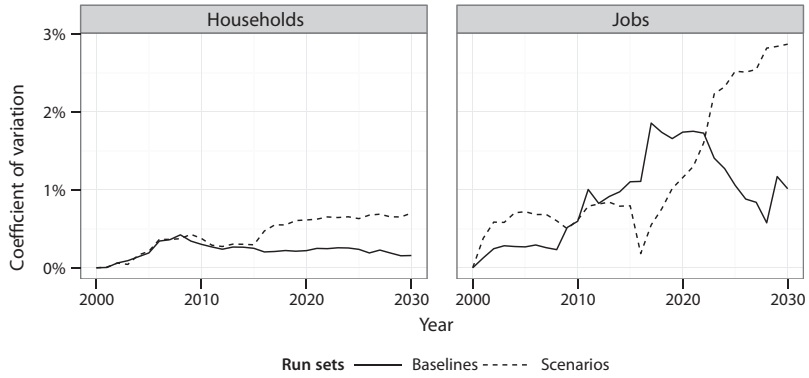


Figure 21.12 Yearly coefficients of variation of a run set of baselines and of all scenarios (including baseline).

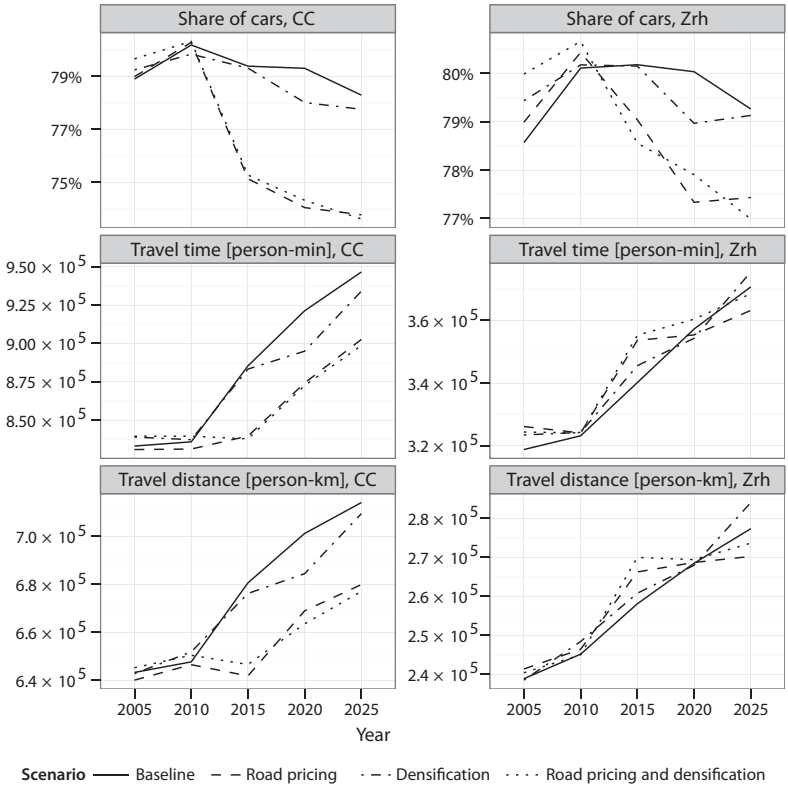


Figure 21.13 Car share, travel time and travel distance for cordon crossing traffic (CC) and within-cordon residents (Zrh).

Consequently, there is a reduction of travel time (middle left) and travel distance (lower left) by car in cordon crossing traffic. Interestingly, we find an opposite effect for residents of Zürich for which car usage and consequently travel time (Fig. 21.13, middle right) and travel distance (lower right) increase with pricing applied. Car share reduces further because the simulated agents do not face additional costs in public transport whereas travel time costs on the road increase with population growth (upper right plot).

The reaction on the land use side is twofold (Fig. 21.14). Households avoid locating inside the cordon area (–0.7%), whereas jobs do the contrary and tend to locate more inside the cordon area (4.9%). A differentiation of the different sectors in this overall tendency has not been done.

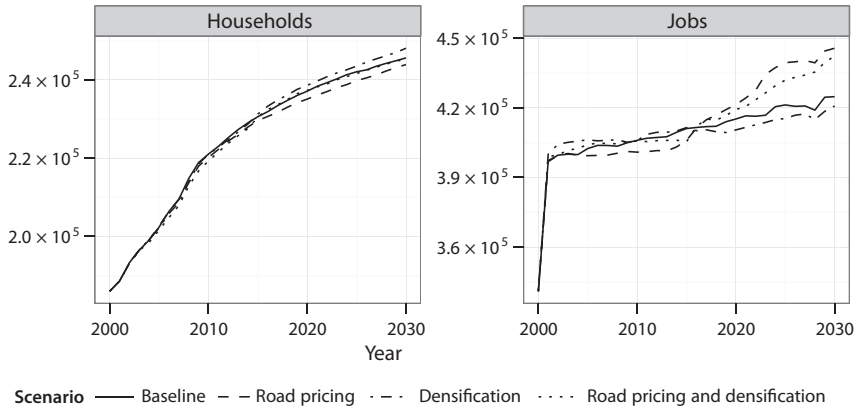


Figure 21.14 Number of households and jobs in the municipality of Zürich.

21.4.3 Scenario 2: densification

A recent referendum in the canton Zürich expresses a common disapproval of urban sprawl. The acceptance of the *Kulturlandinitiative* (Staatskanzlei des Kantons Zürich 2012) forces the administration to reduce sprawl and lead towards densification of settlements. The translation of this limit on sprawl into spatial regulations is not yet specified in detail. However, the claims are precise and demand the protection of all natural resources and cultivated land. The expected population growth has to be allocated in existing building zones which imposes a densification of settlement areas. Our densification scenario uses eleven zones that are marked as central settlement areas in geodata of the cantonal *Richtplan* and implies an increase of allowed floor area in 2015. The zones for densification are depicted in Figure 21.15. They contain 11.4% and 32.5% of the households and jobs in the study area respectively.

The densification scenario is implemented by providing the simulation with an alternatively specified parcels table in 2015. The plan types of the parcels

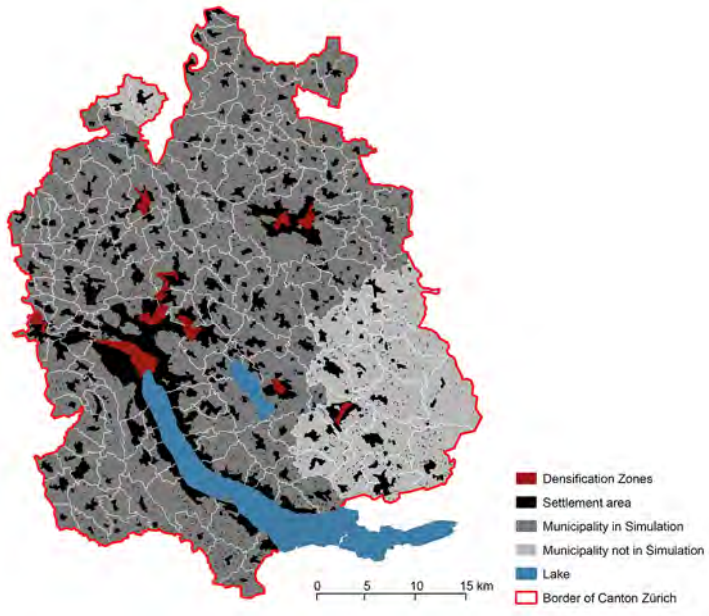


Figure 21.15 Zones with increased allowed floor area ratio (©2010 Amt für Raumordnung und Vermessung (GV2113) and ©2011 Swisstopo (JA100120)).

located within densification zones are given new plan types which allow for the same use but higher floor area ratio. The measure increases the remaining floor capacity of the study area by 5.3%.

The impact to the simulation is depicted in Figure 21.16. There is a clear peak of newly constructed buildings inside densification zones in 2015, but we also note a saturation effect: capacity for new built space is filled up and new construction inside densification zones reduces again. In consequence, we also see an increase in the number of households locating inside the zones of densification. There is no reaction visible on the jobs' side which is probably because of higher non-residential vacancies still allowing jobs to locate elsewhere. With the current simulation it is not possible to evaluate where built space grows in the third or in the second dimension because it only controls for densities.

Travel time increases more rapidly in scenarios with densification (Fig. 21.17). The concentration of households leads to more congestion and thus longer travel times. Another reason for the increase in travel time is the mode share shift towards public transport. A comparison with the baseline scenario shows some delay because persons first have to move to densification zones before more congestion occurs (Fig. 21.13, Share of cars, Zrh). In the workplace choice model persons choose their job based on euclidean distance only, which is a bias with consequences: if persons would choose their jobs depending on travel time to

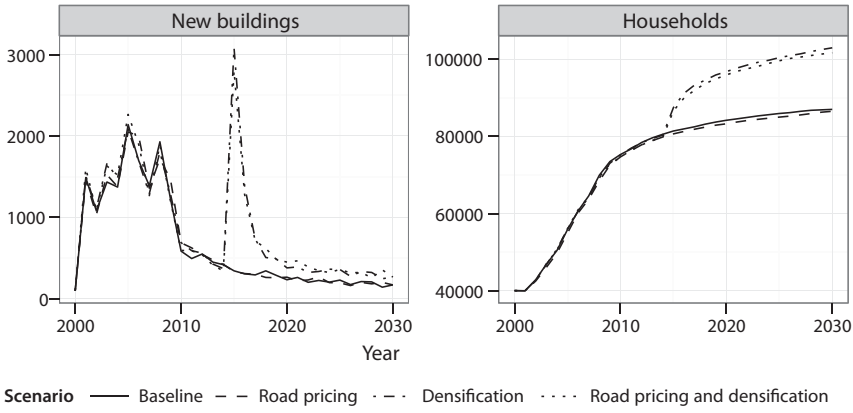


Figure 21.16 Number of new buildings and households in densification zones when measure gets effective.

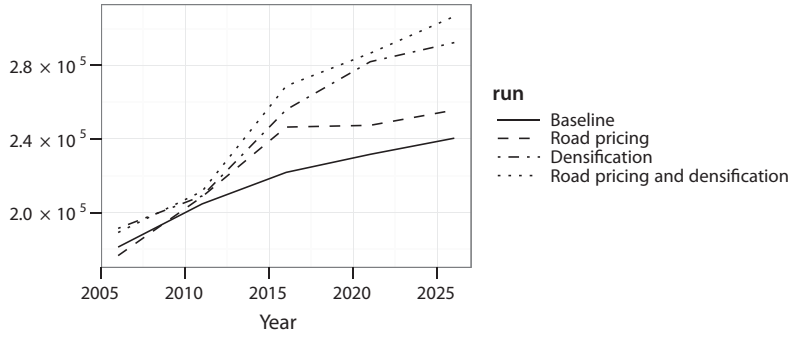


Figure 21.17 Travel time for residents of densification zones.

that one, we would probably see more persons adapting with job switching.

21.4.4 Scenario 3: road pricing and densification

The third simulated scenario is the combination of the two previous ones. With the specification at hand we apply both policies to the municipality of Zürich. The area is subject to cordon pricing and increase of floor area ratios.

For cordon crossing traffic the simulation shows a reduction of travel time by car in comparison to the baseline. This reduction has a similar dimension as the road pricing scenario (Figure 21.13, Travel time [person-min], CC). The same can be observed for traveled distance by car (Fig. 21.13, Travel distance [person-km], CC). We note similar but slightly stronger effects only considering travelling of residents of Zürich (Fig. 21.13, right-hand side).

The number of households in the city of Zürich, where the pricing and densification policy apply, is very close to the baseline (Fig. 21.14, Households). The effects of more expensive travel to Zürich and the increased number of living units balance each other. Jobs, however, seem to be somewhat repelled by densification but definitely more attracted toward the cordon area (Fig. 21.14, Jobs).

21.4.5 Scenario effects in 2030

The quantitative effects of the policies are shown in Table 21.15 in difference to the baseline scenario. Starting with the road price scenario, we compare the main indicators of the policy runs to the baseline and list the relative deviation in percentages.

Road pricing scenario

These quantitative analyses show that the policy effects are higher for cordon crossing traffic than for traffic of within-cordon residents. The reason is that commuting inside the study area is not effected by the cordon pricing. However, the use of car also decreases here (2.4%) in comparison to the baseline and leads to less travel distance (−1.9%) and travel time (−1.7%) by car.

Variable	Road pricing	Densification	Road pricing and densification
Travel indicators			
Travel time in study area	5.1	0.3	4.3
Travel time in cordon crossing traffic	5.3	0.3	4.6
Travel time of inhabitants of densification zones	10.2	24.0	30.2
Distance travelled by car in study area	-1.0	-0.7	-1.9
Distance travelled by car in cordon crossing traffic	-3.2	0.0	-3.5
Distance travelled by car by residents of Zürich	-1.9	3.1	-1.4
Travel time by car in study area	-1.3	-1.0	-1.9
Travel time by car in cordon crossing traffic	-3.5	-0.2	-3.6
Travel time by car by residents of Zürich	6.9	2.8	-0.6
Car share in study area	-4.0	-0.6	-4.1
Car share in cordon crossing traffic	-4.5	-0.2	-4.5
Car share of residents of Zürich	-2.4	-0.2	-2.7
Land use indicators			
Number of households in Zürich	-0.7	1.0	-0.1
Number of jobs in Zürich	4.9	-0.9	4.1
Number of living units in Zürich	-0.6	0.9	-0.1
Number of households in densification zones	-0.6	18.3	16.8
Number of jobs in densification zones	1.1	0.8	2.3

Table 21.15 Relative deviation [%] in main indicators of scenarios compared to baseline in year 2030.

We note an overall increase of 5.1% in travel time in the road pricing scenario, for cordon crossing traffic even an increase of 5.3%, which is caused through a shift towards public transport use. This fits our expectation as the road network does not change and a high number of new jobs are located in the study area. In consequence the network suffers from additional traffic loadings and congestion. This also finds its expression in a decrease of car traffic (–4% respectively –4.5%) and a reduction of travel distance. We see a large difference when comparing travel distance and travel time by car for the study area and the one for cordon crossing traffic (reduction of about 1% for study area and 3.5% for cordon crossing). Road pricing has a stronger effect on the share of cars than on the distance traveled. We conclude that primarily short car trips are replaced with public transport. As only commuting trips are represented in our transport model, we would expect that the effect will be even stronger when including leisure trips which are easy to replan. Due to the additional monetary costs imposed on car usage across there is a mode share shift towards public transport use. The effects in travel time and traveled distance are largely a consequence of that shift.

The effects of road pricing on land use are a reduction of households inside the cordon area (–0.7%) combined with an increase in the number of jobs (+4.9%). Through the reduction of car traffic, congestion is reduced, so that we observe road pricing to increase the workplace accessibility inside the cordon area. This reduces the probability for choosing these locations for households with cars, as the variable for car accessibility has a negative sign in the household location choice model. Jobs have a positive sign for car accessibility and are consequently attracted.

The same effect can be found for living units. The building location choice models of single family buildings and multi family buildings include negative estimates for car accessibility (Table 21.5) resulting in a lower construction activity (–0.6% in the city of Zürich). The opposite effect can be observed for job location choice models. Since car accessibility of workplaces has a positive impact in the job location choice models we see an increasing number of jobs locating inside the cordon area. The increase of car accessibility can thus be put down to two effects: on the one side more jobs relocate into the cordon area, on the other side we find a reduced usage of cars within the cordon area.

Densification scenario

Comparing the scenarios of road pricing and the baseline with the densification scenario offers significant differences in the simulation outcome. The number of households is strongly increasing in the zones of densification (+18.3%) through the availability of more living units in these zones. Surprisingly, this does not happen to the same extent for the municipality of Zürich (+1%), although this is also a zone of densification. A reason might be the higher accessibility in Zürich that reduces the utility for living units (see Sect. 21.3.4). We note here

that we were not able to segment the population. The average tendency hides the segment searching for urban and denser living very evident in the recent upswing in Zürich's population.

The overall travel time of the densification scenario is slightly increased (0.3%) in comparison with the baseline. This is in accordance with a reduced car share (–0.6%), but not substantially different for the city of Zürich (–0.2%). The travel time by car is slightly increased for residents of Zürich (+2.8%) which indicates that densification leads to higher congestion. However, the travel time inside the densification zone increases by even 24%, most of this one is probably added by additional inhabitants causing more congestion. Partly this might also be put down to the increased use of public transport systems, that are slower in comparison to today's car speeds.

Road pricing and densification

When looking at the combination of both policies the municipality of Zürich is of special interest because both measures apply to this area even though densification also applies to other regions. The reduction of car share is higher than in any other example (2.7%). Notably, the distance traveled by car is reduced for city residents (–1.4%) which is less than with road pricing alone. The reason for this outcome is the higher amount of households located in Zürich which produce more congestions.

Concerning the land use indicators we see that the effects of road pricing and densification cancel each other out in respect to the number of living units and households in the city of Zürich. Meanwhile the relocation of jobs is dominated by the road pricing effect and results in an increase of jobs in the city of Zürich (+4.1%).

Comparison of scenarios

Comparing the effects across the scenarios we see that the combination of road pricing and densification reduces travel distance by car most (–1.9%). Acknowledging, that car travel distance is an indicator for emissions of pollutants the combination of both measures is most advisable from the ecological point of view. Travel time in the study area is higher than in the baseline for every scenario due to a shift of mode share to public transport. The car share indicates that mode share shift is after all a consequence of road pricing.

Similar to households we see the effects of road pricing and densification canceling each other out in respect to the total number of living units. The number of jobs in densification zones shows that the tendency of jobs to cluster reinforced most with the combination of road pricing and densification which contributes as well to a reduction of distance traveled. The opportunities in terms of dedicated floor area offered by higher densities are taken advantage of.

21.5 Summary

The case study presented combines three agent based simulations on the level of individual persons: a demographic development model, a transport model, and the land use simulation. These modules interact by using the same set of agents, which allows the maintenance of a consistent level of detail throughout the simulations.

With the use of parcels as primary spatial units, also the spatial level of detail is very high, making this case study the first parcel based implementation of UrbanSim in Europe. As parcels form the object of regulations and are object to land ownership and development, this approach is to be favored over a gridcell or zonal representation of spatial locations.

Furthermore the availability of the Federal Building and Dwelling Register allowed to model the residential locations on the level of single residential units, i.e. dwellings and single family houses. This disaggregation was to be highly advantageous for the modeling of hedonic prices and the residential location choice. To the authors' knowledge this is the first case study of UrbanSim to implement that level of detail.

The drawback of such a disaggregated simulation is the exponential increase of work in data processing. While aggregated data is mainly homogenized by administrations, disaggregated data has to be collected from various sources and lacks of consistency across the datasets. The use of PostgreSQL as database allowed to script the necessary data processing and proved to be of high value: the quality of the simulation data could be increased in an iterative work flow with a simultaneous setup of the modeling and simulation. Furthermore a scripted processing of data forms a good documentation of the work undertaken. The data preparation resulted in a base year of relatively high quality. Only the distribution of jobs demanded a high level of imputations and corrections, due to headquarters reported in the enterprise census, that do not match the available building space.

A model can be estimated inside and outside of UrbanSim. The base year of the simulation represents cross sectional information, i.e. the distribution at a given point in time. Instead of using this distribution, the estimation can be performed on survey data, e.g. on recent movers. The integration of these surveys into the simulation environment can be complicated if not impossible (see Sect. 21.3.1) and can enforce the estimation outside of UrbanSim. To verify the validity of the resulting model inside UrbanSim, we estimated a model with the same specification in UrbanSim and compared the models. If a model misspecification is detected, the entire process needs to be repeated. While this double estimation is a possibility to find errors, it is very time-consuming due to its iterative nature. Therefore, if the estimation data was available inside UrbanSim, model estimation has been performed inside UrbanSim as well.

The models in this simulation do have different levels of quality and a differential impact. While the workplace choice model is very simplified and will need further improvements, it has only an impact on the residential location choice of relocating persons. For longterm simulation the model having the strongest impact is expected to be the land development model that provides new space and thus has an impact on the employment location and residential location choice options. However, all models included have limitations and will be further improved in ongoing work.

The networks, parcels, and regulations are expected to stay constant throughout the simulated period, unless the changes are part of a scenario. It would be possible to implement changes through transition probabilities, e.g. to include a change of zoning regulations. However, in order to allow the analyses of policies, this seemed not an appropriate approach.

Four scenarios have been tested in this case study: the baseline without any intervention, the road-pricing scenario as cordon toll, the densification of center areas, and a combination of road pricing and densification. The integration of these scenarios could be realized with small effort, due to the database used as back-end. Further scenarios on new zoning definitions, new network links and new developments of population and jobs can be included in the same way and will be part of the ongoing work.

The baseline scenario is compared with evaluation data available on the level of municipalities after 10 years of simulation (2010). Although we have high relocation dynamics, the simulation outcome proves to be a good representation of urban dynamics in our simulation. This is true especially for the larger municipalities in the simulation, while very small municipalities have a larger percental deviation due to the large impact of minor changes. We find the number of persons in a municipality to be generally simulated more precisely than the number of jobs. We expect that this is an effect of the data processing needed for the jobs (see Section 21.2.3). The simulation shows an overestimation of building stock, which can be traced down to a wrong vacancy definition and the smaller development units due to the limitation to monofunctional buildings. We can observe that this land development model has an impact on the residential and the employment location choices, as it provides the options for their relocation.

The policy scenarios are compared to this baseline scenario and show identifiable effects on the simulation that follow our assumptions. In the road pricing scenario, jobs are attracted by areas with road pricing whereas households avoid these areas. The densification scenario shows that allowing for higher densities in central zones leads to more real estate development in the affected zones and forms options to households and jobs to locate here. On the transport side the increase in density of activities leads to a mode share shift towards public transport while car travel is reduced. The combination of road

pricing and densification reduces car travel most. The policies cancel each other out regarding the total number of households relocating here whereas the attraction of jobs towards priced areas dominates the repelling effect of densification.

For evaluation all simulation data has been exported to the PostgreSQL database. The spatial distribution was analyzed in maps using *QGIS* and the aggregated information on a municipal level. Additionally longitudinal analyses and distributions have been analyzed using the software *R*. To take advantage of the agent-based approach, various alternative analyses could be performed on the simulation results with the same methodology, e.g. the distribution of certain job-types or social groups. However, as no validation data is available on these, we did not report on these findings in this paper and this will be the content of ongoing work.

21.6 Discussion

Although the simulation has reached a level which allows to clearly identify trends, the models' components and calibration will need further improvements before a fully reliable quantification of the effects can be achieved. Furthermore we haven't elaborated the robustness of the model outcomes: while we tested the scenarios in different runs and found comparable results, changing individual parameters in these runs will allow to quantify the effect of certain variables and models.

The models react to the transport simulation mainly through implementing the accessibilities of jobs. It would be of advantage to have the travel times and mode choice as outcome of the transport model for use in the land use simulation. As we ran the transport model only on a 10% sample of the population and only every 5 years, this was not feasible in the current case study.

Currently, the simulation comprises only a residential rent price model. Considering additional sub-markets would be beneficial for the simulation, as well as including a tenure choice model. A present model for office rent prices was not included, as these are expected to be strongly varying, depending on the contract (Haase 2011). Furthermore we expect that these do have less effect on the job location choice (Bodenmann 2011). Meanwhile further work will have to include tax rates in the employment location choice.

Spatial allocation of jobs is modeled in a job location choice model which locates jobs individually. An improvement to this model will be the grouping of jobs into firms through a new dataset. This will allow the representation of characteristics of the enterprise and its different job types. In this realm not only location choice can be improved but also the evolution of the population

of firms. A more complete model of firmography thus can consider creation of firms, their growth, relocation and disappearance.

Relocation decisions are modelled by applying rates for certain categories. Up to now these do not rely on individual events as the birth of a child in a household or the formation of a couple, while our survey suggests that these are often the reason for relocation.

Besides the amount and distribution of agents and object, identically and transparently defined indicators can allow a comparison of scenario outputs. A first step in this directions is the indicator module applied (cf. Chap. 15). Local guidelines and policies (Bundesamt für Statistik BfS 2009) which are politically accepted give the direction towards which the indicator set can be expanded, but they have not been used for the evaluation reported in this paper. Unfortunately not all the attributes of this one have been evaluated in the simulation and some approximations needed to be done.

In terms of usability as a decision support system, we demonstrated the setup of four different scenarios. Defining scenarios as these, i.e. the change of regulations, new network links or a change of population can easily be integrated, as the models do not need to be reestimated. The outcome of such a simulation is of high value for policy decision on large scale projects, as these mainly rely on these kind of scenarios and have a very long time horizon.

Theoretically the resolution of the simulation would allow to analyze any scale from a building to a region. However, our simulation does not include a setup to analyze the outcomes on a parcel level and we can question whether this is possible at all. This especially applies to the land use development, that relies to zonal definitions of allowed densities and does not include specific information on the scale of parcels, e.g. building lines, position of garages, or the distance to a neighbor. Besides, such a high level of resolution would need to include further aspects in the location choice models, e.g. the spatial quality of the public space, access to open areas, or the position of trees. All these are not part of our simulation framework at present. Meanwhile the advantage of the agent-based approach taken, is the analysis of individuals decisions and their interaction in a complex system.

This case study also showed, that the setup of the base year is the biggest workload for such an agent-based simulation and that the data needed is facing important privacy constraints. It seems impractical to set up such a simulation for short term projects, unless they are simplified and data becomes more easily available. A well maintained and documented data warehouse would help a lot.

Although the data preparation has been scripted, our case study has a strong data dependency on data that are specific to the context of Zürich, which prevents the direct implementation in another study area. While individual data is needed for model estimation, the simulation might be setup easier using synthetic data (cf. Chap. 5).

Part V

Conclusion

Chapter 22

Future challenges in transport and land use modeling

André de Palma, Michel Bierlaire, Ricardo Hurtubia, and Paul Waddell

22.1 Introduction

Researchers in transport and urban planning have historically worked independently. For example congestion was ignored in many urban studies, and conversely, long-term induced demand from transport projects was ignored in many transportation plans. Fortunately, more and more researchers and decision makers now understand the need to model land use and transport in an integrated way rather than in isolation. Transport and Land Use Interaction (LUTI) models have been developed for this purpose.

Transport models require inputs such as firm location and residential location, which are generated by land use models. Similarly, land use models need updated information about traffic conditions for private and public transportation in order to compute accessibility, a fundamental explanatory variable in residential location and firm location models. The main aim of SustainCity was to use an integrated approach, which takes these essential interactions into account.

In this project,[58] considerable effort was made to not only integrate the UrbanSim land use model with the MATSim and METROPOLIS travel models, and to use the integrated models to evaluate sustainability scenarios in Brussels, Paris, and Zürich, but also to consider technical innovations that might improve the combined models, and to consider adaptations to the European context. Although the initial expectation was that such adaptations

[58] We thank all the participants of the SustainCity project. This last chapter is a true outcome of collaborative research. Many of the ideas and part of this chapter were produced collectively during the project, and appeared in several SustainCity working papers, available online at http://www.sustaincity.org/publications

would require the implementation of a new platform, it became clear during the progress of the project that the adaptations needed for representing European peculiarities could be implemented effectively within the existing UrbanSim software platform, as discussed in Chapter 16, so there was in the end no need to develop a separate platform.

Are urban systems predictable?

The reader may wonder why it is possible to model mathematically such complex systems as urban systems. One reason is that there exists strong regularities in the internal structure of the city, as well as in the spatial distribution of cities over space and time.

Strong regularities concerning urban development have been documented for several decades, especially by geographers. A scaling law, for example, specifies by how cities' characteristics increase or decrease with the city size. There exists a scaling law between city size and (a) energy consumption, (b) level of crime, (c) revenues, or (d) urban sprawl (urban economics literature studies for example by how much, on average, urban sprawl increases with city size, Fuller & Gaston 2009.) These strong correlations do not imply, but suggest causalities. At a lower scale level, there are regularities such as the constant share of travel expenditures in total revenues (Schafer 2000). It is not surprising to find also regularities at the scale level of the system of cities: the rank-size rule, which relates the population size of a city to its rank in the country's or world's ranking of cities by population is one of the most stable and robust law in Social Science. Zipf's law, for example (also known as the law of minimum effort) postulates that the city growth rate is independent of city size so that the distribution of city sizes remain Pareto over time (see Gabaix 1999*a*,*b*). Other laws explain the growth rate (or the decline rate) and scaling of cities (Bettencourt et al. 2007), their networks and supply systems (Kühnert et al. 2006).

22.2 Setting the stage

The new ways of modeling large urban systems is in fact grounded in a tradition of theoretical reasoning, the main thrusts of which are briefly outline below. This is not an historical approach (see Lafourcade et al. 2011, Thisse 2012), but such large-scale transport-urban models can be used to reconstitute the pathways describing, explaining, and forecasting the growth and evolution of large cities (de Palma & Guimard 2013).

From an agricultural economy to an urban economy

Some short history may help the reader at this point. A number of the key concepts have been examined in works published since the beginning of the 20th

century on spatial organization in connection with farming operations and the distribution of urban hierarchies (von Thünen 1826, Losch 1940, Christaller 1933). von Thünen uses a system of concentric rings to describe the mechanisms of land allocation through a bidding process (land being allocated to the highest bidder in a context of perfect competition). Lösch begins with an isotropic plane in which self-sufficient entities maximise their accessibility to certain goods and services, and then explore the gradual organisation of the circulation of goods, services and persons. These models, although very simple, are at the root of urban economics.

The monocentric city

The application of these ideas of bidding and optimisation to the framework of urban economics can be associated with Alonso (1964), who (along with Richard Muth and later on with Edwin Mills) helped establish the monocentric model of the city, in which all workers travel each morning to a single working location known as the CBD or the central business district (see the review and references in Fujita & Thisse 2013). This model, which assigns jobs to the city centre, therefore describes only the residential choices of households. Monocentric cities are idealized concepts, but nonetheless are still useful for understanding household location and urban decentralization. Many cities, such as the *Ile-de-France* area, are better represented by polycentric framework, which are more effectively examined with computational models such as UrbanSim.

Short term versus long term

Transport and urban development do not evolve on the same time scales. Recent literature shows, for example, how the introduction of cordon pricing in le-de-France can affect congestion in the short run, and in the long run may influence the spatial structure of the city. The simple monocentric approach with automobile traffic congestion was explored by De Lara et al. (2013). Their model was calibrated with the data available from le-de-France, and able to quantify urban contraction with various types of road pricing. The same problem was studied in Brussels, with UrbanSim and MATSim. Simulations were conducted over the short term (holding household residence constant) and over the long term (allowing residence to vary). Such studies show that pricing effects in the short term are about half of those in the long run. This figure provides only a rough approximation, yet it indicates the utility of factoring in the relocation processes of households and businesses.

Incorporating major policy choices entails detailed modeling of transport systems as well as business relocation modeling (which are disregarded in the monocentric model but included in the polycentric, more complex models).

This militates for the integrated transport/land use models discussed in more details later on.

In a more elaborate model, it is necessary to introduce the spatial externalities that affect households and businesses. Symmetrically, the mobility decisions of households and businesses effectively alter (and are affected by) density values and thus agglomeration economies and other spatial externalities. These externalities, which are difficult to quantify (except for congestion), are to a great extent disregarded by too many theoretical and applied alike models.

We should stress that these temporal dynamics and spatial externalities are important in the construction of integrated land use/transport models, and useful in the interpretation of their results.

Spatial competition. Monopolistic and Oligopolistic competition

The city and its markets for real estate, transport, and labor is far from a textbook representation of perfectly competitive markets. Urban economic models have long disregarded competition with differentiated goods and services. And yet, imperfect competition was factored into an equation nearly a century ago by Hotelling (1929). His highly simplified model was explicitly spatial, although the beach on which his ice cream vendors moved about could be interpreted (as was pointed out by Hotelling himself) as a range of quality differentiation amongst products.

Product differentiation modeling has found an empirical counterpart in discrete choice models, the estimation of which is now operational. This approach achieved clear-cut success in residential location choice applications: see Anderson et al. (1992) who advocate the use of discrete choice models in the transport, and industrial organization. At the same time, differentiation concepts have come (in a modest manner) into their own in the field of urban economics thanks to general equilibrium approaches (see the seminal contribution of Anas 1982). Despite this, initial differentiation-based approaches, which have become central to a majority of applied urban economic models, largely disregard market power and market distortions, monopoly power, entry barriers, and remain closed to the ideal world on perfect competition, helped by the action of the invisible hand. Imperfect competition, largely absent from urban economics, entered in the form of monopolistic competition (an intermediate situation between perfect competition and a monopoly situation) as introduced by Chamberlin (1933) and analyszed by Dixit & Stiglitz (1977). These ideas were taken up later on by the new geographical economics à la Krugman and Fujita (2001), as well as with the key concept of increasing returns to scale.

This description includes enterprises that have only slight market power, and it is well suited to a great number of small businesses, although less so to

a number of large enterprises, each having local and non-local monopoly power (see also our discussion on firmography, below). Whether this approach can be used to describe competition between businesses within a major metropolitan area such as Paris, London, or New York is open to question. We feel that, in a sense, the choice of monopolistic competition is tantamount to overconfidence in the market. Several aspects of political economy, lobbies, and political forces are ignored. All these issues remain empirical as well as theoretical, and new generations of researchers will have to handle them.

Lastly, competition between large metropolitan areas, which is often forgotten by the urban economist, is probably another instance of imperfect competition. It would be grossly oversimplified to assume that competing cities such as London, New York, Singapore, or Paris constitute uniform options for multinationals wishing to establish their headquarters.

Definition of the city and the geography for modeling

The project focused on three cities: Brussels, Zürich, and Paris. However, our ambition is to apply this integrated model (UrbanSim-Metropolis-MATSim) to other cities. For that, it was necessary to have as much as possible an agreement on a common definition of what a city is in order to compare the results and the policy insights. This is a difficult task also related to the geographical scales (hierarchy).

According to Mumford (1938), a city is 'a point of maximum concentration for the power and culture of a community'. Marshall (2007) recalls the definition of urban areas by the US Census Bureau: 'The US Census delineates urban areas as areas based on the population density of a census block or block group being greater than 1000 people per square mile and the population density of surrounding census blocks being greater than 500 people per square mile (US Census, 2004). Urban areas have a population size of at least 50 000 inhabitants. Urban areas are delineated at the start of each decade' (see de Palma and Guimard, 2013).

In France, continuity of settlements and concentration of individuals are the key figures of cities. According to the French bureau of statistics (INSEE), 'an urban unit is a municipality or a set of municipalities presenting a zone of built up continuous (no cut of more than 200 meters between two constructions) and with at least 2,000 inhabitants'. If the definition of cities varies over space, if also varies over time. It has changed, for example in France with the introduction of new concepts, such as urban areas. In 1997 INSEE introduced the concept of urban areas which allows remote settlements to be part of a larger urban core. However, since the core must have a large labor market (10 000 employees for large urban areas) several remote settlements are still excluded because they are not included in an urban core big enough as far as the total employees.

The proportion of urban areas is at its maximum level even if more remote 'rural' areas were connected to urban cores (peri-urbanization). One major issue of SustainCity, was to use the appropriate definitions of space. Although UrbanSim has been used widely in the U.S., Europe, and Asia at a variety of geographic levels, the emerging standard in the U.S. is to use individual land parcels, which are the actual unit of land owned by a firm or household, and on which one or more buildings are located. Other, earlier projects have used gridcells or zones of differing sizes, but the most widely used geography for UrbanSim in the U.S. is the parcel. In the European context, data availability varies, and of the three sites to which UrbanSim was applied, only Zürich used the parcel geography. In Paris, a zonal configuration was used that was based on the Commune administrative geography, and similar choices were made in the Brussels application. As parcel data becomes more widely available in Europe, and researchers gain more access and experience in working with this level of data, it is likely that parcel-level modeling will be the preferred geography.

22.3 Collecting and processing data

We have met the main objective of this project: elaborate a LUTI model perfectly adapted to the European specificities, and apply it to three case studies (Brussels, Paris, and Zürich). We have developed and run UrbanSim with specific configurations consistent with the European context. It has provided an operational tool for policy evaluation. Several lessons were learned. The first major task was to collect and process the necessary data. This was a very substantial task, which needed the collaboration of the different governmental and academic partners.

Data requirements

One major issue in the models developed in SustainCity was data acquisition and validation. We all know that a major issue in transport and urban modeling is data collection, storing, and updating. Note that the situation is changing very rapidly, since data are collected and stored more and more easily (via GPS positioning systems, mobile phone, etc.). The use of disaggregate information available via Web should be considered as major future development challenges. This revolution in data availability will undoubtedly change the way models are built and used, though academic use of these data for integrated modeling has much to gain from recent advances in the use of such data for business applications.

We were able to overcome usual difficulties in data collection and processing since we could (1) update and run the models in spite of imperfect (or incomplete) data (2), share data between different models, and (3) develop a

common set of MOE (Measure of Effectiveness) to compare the results of the different models. These measures are based on effectiveness, equity, individual costs, social cost, and external effects. They are either individual specific or aggregate (geographical or temporal aggregation).

Even though the SustainCity consortium was very well designed, it still became apparent that data was a major challenge. While it was still possible to develop these models for three large and well-developed European cities, it is still open to consider how one might be able to develop similar, 'data-hungry' models in other regions of Europe, where data are harder to collect.

Integrating some of the advanced econometric models (such as spatial models or modeling of intra-household decisions) that were estimated outside of UrbanSim in this research, into UrbanSim will provide a considerable boost for Land Use Transport Interaction modeling in the future.

Two major limitations of our research deserve attention. The first one is related to the lack of relevant data, especially panel data, which made it impossible to estimate and implement the complex dynamic models which could explain, measure and predict life cycle effects in car ownership or residential location decisions, inter alia. The second one is that some specific econometric models (such as modeling of intra-household decisions) that were estimated outside of UrbanSim are not yet implemented in UrbanSim due to the limited time and budget to accomplish this within the current project. As these innovations are evaluated for potential to impact the outcomes of integrated models, such experimental models may be integrated into UrbanSim over time.

Estimation

One element of the project was to provide econometric guidance to the Case Studies, in order to achieve consistent results across them. Even though the nature of each case study is very different, in terms of city characteristics, level of aggregation for running UrbanSim (i.e. zone or parcel), and availability of data, the analysis provided a set of imperfect standardized views, which were used to present the model results. Following these templates, it is possible to have a coherent view of the outcome of the case studies. This becomes particularly important in the context of developing policy assessments based on these outputs. It is expected that these 'standardized views' and the experience gained through this process may be valuable beyond SustainCity.

Furthermore, a number of 'lessons learned' from the runs of the case studies have been formulated, which can both assist others in designing and performing similar studies in a better way, but also prevent them from making some of the more common missteps. One of the main findings of the econometric modeling aspect was that considering spatial models has a great potential in improving the accuracy of land use and transport interaction models. Preliminary results provided encouraging results.

Computational issues

MATSim

Concerning the integration of UrbanSim and MATSim, coupling Java-based with Python, C and C++ based software packages remains problematic. Quite some time was invested in order to achieve an architecture that is still less than optimal in particular, it is impossible for simulated 'persons' making choices in UrbanSim to query the MATSim model directly since the models do not share objects in memory and there is no API implemented at this time to enable such queries across computers. Since there is also no clear reason to move the whole community either to Python, C and C++ or to Java, this means that there will be separate streams of software that will remain difficult to tightly integrate for some time.

The computation of high resolution accessibilities proved rather fruitful, and it is advised to use this in the future, for various scenarios and also for various trip purposes. This approach should, for example, help the planners and the decision makers with the identification of regions which may be underequipped with certain services.

METROPOLIS

METROPOLIS, the dynamic mesoscopic simulator, needs two extensions: first the travel time for public transportation was exogenously given. This is a poor hypothesis if the congestion in public transportation is high (and varying), as it is the case in Paris. The Paris team is developing a dynamic public transportation model with an endogenous level of congestion. This tool will be essential in European applications, where public transport are very much developed. Moreover, heterogeneity is not well taken into account in METROPOLIS. To improve this situation, information about the joint distribution of the individual preference parameters is needed: value of time, unit cost for early arrivals, unit time for late arrivals, desired arrival time at destination. The focus of METROPOLIS was predominantly on commuting. This makes sense, since congestion affects mostly commuters and value of time is the largest for work related trips. However, this is less and less true over time. One the one hand, more people telecommute, use flexible schedule, work part time, etc. and on the other, trip chaining is playing a more and more important role, so that the frontier between commuting and non-commuting tends to decrease.

Further, we discuss the problem of integration of UrbanSim with a transportation model. This was easy conceptually, but appeared to be less trivial from the practical and operational points of view. This limitation is likely pragmatic, since UrbanSim has been coupled successfully in the U.S. with numerous travel modeling software platforms such as VISUM from PTV, Emme/3 from INRO, TransCad from Caliper, and a variety of others, including activity-based

travel models such as CT-RAMP and OpenAMOS, and dynamic assignment models such as DynusT and MALTA.

22.4 Agent behavior

Firmography

The Parisian team, which developed a firm location-birth-and-death model, focused mainly on modeling firmographic events in *Ile-de-France* including the differences by activity sectors (see Buczkowska et al. 2013). To describe the life cycle of the establishments, the authors propose a threefold firmographic model, which explains the disappearance, evolution, and location choice of the business units. They also compute the creation rates across various activity sectors and all counties of Paris Region to get possibly the most detailed overview of changes of the business units.

Comparison of the estimation results of three firmographic models show large differences in the establishments' behavior across different activity sectors. However, after re-estimating all the models separately for different establishments' size-types, the authors could draw the conclusion that the behavior of the business unit depends on its size. With the variables that we use to run the models, it turned out that the large establishments' behavior cannot be explained. Different approaches should be tested to analyze the behavior of big-sized establishments.

As for now, the behavior of the firm is not described in this experimental firmographic model, and the analysis is not based on a micro-economic model. Yet, firms are optimizing agents (maximizing revenue, profit, or market share, or a combination of those). This will be the challenge, needed in order have a real partial equilibrium model. The experience of general spatial equilibrium models will be useful as a guideline to conduct such research.

Household location models

The economics of the family has a range of impacts in the modeling of transport and residential-job location decisions.

The current version of UrbanSim distinguishes between individuals and households, but few models have been developed to exploit the availability of this rich microsimulation data structure. The theoretical model is one such innovation that considers the case of households with husband and wife (Chiappori 1997). The bargaining power of the household members is considered as an important factor for the decision of selecting residential location. The spouses are both active and have different residential location, which explains that each spouse may prefer a specific residential location. These authors show that it could have important consequences in economic analysis (for example,

neglecting the bargaining power within couples may change significantly the value of time). A new method has been suggested to calculate values of time in the context of couples. It is shown that age of the women as well as nationality of men play a crucial role in determining bargaining power.

Demography

A specific demographic model has been derived from an analysis of the state-of-the-art in demographic behavior microsimulation, based on an overview of the issues concerning the integration between demographic modeling of individual and household behavior, and urban development models. This module was built using the MODGEN language developed and maintained by Statistics Canada for creating micro-simulation models, as described in Chapter 7.

Many individual demographic events were simulated, based on annual transition probabilities exits from the population by death or emigration; entries by immigration or birth of a child; exits from the parental home, union formation, marriage, and dissolution; entry into a household (home mates) and exits from complex households; entries in and exits from communal establishments. All these transitions were dependent on sex, age, and specific additional covariates for which parameters estimates were produced; education attainment and labor force participation were also simulated, allowing using enrolment and working status as optional covariates for demographic behavior; micro simulation methods allowed using specific rules of mate matching.

This approach was developed as an independent module and used to simulate the evolution of the population and households in the Paris region, for which available data were found and used as a basis for estimating the parameters of demographic models. The main limitation of the approach taken was that it was a completely separate application from UrbanSim, and was run once for the full time period, and its results imported into UrbanSim. This prevented any opportunity to exploit events being modeled in UrbanSim, or the spatial context of households within the model, to influence demographic events in the demographic simulation model. Independently from this project, demographic event models have been incorporated into UrbanSim directly, which should provide more capacity to address these limitations.

Dynamics

Studies to ascertain the impact of major infrastructure have a 10-, 30- or even 50-year time frame. Long-term forecasting therefore plays an essential role here (see de Palma, 2012).

The main considerations are summarized below.

- Economic trends in terms of growth
- Shifting preferences: How could consumers' craving for utility vehicles have been foreseen?

- Changing technology. For example, what will the consequences of new information and communications technologies be on working conditions, mobility and agglomeration economies?
- Demographic changes, especially as shown by the long-term (from 30 to 50 years).

Factoring in the long term also requires projected rates of economic growth, employment, unemployment and so forth. In both cases, projections should ideally provide multiple scenarios providing inputs for the modeler.

As for modeling urban forms, different agents have different speed of adjustments (households, firms, developers, governments) and different levels of information. See Simmonds et al. (2013) for an extensive assessment of dynamics vs equilibrium assumptions in integrated models, and key arguments in favor of a dynamic approach.

De Palma, Proost and van der Loo (Chap. 14) studied the role of the equilibrium mechanism by developing a simple theoretical model. Their work investigate the role of anticipations in the housing market and the possible interplay with transport investments. The key agents in this model are the government, and the housing entrepreneurs (besides households, and firms). A simple two region model was used, where one region is subject to an unanticipated productivity shock that increases the demand for labor in that region. The demand for labor can be met by additional commuting, by more housing or by additional investment in transport. The results depend on the expectations of the developers, when there is no transport investment. Depending on the specific expectation of the agents (myopic, perfect foresight, imperfect) over or under investments in transport infrastructure has been observed. The model is a simplistic two region model with only one mode of transport. Potentially useful extensions include the introduction of regional governments that can tax income and commuting flows, decide on the building permits etc.

It is also worth further exploring the interplay between slow and fast dynamics. Many results have been derived in physics in this domain, and it remains to explore how such complex dynamic process can be applied in the context of urban modeling.

One approach that may hold promise is the use of regime switching models, such as the model of the real estate market by Maurin et al. (2013). This technique allows taking into account the existence of different regime which explain housing prices. The regime and the switching probabilities between the regimes have been determined endogenously. The study was based on data from Paris region. The proposed technique potentially enables better prediction of the distribution of the future housing prices. However, it remains to be seen, as stressed by Stef Proost and his team in Leuven, how these estimates change the behavior of the different stakeholders (households, firms, real estate developers) and the different levels of governments. Turning to real estate supply,

the work of Kay Axhausen and his team in Zürich has conducted in-depth interviews of developers, which could provide fertile ground for modeling. Zöllig & Axhausen (2012*b*) study the qualitative study of real estate developers in the canton Zürich. The working hypothesis is that heterogeneous developers are present and that their characteristics help to explain development events. The literature review concludes that there is a lack of knowledge in respect of the supply side of land development, that heterogeneity among real estate developers is shown in previous studies and that there is no established typology. Their paper tests the hypothesis that developer types do have an influence on the development events that eventually occur. The study was based on a survey where 11 real estate developers are interviewed. The authors found that the behavior of developers varies in terms of decision criteria, considered information base, search space and fulfilled tasks. Clearly, the understanding of the behavior of real estate agent is essential. In some cities, they are basically the one who set the tone for urban development. Other cities, more market-oriented, they follow more visible strategies amenable to quantitative models. These analysis require estimation of political economic model, based on discrete choice models. This is also left for future research.

22.5 Integration

Agglomeration and de-agglomeration

Ever since the classics of Christaller in 1933 and Lösch in 1940, a series of fundamental questions about the nature, the morphogenesis and the morphology of concentration of human activities and the regularities of central places has remained only partially answered. The growing literature on agglomeration modeling is fragmented, however. As the seminal volume of Fujita et al. (2003) on agglomeration clearly indicates, the complexity of this subject does not allow working on a single, comprehensive model at present. However, despite of its fragmentation, two themes run across the gallery of existing agglomeration models, with a few exceptions.

First, the space on which such models produce agglomeration is typically non-homogeneous in the sense that it contains features, such as boundaries, which implicitly order locations according to centrality or according to distance. Spatial heterogeneity, which permeates the real world, induces agglomerations that reflect the particular locational constraints imposed.

Second, the mechanisms underlying agglomeration apply to modern conditions rather than to conditions at the time when the first cities appeared. For example, models of the new economic geography, typically emphasize the role of market power, monopolistic competition, increasing returns in production and product variety on agglomeration. Firms share knowledge that are not internalized. A polycentrism pattern may appear if transportation costs are high

enough relative to the degree of knowledge spillovers (Fujita et al. 2001). But still, in all cities, as seen previously, the city center is denser, and population density decreases with distance from the city hall (the same appears for rents).

The concentration phenomenon of individuals and businesses is the outcome of two opposite forces: agglomeration forces and dispersion forces. Several comments are in order.

First, low transportation costs associated with high economies of scale encourage firms to be concentrated. A wider market access at the city core encourages firms' concentration.

Second, at the city core, firms benefit from agglomeration economies through the effects of sharing and/or matching and/or learning due to more input suppliers and a larger labor market. For mature cities, knowledge spillovers are crucial. As explained in Glaeser et al. (1991), knowledge spillovers are promoted by diversification (Jacobs 1961) or specialization of industries (see Romer 1986 and Porter 2011), and by either competition or monopoly. By contrast, competition, high rents or high wages all encourage firms to locate farther away. For households, agglomerations forces are related to high wages, the variety of the labor market, social amenities, the variety of goods and high transportation costs. The living conditions within cities are on average better than in rural areas, which may explain why people are often leaving rural areas for cities.

Third, with new and faster modes of travel, households have incentives to relocate farther away from city center. When the urban developments are neglected, developers expand the city in the surroundings which fosters urban sprawl. Market failures lead some researchers to think that urban sprawl is naturally not optimal (Brueckner 2000). The questions of teleshopping, and telecommuting are (too) slowly coming in the picture.

These interplays are very complex to study analytically, so that only simulation models, such as UrbanSim, which capture these complex set of behaviors, externalities, and dynamic can provide an essential tool to study the interplay of all these forces and to simulate agglomeration patterns.

Transport and land use

LUTI models integrate land use model and transportation, by definition. In the case of MATSim, integration with UrbanSim was achieved by linking the models at the person level, which required the generation of special code for both softwares to run in a coordinated way (see Chapter 17). In the case of METROPOLIS, we have opted for the solution of a loose-coupling interface. In this case there was no need to modify either of models and the interface was as much as possible independent of the future evolution of both of them. The data exchange is made by the intermediary of cache files that are directly read by two models. The independent interface makes possible to include future

developments in travel demand computations. The interface is developed in such a manner to be able to manage the models running on different remote servers.

The interface has three major tasks: (1) to manage the integrated simulation and data exchange between models as well as running them, (2) to compute the necessary O-D matrix to feed METROPOLIS from UrbanSim results for population and employment distribution, and (3) to compute necessary accessibility measures as needed in UrbanSim from METROPOLIS outputs. The interface is designed to take advantage of all possible micro-data available in modeling process, in particular the workplace of the individuals. In this case the O-D matrix for home to work and work to home trips is based on known workplace information.

Although achieved to some extent with MATSim, the exact procedure to update the accessibility in the land use model and to update the residential and firm location in the transport model remains an open issue. In particular, the optimal frequency of update of the information on congestion in UrbanSim needs more exploration, both at the empirical level and at the theoretical levels. For example, a rational agent, when buying a house, should be able to forecast its future price, which depends on the forecast of accessibility. This task is very difficult, and even the best experts would make mistakes when providing such forecasts. So agents resort to bounded rational estimates, which need to be better understood. Serious research effort should be devoted to this dimension of human behavior.

Land use and Demographics

The demography is an essential part of an urban model. Unfortunately, very little attention is devoted to this essential component. The task achieved in SustainCity was very preliminary, and much more effort should be devoted to integrate a demographic model to a land use model.

This project has made the first step in such integration. It allows tracking of the household moving and location behavior. Household member's age is updated in the model, so the aging of the resident population can be explicitly modeled. The lifecycle events affect the moving probability to obtain more realistic results. Last but not least, the household location choice history is taken into account, especially through a dummy variable (which happens to be the most significant in household location choice model) indicating that an alternative is located in the same county as the previous residence.

The demography model has been initially tested and calibrated for the Paris case study and is directly integrated in Paris case study simulations. The Paris case study estimations provided a substantial part of the insights in the econometric guidance. Firmography processes were implemented as new models in UrbanSim. Such integration is necessary, when assessing the costs

and the benefits of large-scale systems, such the Grand Paris project, which has to be evaluated over a period of several decades (50 to 75/100 years). In this case, it is no longer valid to assume that demography is independent of urban condition. There is a closed loop feed-back, which is currently totally ignored but needs more attention.

Consistency

Consistency plays an important role in modeling and evaluation. Several issues, central to the projects, were discussed. They include the micro-economic foundations of agent behavior (households and firms). We highlight the micro-foundation of the different models to make the welfare analysis more consistent.

Double counting

The standard example concerns travel time saving and land value in project evaluation. Consider a given infrastructure improvement. If the increased value of property is included in the cost-benefit analysis, it is not necessary to include the value of the travel time saving induced by this change of infrastructure. Such double counting was avoided in our Cost-Benefit analysis.

22.6 Policy Insights

The policy insights, summarized below, were developed by Stef Proost and Saskia van der Loo. A unified evaluation framework has been constructed that trades off social, ecological, and economic objectives. These trade-offs are conveniently expressed in terms of a social welfare function. Such a social welfare function also allows us to incorporate equity issues. The social welfare function computes indicators such as: the utility of the residents which is amongst others a function of the housing, transport costs and income, the utility of commuters and non-residents, the value of the local stock at the end of the horizon, the implementation costs and the collected revenues. The social welfare function is computed as a discounted (weighted) sum of these six components.

Policies can then be evaluated by comparing the social welfare outcome when the policy is implemented with the business as usual scenario. This approach has been operationalized by an extra module to be integrated into UrbanSim.

The three policy schemes that have been selected and implemented in the three case studies are the following.

The first policy concerns road pricing, this can be a cordon pricing or parking fees in the inner city. A cordon toll around the inner city or an increase in parking fees in this zone will decrease the number of trips to the inner city

and one can expect a reduction of congestion, accidents and emissions, together with an increase of the modal share of public transport. In the long run we can expect changes in housing prices and location of businesses and decentralization could occur.

The second policy is a large investment in either the road network (such as a bypass) or public transport network. The effects of an investment in the transport network largely depends on the pricing of the new infrastructure. Building a bypass leads to a reduction of the inner city congestion and to an improvement of air quality. If the bypass is not tolled, the additional road capacity attracts higher traffic volumes. In the long run we can expect relocation of businesses which could affect the location of business centers (decentralization). An investment in the public transport system also increases the use of public transport and reduces congestion. In addition, average trip length may increase (for both car and public transport) leading to city sprawl. We can also expect that large investments will influence housing prices due to better accessibility and relocation of households. Investments in public transport are also likely to boosts the benefits of road pricing.

The last policy is a land use regulation reducing the density of the city. The expected outcome of densification is that it reduces commuting times, although the link between residential density and travel distance is not as strong as it appears at first sight. The most important results are the definition and operationalization of a consistent urban sustainability assessment module. Moreover a set of test policies have been proposed.

This approach adopted in SustainCity differs from the more common sustainable indicators approach in several ways. First it aggregates all values in one function that summarizes the different dimensions. This has the advantage of valuing in a transparent way the different dimensions. A second important characteristic of the economic approach is that attention is paid to the incentives of the local decision maker. The local decision maker does not have the incentives to take all sustainability incentives on board and this may lead to inefficient local policies.

Available evidence on the costs and benefits of several urban sustainability policies suggest that for some policies like road pricing, there is robust evidence on its effects and its potential. For other policies, like densification and clean vehicles, the benefits are less clear and for these policies there is a need to generate more precise insights.

22.7 Closing

Several important tasks are involved in a project of this scale and scope: modeling, econometrics, calibration and simulation. Division of labor was clearly needed since different researchers in the project had substantially different

skills. However, efficient coordination is essential and should not be underestimated. The planning of the timing was difficult in the SustainCity project, since many challenges were encountered, such as problems with the data, or the need to implement new models in software, and adequately debugging and testing these. Support was absolutely essential. The project involved two open source software platforms: MATSim and UrbanSim, and one commercial software, METROPOLIS. Regardless of whether software for land use and transport modeling is open or closed, substantial expertise is still needed for developing successful applications with such tools, and the importance of high quality software, documentation, and professional support should not be underestimated. These issues are being handled in different ways among the three software platforms used in this project, including the creation of private enterprises to provide high-quality professional support for open source modeling platforms such as MATSim (Senozon) and UrbanSim (Synthicity).

LUTI models allow exploring policies under consideration by policy makers or posed by researchers. These policy proposals need to be translated by researchers and urban planners working with modelers to create appropriate inputs to the models that will allow their evaluation. LUTI models are not plug and play tools, but require a high degree of expertise and critical judgment. The interaction between the different participants of this experience and the main stakeholders remains somewhat limited, for the time being, due to difficulties in obtaining and processing urban data, coordination challenges, and limited prior experience of many of the researchers in this project with the development and application of urban models such as UrbanSim. See Waddell & Ulfarsson (2004) for a detailed analysis of the challenges in integrating land use and transportation planning and modeling, and tangible steps that have been taken to address these challenges. We remain convinced that LUTI models will continue to become more useful, more sensitive to a range of policies, and easier to use for research and for operational applications in urban and transport planning, as has been demonstrated by the wide adoption of UrbanSim in the U.S. for operational planning by Metropolitan Planning Organizations, and its use by the Société du Grand Paris to analyze the impacts of alternative designs and phasing of the Grand Paris Express project. Substantial progress has clearly been made in the advancement and application of integrated models, and more innovative developments on these interdisciplinary frontiers are expected in the coming years. However, in order to make research relevant to addressing the massive challenges of urban sustainability, there is some urgency for researchers to engage more actively with practice to ensure that the research is relevant and prepared to answer questions that are posed by policy-makers, and in order to have some opportunity to advance the state of the practice and influence the adoption of sustainability policies.

Bibliography

9th Conference on Computers in Urban planning and urban management (CUPUM) (2005), University College London, UK. See www.cupum.org.

Abdel-Aty, M. & Abdalla, M. F. (2004), 'Modeling drivers' diversion from normal routes under ATIS using generalized estimating equations and binomial probit link function', *Transportation* **31**(3), 327–348.

Abdel-Aty, M., Kitamura, R. & Jovanis, P. (1997), 'Using stated preference data for studying the effect of advanced traffic information on drivers' route choice', *Transportation Research Part C* **5**(1), 39–50.

Abello, A., Kelly, S. & King, A. (2002), Demographic projections with DYNAMOD-2, Technical Report 21, National Centre for Social and Economic Modelling.

Abraham, J. E. & Hunt, J. D. (1997), 'Specification and estimation of nested logit model of home, workplaces, and commuter mode choices by multiple-worker households', *Transportation Research Record* **1606**, 17–24.

Abraham, J. E. & Hunt, J. D. (1999), 'Policy analysis using the Sacramento MEPLAN land use-transportation interaction model', *Transportation Research Record* **1685**, 199–208.

Abraham, J. E. & Hunt, J. D. (2007), 'Random utility location, production, and exchange choice; additive logit model; and spatial choice microsimulations', *Transportation Research Record* **2003**, 1–6.

Abraham, J. E., Stefan, K. J. & Hunt, J. D. (2012), Population synthesis using combinatorial optimization at multiple levels, *in* '91st Annual Meeting of the Transportation Research Board'.

Abraham, J., Weidner, T., Gliebe, J., Willison, C. & Hunt, J. (2005), 'Three methods for synthesizing base-year built form for integrated land use-transport models', *Transportation Research Record* **1902**, 114–123.

Ahlfeldt, G. (2011), 'If Alonso was right: modeling accessibility and explaining the residential land gradient', *Journal of Regional Science* **51**, 318–338.

Alamá-Sabater, L., Artal-Tur, A. & Navarro-Azorín, J. M. (2011), 'Industrial location, spatial discrete choice models and the need to account for neighbourhood effects', *The Annals of Regional Science* **47**(2), 393–418.

Alberti, M. & Waddell, P. (2000), 'An integrated urban development and ecological simulation model', *Integrated Assessment* **1**(3), 215–227.

Alonso, W. (1964), *Location and Land Use: Toward a General Theory of Land Rent*, Harvard University Press.

Amemiya, T. (1984), 'Tobit models: A survey', *Journal of Econometrics* **24**(1-2), 3–61.

Amrhein, C. G. (1995), 'Searching for the elusive aggregation effect: evidence from statistical simulations', *Environment and planning A* **27**, 105–119.

Amstrong, A. (1986), 'Road pricing and user retraint: opportunities and constraints in developing countries', *Transportation Research* **20A**(2), 123–127.

Anas, A. (1982), *Residential location markets and urban transportation: Economic theory, econometrics, and policy analysis with discrete choice models*, Academic Press, London.

Anas, A. & Arnott, R. (1991), 'Dynamic housing market equilibrium with taste heterogeneity, idiosyncratic perfect foresight and stock conversions', *Journal of Housing Economics* **1**, 2–32.

Anas, A. & Arnott, R. (1993), 'Technological progress in a model of the housing-land cycle', *Journal of Urban Economics* **34**(2), 186–206.

Anas, A., Arnott, R. & Small, K. A. (1998), 'Urban spatial structure', *Journal of Economic Literature* **36**(3), 1426–1464.

Anas, A. & Chu, C. (1984), 'Discrete choice models and the housing price and travel to work elasticities of location demand', *Journal of Urban Economics* **15**(1), 107–123.

Anas, A. & Liu, Y. (2007), 'A regional economy, land use and transportation model (relu-tran): Formulation, algorithm design, and testing', *Journal of Regional Science* **47**(3), 415–455.

Anderson, S. P., de Palma, A. & Thisse, J.-F. (1992), *Discrete choice theory of product differentiation*, MIT Press, Cambridge, MA, USA.

Anderstig, C. & Mattsson, L. G. (1991), 'An integrated model of residential and employment location in a metropolitan region', *Papers in Regional Science* **70**(2), 167 – 184.

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Vol. 4 of *Studies in Operational Regional Science*, Kluwer Academic Publishers, Dordrecht.

Anselin, L. (1995), 'Local indicators of spatial association - LISA', *Geographical Analysis* **27**, 93–115.

Anselin, L. (2001), Spatial econometrics, *in* B. Baltagi, ed., 'A Companion to Theoretical Econometrics', Basil Blackwell, Oxford, pp. 310–330.

Anselin, L. & Lozano-Gracia, N. (2008), 'Errors in variables and spatial effects in hedonic house price models of ambient air quality', *Empirical Economics* **34**, 5–34.

Anselin, L. & Rey, S. (2007), 'PySAL: A python library of spatial analytical methods', *The Review of Regional Studies* **37**(1), 5–27.

Anselin, L., Syabri, I. & Kho, Y. (2006), 'GeoDa: An introduction to spatial data analysis', *Geographical Analysis* **38**(1), 5–22.

APUR (2001), Participation de l'APUR à la 7^{e} enquête régionale sur l'emploi, Technical report, Atelier parisien d'urbanisme.

Arbia, G. (1989), Statistical effect of spatial data transformations: A proposed general framework, *in* M. F. Goodchild & S. Gopal, eds, 'Global Accuracy of Spatial Datasets', Taylor and Francis, London.

Ardilly, P., Tillé, Y. & Jang, L. (2006), *Sampling methods: Exercises and solutions*, Springer.

Arentze, T., van den Berg, P. & Timmermans, H. (2012), 'Modeling social networks in geographic space: approach and empirical application', *Environment and Planning A* **44**(5), 1101–1120.

Arnott, A. & McMillan, D. (2006), *A companion to Urban Economics*, Blackwell publishing.

Arnott, R., de Palma, A. & Lindsey, R. (1993), 'A structural model of peak-period congestion: A traffic bottleneck with elastic demand', *The American Economic Review* pp. 161–179.

Arrondel, L. & Lefebvre, B. (2001), 'Consumption and investment motives in housing wealth accumulation: A french study', *Journal of Urban Economics* **50**(1), 112–137.

Arrow, K., Dasgupta, P., Goulder, L., Daily, G., Elhrlich, P., Heal, G., Levin, S., Mailer, K. G., Schneider, S., Starrett, D. & Walker, B. (2004), 'Are we consuming too much?', *Journal of Economic Perspectives* **18**(3), 147–172.

Ashworth, J. S. & Ulph, D. (1981), Household models, *in* C. Brown, ed., 'Taxation and labor supply', George Allen and Unwin, London.

Atkinson, B. A. (1970), 'On the measurement of inequality', *Journal of Economic Theory* **2**, 244–263.

Auld, J. & Mohammadian, A. K. (2010), 'Efficient methodology for generating synthetic populations with multiple control levels', *Transportation Research Record* **2183**, 19–28.

Auld, J., Mohammadian, A. K. & Wies, K. (2008), Population synthesis with region-level control variable aggregation, *in* TRB, ed., '87th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Axhausen, K. W. (1998), Can we ever obtain the data we would like to have?, *in* T. Gärling, ed., 'Theoretical Foundations of Travel Choice Modelling', Elsevier, Oxford.

Babin, A., Florian, M., James-Lefebvre, L. & Spiess, H. (1982), 'EMME/2: Interactive graphic method for road and transit planning', *Transportation Research Record* **866**, 1–9.

Ball, M. (2003), 'Markets and the structure of the housebuilding industry: An international perspective', *Urban Studies* **40**(5-6), 897 –916.

Balmer, M. (2007), Travel Demand Modeling for Multi-Agent Traffic Simulations: Algorithms and Systems, PhD thesis, ETH Zurich, Zurich, Switzerland.

Balmer, M., Meister, K., Rieser, M., Nagel, K. & Axhausen, K. W. (2008), Agent-based simulation of travel demand: Structure and computational performance of MATSim-T, *in* TRB Committee: Transportation Demand Forecasting, ed., 'Innovations in Travel Modeling (ITM'08)', Transportation Research Board, Portland.
URL: http://www.trb-forecasting.org/innovationsConference2008.html

Balmer, M., Raney, B. & Nagel, K. (2005), Adjustment of activity timing and duration in an agent-based traffic flow simulation, *in* H. Timmermans, ed., 'Progress in activity-based analysis', Elsevier, Oxford, UK, pp. 91–114.

Balmer, M., Rieser, M., Meister, K., Charypar, D., Lefebvre, N., Nagel, K. & Axhausen, K. (2009), MATSim-T: Architecture and simulation times, *in* A. Bazzan & F. Klügl, eds, 'Multi-Agent Systems for Traffic and Transportation', IGI Global, pp. 57–78.

Bar-Gera, H., Konduri, K., Sana, B., Ye, X. & Pendyala, R. M. (2009), Estimating survey weights with multiple constraints using entropy optimization methods, *in* TRB, ed., '88th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Barra, T. l. d., Pérez, B. & Vera, N. (1984), 'TRANUS-J: putting large models into small computers', *Environment and Planning B: Planning and Design* **11**(1), 87 – 101.
URL: http://www.envplan.com/abstract.cgi?id=b110087

Barthelemy, J. & Toint, P. L. (2013), 'Synthetic population generation without a sample', *Transportation Science* **47**(2), 266–279.

Beaujouan, É. & Toulemon, L. (2009), Family arrangements in france: Toward which trends for the near future?, *in* 'Session 1902, Population forecasting: new approaches and uses in policy, XXVI IUSSP International Population Conference, Marrakech, 27 September – 2 October 2009'.

Becker, G. (1965), 'A theory of the allocation of time', *The Economic Journal* **75**, 493–517.

Becker, G. (1973), 'A theory of marriage: Part I', *Journal of Political Economy* **81**, 813–846.

Becker, G. (1974*a*), 'A theory of marriage: Part II', *Journal of Political Economy* **82**(2), S11–S26.

Becker, G. (1974*b*), 'A theory of social interactions', *Journal of Political Economy* **82**(6), 1063–1093.

Becker, G. (1991), *A Treatise on the Family*, Harvard University Press.

Beckman, R. J., Baggerly, K. A. & McKay, M. D. (1996), 'Creating synthetic baseline populations', *Transportation Research, Part A* **30**(6), 415–429.

Beckmann, K. J., Brüggemann, U., Gräfe, J., Huber, F., Meiners, H., Mieth, P., Moeckel, R., Mühlhans, H., Rindsfüser, G., Schaub, H., Schrader, R., Schürmann, C., Schwarze, B., Spiekermann, K., Strauch, D., Spahn, M. & und Michael Wegener, P. W. (2007), 'ILUMASS integrated land-use modelling and transportation system simulation'. Endbericht.

Beckx, C., Arentze, T. A., Int Panis, L., Janssens, D., Vankerkorn, J. & Wets, G. (2009), 'An integrated activity-based modelling framework to assess vehicle emissions: Approach and application', *Environment and Planning B* **36**(6), 1086–1102.

Beharry-Borg, N., Hensher, D. & Scarpa, R. (2009), 'An analytical framework for joint vs separate decisions by couples in choice experiments: The case of coastal water quality in Tobago', *Environmental and Resource Economics* **43**(1), 95–117.

Beige, S. (2008), Long-term and mid-term mobility decisions during the life course, PhD thesis, ETH Zurich, Zurich.
URL: http://www.ivt.ethz.ch/vpl/publications/reports/ab510.pdf

Beimborn, E., Kennedy, R. & Schaefer, W. (1996), 'Inside the blackbox: Making transportation models work for livable communities', Citizens for a Better Environment and Environmental Defense Fund. Available from Environmental Defense Fund, Washington D.C., http://www.edf.org.

Belart, B. C. (2011), Wohnstandortwahl im grossraum zürich, Master's thesis, ETH Zurich, Zurich.

Ben-Akiva, M. (1973), Structure of Travel Passenger Demand Models, PhD thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Ben-Akiva, M. & Bowman, J. (1998), Activity based travel demand model systems, *in* P. Marcotte & S. Nguyen, eds, 'Equilibrium and Advanced Transportation Models', Kluwer Academic Publishers.

Ben-Akiva, M. & de Palma, A. (1986), 'Analysis of a dynamic residential location choice model with transaction costs', *Journal of Regional Science* **26**(2), 321–341.

Ben-Akiva, M. E., Bierlaire, M., Koutsopoulos, H. & Mishalani, R. (2002), Real time simulation of traffic demand-supply interactions with DynaMIT, *in* M. Gendreau & P. Marcotte, eds, 'Transportation and Network Analysis: Current Trends', Kluwer, Dordrecht, pp. 19–36.

Ben-Akiva, M. E. & Lerman, S. R. (1985), *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA, USA.

Berndt, E. K., Hall, B. H. & Hall, R. E. (1974), Estimation and inference in nonlinear structural models, *in* S. V. Berg, ed., 'Annals of Economic and Social Measurement, Volume 3, Number 4', National Bureau of Economic Research, pp. 653–665. **URL:** http://www.nber.org/chapters/c10206.pdf

Besley, T. (2006), *Principled Agents: The political economy of good goverment*, Oxford University Press.

Bettencourt, L., Lobo, J., Helbing, D., Kühnert, C. & West, G. (2007), 'Growth, innovation, scaling, and the pace of life in cities', *PNAS* **104**(17), 7301–7306.

Bhat, C. R., Guo, J. Y., Srinivasan, S. & Sivakumar, A. (2004), 'A comprehensive econometric microsimulator for daily activity-travel patterns (CEMDAP)', *Transportation Research Record* **1894**, 57–66.

Bhat, C. R. & Pendyala, R. M. (2005), 'Modeling intra-household interactions and group decision-making', *Transportation* **32**, 443–448.

Bierlaire, M. (2003), Biogeme: a free package for the estimation of discrete choice models, *in* 'Proceedings of the Swiss Transport Research Conference', Ascona, Switzerland.

Blanchet, D. (2014), La microsimulation dynamique : principes généraux et exemples en langage R, Technical report, Provisional Draft.

Blanchet, D. & Le Minez, S. (2007), The DESTINIE II model: a progress report', *in* '1st General Conference of the International Microsimulation Association, Vienna, Austria'.

Blondel, V., Krings, G. & Thomas, I. (2010), 'Régions et frontières de téléphonie mobile en Belgique et dans l'aire métropolitaine bruxelloise', *Brussels Studies* **42**, 12.

Blundell, R., Chiappori, P.-A., Magnac, T. & Meghir, C. (2007), 'Collective labour supply: Heterogeneity and non-participation', *The Review of Economic Studies* **74**(2), 417–445.

Bode, E. (2008), 'Delineating metropolitan areas using land prices', *Journal of Regional Science* **48**(1), 131–163.

Bodenmann, B. & Axhausen, K. (2010), Synthesis report on the state of the art on firmographics. SustainCity Working Paper 2.3, IVT, ETH Zürich.

Bodenmann, B. R. (2011), Location choice of firms with special emphasis on spatial accessibility, PhD thesis, ETH Zurich, Zurich.

Bodeux, F. (2002), 'La mobilité résidentielle en Belgique: perspectives d'études sociologiques', *Recherches Sociologiques* **XXXIII**(3), 25–45.

Boon, F. & Gayda, S. (2008), Quelques impacts du futur R.E.R. sur la Région bruxelloise et sa périphérie, *in* '14ième Congrès des Economistes Belges de Langue Francaise'. see http://www.stratec.be.

Borning, A., Ševčíková, H. & Waddell, P. (2008), A domain-specific language for urban simulation variables, *in* 'Proceedings of the 2008 international conference on digital government research', Digital Government Society of North America, Montréal, Canada, p. 207–215.

Borzacchiello, M. T., Nijkamp, P. & Koomen, E. (2010), 'Accessibility and urban development: a grid-based comparative statistical analysis of Dutch cities', *Environment and Planning B: Planning and Design* **37 (1)**, 148–169.

Bosredon, M., Dobson, A., Simmonds, D., Minta, P., Simpson, T., Andrade, K., Gillies, H. & Lumsden, K. (2009), Transport/economic/land-use model of scotland: Land-use modeling with DELTA, *in* '11th Conference on Computers in Urban planning and urban management (CUPUM)', Hong Kong.

Bowman, J. L. & Ben-Akiva, M. E. (2001), 'Activity-based disaggregate travel demand model system with activity schedules', *Transportation Research Part A* **35**(1), 1–28.

Box, G. & Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day.

Boyce, D. & Bar–Gera, H. (2003), 'Validation of multiclass urban travel forecasting models combining Origin–Destination, mode, and route choices', *Journal of Regional Science* **43**(3), 517–540.
URL: http://onlinelibrary.wiley.com/doi/10.1111/1467-9787.00309/abstract

Boyce, D. E. & Zhang, Y.-F. (1997), 'Calibrating combined model of trip distribution, modal split, and traffic assignment', *Transportation Research Record* **1607**, 1 – 5.
URL: http://trb.metapress.com/content/C322X011245X0L3X

Boyle, M. A. & Kiel, K. A. (2001), 'A survey of house price hedonic studies of the impact of environmental externalities', *Journal of Real Estate Literature* **9**, 117–144.

BPR (1964), *Traffic Assignment Manual*, U.S. Dept. of Commerce, Urban Planning Division, Bureau of Public Roads, Washington D.C., USA.

Bradley, M. A., Bowman, J. L. & Griesenbeck, B. (2010), 'SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution', *Journal of Choice Modelling* **3**(1), 5–31.

Bradley, M. & Vovsha, P. (2005), 'A model for joint choice of daily activity pattern types of household members', *Transportation* **32**, 545–571.

Brand, R. (2002), 'Microdata protection through noise addition, inference control in statistical databases: From theory to practice', *Lecture Notes in Computer Science* **2316**.

Brasington, D. M. & Hite, D. (2005), 'Demand for environmental quality: a spatial hedonic analysis', *Regional Science and Urban Economics* **35**, 57–82.

Breusch, T. & Pagan, A. (1979), 'Simple test for heteroscedasticity and random coefficient variation', *Econometrica* **47**(5), 1287–1294.

Briant, A., Combes, P. & Lafourcade, M. (2010), 'Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?', *Journal of Urban Economics* **67**(3), 287–302.

Briassoulis, H. (2000), Analysis of land use change: Theoretical and modeling approaches, *in* S. Loveridge, ed., 'The Web Book of Regional Science', Regional Research Institute, West Virginia University, Morgantown.
URL: http://www.rri.wvu.edu/WebBook/Briassoulis/contents.htm

Brueckner, J. (1997), 'Consumption and investment motives and the portfolio choices of homeowners', *The Journal of Real Estate Finance and Economics* **15**(2), 159–180.

Brueckner, J. K. (2000), 'Urban sprawl: diagnosis and remedies', *International regional science review* **23**(2), 160–171.

Brueckner, J., Thisse, J. & Zenou, Y. (1999), 'Why is central Paris rich and downtown Detroit poor? An amenity-based theory', *European Economic Review* **43**, 91–107.

Brunsdon, C., Fotheringham, A. & Charlton, M. (1996), 'Geographically weighted regression: A method for exploring spatial nonstationarity', *Geographical Analysis* **28**(4), 281–298.

Buczkowska, S., de Palma, A., Motamedi, K. & Picard, N. (2013), Firmographics: Guideline for implementation in UrbanSim and for estimation. sustaincity working paper 3.7, Technical report, Ecole Normale Supérieure de Cachan.
URL: http://www.sustaincity.org/publications

Buergle, M., Loechl, M., Waldner, U. & Axhausen, K. W. (2005), Land use and transport simulation: applying UrbanSim in the greater Zurich area, *in* 'Proceedings of the 9th Computers in Urban Planning and Urban Management Conference (CUPUM)', London, UK.

Buisson, G. & Daguet, F. (2012), 'Qui vit seul dans son logement? Qui vit en couple?', *Insee première* **1392**.

Bundesamt für Statistik BfS (2009), MONET – Nachhaltige Entwicklung messen – Indikatorensystem zur Nachhaltigen Entwicklung, Technical report, Bundesamt für Statistik BFS, Neuchâtel.

Bureau, B. & Glachant, M. (2010), 'Evaluation de l'impact des politiques "quartiers verts" et "quartiers tranquilles" sur les prix de l'immobilier à Paris', *Economie et Prévision* **192**, 27–44.

Caindec, E. & Prastacos, P. (1995), A description of POLIS. the projective optimization land use information system, Technical report, Working Paper 95-1, Oakland: Association of Bay Area Governments.

Cantonal Statistical Office of Zürich (2013), 'Population forecast 2013'.
URL: http://www.zh.ch/internet/justiz_inneres/statistik/de/dienstleistungen/analysen/bevoelkerungsprognosen.html

Carey, M. & Watling, D. (2003), 'Introduction to a special issue "dynamic traffic assignment ii"', *Networks and Spatial Economics* **3**, 403–406.

Carod, J. M. A. (2008), 'Industrial location at a local level: comments on the territorial level of the analysis', *Tijdschrift voor economische en sociale geografie* **99**(2), 193–208.

Carod, J. M. A. & Antolín, M. C. M. (2004), 'Firm size and geographical aggregation: an empirical appraisal in industrial location', *Small Business Economics* **22**(3-4), 299–312.

Cassan, F., Héran, F. & Toulemon, L. (2000), 'Study of family history. France's 1999 Family Survey', *Courrier des statistiques, English series* **6**, 7–19.

Cassells, R., Harding, A. & Kelly, S. (2006), 'Problems and prospects for dynamic microsimulation: A review and lessons for APPSIM', *Discussion paper* **93**. 50p.

CEPA (2006), 'Assembly bill 32', *California Environmental Protection Agency, California Health and Safety Code (HSC)* pp. 38500–38598.

CEPA (2008), 'Senate bill 375', *California Environmental Protection Agency, California Health and Safety Code (HSC)* .
URL: http://www.arb.ca.gov/cc/sb375/sb375.htm

Cérani, N. & Camus, M. (2004), *Le budget des familles en 2001*, Vol. 29 of *INSEE Résultats - Société*, Institut national de la statistique et des études économiques.

Cetin, N., Burri, A. & Nagel, K. (2003), A large-scale agent-based traffic microsimulation based on queue model, *in* 'Proceedings of the Swiss Transport Research Conference (STRC)', Monte Verita, CH.

Cetin, N., Nagel, K., Raney, B. & Voellmy, A. (2002), 'Large-scale multi-agent transportation simulations', *Computer Physics Communications* **147**(1), 559–564.

Chamberlain, G. (1980), 'Analysis of covariance with qualitative data', *Review of Economic Studies* **47**, 225–238.

Chamberlin, E. (1933), *Theory of monopolistic competition*, Harvard University Press, Cambridge, MA.

Charypar, D. & Nagel, K. (2005), 'Generating complete all-day activity plans with genetic algorithms', *Transportation* **32**(4), 369–397.

Cheshire, P. (2010), *Why Brussels needs a City-Region for the City?*, Mimeo.

Cheshire, P. & Gornostaeva, G. (2002), 'Cities and regions: comparable measures require comparable territories', *Les cahiers de l'institut d'aménagement et d'urbanisme de la region d'Ile de France* **135**, 13–32.

Chiappori, P.-A. (1988), 'Rational household labor supply', *Econometrica* **56**(1), 63–90.

Chiappori, P.-A. (1992), 'Collective labor supply and welfare', *Journal of Political Economy* **100**(3), 437–467.

Chiappori, P.-A. (1997), 'Introducing household production in collective models of labor supply', *Journal of Political Economy* pp. 191–209.

Chiappori, P.-A., de Palma, A., Picard, N. & Inoa, I. A. (2013), Couple residential location and spouses workplaces, Mimeo, Université de Cergy Pontoise, THEMA.

Chiappori, P., de Palma, A., Inoa, I. & Picard, N. (2013), Couple residential location and spouses' workplaces, Mimeo, Université de Cergy-Pontoise.

Cho, C. (1997), 'Joint choice of tenure and dwelling type: a multinomial Logit analysis for the city of Chongju', *Urban Studies* **34**(9), 1459–1473.

Christaller, W. (1933), *Central places in southern Germany*, Prentice-Hall.

Cochran, W. G. (2007), *Sampling techniques*, John Wiley & Sons.

Coiacetto, E. (2001), 'Diversity in real estate developer behaviour: A case for research', *Urban Policy and Research* **19**(1), 43–59.

Coiacetto, E. (2009), 'Industry structure in real estate development: Is city building competitive?', *Urban Policy and Research* **27**(2), 117.
URL: http://www.informaworld.com/10.1080/08111140802499080

Combes, P. P., Lafourcade, M., Thisse, J.-F. & Toutain, J.-C. (2011), 'The rise and fall of spatial inequalities in france: A long run perspective', *Explorations in Economic History* **48**, 243–271.

Cörvers, F., Hensen, M. & Bongaerts, D. (2009), 'Delimitation and coherence of functional and administrative regions', *Regional Studies* **43**(1), 19–31.

Cressie, N. a. C. & Whitford, H. J. (1986), 'How to use the two sample t-test', *Biometrical Journal* **28**(2), 131–148.
URL: http://onlinelibrary.wiley.com/doi/10.1002/bimj.4710280202/abstract

Dalenius, T. & Reiss, S. P. (1982), 'Dataswapping: A technique for disclosure limitation', *Journal of Statistical Planning and Inference* **6**, 73–85.

Daly, A. & Hess, S. (2010), Simple approaches for random utility modelling with panel data, *in* 'European transport conference, Glasgow'.

Dantan, S., de Palma, A., Motamedi, K. & Picard, N. (2010), Inequalities in accessibility, local amenities and school quality: A study on real estate price and household location choice in Paris area, Mimeo, Université de Cergy-Pontoise.

Dantan, S. & Picard, N. (2013), The role of preferences and constraints in residential segregation: Methodology and empirical evidence, Mimeo, Université de Cergy-Pontoise.

de Bok, M. & Bliemer, M. (2005), Using firmdemographic microsimulation for land use and transport scenario evaluation: model calibration, *in* '45th ERSA Congress'.

de Bok, M. & Sanders, F. (2005), Firm location and the accessibility of locations: Empirical results from the netherlands, *in* 'TRB Annual Meeting', Transportation Research Board.

de Jong, G., Daly, A., Pieters, M. & van der Hoorn, T. (2007), 'The logsum as an evaluation measure: Review of the literature and new results', *Transportation Research Part A* **41**(9), 874–889.

de la Barra, T. (1989), *Integrated land use and transport modelling: decision chains and hierarchies*, Cambridge University Press, Cambridge; New York.

de la Barra, T. (1995), *Integrated Land Use and Transportation Modeling: Decision Chains and Hierarchies*, Cambridge University Press.

De Lara, M., De Palma, A., Kilani, M. & Piperno, S. (2013), 'Congestion pricing and long term urban form: Application to paris region', *Regional Science and Urban Economics* **43**(2), 282–295.

de Palma, A. & Fontan, C. (2001), 'Choix modal et valeurs du temps en ile-de-france', *Recherche Transports Sécurité,* **71**, 24–48.

de Palma, A. & Guimard, A. (2013), Urbanization, an overview, Technical report, Ecole Normale Supérieure de Cachan.

de Palma, A. & Lefevre, C. (1985), 'Residential change and economic choice behavior', *Regional Science and Urban Economics* **15**(3), 421–434.

de Palma, A. & Lindsey, R. (2006), 'Modelling and evaluation of road pricing in paris', *Transport Policy* **13**(2), 115–126.

de Palma, A., Lindsey, R., Quinet, E. & Vickerman, R. (2011*a*), *Handbook in Transport Economics*, Edgar Elgard.

de Palma, A., Lindsey, R., Quinet, E. & Vickerman, R., eds (2011*b*), *Traffic Assignment Manual*, edward elgar publishing edn, U.S. Dept. of Commerce, Urban Planning Division, Bureau of Public Roads, Cheltenham, UK.

de Palma, A. & Marchal, F. (2002), 'Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for global large-scale mesoscopic transportation systems', *Networks and Spatial Economics* **2**(4), 347–369.

de Palma, A., Marchal, F. & Nesterov, Y. (1997), 'METROPOLIS: Modular system for dynamic traffic simulation', *Transportation Research Record* **1607**(1), 178–184.

de Palma, A., Motamedi, K., Picard, N. & Nguyen-Luong, D. (2005), An integrated land use - transportation model for the Paris area, ERSA conference papers ersa05p421, European Regional Science Association.

de Palma, A., Motamedi, K., Picard, N. & Waddell, P. (2005*a*), 'Incorporating land use in metropolitan transportation planning', *European Transport* **31**, 67–82.

de Palma, A., Motamedi, K., Picard, N. & Waddell, P. (2005*b*), 'A model of residential location choice with endogenous housing prices and traffic for the Paris region', *European Transport* **31**, 67–82.

de Palma, A., Motamedi, K., Picard, N. & Waddell, P. (2007), 'Accessibility and environmental quality: Inequality in the Paris housing market', *European Transport* **36**, 47–64.

de Palma, A. & Picard, N. (2005), 'Route choice decision under travel time uncertainty', *Transportation Research Part A* **39**(4), 295–324.

de Palma, A., Picard, N. & Inoa, I. A. (2013), Discrete choice decision-making with multiple decision makers within the household, *in* S. Hess & A. Daly, eds, 'Choice Modelling: the state of the art and the state of practice', Edgwar Elgar Publishing.

de Palma, A., Picard, N. & Waddell, P. (2005), 'Residential location choice with endogenous prices and traffic in the Paris metropolitan region', *European Transport* **31**, 67–82.

de Palma, A., Picard, N. & Waddell, P. (2007), 'Discrete choice models with capacity constraints: an empirical analysis of the housing market of the greater Paris region', *Journal of Urban Economics* **62**(2), 204–230.

Debreu, G. (1960), 'Review of D. Luce Individual Choice Behavior: A Theoretical Analysis', *American Economic Review* **50**, 186–188.

Deming, W. E. & Stephan, F. F. (1940), 'On the least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals of Mathematical Statistics* **11**(4), 427–444.

Devroye, L. (1986), *Non-Uniform Random Variate Generation(originally published with*, Springer-Verlag.
URL: http://cg.scs.carleton.ca/~luc/rnbookindex.html

Di Salvo, P. & Ermisch, J. (1997), 'Analysis of the dynamics of housing tenure choice in Britain', *Journal of Urban Economics* **42**(1), 1–17.

DiPasquale, D. & Wheaton, W. C. (1996), *Urban economics and real estate markets*, Prentice Hall.

Dixit, A., Grossman, G. & Helpman, E. (1997), 'Common agency and coordination: General theory and application to government policy making', *Journal of Political Economy* **105**, 752–769.

Dixit, A. K. & Stiglitz, J. E. (1977), 'Monopolistic competition and optimum product diversity', *The American Economic Review* pp. 297–308.

Docu Media Schweiz GmbH (2013), 'Home | Documedia'. Accessed December 2013.
URL: http://www.docu.ch/

Domencich, T. A. & McFadden, D. (1975), Urban travel demand, *in* D. Jorgenson & J. Waelbroeck, eds, 'Urban travel demand', number 93 *in* 'Contributions to Economic Analysis', North-Holland and American Elsevier.

Dong, H. & Gliebe, J. (2010), Exploring the taste heterogeneity in home developers' location choice, *in* 'Proceedings of the 57th annual north american meetings of the regional science association international', Denver.

Dong, H. & Gliebe, J. (2011), 'Forecasting location of new housing in integrated models of land use', *Transportation Research Record* **2255**, 79–90.

Donnay, J.-P. & Lambinon, M. (1997), Détermination des limites d'agglomération par télédetection: discussion méthodologique et application au cas de Huy (Belgique), *in* 'Actes des journées scientifiques du réseau de télédétection de l'UREF: Télédétetection des espaces urbains et périurbains', AUPELF, Liège, 1995, Montréal, pp. 239–246.

Donnelly, R., Erhardt, G. D., Moeckel, R. & Davidson, W. A. (2010), Advanced practices in travel forecasting, nchrp synthesis 406, Technical report, National Cooperative Highway Research Program.

D'Orazio, M., Zio, M. D. & Scanu, M. (2006), *Statistical Matching: Theory and Practice*, John Wiley & Sons, Ltd.

Dowling, R. G. (2005), *Predicting air quality effects of traffic-flow improvements: final report and user's guide*, Vol. 535, Transportation Research Board National Research.

Dowling, R., Ireson, R., Skabardonis, A., Gillen, D. & Stopher, P. (2005), Predicting air quality effects of traffic-flow improvements: Final report and user's guide, Technical Report 535, National Cooperative Highway Research Program, Transportation Research Board, National Research Council.

DREIF (2005), Enquête globale de transport 2001, Technical report, DREIF.

DREIF (2008), Modus v2.1 documentation détaillée du modèle de déplacements de la dreif, Technical report, DREIF (Direction Régionale de l'Équipement d'Île-de-France).

Dujardin, C., Selod, H. & Thomas, I. (2008), 'Unemployment and urban structure for young adults. the case of brussels', *Urban Studies* **45**(1), 89–113.

Dujardin, C., Thomas, I. & Tulkens, H. (2007), 'Quelles frontières pour Bruxelles? Une mise à jour', *Reflets et perspectives de la vie économique* **46**(2), 155–176.

Echenique, M., Flowerdew, A., Hunt, J., Mayo, T., Skidmore, I. & Simmonds, D. (1990), 'The MEPLAN models of Bilbao, Leeds and Dortmund', *Transport Reviews* **10**(4), 309–322.

Efthymiou, D. & Antoniou, C. (2013), 'How do transport infrastructure and policies affect house prices and rents? evidence from Athens, Greece', *Transportation Research Part A* **52**, 1–22.

Efthymiou, D. & Antoniou, C. (2014), 'Measuring the Effects of Transportation Infrastructure Location on Real Estate Prices and Rents. Investigating the Current Impact of a Planned Metro Line', *EURO Journal on Transportation and Logistics* . forthcoming.

Efthymiou, D., Antoniou, C. & Tyrinopoulos, I. (2012), 'Spatially-aware optimal site selection. Method and application in mobility center in Greece', *Transportation Research Record: Journal of the Transportation Research Board* **2276**, 146–155.

Efthymiou, D., Hurtubia, R., Bierlaire, M. & Antoniou, C. (2013), The integrated land use and transport model of brussels, *in* 'Proceedings of the Swiss Transport Research Conference', Ascona, Switzerland.

Einar, H., Holme, K., Mäkilä, K., Mattsson-Kauppi, M. & Mörtvik, G. (2002), *The SVERIGE spatial microsimulation model: content, validation, and example applications*, Kiruna,Sweden: Spatial Modelling Centre.

Eliasson, J. & Lundberg, M. (2003), Road pricing in urban areas, Technical report, Swedish National Road Administration.

Ernst, F. A. (1997), Die Integration von Unternehmens- und personenbezogenen Lebenszyklen, PhD thesis, Universitaet St. Gallen, Difo-Druck GmbH, Bamberg.

ESPON (2006), 'Mapping the structure of European territory. ESPON project 3.1, ESPON atlas', webpage.
URL: http://www.espon.eu/export/sites/default/Documents/Projects/ESPON2006Projects/CoordinatingCrossThematicProjects/Coordination/fr-3.1-full.pdf

Estrella, A. (1998), 'A new measure of fit for equations with dichotomous dependent variables', *Journal of Business and Economic Statistics* **16**, 198–205.

Ettema, D., Jong, K., Timmermans, H. & Bakema, A. (2007), 'PUMA: multi-agent modelling of urban systems', *Modelling Land-Use Change* pp. 237–258.

Evandrou, M., Falkingham, J., Johnson, P., Scott, A. & Zaidi, A. (2007), ''SAGE dynamic microsimulation model for Britain", *Modelling our Future: Population Ageing, Health and Aged Care, North-Holland, Amsterdam* .

Farmer, C. J. & Fotheringham, A. S. (2012), 'Network-based functional regions', *Environment and Planning A* **43**(11), 2723–2741.

Farooq, B., Bierlaire, M., Hurtubia, R. & Flötteröd, G. (2013), 'Simulation based population synthesis', *Transportation Research Part B* **58**, 243–263.

Farooq, B. & Miller, E. J. (2012), 'Towards integrated land use and transportation: A dynamic disequilibrium based microsimulation framework for built space markets', *Transportation research part A* **46**(7), 1030–1053.

Farooq, B., Salvini, P. J. & Miller, E. J. (2008), Development of an operational integrated urban model system: Software documentation, Technical report, Urban Transportation Research and Advancement Centre, University of Toronto.

Federal Statistical Office FSO, N. (2005), 'Statistik zur unternehmensdemografie (UDEMO): Grundlagen und methoden', *Statistik der Schweiz: 6 Industrie und Dienstleistungen* .

Fischer, M. M. & Getis, A. (2010), *Handbook of applied spatial analysis: software tools, methods and applications*, Springer.

Flemming, M. (2001), Techniques for estimating spatially dependent discrete choice models, *in* L. Anselin, R. Florax & S. Rey, eds, 'Advances in Spatial Econometrics. Methodology, Tools and Applications', Springer–Verlag, Berlin Heidelberg, pp. 145–166.

Fleurbaey, M. (2010), 'Beyond GDP: The quest for a measure of social welfare', *Journal of ecnomic literature* **47**(4), 1029–1075.

Fontaine, N. (2010), ABC, un outil d'aide à la décision territoriale, Technical report, Ministère des Affaires municipales, des Régions et de l'Occupation du territoire. See http://www.mamrot.gouv.qc.ca.

Fotheringham, A. S., Brunsdon, C. & Charlton, M. (2000), *Quantitative geography: perspectives on spatial data analysis*, SAGE Publications Limited.

Fotheringham, A. S. & Curtis, A. (1992), Encoding spatial information: The evidence for hierarchical processing, *in* A. U. Frank, I. Campari & U. Formentini, eds, 'Theories and Methods of Spatio-Temporal Reasoning in Geographic Space', number 639 *in* 'Lecture Notes in Computer Science', Springer Berlin Heidelberg, pp. 269–287.

Fotheringham, A. S. & Rogerson, P. A. (2009), *The SAGE Handbook of Spatial Analysis*, SAGE, London.

Fotheringham, A. & Wong, D. (1991), 'The modifiable areal unit problem in multivariate statistical analysis', *Environment and Planning A* **23**(7), 1025–1044.

Foti, F., Waddell, P. & Luxen, D. (2012), A generalized computational framework for accessibility: From the pedestrian to the metropolitan scale, *in* 'Proceedings of the 91th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, DC.

Franklin, J. P. & Waddell, P. (2003), A hedonic regression of home prices in King County, Washington, using activity-specific accessibility measures, *in* 'Proceedings of the 82nd Annual Meeting of the Transportation Research Board', Washington, D.C.

Frazier, T. & Alfons, A. (2012), Generating a close-to-reality synthetic population of Ghana, Open access publications from katholieke universiteit leuven, Katholieke Universiteit Leuven.

Fredriksen, D. & Stølen, N. (2007), 'MOSART modelling our future: Population, ageing, health and aged care', *North-Holland, Amsterdam* .

Freedman, O. & Kern, C. R. (1997), 'A model of workplace and residence choice in two-worker households', *Regional Science and Urban Economics* **27**(3), 241–260.

Friedrich, S. (2004), *Stadtumbau Wohnen*, 1 edn, Vdf Hochschulverlag, Zurich.

Froehlich, P., Frey, T., Reubi, S. & Schiedt, H.-U. (2004), 'Entwicklung des transitverkehrs-systems und deren auswirkung auf die raumnutzung in der schweiz (cost 340): Verkehrsnetz-datenbank,', *Arbeitsberichte Verkehrs- und Raumplanung, Institut fuer Verkehrsplanung und Transportsysteme (IVT), ETH Zuerich* **208**(208).

Fujita, M., Krugman, P. & Venables, A. (2001), *The Spatial Economy: Cities, Regions and International Trade*, MIT Press, Cambridge, Massachussets.

Fujita, M. & Thisse, J.-F. (2013), *Economics of Agglomeration: Cities, Industrial Location, and Globalization*, Cambridge University Press.

Fujita, M., Thisse, J. F. et al. (2003), Agglomeration and market interaction, *in* M. Dewatripont, H. L. Peter & S. J. Turnovsky, eds, 'Advances in Economics and Econometrics', Vol. 1, Cambridge University Press, pp. 302–338. Cambridge Books Online.
URL: http://dx.doi.org/10.1017/CBO9780511610240.011

Fuller, R. A. & Gaston, K. J. (2009), 'The scaling of green space coverage in european cities', *Biology letters* **5**(3), 352–355.

Gabaix, X. (1999*a*), 'Zipf's law and the growth of cities', *American Economic Review* pp. 129–132.

Gabaix, X. (1999*b*), 'Zipf's law for cities: an explanation', *Quarterly journal of Economics* pp. 739–767.

Gallay, O. (2010), Starting with UrbanSim: On the creation of an introductory project, Technical report, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Gardner, M. (1970), 'Mathematical games: The fantastic combinations of John Conway's new solitaire game "life"', *Scientific American* **223**(4), 120–123.

Gargiulo, F., Ternes, S., Huet, S. & Deffuant, G. (2010), 'An iterative approach for generating statistically realistic populations of households', *PloS One* **5**(1), e8828. PMID: 20107505.

Garret, M. & Wachs, M., eds (1996), *Transportation Planning on Trial: The Clean Air Act and Travel Forecasting*, Sage Publications, Thousand Oaks, CA.

Gawande, K. & Jenkins-Smith, H. (2001), 'Nuclear waste transport and residential property values: Estimating the effects of perceived risks', *Journal of Environmental Economics and Management* **42**, 207–233.

Gawron, C. (1998), 'An iterative algorithm to determine the dynamic user equilibrium in a traffic simulation model', *International Journal of Modern Physics C* **9**(3), 393–407.

Gehlke, C. & Biehl, K. (1934), 'Certain effects on grouping upon the size of the correlation coefficient in census tract material', *Journal of the American Statistical Association* **29**(185A), 169–170.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **PAMI-6**(6), 721 –741.

General Transit Feed Specification www pages (2012), 'General transit feed specification', https://developers.google.com/transit/gtfs/. Accessed April 2013.
URL: https://developers.google.com/transit/gtfs/

Geoghegan, J., Wainger, L. & Bockstael, N. (1997), 'Spatial landscape indices in a hedonic framework: an ecological economics analysis using gis', *Ecological Economics* **23**, 251–264.

Geurs, K. & Ritsema van Eck, J. (2001), Accessibility measures: review and applications, Technical Report 408505, National Institut of Public Health and the Environment, RIVM, P.O. Box 1, 3720 BA Bilthoven. See http://www.rivm.nl/bibliotheek/rapporten/408505006.html.

Geurs, K. T. & van Wee, B. (2004), 'Accessibility evaluation of land-use and transport strategies: review and research directions', *Journal of Transport Geography* **12**(2), 127–140.

Gilbert, N. G. & Troitzsch, K. G. (1999), Simulation as a method, *in* N. G. Gilbert & K. G. Troitzsch, eds, 'Simulation for the Social Scientist', 1st edn, Open University Press, Buckingham, pp. 14–26.

Glaeser, E. (2008), *Agglomeration and Spatial Equilibrium*, Oxford University Press.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A. & Shleifer, A. (1991), Growth in cities, Technical report, National Bureau of Economic Research.

Gläser, J. & Laudel, G. (2004), *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonsturierender Untersuchungen*, 2 edn, VS Verlag für Sozialwissenschaften, Wiesbaden.

Gliebe, J. P. & Koppelman, F. S. (2002), 'A model of joint activity participation between household members', *Transportation* **29**, 49–72.

Gobillon, L. & Le Blanc, D. (2004), 'L'impact des contraintes d'emprunt sur la mobilité résidentielle et les choix entre location et propriété', *Annales d'Economie et de Statistique* **74**, 15–45.

Goffette-Nagot, F., Reginster, I. & Thomas, I. (2011), 'Spatial analysis of residential land prices in belgium: Accessibility, linguistic border, and environmental amenities', *Regional Studies* **45**(9), 1253–1268.

Goffette-Nagot, F., Thomas, I., Zenou, Y. et al. (2000), 'Structure urbaine et revenu des ménages', *Economie géographique: Approches théoriques et empiriques* pp. 276–302.

Gort, M. & Klepper, S. (1982), 'Time paths in the diffusion of product innovations', *The economic journal* **92**, 630–653.

Gourieroux, C., Monfort, A., Renault, E. & Trognon, A. (1987), 'Generalised residuals', *Journal of Econometrics* **34**(1–2), 5–32. ISSN 0304–4076.

Greene, W. H. (2002), *Econometric Analysis*, 5th edn, Pearson Education.

Gregor, B. (2007), 'Land use scenario DevelopeR: practical land use model using a stochastic microsimulation framework', *Transportation Research Record: Journal of the Transportation Research Board* **2003**, 93–102.
URL: http://dx.doi.org/10.3141/2003-12

Greiner, L. E. (1972), 'Evolution and revolution as organizations grow. 1972.', *Harvard business review* **76**(3), 55.
URL: http://raj.rajaditi.com/management/Strategy/grainer.pdf

Grether, D., Chen, Y., Rieser, M. & Nagel, K. (2009), Effects of a simple mode choice model in a large-scale agent-based transport simulation, *in* A. Reggiani & P. Nijkamp, eds, 'Complexity and Spatial Networks. In Search of Simplicity', Advances in Spatial Science, Springer, chapter 13, pp. 167–186.

Gribble, S. (2000), Lifepaths: A longitudinal microsimulation model using a synthetic approach, *in* A. Gupta & V. Kapur, eds, 'Microsimulation in Government Policy and Forecasting', Elsevier Science, Netherlands, pp. 383–395.

Gribble, S., Hicks, C. & Rowe, G. (2003), The lifepaths microsimulation model, *in* 'International Microsimulation Conference on Population Ageing and Health: Modelling Our Future, Canberra', pp. 7–12.

Gruber, R., Zbinden, R. & Schmid, W. A. (2000), Räumliche Effekte von Swissmetro, Modellsimulation der Auswirkungen von Verkehrsinfrastrukturänderungen, Teil ORL-ETHZ von Projekt F5, Technical Report F5b, BBL/EDMZ, Bern.

Guevara, C. A. & Ben-Akiva, M. (2006), 'Endogeneity in residential location choice models', *Transportation Research Record* **1977**, 60–66.

Guo, J. Y. & Bhat, C. R. (2004), 'Modifiable areal units: Problem or perception in modeling of residential location choice?', *Transportation Research Record* **1898**(1), 138–147.

Guo, J. Y. & Bhat, C. R. (2007), 'Population synthesis for microsimulating travel behavior', *Transportation Research Record* **2014**(12), 92–101.

Haag, G. (1990), Master equations, *in* C. S. Bertuglia, G. Leonardi & A. G. Wilson, eds, 'Urban dynamics: Designing an integrated model', Routledge, London, pp. 69–83.

Haase, R. (2011), Ertragspotenziale – Hedonische Mietpreismodellierungen am Beispiel von Büroimmobilien, PhD thesis, ETH Zürich, Zürich.

Haider, M. & Miller, E. (2004), 'Modeling Location Choices of Housing Builders in the Greater Toronto, Canada, Area', *Transportation Research Record* **1898**, 148–156.

Haining, R. (1993), *Spatial data analysis in the social and environmental sciences*, Cambridge University Press.

Hammel, E. & Mason, C. (1990), *SOCSIM II: A sociodemographic microsimulation program. Revision 1.0: Operating manual*, University of California, Berkeley. Special Publication of the Graduate Group in Demography and Program in Population Research, Working Paper No. 29.

Hansen, W. (1959), 'How accessibility shapes land use', *Journal of the American Planning Association* **25**(2), 73 – 76.

Hao, J. Y., Hatzopoulou, M. & Miller, E. J. (2010), 'Integrating an Activity-Based Travel Demand Model with Dynamic Traffic Assignment and Emission Models – Implementation in the Greater Toronto, Canada, Area', *Transportation Research Record* **2176**, 1–13.

Harding, A. (1993), *Lifetime income distribution and redistribution: Applications of a microsimulation model*, North-Holland Amsterdam.

Harland, K., Heppenstall, A., Smith, D. & Birkin, M. (2012), 'Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques', *Journal of Artificial Societies and Social Simulation* **15**(1), 1–15.

Hassan, M. M. & Jun, C. (2011), 'A microscale simulation of land prices, household and employment growth using UrbanSim in Yongsan-Gu, Seoul', *International Journal of Urban Sciences* **15**(3), 201–213.

Hastings, W. K. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**(1), 97–109.

Hatzopoulou, M. & Miller, E. (2010), 'Linking an activity-based travel demand model with traffic emission and dispersion models: Transport's contribution to air pollution in Toronto', *Transportation Research Part D* **15**(6), 315–325.

Hausman, J. (1978), 'Specification tests in econometrics', *Econometrica* **46**(6), 1251–1271.

Healey, P. (1991), 'Models of the development process: A review', *Journal of Property Research* **8**(3), 219–238.
URL: http://dx.doi.org/10.1080/09599919108724039

Heckman, J. (1981), The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process, *in* C. Manski & D. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', MIT Press, Cambridge, MA.

Hediger, W. (2000), 'Sustaible development and social welfare', *Ecological economics* **32**, 481–492.

Henderson, J. & Ioannides, Y. (1986), 'A model of housing tenure choice', *American Economic Review* **73**, 98–113.

Hensher, D. A. & Ton, T. (2002), 'Tresis: A transportation, land use and environmental strategy impact simulator for urban areas', *Transportation* **29**(4), 439–457.

Hepinstall, J., Alberti, M. & Marzluff, J. (2008), 'Predicting land cover change and avian community responses in rapidly urbanizing environments', *Landscape Ecology* **23**, 1257–1276.

Hérault, N. (2005), A micro-macro model for South Africa: building and linking a microsimulation model to a CGE model, Technical report, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.

Hermes, K. & Poulsen, M. (2012), 'A review of current methods to generate synthetic spatial microdata using reweighting and future directions', *Computers, Environment and Urban Systems* **36**(4), 281–290.

Hotelling, H. (1929), 'Stability in competition', *Economic Journal* **39**(41), 57.

Hsiao, C. (1986), *Analysis of Panel Data.*, Cambridge University Press, Cambridge, U.K.

Hubert, J. P. & Toint, P. L. (2002), *La Mobilité Quotidienne des Belges*, Presses Universitaires de Namur.

Hunt, J. D. & Abraham, J. E. (2003), Design and application of the pecas land use modelling system, *in* '8th Conference on Computers in Urban planning and urban management (CUPUM)', Sendai, Japan.

Hunt, J. D. & Abraham, J. E. (2005), Design and Implementation of PECAS: A Generalised System for Allocating Economic Production, Exchange and Consumption Quantities, *in* M. Lee-Gosselin & S. T. Doherty, eds, 'Integrated Land-Use and Transportation Models: Behavioural Foundations', Elsevier, Amsterdam, pp. 253–274.

Hunt, J. D., Kriger, D. S. & Miller, E. J. (2005), 'Current operational urban land-use–transport modelling frameworks: A review', *Transport Reviews* **25**(3), 329–376.

Hurtubia, R., Bierlaire, M. & Martínez, F. (2012), Dynamic microsimulation of location choices with a quasi-equilibrium auction approach, *in* 'Swiss Transport Research Conference', Ascona.

Hurtubia, R., Gallay, O. & Bierlaire, M. (2010), Attributes of households, locations and real-estate markets for land use modeling (sustaincity deliverable), Technical Report TRANSP-OR 100419, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Hutter, T. (2005), Neugruendungen und ueberlebensraten von unternehmen: Der kanton st.gallen im inter- und intrakantonalen vergleich, 1996 bis 2002, *in* 'Statistik aktuell, 7', Fachstelle fuer Statistik Kanton St.Gallen, St.Gallen.

Iacono, M. & Levinson, D. (2008), Review of methods for estimating the economic impact of transportation improvements, *in* 'Proceedings of the Transportation Research Board 88th Annual Meeting', Washington D.C.

Iacono, M., Levinson, D. & El-Geneidy, A. (2008), 'Models of transportation and land use change: a guide to the territory', *Journal of Planning Literature* **22**(4), 323–340.

IAU-IdF (2012), 'Mode d'occupation du sol (mos)', http://www.iau-idf.fr/cartes/mode-doccupation-du-sol-mos.html. Accessed: 2013-12-31.

Ihaka, R. & Gentleman, R. (1996), 'R: a language for data analysis and graphics', *Journal of computational and graphical statistics* **5**(3), 299–314.

Inoa, I. A., Picard, N. & de Palma, A. (2013), 'Commuting time and accessibility in a joint residential location, workplace, and job type choice model'.

Inoa, I., Picard, N. & de Palma, A. (forthcoming), 'Effect of an accessibility measure in a model for choice of residential location, workplace, and type of employment', *Mathematical Population Studies* .

Institute of Transport and Logistics Studies (2009), 'TRESIS v1.10 - Institute of Transport and Logistics Studies - The University of Sydney'. Accessed December 2013.
URL: http://sydney.edu.au/business/itls/tresis

Ioannides, Y. & Rosenthal, S. (1974), 'Estimating the consumption and investment demands for housing and their effect on housing tenure status', *Review of Economics and Statistics* **76**(1), 127–141.

Jacobs, J. (1961), *The death and life of great American cities*, Modern Library, New York.

JEPP www pages (Accessed June 2013), 'Jepp – java embedded python', http://jepp.sourceforge.net/.
URL: http://jepp.sourceforge.net/

Jones, J., Peeters, D. & Thomas, I. (2012), Monocentric city, polycentric city: an empirical comparison through scales, *in* 'Proceedings of 52nd European Congress of the RSAI', Bratislava.

Jones, J., Peeters, D. & Thomas, I. (2013), Does size matter? LUTI models, policy evalutations, and the scale effect, *in* 'Proceedings of the European Colloqium of Theoretical & Quantitative Geography', Dourdan, France.

Jones, P. M., Dix, M. C., Clarke, M. I. & Heggie, I. G. (1983), *Understanding Travel Behaviour*, Gower, Aldershot.

Joo, Y., Hassan, M. M. & Jun, C. (2011), 'An application of the UrbanSim land price model in yongsan-gu, seoul, korea', *International Journal of Urban Sciences* **15**(1), 15–24.

Joshi, H., Guhathakurta, S., Konjevod, G., Crittenden, J. & Li, K. (2006), 'Simulating the effect of light rail on urban growth in phoenix: An application of the UrbanSim modeling environment', *Journal of Urban Technology* **13**(2), 91–111.

JPype www pages (accessed june 2013), 'JPype – Java to Python integration', http://jpype.sourceforge.net/.
URL: http://jpype.sourceforge.net/

Kain, J. F. (1985), *Housing and Neighborhood Dynamics: A Simulation Study*, Harvard University Press.

Karniouchina, E. V. C., J., S., Short, J. C. & Ketchen jr., D. (2013), 'Extending the firm vs. industry debate: does industry life cycle stage matter?', *Strategic Management Journal, Research Notes and Commentaries, published online EarlyView DOI: 10.1002/smj.2042* .

Keiding, N. (1990), 'Statistical inference in the lexis diagram', *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences* **332**(1627), 487–509.

Kickhöfer, B., Grether, D. & Nagel, K. (2011), 'Income-contingent user preferences in policy evaluation: application and discussion based on multi-agent transport simulations', *Transportation* **38**, 849–870.

Kickhöfer, B. & Nagel, K. (2013, forthcoming), 'Towards high-resolution first-best air pollution tolls: An evaluation of regulatory policies and a discussion on long-term user reactions', *Networks and Spatial Economics* **In press**. Published online: September 2013.

Kilani, A. M. & Lindsay, R. (2005), 'Congestion pricing on the road network: A study using the dynamic equilibrium simulator metropolis', *Transportation Research A* **39**, 588–611.

Kim, C., Phipps, T. & Anselin, L. (2003), 'Measuring the benefits of air quality improvement: a spatial hedonic approach', *Journal of Environmental Economics and Management* **45**(1), 24–39.

Kim, J., Pagliara, F. & Preston, J. (2005), 'The intention to move and residential location choice behaviour', *Urban Studies* **42**(9), 1621–1636.

Kim, T. J., Rho, J. H. & Suh, S. (1989), *Integrated urban systems modeling: theory and applications*, Kluwer Academic Publishers Dordrecht/Boston/London.

Klepper, S. & Graddy, E. (1990), 'The evolution of new industries and the determinants of market structure', *RAND Journal of Economics* **21**(1), 27–44.

Kriplean, T., Borning, A., Waddell, P., Klang, C. & Fogarty, J. (2010), Supporting agile modeling through experimentation in an integrated urban simulation framework, *in* 'Proceedings of the 10th Annual International Conference on Digital Government Research', Puebla, Mexico, pp. 112–121.

Kryvobokov, M., Ovtracht, N. & Thiebaut, V. (2009), Analysis and prediction of household location choice in Grand Lyon with urban land use simulation tool UrbanSim, *in* 'Proceedings of the International Conference on Territoral Intelligence', Salerno, Italy.

Kühnert, C., Helbing, D. & West, G. B. (2006), 'Scaling laws in urban supply networks', *Physica A: Statistical Mechanics and its Applications* **363**(1), 96–103.

Lafourcade, M., Thisse, J.-F. et al. (2011), 'New economic geography: the role of transport costs', *A. De Palma, R. Lindsey, E. Quinet, Vickerman (Eds.), Handbook of Transport Economics, Edward Elgar Publishing Ltd, Cheltenham* pp. 67–96.

Landis, J. D. (1994), 'The California Urban Futures Model: a new generation of metropolitan simulation models', *Environment and Planning B: Planning and Design* **21**, 399–422.

Landis, J. & Zhang, M. (1998*a*), 'The second generation of the California urban futures model. Part 1: Model logic and theory', *Environment and Planning B: Planning and Design* **25**(5), 657–666.
URL: http://www.environmentandplanning.com/abstract.cgi?id=b250657

Landis, J. & Zhang, M. (1998*b*), 'The second generation of the California urban futures model. Part 2: Specification and calibration results of the land-use change submodel', *Environment and Planning B: Planning and Design* **25**(6), 795–824.
URL: http://envplan.com/abstract.cgi?id=b250795

Larsen, O. (1995), 'The toll cordons in Norway: an overview', *Journal of Transport Geography* **3**(3), 187–197.

Lautso, K., Spiekerman, K., Wegener, M., Sheppard, I., Steadman, P., Martino, A., Domingo, R. & Gayda, S. (2004), Propolis. planning and research of policies for land use and transport for increasing urban sustainability. final report, Technical report, LT Consultants, Helsinki.

Le Bras, H. (1973), 'Parents, grand-parents, bisaïeux', *Population* **28**(1), 9–38.

Lee, B. & Waddell, P. (2010), 'Residential mobility and location choice: a nested logit model with sampling of alternatives', *Transportation* **37**(4), 587–601.

Lee, D. B. (1973), 'Requiem for large-scale models', *Journal of the American Institute of Planners* **39**, 163–178.

Lee, D. B. (1994), 'Retrospective on large-scale urban models', *Journal of the American Planning Association* **60**(1), 35–40.

Lee, D.-H. & Fu, Y. (2011), 'Cross-entropy optimization model for population synthesis in activity-based microsimulation models', *Transportation Research Record* **2255**(2), 20–27.

Lee, D. J. H. (2009), '2009 TMA/MPO modeling activity survey'. Presentation.
URL: \url{http://www.urbansim.org/pub/Main/WebHome/2009_MPO_Modeling_Survey_RESULT_TMIP.pdf}

Lefebvre, N. & Balmer, M. (2007), Fast shortest path computation in time-dependent traffic networks, *in* 'Proceedings of the Swiss Transport Research Conference (STRC)', Monte Verita, CH. See www.strc.ch.

Lefèvre, C. & Filon, A. (2005), Histoires de familles, histoires familiales. les résultats de l'enquête Famille de 1999, number 156 *in* 'Les Cahiers de l'INED', Ined.

Lenormand, M. & Deffuant, G. (2012), 'Generating a synthetic population of individuals in households: Sample-free vs sample-based methods', *arXiv:1208.6403* .
URL: \url{http://arxiv.org/abs/1208.6403}

Lesage, J. & Pace, R. (2009), *Introduction to Spatial Econometrics*, Chapman and Hall, Boca Raton, Florida.

Lester, D. L., Parnell, J. A. & Carraher, S. (2003), 'Organizational life cycle: a five stage empirical scale', *The International Journal of Organizational Analysis* **11**(4), 339–354.

Leuthold, J. H. (1968), 'An empirical study of female income transfers and the work decision of the poor', *Journal of Human Resources* **3**, 312–323.

Li, J. & O'Donoghue, C. (2013), 'A survey of dynamic microsimulation models: uses, model structure and methodology', *International Journal of Microsimulation* **6**(2), 3–55.

Liang, S. (1999), *The Java Native Interface: Programmer's Guide and Specification*, Addison-Wesley Professional.

Linneman, P. & Wachter, S. (1989), 'The impacts of borrowing constraints on homeownership', *American Real Estate and Urban Economics Association Journal* **17**(5), 389–402.

Linstone, H. A., Turoff, M. et al. (1975), *The Delphi method: Techniques and applications*, Vol. 29, Addison-Wesley Reading, MA.

Löchl, M. (2010), Application of spatial analysis methods for understanding geographic variation of prices, demand and market success, Dissertation, ETH Zurich, Zurich.

Löchl, M. & Axhausen, K. W. (2010), 'Modelling hedonic residential rents for land use and transport simulation while considering spatial effects', *Journal of Transport and Land Use* **3**(2), 39–63.

Löchl, M., Bürgle, M. & Waldner, U. (2007), Handbuch simulationsmodell grossraum zürich, Technical Report 10, NSL, ETH Zürich, Zürich.
URL: http://www.ivt.ethz.ch/vpl/research/iasp

Loechl, M., Buergle, M. & Axhausen, K. (2007), 'Implementierung des integrierten flaechennutzungsmodells UrbanSim fuer den grossraum Zuerich', *DISP - The Planning Review* **168**(1), 13–25.

Losch, A. (1940), 'The economics of location', *New Haven, CT: Yale UP* .

Louchart, P. (2010), 'Près d'un million de ménages franciliens en plus à l'horizon 2030', *Note rapide IAU-ÎdF* **495**.

Louviere, J., Hensher, D. & Swait, J. (2000), *Stated Choice Methods: Analysis and Application*, Cambridge University Press, Cambridge, UK.

Lowry, I. S. (1964), A model of metropolis, Technical Report RM-4035-RC, RAND Corporation, Santa Monica.

Luce, R. (1959), *Individual Choice Behavior*, Wiley, New York.

Luna, F. & Stefannson, B., eds (2000), *Economic Simulations in Swarm: Agent-Based Modelling and Object-Oriented Programming*, Kluwer Academic Publishers.

Lundberg, S. & Pollak, R. A. (1993), 'Separate spheres bargaining and the marriage market', *Journal of Political Economy* **101**(6), 988–1010.

Ma, L. (2011), *Generating disaggregate population characteristics for input to travel-demand models*, Dissertation, University of Florida.
URL: \url{http://gradworks.umi.com/35/14/3514962.html}

Mackett, R. L. (1991), 'LILT and MEPLAN: a comparative analysis of land-use and transport policies for leeds', *Transport Reviews* **11**(2), 131–154.

Magrini, S. (2004), Regional (di)convergence, *in* J. V. Henderson & J. F. Thisse, eds, 'Handbook of Regional and Urban Economics, Volume 4', Elsevier, Amsterdam, pp. 2741–2796.

Mahmassani, H. S., Hu, T. & Jayakrishnan, R. (1995), Dynamic traffic assignment and simulation for advanced network informatics (DYNASMART), *in* N. H. Gartner & G. Improta, eds, 'Urban traffic networks: dynamic flow modeling and control', Springer, Berlin.

Manski, C. F. & Lerman, S. R. (1977), 'The estimation of choice probabilities from choice based samples', *Econometrica* **45**(8), 1977–1988.

Maoh, H. F. & Kanaroglou, P. S. (2004), *Economic Clustering and Urban Form: The Case of Hamilton, Ontario,, 003, McMaster University*, Hamilton.

Maoh, H. F. & Kanaroglou, P. S. (2005), Agent-based firmographic models: A simulation framework for the city of hamilton, vortrag,, *in* 'PROCESSUS Second International Colloquium on the Behavioural Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications, University of Toronto', Juni 2005.

Maoh, H. F. & Kanaroglou, P. S. (2007), 'Business establishment mobility behavior in urban areas: A microanalytical model for the city of hamilton in ontario, canada', *Journal of Geographical Systems* **9**(3), 229–252.

Maoh, H. F. & Kanaroglou, P. S. (2009), 'Intrametropolitan location of business establishments: Microanalytical model for hamilton, ontario, canada', *Transportation Research Record: Journal of the Transportation Research Board* **2133**, 33–45.

Marshall, J. D. (2007), 'Urban land area and population growth: a new scaling relationship for metropolitan expansion', *Urban Studies* **44**(10), 1889–1904.

Martens, M. & Griethuysen, S. (1999), The ABC location policy in the netherlands, the right business at the right place, Technical report, TNO-INRO.

Martinez, F. & Hurtubia, R. (2006), 'Dynamic model for the simulation of equilibrium status in the land use market', *Network and Spatial Economics* **6**, 55–73.

Martínez, F. J. (1992), 'The bid-choice land-use model: an integrated economic framework', *Environment and Planning A* **24**(6), 871 – 885.

Martinez, F. J. (1997*a*), MUSSA: A land Use Model for Santiago City, Technical report, University of Chile, Department of Civil Engineering, Santiago de Chile.

Martinez, F. J. (1997*b*), Towards a microeconomic framework for travel behavior and land use interactions, *in* '8th Meeting of the International Association of Travel Behavior Research', Austin.

Martinez, F. J. (2000), 'Towards a land-use and transport interaction framework', *Handbook of transport modelling* **1**, 393–407.

Martínez, F. J. & Donoso, P. (2010), The MUSSA II Land Use Auction Equilibrium Model, *in* F. Pagliara, J. Preston & D. Simmonds, eds, 'Residential Location Choice', Advances in Spatial Science, Springer Berlin Heidelberg, pp. 99–113.

MATSim (accessed 2014), 'Multi-Agent Transportation Simulation', webpage.
URL: http://www.matsim.org

MATSim extensions (accessed 2014), 'MATSim extensions', webpage.
URL: http://www.matsim.org/extensions

Maurin, C., de Palma, A. & Picard, N. (2013), Regime switching models: An application to the real estate market in ile de france. sustaincity working paper 3.5, Technical report, Ecole Normale Supérieure de Cachan and THEMA-UCP.
URL: http://www.sustaincity.org/publications

McElroy, M. & Horney, M. (1981), 'Nash-bargained decisions: Toward a gen- eralization of the theory of demand', *International Economic Review* **22**, 333–349.

McFadden, D. (1974), Conditional logit analysis of qualitative choice behavior, *in* P. Zarembka, ed., 'Frontiers in Econometrics', Academic Press, New York, pp. 105–142.

McFadden, D. (1978), Modeling the choice of residential location, *in* A. Karlqvist, L. Lundqvist, F. Snickars & J. Wiebull, eds, 'Spatial Interaction Theory and Planning Models', North Holland, Amsterdam, pp. 75–96.

McFadden, D. (1981), Econometric models of probabilistic choice, *in* C. Manski & D. McFadden, eds, 'Structural Analysis of Discrete Data with Econometric Applications', Academic Press, MIT Press, Cambridge, MA, pp. 198–272.

McFadden, D. & Train, K. (2000), 'Mixed MNL models of discrete response', *Journal of Applied Econometrics* **15**, 447–470.

Meister, K., Rieser, M., Ciari, F., Horni, A., Balmer, M. & Axhausen, K. W. (2009), 'Anwendung eines agentenbasierten Modells der Verkehrsnachfrage auf die Schweiz', *Straßenverkehrstechnik* **53**(5), 269–280.

Mercenier, J. (2013), Un modèle spatio-temporel d'équilibre général pour l'évaluation des effets des investissements d'infrastructure liés au grand paris, Technical report, University Paris 2l.

Metro Regional Government (2013), 'Publications and technical papers', http://www.oregonmetro.gov/index.cfm/go/by.web/id=24906. Accessed March 2013.

Metropolitanraum (2013), 'Metropolitanraum Zürich – Porträt', webpage.
URL: http://www.metropolitanraum-zuerich.ch/fileadmin/user_upload/downloads/Portrait_Metropolitanraum_ZH.pdf

Meyer, R. (2004), 'Der beitrag der unternehmensgruendungen zum strukturwandel in der schweizer wirtschaft', *Reihe A: Discussion Paper DPW, Fachhochschule Solothurn Nordwestschweiz* **2004-08**.

Miller, E., Hunt, J., Abraham, J. & Salvini, P. (2004), 'Microsimulating urban systems', *Computers, Environment and Urban Systems* **28**(1-2), 9–44.

Miller, E. J., Noehammer, P. & Ross, D. R. (1987), "a micro-simulation model of residential mobility", *in* '"International Symposium on Transport, Communication and Urban Form: 2, Analytical Techniques and Case Studies"'.

Miller, E. J. & Salvini, P. A. (2001), The Integrated Land Use, Transportation, Environment (ILUTE) Microsimulation Modelling system: Description and current status., *in* D. Henscher, ed., 'The Leading Edge in Travel Behaviour Research', selected papers from the 9th International Association for Travel Behaviour Research Conference, Elsevier, Gold Coast, Queensland, Australia, chapter 41.

Miller, E. & Roorda, M. (2003), 'A prototype model of household activity/travel scheduling', *Transportation Research Record* **1831**, 114–121.

Mills, E. (1990), 'Housing tenure choice', *The Journal of Real Estate Finance and Economics* **3**(4), 323–331.

Mills, E. S. (1967), 'An aggregative model of resource allocation in a metropolitan area', *The American Economic Review* **57**(2), 197–210.
URL: http://www.jstor.org/stable/1821621

Minar, N., Burkhart, R., Langton, C. & Ashkenazi, M. (1996), The swarm simulation system: A toolkit for building multi-agent simulations, Technical report, Santa Fe Institute.

Miyamoto, K., Udomsri, R., Sathyaprasad, S. & Ren, F. (1996), 'A decision support system for integrating land use, transport and environmental planning in developing metropolises', *Computers, environment and urban systems* **20**(4), 327–338.

MODELISTICA (2013), 'TRANUS'. Accessed December 2013.
URL: http://www.tranus.com/

Moeckel, R. (2005), Microsimulation of firm location decision, *in* 'In the proceedings of the 9th International Conference on Computers in Urban Planning and Urban Management, London, England'.

Moeckel, R. (2006), Business Location Decisions and Urban Sprawl A Microsimulation of Business Relocation and Firmography, PhD thesis, University of Dortmund.

Moeckel, R., Spiekermann, K. & Wegener, M. (2003), Creating a synthetic population, *in* 'Proceedings ot the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)', Sendai, Japan.

Moniteur belge (2004), 'Ordonnance portant assentiment à la Convention du 4 avril 2003 entre l'Etat Fédéral, la Région flamande, la Région wallone et la Région de Bruxelles-Capitale, visant à mettre en oeuvre le programme du Réseau Express Régional de, vers, dans et autour de Bruxelles'.

Montero, J., Larraz, B. & Chasco, C. (2008), *Building an Environmental Quality Index for a big city: a spatial interpolation approach with DP2*, MPRA Paper 10736.

Morand, E., Toulemon, L., Pennec, S., Baggio, R. & Billari, F. (2010), 'Demographic modelling: the state of the art', *FP7-244557 Projet SustainCity* .
URL: http://www.sustaincity.org/publications

Morrison, R. (2008), 'Validation of longitudinal microsimulation models: Dynacan II practices and plans', *Online Working Paper* **8**. NATSEM, University of Canberra.

Müller, K. & Axhausen, K. W. (2010), Population synthesis for microsimulation: State of the art, *in* 'Proceedings of the 10th Swiss Transport Research Conference', Ascona, Switzerland.

Müller, K. & Axhausen, K. W. (2011*a*), Hierarchical IPF: Generating a synthetic population for Switzerland, *in* ERSA, ed., '51st Congress of the European Regional Science Association', University of Barcelona, Barcelona.

Müller, K. & Axhausen, K. W. (2011*b*), Population synthesis for microsimulation: State of the art, *in* TRB, ed., '90th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Müller, K. & Axhausen, K. W. (2011*c*), 'Population synthesis for microsimulation: State of the art', *Proceeding of Transportation Research Board 90th Annual Meeting* .

Müller, K. & Axhausen, K. W. (2012), Preparing the Swiss Public-Use Sample for generating a synthetic population of Switzerland, *in* STRC, ed., '12th Swiss Transport Research Conference', Ascona.

Mumford, L. (1938), 'The culture of cities'.

Muth, R. F. (1969), *Cities and Housing*, University of Chicago Press, Chicago.
URL: http://trid.trb.org/view.aspx?id=545388

Nagel, K. (2008), Towards simulation-based sketch planning: Some results concerning the Alaskan Way viaduct in Seattle WA, VSP Working Paper 08-22, TU Berlin, Transport Systems Planning and Transport Telematics. See www.vsp.tu-berlin.de/publications.

Nagel, K. (2011), Towards simulation-based sketch planning, part II: Some results concerning a freeway extension in Berlin, VSP Working Paper 11-18, TU Berlin, Transport Systems Planning and Transport Telematics. See www.vsp.tu-berlin.de/publications.

Nagel, K. & Flötteröd, G. (2012), Agent-based traffic assignment: Going from trips to behavioural travelers, *in* R. Pendyala & C. Bhat, eds, 'Travel Behaviour Research in an Evolving World – Selected papers from the 12th international conference on travel behaviour research', International Association for Travel Behaviour Research, chapter 12, pp. 261–294.
URL: http://www.lulu.com/shop/ram-m-pendyala-and-chandra-r-bhat/travel-behaviour-research-in-an-evolving-world/hardcover/product-18838195.html

Nagel, K., Grether, D., Beuck, U., Chen, Y., Rieser, M. & Axhausen, K. (2008), 'Multi-agent transport simulations and economic evaluation', *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)* **228**(2+3), 173–194. Special issue on agent based models for economic policy advice, edited by B. LeBaron and P. Winker.

Nechyba, T. J. & Strauss, R. P. (1998), 'Community choice and local public services: A discrete choice approach', *Regional Science and Urban Economics* **28**(1), 51–73.

Neumann, A., Balmer, M. & Rieser, M. (2012), Converting a static macroscopic model into a dynamic activity-based model for analyzing public transport demand in Berlin, *in* 'Proceedings of the 13th Conference of the International Association for Travel Behaviour Research (IATBR)', Toronto, Canada.

Nguyen-Luong, D. (2008), An integrated land use-transport model for the Paris region (SIMAURIF): Ten lessons learned after four years of development, Technical report, Institut d'Aménagement et d'Urbanisme de la Région d'Ile-de-France, Paris.

Nicolai, T. W. & Nagel, K. (2010), Coupling MATSim and UrbanSim: Software design issues, SustainCity Working Paper 6.1. Also VSP WP 10-13, see www.vsp.tu-berlin.de/publications.
URL: http://www.sustaincity.org/publications

Nicolai, T. W. (2012), Using MATSim as a travel model plug-in to UrbanSim, VSP Working Paper 12-29, TU Berlin, Transport Systems Planning and Transport Telematics. See www.vsp.tu-berlin.de/publications.

Nicolai, T. W. (2013), Investigating the MATSim warm and hot start capability, SustainCity Working Paper 6.1. Also VSP WP 13-06, see www.vsp.tu-berlin.de/publications.
URL: http://www.sustaincity.org/publications

Nicolai, T. W. & Nagel, K. (in press), High resolution accessibility computations, *in* A. Condeço, A. Reggiani & J. Gutiérrez, eds, *Accessibility and spatial interaction*, Edward Elgard. Also VSP WP 13-02, see www.vsp.tu-berlin.de/publications.

Nicolai, T. W. & Nagel, K. (2012), Sensitivity tests with high resolution accessibility computations, VSP Working Paper 12-22, TU Berlin, Transport Systems Planning and Transport Telematics. See www.vsp.tu-berlin.de/publications.

Nicolai T. W., Wang L., Nagel K. & Waddell, P. (2011), Coupling an urban simulation model with a travel model - A first sensitivity test, *in* E. Hope, J. Bennett & A. Stuart, eds, 'Proceedings of the 12th International Conference on Computers in Urban Planning and Urban Management (CUPUM)', Lake Louise, Canada. Also VSP WP 11-07, see www.vsp.tu-berlin.de/publications

Nischalke, P. (2006), Die Organisation wachsender Unternehmen: Eine Entwicklung idealtypischer Gestaltungsalternativen auf system- und kontingenztheoretischer Basis, PhD thesis, Handelshochschule Leipzig, Dissertation 2005, Hampp, Mering. Schriften zu Organisation und Information.

Noth, M., Borning, A. & Waddell, P. (2003), 'An extensible, modular architecture for simulating urban development, transportation, and environmental impacts', *Computers, Environment and Urban Systems* **27**(2), 181–203.

Openshaw, S. & Rao, L. (1995), 'Algorithms for reengineering 1991 census geography', *Environment and Planning A* **27**(3), 425–446.

Openshaw, S. & Taylor, P. J. (1979), A million or so correlation coefficients: Three experiments on the modifiable areal unit problem, *in* N. Wrigley, ed., 'Statistical Applications in the Social Sciences', Plon, London.

Openshaw, S. et al. (1984), 'Ecological fallacies and the analysis of areal census data', *Environment and Planning A* **16**(1), 17–31.

OPUS User Guide (2011), *The Open Platform for Urban Simulation and UrbanSim Version 4.3*, University of California Berkeley and University of Washington.
URL: http://www.urbansim.org

Orcutt, G., Caldwell, S. & Wertheimer, R. (1976), *Policy Exploration Through Microanalytic Simulation*, The Urban Institute, Washington, DC.

Orcutt, G. H. (1957), 'A new type of socio-economic system', *The Review of Economics and Statistics* **39**(2), 116–123.

Orcutt, G. H. (1960), 'Simulation of economic systems', *The American Economic Review* **50**(5), 894–907.
URL: http://www.jstor.org/stable/1810951

Orcutt, G. H., Greenberger, M., Korbel, J. J., Rivlin, A. M. et al. (1961), *Microanalysis of socioeconomic systems: A simulation study*, Vol. 6, Harper New York.

Ortúzar, J. d. D. & Willumsen, L. (2001), *Modelling transport*, 3. edn, John Wiley Sons Ltd, Chichester.

Otani, N., Sugiki, N., Vichiensan, V. & Miyamoto, K. (2012), Modifiable attribute cell problem and a method of solution for population synthesis in land-use microsimulation, *in* 'Transportation Research Board 91st Annual Meeting'.

Parr, J. B. (2007), 'Spatial definitions of the city: four perspectives', *Urban Studies* **44**(2), 381–392.

Patterson, Z. & Bierlaire, M. (2010), 'Development of prototype UrbanSim models', *Environment and Planning B* **37**(2), 344–366.

Patterson, Z. & Hurtubia, R. (2008), A prototype integrated transportation land-use model for the lausanne region, Technical report, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.

Patterson, Z., Kryvobokov, M., Marchal, F. & Bierlaire, M. (2010), 'Disaggregate models with aggregate data: Two UrbanSim applications', *Journal of Transport and Land Use* **3**(2), 5–37.

Pellenbarg, P. H. (2005), Firm migration in the netherlands, *in* '45th ERSA congress, Amsterdam'.

Peltoniemi, M. (2011), 'Reviewing industry life-cycle theory: avenues for future research', *International Journal of Management Reviews* **13**, 349–375.

Pendyala, R., Bhat, C., Goulias, K., Paleti, R., Konduri, K., Sidharthan, R., Hu, H.-H., Huang, G. & Christian, K. (2012), 'Application of socioeconomic model system for activity-based modeling', *Transportation Research Record* **2303**, 71–80.

Pennec, S. (1997), 'Four-generation families in France', *Population an English Selection* **9**, 75–100.

Pennec, S., Bacon, B. & Gong, C. (2009), The Challenges of Modelling Demographics in Dynamic Microsimulation Models: An Examination of Fertility Modelling and the Australian APPSIM Model, *in* '2nd international conference of the International Microsimulation Association', Ottawa (Canada), 8-11 juin 2009. Oral Presentation.

Pezzey, J. C. V. & Toman, M. A. (2002), The economics of sustainability: A review of journal articles, Technical report, Ahsgate, U.K.

Pfingsten, F. (1998), Shareholder-Value im Lebenszyklus: Methoden einer marktwertorientierten Unternehmensfuehrung, PhD thesis, Univ. Muenchen, Deutscher Universitaetsverlag, Wiesbaden.

Pholo Bala, A., Peeters, D. & Thomas, I. (2011), Delineation of Brussels, Paris and Zuerich: Towards functional urban regions, *in* 'SustainCity Working Paper', www.sustaincity.org.

Pholo Bala, A., Peeters, D. & Thomas, I. (2013), *Spatial Issues on a Hedonic Estimation of Rents in Brussels*, CORE Discussion Paper 2013/36.

Picard, N. & Antoniou, C. (2011), Econometric guidance, Technical report, SustainCity European Project.

Picard, N., de Palma, A. & Dantan, S. (2013), 'Intra-household discrete choice models of mode choice and residential location', *International Journal of Transport Economics* **XL**(3), 419–445.

Pinjari, A. R. & Bhat, C. R. (2010), 'A multiple discrete-continuous nested extreme value (MDCNEV) moddel: Formulation and application to non-worker activity time-se and timing behavior on weekdays', *Transportation Research Part B* **44**(4), 562–583.

Polebitski, A. S., Palmer, R. N. & Waddell, P. (2011), 'Evaluating water demands under climate change and transitions in the urban environment', *Journal of Water Resources Planning and Management* **137**(3), 249–257.

Poole, D. & Raftery, A. E. (2000), 'Inference for deterministic simulation models: the bayesian melding approach', *Journal of the American Statistical Association* **95**(452), 1244–1255.

Porter, M. E. (1999), *Wettbewerbsstrategie: Methoden zur Analyse von Branchen und Konkurrenten, uebersetzt von V. Brandt und T.C. Schwoerer, Originaltitel: Competitive strategy, 10*, Vol. 10., Campus Verlag, Frankfurt a.M., Brandt und T. C. Schwoerer, Originaltitel.

Porter, M. E. (2011), *Competitive advantage of nations: creating and sustaining superior performance*, Simon and Schuster.

Pritchard, D. R. (2008), Synthesizing agents and relationships for land use/transportation modelling, Master's thesis, University of Toronto.
URL: \url{http://hdl.handle.net/1807/17214}

Pritchard, D. R. & Miller, E. J. (2012), 'Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously', *Transportation* **39**(3), 685–704.

PSRC (2012), 2012 land use forecast, Technical Report White Paper 2936, Puget Sound Regional Council.

PTV (2012), *PTV VISUM 12.5, Fundamentals.*, PTV, Karlsruhe, Germany.

PTV AG (2009*a*), *VISUM 11.0 Benutzerhandbuch*, 76131 Karlsruhe.

PTV AG (2009*b*), *VISUM 11.0 Grundlagen*, 76131 Karlsruhe.

Pumain, D. & Haag, G. (1991), 'Urban and regional dynamics – towards an integrated approach', *Environment and Planning A* **23**(9), 1301–1313.
URL: http://www.envplan.com/abstract.cgi?id=a231301

Putman, S. (1983), *Integrated Urban Models: Policy Analysis of Transportation and Land Use*, Pion, London.

Putman, S. H. (1974), 'Preliminary results from an integrated transportation and land use models package', *Transportation* **3**(3), 193–224.

Putman, S. H. (1996), 'Extending dram model: Theory-practice nexus', *Transportation Research Record* **1552**, 112–119.

Quigley, J. M. (1985), 'Consumer choice of dwelling, neighborhood and public services', *Regional Science and Urban Economics* **15**(1), 41–63.

R Development Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.
URL: \url{http://www.r-project.org}

Raftery, A. E., Givens, G. H. & Zeh, J. E. (1995), 'Inference from a deterministic population dynamics model for bowhead whales', *Journal of the American Statistical Association* **90**(430), 402–416.

Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92**(437), 179–191.

Rahman, A., Harding, A. M., Tanton, R. & Liu, S. (2010), 'Methodological issues in spatial microsimulation modelling for small area estimation', *International Journal of Microsimulation* **3**(2), 3–22.

Raney, B. & Nagel, K. (2006), An improved framework for large-scale multi-agent simulations of travel behaviour, *in* P. Rietveld, B. Jourquin & K. Westin, eds, 'Towards better performing European Transportation Systems', Routledge, London, pp. 305–347.

Redfearn, C. L. (2007), 'The topography of metropolitan employment: Identifying centers of employment in a polycentric urban area', *Journal of Urban Economics* **61**(3), 519–541.

Rho, J. H. & Kim, T. J. (1989), 'Solving a three-dimensional urban activity model of land use intensity and transport congestion', *Journal of Regional Science* **29**(4), 595–613.
URL: http://www3.interscience.wiley.com/journal/120004339/references

Rich, J. & Mulalic, I. (2012), 'Generating synthetic baseline populations from register data', *Transportation Research Part A* **46**(3), 467–479.

Rieser, M. (2010), Adding transit to an agent-based transportation simulation concepts and implementation, PhD thesis, TU Berlin. Also VSP WP 10-05, see www.vsp.tu-berlin.de/publications.

Rieser, M., Dobler, C., Dubernet, T., Grether, D., Horni, A., Lämmel, G., Waraich, R., Zilske, M. & Axhausen, K. W. (2013), MATSim user guide, Technical report, VSP, TU Berlin and IVT, ETHZ and Senozon.
URL: http://www.matsim.org/userguide

Rieser, M. & Nagel, K. (2009), Combined agent-based simulation of private car traffic and transit, *in* 'Proceedings of The 12th Conference of the International Association for Travel Behaviour Research (IATBR)', Jaipur, India. Also VSP WP 09-11, see www.vsp.tu-berlin.de/publications.

Rieser, M., Nagel, K., Beuck, U., Balmer, M. & Rümenapp, J. (2007), 'Agent-oriented coupling of activity-based demand generation with multiagent traffic simulation', *Transportation Research Record* **2021**, 10–17.

Riguelle, F., Thomas, I. & Verhetsel, A. (2007), 'Urban polycentrism: a measurable reality. the case of brussels', *Journal of Economic Geography* **7**, 193–215.

Robinson, W. S. (1950), 'Ecological correlations and the behavior of individuals', *American Sociological Review* **15**(3), 351–357.

Röder, D., Cabrita, I. & Nagel, K. (2013), Simulation-based sketch planning, part III: Calibration of a MATSim-model for the greater Brussels area and investigation of a cordon pricing for the highway ring, VSP working paper 13-16, TU Berlin, Berlin, Germany. See www.vsp.tu-berlin.de/publications.

Rodrigue, J.-P., Comtois, C. & Slack, B. (2009), *The geography of transport systems*, Routledge.

Romer, P. M. (1986), 'Increasing returns and long-run growth', *The journal of political economy* pp. 1002–1037.

Roorda, M. J., Carrasco, J. A. & Miller, E. J. (2009), 'An integrated model of vehicle transactions, activity scheduling and mode choice', *Transportation Research Part B* **43**(2), 217–229.
URL: http://www.sciencedirect.com/science/article/pii/S0191261508000623

Roorda, M. J., Miller, E. J. & Habib, K. M. N. (2008), 'Validation of TASHA: A 24-h activity scheduling microsimulation model', *Transportation Research Part A* **42**(2), 360–375.

Rosen, S. (1974), 'Hedonic prices and implicit markets: Product differentiation in pure competition', *Journal of Political Economy* **82**(1), 34–55.

Rosenbaum, A. M. & Koenig, B. E. (1997), Evaluation of the modelling tools for assessing land use policies and strategies, Technical report, Technical report prepared for the transportation and market incentives group, United States Environmental Protection Agency.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Ruggles, S. (1987), *Prolonged connections: the rise of the extended family in nineteenth-century England and America*, Social demography, University of Wisconsin Press. http://www.hist.umn.edu/ ruggles/Articles/ISBNO-299-11034-6.pdf.

Ruming, K. (2010), 'Developer Typologies in Urban Renewal in Sydney: Recognising the Role of Informal Associations between Developers and Local Government', *Urban Policy and Research* **28**(1), 65–83.

Ryan, J., Maoh, H. & Kanaroglou, P. (2009), 'Population synthesis: Comparing the major techniques using a small, complete population of firms', *Geographical Analysis* **41**(2), 181–203.

Salvini, P. & Miller, E. J. (2005), 'ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems', *Networks and Spatial Economics* **5**(2), 217–234.

Schaetzl, L. (2003), 'Wirtschaftsgeographie 1: Theorie,', *Uni-Taschenbuecher UTB, Schoeningh, Paderborn* **782, 9. Auflage**.

Schafer, A. (2000), 'Regularities in travel demand: an international perspective', *Journal of transportation and statistics* **3**(3), 1–31.

Schelling, T. (1978), Micromotives and macrobehaviour, Technical report, W. W. Norton, N-Y.

Schelling, T. C. (1971), 'Dynamic models of segregation', *Journal of mathematical sociology* **1**(2), 143–186.

Schirmer, P. M., van Eggermond, M. A. B. & Axhausen, K. W. (2013), Measuring location in residential location choice: An empirical study on the canton of Zurich, *in* CUPUM, ed., '13th International Conference on Computers in Urban Planning and Urban Management (CUPUM)', University of Utrecht, Utrecht.

Schirmer, P., Van Eggermond, M. & Axhausen, K. (2012), Reviewing measurements in residential location choice models, *in* IATBR, ed., '13th International Conference on Travel Behaviour Research (IATBR)', Toronto.

Schirmer, P., Zöllig, C., Müller, K., Bodenmann, B. R. & Axhausen, K. W. (2011), The Zurich case study of UrbanSim, *in* ERSA, ed., 'Proceedings of the 51th Congress of the European Regional Science Association', Barcelona.
URL: http://e-collection.library.ethz.ch/eserv/eth:3085/eth-3085-01.pdf

Schüssler, R. & Thalmann, P. (2005), Was treibt und hemmt den Wohnungsbau? Ergebnisse einer Bauträger- und Investorenbefragung, Technical Report 76, Bundesamt für Wohnungswesen (BWO), Grenchen.

Scott, A., Evandrou, M., Falkingham, J., Johnson, P. & Zaidi, A. (2003), 'Beyond 2020: Lessons from the development of the sage dynamic microsimulation model for britain', *ESRC-Sage Discussion Paper* **13**.

Scott, D. M. & Kanaroglou, P. S. (2002), 'An activity-episode generation model that captures interactions between household heads: development and empirical analysis', *Transportation Research Part B* **36**(10), 875–896.

SDRIF (2008), Schéma directeur de la région île-de-france (projet adopté), Technical report, Conseil régional d'Île-de-France.

Sener, I. N., Pendyala, R. M. & Bhat, C. R. (2011), 'Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior', *Journal of Transport Geography* **19**(2), 294–303.

Sermons, M. W. & Koppelman, F. S. (2001), 'Representing the differences between female and male commute behavior in residential location choice models', *Journal of Transport Geography* **9**(2), 101–110.

Ševčíková, H., Raftery, A. & Waddell, P. (2007), 'Assessing uncertainty in urban simulations using Bayesian melding', *Transportation Research B* **41**(6), 652–669.

Ševčíková, H., Raftery, A. & Waddell, P. (2011), 'Uncertain benefits: Application of Bayesian melding to the Alaskan Way Viaduct in Seattle', *Transportation Research A* **45**, 540–553.

SGP (2011), Schéma d'ensemble, Technical report, Société du Grand Paris.

SGP (2012), 'Société du grand paris, english version'.
URL: http://www.societedugrandparis.fr/english

Shen, L. Y., Jorge Ochoa, J., Shah, M. N. & Zhang, X. (2011), The application of urban sustainability indicators–a comparison between various practices., 35 1, habitat International, 17-29.

Side Effects Software Inc. (2013), 'Side Effects Software – Houdini 3d Animation Tools'. Accessed December 2013.
URL: http://www.sidefx.com/

Simmonds, D. (2001), The objectives and design of a new land-use modelling package: DELTA, *in* G. P. Clarke & M. Madden, eds, 'Regional Science in Business', Springer, Berlin/Heidelberg, pp. 159 – 188.

Simmonds, D. C. (1999), 'The design of the DELTA land-use modelling package', *Environment and Planning B* **26**(5), 665–684.

Simmonds, D. & Feldman, O. (2005), Land-use modelling with delta: Update and experience,, in *9th Conference on Computers in Urban planning and urban management (CUPUM) 9th Conference on Computers in Urban planning and urban management (CUPUM)* (2005). See www.cupum.org.

Simmonds, D., Waddell, P. & Wegener, M. (2013), 'Equilibrium versus dynamics in urban modelling', *Environment and Planning B: Planning and Design* **40**(6), 1051–1070.

Skaburskis, A. (1999), 'Modelling the choice of tenure and building type', *Urban Studies* **36**(13), 2199–2215.

Smith, L., Beckman, R., Baggerly, K., Anson, D. & Williams, M. (1996), 'TRANSIMS TRansportation ANalysis and SIMulation system project summary and status', NASA Open Source Agreement Version 1.3.
URL: \url{http://ntl.bts.gov/DOCS/466.html}

Spiekerman, K. & Wegener, M. (2004), 'Evaluating urban sustainability using land-use transport interaction models', *European Journal of Transport and Infrastructure research* **4**(3), 251–272.

Srinivasan, S. & Bhat, C. R. (2005), 'Modeling household interactions in daily in-home and out-of-home maintenance activity participation', *Transportation* **32**, 523–544.

Srinivasan, S. & Bhat, C. R. (2006), 'A multiple discrete-continuous model for independent- and joint-discretionary-activity participation decisions', *Transportation* **33**, 497–515.

Staatskanzlei des Kantons Zürich (2012), Kantonale Volksinitiative zum Erhalt der landwirtschaftlich und ökologisch wertvollen Flächen (Kulturlandinitiative). Abstimmungsergebnis, *in* 'Zürcher Amtsblatt', Kanton Zürich, Zürich, pp. 1352–1356.

STASA (2013), 'STASA - publications'. Accessed December 2013.
URL: http://www.stasa.de/en/Veroeffentlichungen.shtml

Storey, D. & Tether, B. (1996), 'A review of the empirical knowledge and an assessment of statistical data on the economic importance of new technology based firms (ntbfs) in europe', *EIMS Publication, The European Innovation Monitoring System (EIMS), Luxemburg* **44**.

Stratec (2003), Evaluation et optimisation des mesures d'accompagnement du R.E.R desservant l'agglomration centre sur la Région de Bruxelles-Capitale, Technical report, Stratec, Brussels, Belgium. see http://www.stratec.be.

Stratec (2005), FP5 European project SCATTER - Sprawling cities and transport: from evaluation to recommandations (2002-2005) - Case study Brussels, Technical report, Stratec, Brussels, Belgium. see http://www.stratec.be.

Stratec (2006), Mise à jour et adaptation du Plan des Déplacements Urbains de la Région de Bruxelles-Capitale, Technical report, Bruxelles-Mobilité, Stratec, Brussels, Belgium. see http://www.stratec.be.

Strauch, D., Moeckel, R., Wegener, M., Gräfe, J., Mühlhans, H., Rindfuesser, G. & Beckmann, K.-J. (2005), Linking transport and land use planning: The microscopic dynamic simulation model ILUMASS, *in* P. M. Atkinson, G. M. Foody, S. E. Darby & F. Wu, eds, 'Geodynamics', CRC Press, Boca Raton, Florida, pp. 295–311.

Sundberg, O. (2007), Model 5: SESIM (longitudinal dynamic microsimulation model), *in* 'International Symposia in Economic Theory and Econometrics', Vol. 16, Emerald Group Publishing Limited, pp. 453–460.

Sweeney, L. (2002), 'Achieving k-anonymity privacy protection using generalization and suppression', *Journal on Uncertainty, Fuzziness, and Knowledge-based Systems* **10**(5), 571–588.

Swiss Federal Statistical Office (BFS) (2000*a*), 'Eidgenössische Volkszählung 2000'. **URL:** \url{http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/vz/uebersicht.html}

Swiss Federal Statistical Office (BFS) (2000*b*), 'Public use samples (PUS): Excerpts for general use from the Swiss federal population censuses 1970-2000'. **URL:** \url{http://www.portal-stat.admin.ch/pus/files/index_e.html}

Swiss Federal Statistical Office (BfS) (2005), 'Mikrozensus Verkehr 2005'. **URL:** http://www.bfs.admin.ch/bfs/portal/de/index/infothek/erhebungen__quellen/blank/blank/mz/01.html

Swiss Federal Statistical Office (BfS) (2008), 'Eidgenössische Betriebszählung 2008'. **URL:** http://www.bfs.admin.ch/bfs/portal/de/index/dienstleistungen/geostat/datenbeschreibung/eidgenoessische_betriebszaehlung2.html

Swiss Federal Statistical Office (BfS) (2011), 'Classifications, inventories – NOGA general classification of economic activities'. **URL:** http://www.bfs.admin.ch/bfs/portal/en/index/infothek/nomenklaturen/blank/blank/noga0/vue_d_ensemble.html

Tannier, C. & Thomas, I. (2013), 'Defining and characterizing urban boundaries: A fractal analysis of theoretical cities and belgian cities', *Computer, Environment and Urban Systems* **41**, 234–248.

Tannier, C., Thomas, I., Vuidel, G. & Frankhauser, P. (2011), 'A fractal approach to identifying urban boundaries', *Geographical Analysis* **43**(2), 211–227.

Te Brömmelstroet, M. & Bertolini, L. (2010), 'Integrating land use and transport knowledge in strategy-making', *Transportation* **37**(1), 85–104.

Thalmann, P. (2012), 'Housing market equilibrium (almost) without vacancies', *Urban Studies* **49**(8), 1643–1658.

The UrbanSim Project (2011), Opus: The open platform for urban simulation and UrbanSim version 4.3 - users guide and reference manual, Technical report, University of California Berkeley and University of Washington, Berkeley, California.

THEMA-TT-R (2002), QUATUOR: Outils dynamiques de simulation pour la gestion des déplacements en Ile-de-France, Technical report, DRAST/ PREDIT.

Thisse, J. (2010), 'Toward a unified theory of economic geography and urban economics', *Journal of Regional Science* **50**, 281–296.

Thisse, J.-F. (2012), Avis du conseil scientifique sur l'évaluation socio-économique du schéma densemble du réseau de transport public du Grand Paris, Technical report, Société du Grand Paris.

Thisse, J.-F. & Thomas, I. (2010), 'Brussels within the belgian economy: a geo-economic approach', *What does geography teach us about the future of Belgium's institutions* .

Thomas, I., Cotteels, C., Jones, J. & Peeters, D. (2012), 'Revisiting the extension of the brussels urban agglomeration: new methods, new dataö new results?', *Belgeo. Revue belge de géographie* **1**(1-2).

Thomas, I., Vanneste, D. & Querriau, X. (2011), *Atlas de Belgique. Partie 4: Habitat*, Academia Press, Ghent, Belgique.

Thompson, W. R. & Barnbrock, J. (1975), Interne und externe faktoren bei der entwicklung der staedtischen wirtschaft, *in* 'Materialien zur Oekonomie der Stadtplanung, Bauwelt Fundament', Vieweg, Braunschweig.

Thunig, T. & Nicolai, T. W. (2013), Spatial interpolation of accessibilities, VSP Working Paper 13-07, TU Berlin, Transport Systems Planning and Transport Telematics. See www.vsp.tu-berlin.de/publications.

Timmermans, H. J. (2006), Analyses and models of household decision making processes. resource paper for the workshop on "group behavior"., *in* 'Proceedings of the 11th International Conference on Travel Behaviour Research, Kyoto, Japan.'.

Timmermans, H. J. & Zhang, J. (2009), 'Modeling household activity travel behavior: Examples of state of the art modeling approaches and research agenda', *Transportation Research Part B* **43**(2), 187 – 190.

Tomassini, C. & Wolf, D. A. (2000), 'Shrinking kin networks in italy due to sustained low fertility', *European Journal of Population/Revue européenne de Démographie* **16**(4), 353–372.

Train, K. E. (2009), *Discrete Choice Methods with Simulation*, 2nd edn, Cambridge University Press, Cambridge.
URL: http://www.econ.berkeley.edu/books/choice2.html

TRB (2007), Metropolitan travel forecasting, current practice and future direction, Technical report, Transportation Research Board, Committee for Determination of the State of the Practice in Metropolitan Area Travel Forecasting.

Turci, L., Bringe, A., Morand, E., Pennec, S., Toulemon, L., Baggio, R. & Billari, F. (2010), 'Provisional demographic outline, SustainCity working paper, 4.1', *FP7-244557 Projet SustainCity* .
URL: http://www.sustaincity.org/publications

Turci, L., Pennec, S., Toulemon, L., Bringe, A., Morand, E. & Baggio, R. (2012), 'Demographic Model User Guide, Sustaincity Working Paper, 4.4,', *FP7-244557 Projet SustainCity* .
URL: http://www.sustaincity.org/publications

UrbanSim (2011), 'Open Platform for Urban Simulation', webpage.
URL: \url{http://www.urbansim.org}

UrbanSim Developers (2013), 'UrbanSim WebHome'. Accessed December 2013.
URL: http://www.urbansim.org/Main/WebHome

UrbanSim Project (2011), *The Open Platform for Urban Simulation and UrbanSim Version 4.3 Users Guide and Reference Manual*, The UrbanSim Project, University of California Berkeley and University of Washington. Available from http://www.urbansim.org.

Vallin, J. & Meslé, F. (2001), *Table de mortalité françaises pour les XIXe et XXe siècles et projection pour le XXIe siècle*, Données Statistiques, Institut National d'Études Démographiques, Paris. 102 p.

van der Vlist, E. (2002), *XML Schema.*, 1 st edn, O'Reilly.

Van Hecke, E., Halleux, J.-M., Decroly, J.-M. & Mérenne-Schoumaker, B. (2009), *Noyaux d'habitat et Régions urbaines dans une Belgique urbanisées, Enquête Socio–Economique 2001*, Monographies, Direction générale Statistique et Information économique.

Van Imhoff, E. & Post, W. (1998), 'Microsimulation methods for population projection', *Population: An English Selection* pp. 97–138.

Van Soest, A. (1995), 'Structural models of family labor supply: A discrete choice approach', *The Journal of Human Resources* **30**(1), 63–88.

van Steen, P. (2005), 'Bedrijvendynamiek onder het vergrootglas, in p.h. pellenbarg, p.j.m. van steen und l.j.g. van wissen', *Ruimtelijke aspecten van de bedrijvendynamiek in Nederland, van Gorcum, Assen* .

van Wezemael, J. E. (2005), *Investieren im Bestand*, Ostschweizerische Geographische Gesellschaft, St.Gallen.

Van Wissen, L. (2000), 'A micro-simulation model of firms: applications of concepts of the demography of the firm', *Papers in Regional Science* **79**(2), 111–134.

Vandenbulcke, G. (2007), *Accessibility indicators to places and transports. Final Report*, Service public federal Mobilité et Transports.

Vandermotten, C. (2008), Bruxelles et son hinterland economiques, *in* 'Colloque du conseil économique et social de la Region de Bruxelles-Capitale', Bruxelles.

Vandermotten, C., Vermoesen, F., De Lannoy, W. & De Corte, S. (1999), 'Villes d'europe. cartographie comparative', *Bulletin du Crédit Communal* pp. 207–208.

Vanegas, C., Aliaga, D., Müller, P., Waddell, P. & Watson, B. (2009), 'Modeling the Appearance and Behavior of Urban Spaces', *COMPUTER GRAPHICS forum* **28**(2), 1 – 18.

Vanneste, D., Thomas, I. & Goossens, L. (2007), Le logement en Belgique, Technical report, Statistics Belgium, Bruxelles, Belgique.

Verburg, P. H., Schot, P. P., Dijst, M. J. & Veldkamp, A. (2004), 'Land use change modelling: current practice and research priorities', *GeoJournal* **61**(4), 309–324. **URL:** http://dx.doi.org/10.1007/s10708-004-4946-y

Verhetsel, A., Van Hecke, E., Thomas, I., Beelen, M., Halleux, J., Lambotte, J., Rixhon, G. & Merenne-Shoumaker, B. (2009), Le mouvement pendulaire en Belgique. les déplacements domicile-lieu de travail. les déplacements domicile-école. Enquête Socio-économique 2001. Monographies (n°10), Technical report, Statistics Belgium, Bruxelles, Belgique.

Vermeulen, F. (2002*a*), 'Collective household models: Principles and main results', *Journal of Economic Surveys* **16**(4), 533–564.

Vermeulen, F. (2002*b*), Where does the unitary model go wrong? simulating tax reforms by means of unitary and collective labour supply models. the case for belgium, *in* F. Laisney, ed., 'Welfare analysis of fiscal and social security reforms in Europe: does the representation of family decision processes matter?', Final report on EU-project VS/2000/0778.

Vermeulen, F. (2006), 'A collective model for female labour supply with non-participation and taxation', *Journal of Population Economics* **19**, 99–118.

Vickrey, W. S. (1969), 'Congestion theory and transport investment', *The American Economic Review* **59**(2), 251–260.

Voas, D. & Williamson, P. (2000), 'An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata', *International Journal of Population Geography* **6**(5), 349–366.

Von Neumann, J. (1966), *Theory of Self-Reproducing Automata*, University of Illinois Press Champaign.

von Thünen, J. (1826), *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationaloekonomie*, Hamburg.

Vovsha, P., Petersen, E. & Donnelly, R. (2003), 'Explicit modeling of joint travel by household members: Statistical evidence and applied approach', *Transportation Research Record: Journal of the Transportation Research Board* **1831**, 1–10.

Vovsha, P., Petersen, E. & Donnelly, R. (2004*a*), 'Impact of intrahousehold interactions on individual daily activity-travel patterns', *Transportation Research Record: Journal of the Transportation Research Board* **1898**, 87–97.

Vovsha, P., Petersen, E. & Donnelly, R. (2004*b*), 'Model for allocation of maintenance activities to household members', *Transportation Research Record: Journal of the Transportation Research Board* **1894**, 170–179.

Vrtic, M., Fröhlich, P., Schüssler, N., Kern, P., Perret, F., Pfisterer, S., Schulze, C., Zimmermann, A., Heidl, U. & Axhausen, K. W. (2005), Verkehrsmodell für den öffentlichen Verkehr des Kantons Zürich, Technical report, Amt für Verkehr des Kantons Zürich, IVT, ETH Zürich, Ernst Basler + Partner AG, Planung Transport Verkehr (PTV), Zürich.

W3C (2008), *eXtensible Markup Language (XML)*, World Wide Web Consortium (W3C). See www.w3.org/XML.

Wachter, K. W., Blackwell, D. & Hammel, E. A. (1997), 'Testing the validity of kinship microsimulation', *Mathematical and Computer Modelling* **26**(6), 89–104.

Waddell, P. (1996), Accessibility and residential location: The interaction of workplace, housing tenure, residential mobility and location choices. Paper presented at the 1996 Lincoln Land Institute TRED Conference.

Waddell, P. (2000), 'A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim', *Environment and Planning B: Planning and Design* **27**, 247 – 263.

Waddell, P. (2002), 'UrbanSim: Modeling urban development for land use, transportation, and environmental planning', *Journal of the American Planning Association* **68**(3), 297–314.

Waddell, P. (2005), 'Towards an open platform for urban simulation'. Presentation at 5th Swiss Transport Research Conference, Ascona.

Waddell, P. (2010), Modeling residential location in UrbanSim, *in* F. Pagliara, J. Preston & D. Simmonds, eds, 'Residential Location Choice, Models and Applications', Advances in Spatial Science, Springer Berlin Heidelberg, pp. 165–180.

Waddell, P. (2011), 'Integrated land use and transportation planning modelling: Addressing challenges in research and practice', *Transport Reviews* **31**(2), 209–229.

Waddell, P., Bhat, C., Eluru, N., Wang, L. & Pendyala, R. (2007), 'Modeling interdependence in household residence and workplace choices', *Transportation Research Record* **2003**, 84–92.

Waddell, P. & Borning, A. (2004), 'A case study in digital government: Developing and applying UrbanSim, a system for simulating urban land use, transportation, and environmental impacts', *Social Science Computer Review* **22**(1), 37–51.

Waddell, P., Borning, A., Noth, M., Freier, N., Becke, M. & Ulfarsson, G. (2003), 'Microsimulation of urban development and location choices: Design and implementation of UrbanSim', *Networks and Spatial Economics* **3**(1), 43–67.

Waddell, P., Ševčíková, H., Socha, D., Miller, E. & Nagel, K. (2005), Opus: An Open Platform for Urban Simulation, *in* '9th Conference on Computers in Urban Planning and Urban Managment (CUPUM)', University College London, UK.
URL: http://128.40.111.250/cupum/searchpapers/papers/paper428.pdf

Waddell, P. & Ulfarsson, G. (2003*a*), Dynamic simulation of real estate development and land prices within and integrated land use and transportation model system, *in* 'Proceedings of the 82nd Annual Meeting of the Transportation Research Board'.

Waddell, P. & Ulfarsson, G. F. (2003*b*), Accessibility and agglomeration: Discrete-choice models of employment location by industry sector, *in* 'TRB Annual Meeting', Transportation Research Board.

Waddell, P. & Ulfarsson, G. F. (2004), Introduction to urban simulation: Design and development of operational models, *in* D. A. Hensher, K. L. Button, K. E. Haynes & P. R. Stopher, eds, 'Handbook of transport geography and spatial systems', Vol. 5 of *Handbook in Transport*, Elsevier Ltd, Oxford, pp. 203–234.

Waddell, P., Ulfarsson, G., Franklin, J. & Lobb, J. (2007), 'Incorporating land use in metropolitan transportation planning', *Transportation Research Part A* **41**(5), 382–410.

Waddell, P., Wang, L. & Charlton, B. (2007), Integration of a parcel-level land use model and an activity-based travel model, *in* 'Proceedings of the 87th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, DC.

Waddell, P., Wang, L., Charlton, B. & Olsen, A. (2010), 'Microsimulating parcel-level land use and activity-based travel: Development of a prototype application in San Francisco', *Journal of Transport and Land Use* **3**(2), 65–84.

Waddell, P., Wang, L. & Liu, X. (2008), UrbanSim: An evolving planning support system for evolving communities, *in* R. Brail, ed., 'Planning Support Systems for Cities and Regions', Cambridge, MA: Lincoln Institute for Land Policy, pp. 103–138.

Wagner, P. & Wegener, M. (2007), 'Urban Land Use, Transport and Environment Models, Experiences with an Integrated Microscopic Approach', *disP* **170**, 45 – 56.

Walker, J. L. (2001), Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables, PhD thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Wang, L. & Waddell, P. (2013), Disaggregated real estate demand model with price formation for integrated land use and transportation modeling, *in* 'Transportation Research Board 92nd Annual Meeting'.

Wang, Y., Kockelman, K. M. & Wang, X. (2011), Anticipating land use change using geographically weighted regression models for discrete response, *in* 'Proceedings of the 90th Annual Meeting of the Transportation Research Board', Washington, D.C.

Washington, S. P., Karlaftis, M. G. & Mannering, F. L. (2010), *Statistical and Econometric Models for Transportation Data Analysis*, Chapman & Hall / CRC Press, Boca Raton, FL.

Wegener, M. (1982*a*), 'Modeling urban decline: a multilevel economic-demographic model of the Dortmund region', *International Regional Science Review* **7**(1), 21 – 41.

Wegener, M. (1982*b*), 'A multilevel economic-demographic model for the Dortmund region', *Sistemi Urbani* **3**(4), 371 – 401.

Wegener, M. (1994), 'Operational urban models: State of the art', *Journal of the American Planning Association* **60**, 17–29.

Wegener, M. (2004), Overview of land-use transport models, *in* D. A. Hensher & K. Button, eds, 'Transport Geography and Spatial Systems. Handbook 5 of Handbook in Transport', Pergamon/Elsevier Science, Kidlington, UK.

Wegener, M. (2007), Themes and relationships, *in* S. Marshall & D. Banister, eds, 'Land Use and Transport. European Research towards Integrated Policies', Elsevier, Oxford, United Kingdom.

Wegener, M. (2011), 'From Macro to Micro – How Much Micro is too Much?', *Transport Reviews* **31**, 161 – 177.

Wegener, M. & Fürst, F. (1999), Land-use transport interaction: State of the art, Technical Report 46, Institut für Raumplanung, Universität Dortmund, Dortmund. **URL:** http://spiekermann-wegener.com/pub/pub2.htm

Weidner, T., Donnelly, R., Freedman, J., Abraham, J. E. & Hunt, J. D. (2007), A summary of the Oregon TLUMIP Model Microsimulation Modules, *in* 'Proceedings of the 86th Annual Meeting of the Transportation Research Board', Washington, D.C.

Weisbrod, G., Lerman, S. & Ben-Akiva, M. (1980), 'Tradeoffs in residential location decisions: transportation vs. other factors', *Transportation Policy and Decision Making* **1**, 13–26.

Wen, C.-H. & Koppelman, F. (1999), 'Integrated model system of stop generation and tour formation for analysis of activity and travel patterns', *Transportation Research Record: Journal of the Transportation Research Board* **1676**, 136–144.

Wen, C.-H. & Koppelman, F. (2000), 'A conceptual and methdological framework for the generation of activity-travel patterns', *Transportation* **27**, 5–23.

Wikipedia (2011), 'Croissance de la population parisienne de 1801 à 2008', http://commons.wikimedia.org/wiki/File:Population_Paris_(1801-2008).png. Accessed: 2013-12-31.

Willekens, F. (2006), 'Description of the micro-simulation model (continuous-time micro-simulation)', *Deliverable D8 (first part), MicMac Bridging the micro-macro gap in population forecasting, NIDI, The Netherlands* .

Williamson, P. (2013), An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation, *in* R. Tanton & K. Edwards, eds, 'Spatial Microsimulation: A Reference Guide for Users', number 6 *in* 'Understanding Population Trends and Processes', Springer Netherlands, pp. 19–47.

Williamson, P., Birkin, M. & Rees, P. H. (1998), 'The estimation of population microdata by using data from small area statistics and samples of anonymised records', *Environment and Planning A* **30**(5), 785–816.

Wooldridge, J. (2002), *Econometrics analysis of cross section and panel data*, The MIT Press, Cambridge, London.

World Commission on Environment and Development (1987), Our common future, Technical report, Report of the World Commission on Environment and Development, Published as Annex to General Assembly document A/42/427, Development and International Co-operation: Environment .

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. A. (2009), A methodology to match distributions of both household and person attributes in the generation of synthetic populations, *in* TRB, ed., '88th Annual Meeting of the Transportation Research Board', Transportation Research Board, Washington, D.C.

Yusuf, A. & Resosudarmo, B. (2009), 'Does Clean Air Matter in Developing Countries' Megacities? A Hedonic Price Analysis of the Jakarta Housing Market, Indonesia', *Ecological Economics* **68**(5), 1398–1407.

Zaidi, A. & Rake, K. (2001), 'Dynamic microsimulation models: a review and some lessons for sage', *Simulating Social Policy in an Ageing Society (SAGE) discussion paper* **2**.

Zhao, F. & Chung, S. (2006), A study of alternative land use forecasting models - final report, Technical Report BD015-10, Florida Department of Transportation, Tallahassee, Florida.

Zhao, F. & Park, N. (2004), 'Using geographically weighted regression models to estimate annual average daily traffic', *Transportation Research Record: Journal of the Transportation Research Board* **1879**, 99–107.

Zhou, B. & Kockelman, K. (2008), 'Microsimulation of Residential Land Development and Household Location Choices: Bidding for Land in Austin, Texas', *Transportation Research Record* **2077**, 106–112.

Zio, S. D., Montanari, A. & Staniscia, B. (2010), 'Simulation of urban development in the city of rome: Framework, methodology, and problem solving', *Journal of Transport and Land Use* **3**(2).

Zitouni, B., Mathus, P. & Olivier, V. (2004), Chapitre V : Comportements résidentiels des ménages. in : Rapport Final de la Subvention 1999, Thème 8.3 Réflexion sur le développement d'un logement durable., Technical report, Ministère de la Région Wallonne. Conférence Permanente de Développement Territorial.

Zöllig, C. & Axhausen, K. W. (2012*a*), Heterogeneity of real estate developers in Canton Zurich, *in* 'Proceedings of the 12th Swiss Transport Research Conference', Ascona.
URL: http://e-collection.library.ethz.ch/eserv/eth:5587/eth-5587-01.pdf

Zöllig, C. & Axhausen, K. W. (2012*b*), Uncovering the heterogeneity of real estate developers in the canton zurich. sustaincity working paper 3.6, Technical Report 3.6, IVT, ETH Zürich, Zurich.
URL: http://www.sustaincity.org/publications

Zöllig, C., Hilber, R. & Axhausen, K. W. (2011), Konzeptstudie Flächennutzungsmodellierung – Grundlagenbericht, Technical report, IVT, ETH Zürich und Mappuls AG, Bern.

Zöllig Renner, C. & Axhausen, K. W. (2013), Comparing Estimation Results of Land Use Development Models Using Different Data Bases Available in Switzerland, *in* 'Proceedings of the Transportation Research Board 92nd Annual Meeting', Washington, D.C.
URL: http://e-collection.library.ethz.ch/eserv/eth:5872/eth-5872-01.pdf

Zondag, B. (2007), Joint modeling of land-use, transport and economy, PhD thesis, TU Delft, Delft. uuid:9378cee6-aeae-4e50-88de-a546681a42b3.

Zondag, B. & Pieters, M. (2005), 'Influence of accessibility on residential location choice', *Transportation Research Record: Journal of the Transportation Research Board* **1902**, 63–70. ,.
URL: http://dx.doi.org/10.3141/1902-08

Zorn, P. (1989), 'Mobility-tenure decisions and financial credit : Do mortgage qualification requirements constrain home ownership?', *American Real Estate and Urban Economics Association Journal* **17**(1), 1–16.

Index

accessibility, 17, 19, 20, 34, 37, 38, 47, 51, 60, 66, 67, 138, 150, 153, 163, 166, 174, 177, 187, 190, 195–197, 199, 229, 232, 233, 242–244, 247, 256, 293, 295, 297, 327, 333, 334, 341–343, 347, 349–355, 357–360, 362–365, 375, 387–391, 393, 394, 399–402, 406–409, 413, 422, 425, 428, 429, 431, 432, 436–438, 443, 450, 453, 456, 464, 476–479, 481–487, 491, 492, 504, 508, 513, 515, 520, 526, 528
agent synthesis, 77
agent-based, 2, 17, 23, 34, 36, 41, 51, 78, 113, 123, 311, 333, 334, 341, 348, 353, 355, 394, 461, 463, 465, 475, 476, 489, 508, 509
aggregation, 23, 30, 62, 64, 67, 80, 86, 87, 93, 96, 220, 238, 239, 247, 250, 258, 414, 425, 467, 468, 493, 494, 519
areal unit, 29, 55, 56, 61, 64, 65, 67, 238

closed cross-sectional, 122
clustering analysis, 71
complex households, 113, 122, 131, 133–135, 475
computational costs, 3, 10, 38, 41, 85, 185, 223, 225, 245, 246, 252, 312, 330, 334, 336, 339, 347, 351, 354, 361, 477, 515, 520

data collection, 42, 45–47, 49, 51, 53, 58, 60, 70, 73, 188, 220, 302, 373, 374, 378, 414, 416, 445, 518
data issues, 225
data processing, 42, 45–48, 53, 63, 68, 325, 329, 379, 465–467, 470, 471, 496, 506, 507
debugging, 50, 51, 529
demographics, 3, 4, 33, 36, 37, 47, 113, 114, 121–125, 131, 134, 135, 311, 328, 329, 357, 414, 432, 457, 482, 489, 494, 506, 522, 523, 526
developer types, 181–184, 186, 188, 191, 192, 197, 199, 524
disaggregation, 3, 47, 48, 54, 185, 322, 467, 506
disequilibrium, 30, 41
dynamics, 2, 21, 23, 47, 111, 113, 137–140, 148, 154, 158, 159, 232, 265, 266, 310, 334, 356, 463, 472, 474, 476, 496, 507, 516, 522, 523

ecological fallacy, 68, 73, 258
econometrics, 3, 4, 12, 22, 36, 37, 52, 62, 63, 69, 70, 72, 73, 140, 154, 159, 176, 177, 219, 220, 225, 227, 230, 234, 235, 237–239, 245–247, 250, 256, 258–260, 414, 519, 526, 528
Employment Location Choice Model, 37, 67, 390, 391, 408, 409, 485, 487, 488
endogeneity, 57, 68, 199, 219, 226, 229–232, 235, 387, 410, 413
equilibrium, 2, 21, 22, 28, 29, 34, 115, 133, 159, 185, 187, 265, 266, 268, 270–274, 276, 277, 279,

284–288, 290, 299, 337, 355, 356, 361, 411, 450, 459, 464, 465, 476, 490, 516, 521, 523
estimation, 3, 38, 39, 41, 42, 44–47, 49, 50, 52, 53, 64, 66, 68, 69, 82, 98, 117, 123, 125, 127, 140, 182, 192–195, 197–199, 218, 223, 224, 227, 234–237, 239, 244–251, 253–260, 262, 263, 316, 317, 319, 320, 328, 329, 331, 374, 377, 381, 386, 387, 390–393, 407, 409, 424, 434, 435, 442, 473, 476, 480, 481, 483, 486, 487, 489, 506, 509, 516, 519, 521, 524, 526
Exploratory Spatial Data Analysis, 63
externalities, 2, 247, 298, 300, 413, 516, 525

fitting and generation, 79, 81, 88, 91
fractal, 58–60

geo-located data, 55, 62
geography, 55, 69, 73, 322, 324, 517, 518, 524
Gibbs Sampling, 84, 85, 239, 245

hedonic models, 44, 63, 64, 68, 185, 226, 227, 231, 233, 235, 237–239, 242–244, 253, 256, 259, 374, 410, 424, 439, 472, 476, 480, 506
heterogeneity, 2, 110, 114, 124, 128, 154, 178, 185, 186, 191–193, 199, 224, 226, 229, 245, 265, 358, 441, 459, 463, 520, 524
housing choices, 137, 148

individuals and households, 122, 135, 521
interval regression, 63, 66, 222, 228, 231, 233, 239, 244–246, 248–256, 260, 263
Iterative Proportional Fitting, 46, 82, 91

land use transport interaction (LUTI), 3, 17–23, 26–28, 35, 55–57, 60, 62–65, 68, 69, 72, 73, 182, 184, 229, 333, 513, 518, 525, 529
location choice, 19, 27, 28, 30, 32, 34, 35, 37, 39, 43, 44, 47, 52, 63, 66, 67, 81, 122, 137–140

Markov Chain Monte Carlo (MCMC), 78, 82, 84, 89, 92–96, 379
model diagnostics, 229
model validation, 3, 9, 10, 50, 373, 374, 377, 380, 385, 395, 410, 411, 463, 494–496, 508, 518
Modgen, 121
Modifiable Areal Unit Problem, 64, 67
Moran I, 71
Multi Agent Systems (MAS), 23, 29, 115

Nested Logit, 140, 155, 159, 163, 177, 184, 192, 219, 223, 224, 226, 232, 317, 319, 359, 366, 433, 472, 482, 489

planning, 1–13, 17, 18, 28, 30, 36, 41, 56, 62, 78, 193, 209, 237, 329, 331, 357, 359, 409, 451, 513, 529
policy, 3, 7, 8, 11, 17, 22, 32, 36, 38, 44, 56, 58, 113–115, 117, 166, 200, 220, 229, 247, 250, 293–295, 298–302, 305, 306, 309, 311, 357, 373–375, 377, 385, 401, 409, 411, 414, 416, 417, 427, 442, 459, 507, 515, 517–519, 527–529
population heterogeneity, 128
Project Location Choice Model, 197, 198, 391, 392

Random Utility, 21, 28, 31, 32, 140, 158, 221, 223, 319
real estate development, 36–38, 44, 47, 49, 181–183, 186, 189–192, 200, 324, 357, 391, 507
real estate price model, 37, 38, 44, 69, 219, 228, 233, 235, 238, 259, 318, 329, 342, 386, 388, 408, 414, 439, 468, 480–483, 492

sampling, 52, 62, 63, 66, 73, 82, 93, 103, 168, 174, 188, 192, 197, 199, 219, 225, 226, 231, 234, 262, 318, 319, 324, 326, 327, 345, 382, 410, 421, 433, 472, 476
scale, 1, 3, 4, 6, 8, 29, 56, 60, 61, 67, 68, 186, 204, 237, 238, 242, 243, 250, 258, 299, 362, 414, 429, 457, 459, 462, 464, 487, 489, 509, 514–517, 525, 527, 528
software integration, 333
spatial autocorrelation, 63, 64, 66, 70, 73, 228, 229, 237–239, 246, 256, 261, 263, 472
spatial biases, 55, 63, 65, 66, 68, 73, 237, 259
spatial dependence, 63, 67, 226, 227, 238, 239, 244, 245, 250, 252–254, 256, 258, 259, 262, 263
spatial disaggregation, 377, 380
spatial econometrics, 55, 56, 63, 73, 219, 226–228, 239, 310
spatial heterogeneity, 66, 103, 226–228, 238, 239, 247, 254, 259, 524
spatial scale, 245, 250, 251, 256, 432
spatial units, 238, 242, 245, 246, 250, 256, 258
sustainability indicators, 293, 295
synthetic city, 56, 65, 66
synthetic population, 35, 48, 51, 65, 68, 78–81, 83–86, 91–93, 95, 96, 103, 106–110, 118, 119, 379, 380, 420

tenure, 35, 120, 137–140, 142, 144, 147–150, 154, 157–159, 223, 224, 231, 232, 420, 423, 427, 447, 448, 508
transport simulation, 45, 218, 333, 461, 508

urban delineation, 57
urban environment, 56, 59, 60, 438
urban system, 17–21, 28–30, 32, 77, 362, 514