Geotechnologies and the Environment

Martin Behnisch
Gotthard Meinel *Editors*

Trends in Spatial Analysis and Modelling

Decision-Support and Planning Strategies

Springer

Geotechnologies and the Environment

Volume 19

The *Geotechnologies and the Environment* series is intended to provide specialists in the geotechnologies and academics who utilize these technologies, with an opportunity to share novel approaches, present interesting (sometimes counter-intuitive) case studies, and most importantly to situate GIS, remote sensing, GPS, the internet, new technologies, and methodological advances in a real world context. In doing so, the books in the series will be inherently applied and reflect the rich variety of research performed by geographers and allied professionals.

Beyond the applied nature of many of the papers and individual contributions, the series interrogates the dynamic relationship between nature and society. For this reason, many contributors focus on human-environment interactions. The series are not limited to an interpretation of the environment as nature per se. Rather, the series "places" people and social forces in context and thus explore the many socio-spatial environments humans construct for themselves as they settle the landscape. Consequently, contributions will use geotechnologies to examine both urban and rural landscapes.

More information about this series at http://www.springer.com/series/8088

Martin Behnisch • Gotthard Meinel
Editors

Trends in Spatial Analysis and Modelling

Decision-Support and Planning Strategies

Springer

Leibniz Institute of
Ecological Urban and
Regional Development

Editors
Martin Behnisch
Leibniz Inst of Eco Urb & Reg Dev (IOER)
Dresden, Germany

Gotthard Meinel
Leibniz Inst of Eco Urb & Reg Dev (IOER)
Dresden, Germany

ISSN 2365-0575 ISSN 2365-0583 (electronic)
Geotechnologies and the Environment
ISBN 978-3-319-52520-4 ISBN 978-3-319-52522-8 (eBook)
DOI 10.1007/978-3-319-52522-8

Library of Congress Control Number: 2017955230

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Land is a limited resource. It must be treated with utmost responsibility in order to minimise unavoidable damage to soils and landscape through urban development as well as the numerous environmental problems associated with such damage.

This book is a result of the first International Land Use Symposium (ILUS), which was held in Dresden from 11 to 13 November 2015. Organised by the Leibniz Institute of Ecological Urban and Regional Development (IOER), the symposium's title was "Trends in Spatial Analysis and Modelling of Settlements and Infrastructure". The book's structure reflects the four core themes of this symposium, namely:

- Towards a better understanding of settlements and infrastructure
- Geographic data mining
- Spatial modelling, system dynamics and geosimulation
- Multi-scale representation and analysis

Leading experts from a wide range of institutions have been commissioned by the editors to discuss the various topics. In addition to this book a selection of post-conference full papers was published in the *ISPRS International Journal of Geo-Information*. The aim of this special issue was to publish original research or review papers in order to stimulate further discussions on recent trends in spatial analysis and modelling of built-environment characteristics. All published papers (11/21 submissions) of the special issue are gathered at http://www.mdpi.com/journal/ijgi/special_issues/Built-Environment2015.

ILUS brought together leading academics and interested parties for presentations, discussions and collaborative networking on the issues of settlements and infrastructures. Clearly, these involve many fields of expertise in the spatial sciences, information sciences, environmental studies, geography, cartography, GIScience, urban planning and architecture. The interdisciplinary nature of the symposium encouraged the cross-fertilisation of new ideas from overlapping fields of studies with the goal of advancing our understanding of built-up areas. In particular, participants considered how recent developments in spatial analysis and modelling can foster the sustainable management of resources, support planning and regional development, enhance spatial information and knowledge as well as optimise strategies, instruments and tools. An investigation of settlements and infrastructures throws up many questions: What are likely to be the most relevant challenges and research questions in this topic over the coming years? What data and analysis strategies do we need? What are the strengths and weaknesses of the current frameworks and methods? In what way are developments in theory supported by the quantitative exploration of spatial and process-related interrelations, structures and patterns?

The symposium also included a presentation of current developments and results of the so-called *Monitoring of Settlement and Open Space Development* (www.ioer-monitor.de). This freely available scientific service run by the IOER provides visual illustration, comparison and statistical analysis of almost 80 indicators of land use structure at spatial levels ranging from the whole of Germany down to individual municipalities and regular grids. One recent innovation has been the development of a new smartphone app *Land Use Monitor DE* which presents high-resolution raster maps of local land use to users while also permitting comparison with previous time periods.

We are grateful to all authors for the excellent collaboration in the editing process. Our particular thanks go to the numerous colleagues at the IOER who have assisted us in realising this book.

Dresden, Germany

Martin Behnisch
Gotthard Meinel

Contents

Part I Towards a Better Understanding of Settlements and Infrastructure

Part II Geographic Data Mining

Part III Spatial Modelling, System Dynamics and Geosimulation

Contributors

Jonatan Almagor The Porter School of Environmental Studies, Tel Aviv University, Tel Aviv, Israel

Jorge Antunes Nova Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal

Fernando José Ferreira Lucas Bação Nova Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal

Itzhak Benenson Department of Geography and Human Environment, Tel Aviv University, Tel Aviv, Israel

Jean-Christophe Castella Institut de Recherche pour le Développement, Montpellier, France

Daniel Czamanski Faculty of Architecture and Town Planning, Technion – Israel Institute of Technology, Haifa, Israel

Ton de Nijs National Institute for Public Health and the Environment, De Bilt, The Netherlands

Zengqiang Duan China Agricultural University, Beijing, China

Eric Fotsing Computer Science, University of Dschang, Dschang, Cameroon

Noah Goldstein Lawrence Livermore National Laboratory, Livermore, CA, USA

Julian Hagenauer Leibniz Institute of Ecological Urban and Regional Development, Dresden, Germany

Jiawei Han Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Marco Helbich Department of Human Geography and Planning, Utrecht University, Utrecht, The Netherlands

Roberto Henriques Nova Information Management School, Universidade Nova de Lisboa, Lisbon, Portugal

Bin Jiang Faculty of Engineering and Sustainable Development, Division of GIScience, University of Gävle, Gävle, Sweden

Andreas Koch Department of Geography and Geology, University of Salzburg, Salzburg, Austria

Kasper Kok Soil Geography and Landscape, Wageningen University, Wageningen, The Netherlands

Thomas H. Kolbe Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Munich, Germany

Eric Koomen Spatial Analysis & Modelling, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Stefan Leyk Department of Geography, University of Colorado, Boulder, CO, USA

Christopher D. Lippitt Geography & Environmental Studies, The University of New Mexico, Albuquerque, NM, USA

Galen Maclaurin Department of Geography, University of Colorado, Boulder, CO, USA

William McConnell Center for Sytems Integration and Sustainability, Michigan State University, East Lansing, MI, USA

Bryan Pijanowski Forestry and Natural Resources, Purdue University, West Lafayette, IN, USA

Robert Gilmore Pontius Jr Clark University, Worcester, MA, USA

Michigan State University, East Lansing, MI, USA

Universiti Putra Malaysia, Serdang, Malaysia

Purdue University, West Lafayette, IN, USA

University of Twente, Enschede, The Netherlands

Denise Pumain Université Paris 1 – Panthéon-Sorbonne, Paris, France

Monika Sester Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Hannover, Germany

Maximilian Sindram Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Munich, Germany

Alias Mohd Sood Universiti Putra Malaysia, Selangor, Malaysia

Frank Thiemann Institute of Cartography and Geoinformatics, Leibniz Universität Hannover, Hannover, Germany

A. Tom Veldkamp Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands

Peter Verburg Vrije Universiteit Amsterdam, Institute for Environmental Studies, Amsterdam, The Netherlands

Bruno Willenborg Department of Civil, Geo and Environmental Engineering, Technical University of Munich, Munich, Germany

Quan Yuan Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Chao Zhang Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

About the Editors

Martin Behnisch received his diploma and doctoral degrees at the Department of Architecture, Karlsruhe Institute of Technology. He also received a diploma degree in wood processing technologies (University of Cooperative Education, Dresden, Germany) and a master's degree in geographical information science (University of Salzburg, Austria) with distinction. He worked in Switzerland as a postdoctoral researcher (2007–2011) at the Institute of Historic Building Research and Conservation (ETH Zurich). He is currently a senior scientist at the Leibniz Institute of Ecological Urban and Regional Development (IOER). His research interests are in spatial analysis and modelling, urban data mining, spatial monitoring, land use science as well as building stock research. He has published numerous refereed articles in international journals, scientific books and conference proceedings in his discipline.

Dr. Gotthard Meinel is specialist in the field of monitoring of land use development. His research interests are indicator development, automated spatial analysis of large datasets and visualisation technologies. Since 1992, he has been acting as a project leader in the field of informatics, GIS and remote sensing at Leibniz Institute of Ecological Urban and Regional Development (IOER). Since 2009, he has been head of the research area "Monitoring of Settlement and Open Space Development" at IOER in Dresden. He received an MS in information technology in 1981 and a PhD degree in image processing at Dresden University of Technology in 1987. Later, he was a postdoctoral researcher in biomathematics and technical mathematics. He has published more than 100 research articles in international journals and refereed conference proceedings.

Part I
Towards a Better Understanding of Settlements and Infrastructure

Chapter 1
Reverse Engineering of Land Cover Data: Machine Learning for Data Replication in the Spatial and Temporal Domains

Galen Maclaurin and Stefan Leyk

Abstract Land cover datasets are generally produced from satellite imagery using state-of-the-art model-based classification methods while integrating large amounts of ancillary data to help improve accuracy levels. The knowledge base encapsulated in this process is a resource that could be used to produce new data of similar quality, more efficiently. A central question is whether this richness of information could potentially be extracted from the underlying remote sensing imagery to then classify an image for a different geographic extent or a different point in time. This chapter summarizes the state of research in this field and highlights the most important insights derived from recent studies. Regional and national land cover datasets exist in many countries and the development of automated, robust methods for spatial extrapolation or temporal extension of such data would benefit the scientific community and planning agencies, and advance similar areas of research in the methodological and applied domains.

Keywords Land cover classification • Spatial data replication • Machine learning • Remote sensing • Information extraction

1.1 Introduction

Land cover data serve an important role in physical and social science research by providing a thematic survey of the Earth's surface at broad scales – from regional to global. Integrated with other data sources, such as population, terrain or climate information, land cover classifications allow researchers to ask questions and test hypotheses over large geographic extents. Such integration facilitates an efficient, data-driven process for research on topics ranging from urbanization to soil

G. Maclaurin (✉) • S. Leyk
Department of Geography, University of Colorado, Boulder, CO, USA
e-mail: galen.maclaurin@colorado.edu; stefan.leyk@colorado.edu

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_1

mapping. National level land cover datasets are generally created from publicly available multispectral imagery (e.g., Landsat) using state-of-the-art classification methods, and often integrate ancillary data sources to improve accuracy (e.g., terrain models, vegetation surveys, agricultural databases, etc.). Landsat 5 TM and Landsat 7 ETM+ imagery has been used widely in land cover mapping efforts due to the extensive temporal extent (starting with the Landsat 5 launch in 1984) and the relatively high spatial resolution (30 m) compared to other publically available image sources. For example, Landsat imagery was used to produce the United Kingdom Land Cover Map (LCM) for 1990, 2000 and 2007 (Morton et al. 2011), the South African National Land Cover dataset for 1994 and 2000 (Van den Berg et al. 2008), and the National Land Cover Database (NLCD) in the U.S. for 1992, 2001, 2006 and 2011 (Jin et al. 2013). Global land cover datasets, such as GlobCover (ESA 2010) or MODIS Global Land Cover (Friedl et al. 2010), are generally produced at a coarser spatial resolution (250 m or coarser) and land cover classes are more generalized to be applicable globally, which makes them less appropriate for regional studies.

As can be seen in this short summary, dataset updates in some countries are released at 5–10 year intervals (e.g., in the U.S. and the UK), which allow for some land cover change monitoring. However, this is not common globally, and many countries only have fine resolution (i.e., 30 m) national land cover data for at most two points in time (e.g., China and South Africa). Coarse temporal resolution and limited temporal extent of existing land cover databases makes characterizing trends across time difficult and can inhibit integration with other data sources. This and the geographically constrained extent of existing land cover datasets can limit research potential, particularly for cross-border studies where data are unavailable or inconsistent between regions or countries.

Overall, there is an emerging need for land cover data at fine spatial and temporal resolutions for regions where such data are not currently available. For example, research addressing urbanization along the U.S.-Mexico border (e.g., Biggs et al. 2010; Norman et al. 2009) would benefit from improved spatial and temporal coverage of land cover data in this region. The North American Land Cover dataset – covering Mexico, the U.S. and Canada – was created at 250 m resolution and is available for two points in time (Latifovic et al. 2010). The National Land Cover Database (NLCD) is available for four points in time at 30 m resolution, but only covers the contiguous U.S. (Jin et al. 2013). Such datasets demonstrate that extensive efforts have been made to create land cover data at various spatial scales and resolutions and for multiple points in time. Given that such regional and national scale land cover datasets are expensive to produce and can take a number of years to complete, the scientific community would benefit from more efficient methods for fulfilling land cover data requirements.

Significant progress has been made in the fields of image classification and information extraction for replicating existing land cover data products in the spatial and temporal domains. This chapter summarizes recent efforts in land cover data replication using machine learning frameworks and outlines the state-of-the-art in the field.

1.2 State-of-Research: Machine Learning for Land Cover Data Replication

With the increasing availability of large-volume geospatial data, machine learning techniques have gained importance in developing efficient and robust analytical procedures to process these data volumes and to generate new knowledge. For example, information extraction from remote sensing data to create land cover or land use classifications supports novel research across a broad range of disciplines and has become increasingly important for many interdisciplinary efforts across the social and environmental sciences. Since numerous high quality land cover databases already exist, a logical question is whether these existing products could be used effectively to increase the coverage of land cover information. This avenue of inquiry requires new methodological approaches that enable successful data replication. This section reflects on these fundamental challenges and summarizes the state of the research in this growing field. Insights into the concept of land cover data replication using existing spatial data is presented, while shedding light on the nature of integrated spatial data products and how they are created to establish valuable knowledge bases. One of the most significant problems in land cover data replication is dataset shift, which is formally defined and discussed within the context of remote sensing image classification. Finally, this section reviews two promising machine learning algorithms demonstrated for information extraction: Supported Vector Machines and Maximum Entropy Classifiers.

1.2.1 Data Replication: Information Extraction from Remote Sensing Imagery Using Existing Spatial Data Products

Research on information extraction in the field of remote sensing has developed various models for interpreting radiation data from airborne or satellite sensors for purposes of classification, temporal change detection, or measuring physical characteristics such as elevation or temperature (Verstraete et al. 1996). In other fields, such as in natural language processing (NLP) or knowledge discovery in databases (KDD), information extraction approaches are used to process and analyze existing digital data products (e.g., text documents or database records) in order to discover patterns and derive semantic meaning from the data that can be applied for other documents (Fayyad et al. 1996). These two approaches towards information extraction take different, yet interrelated, perspectives: The former produces thematic information from raw data (e.g. imagery), while the latter examines existing data products in databases (e.g., published documents) to produce generalizable models for future data characterization. A fusion of these two approaches in the form of data replication has demonstrated potential in remote sensing and GIScience research for extracting information from imagery guided by existing spatial data (Guo and Mennis 2009; Miller and Han 2009). Specific examples include spatial extrapolation of models of ecological processes using satellite imagery as input to other study areas (Miller et al. 2004) and regional models of productivity of soil landscapes derived from Landsat imagery based on and guided by local soil maps (Grinand et al. 2008).

When spatial data layers are created (through measurement or modeling) information is encapsulated in the final data product about processes and properties of the geographic extent they represent. Reality is abstracted to produce thematic layers (e.g., land cover, soil types) and measurements are summarized spatially to represent phenomena and processes (e.g., temperature gradients, terrain). The wealth of information encapsulated in such spatial data layers, which often remains underused, represents an important knowledge base and if accessed and integrated appropriately, has great potential to improve information extraction from raw data sources such as remote sensing imagery. The creation of land cover data based on classifying remote sensing imagery generally integrates large amounts of ancillary data sources (e.g., terrain derivatives, agricultural data, and vegetation surveys) to improve accuracy of land cover classes that are particularly difficult to classify from imagery alone, such as wetland or agriculture. The knowledge base encapsulated through the model-based integration of satellite imagery and ancillary data is a valuable resource that could be extracted and used to replicate the land cover classification efficiently and in an automated manner. From a practical perspective, one central question to be addressed is whether this richness of information could potentially be extracted from the underlying remote sensing imagery alone to then classify an image for a different geographic extent or a different point in time (Fig. 1.1). The global extent and broad temporal coverage of satellite imagery allow for wide-ranging potential application.

1.2.2 *Spatial Data Integration to Create Geographic Knowledge Bases*

Effective data integration is an ongoing challenge where multiple disparate data sources must be combined to provide a unified view of the data (Lenzerini 2002) and potentially create new data products. In the spatial sciences, data integration has played an important role in integrating GIScience and remote sensing research to improve information extraction (Congalton 1991; Harris and Ventura 1995), particularly for work on land cover classification and change detection (e.g., Liu et al. 2003; Stefanov et al. 2001). Here, the integration process is usually implicit and subject to the inclusion of additional ancillary data in the classification process in order to improve extraction results from raw image data. For example, multi-sensor (e.g., Landsat imagery and aerial photography) and multi-source (e.g., multispectral imagery and long-term climate observations) data integration has resulted in the improvement of land cover classifications (Geneletti and Gorte 2003; Liu et al. 2003), estimates of tree cover density (Huang et al. 2001), and land cover change

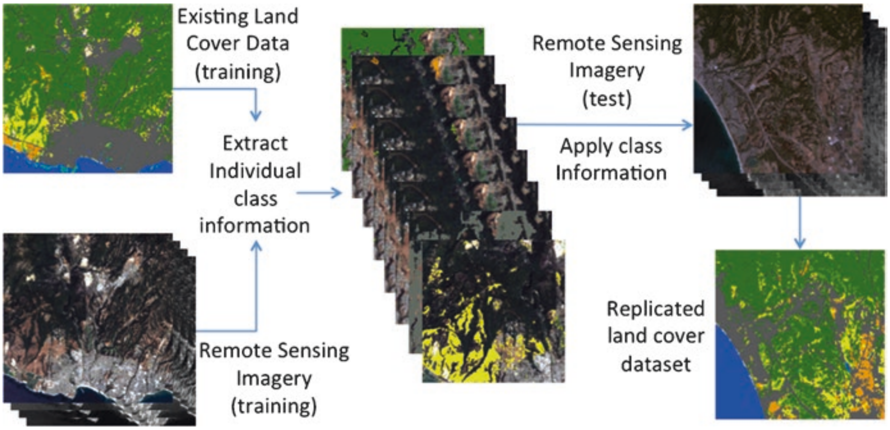

Existing Land Cover Data (training)
Extract Individual class information
Remote Sensing Imagery (training)
Remote Sensing Imagery (test)
Apply class Information
Replicated land cover dataset

Fig. 1.1 Conceptual process diagram of land cover data replication

observations (Petit and Lambin 2001). The production of broad scale land cover datasets (e.g., the NLCD in the U.S. and the CORINE dataset in Europe) using model-based classification approaches has benefitted from the integration of large amounts of ancillary spatial data. For example, in order to create the NLCD several disparate spatial data layers were integrated including forest inventory data, terrain data, and agricultural surveys with Landsat imagery to improve accuracy levels (Homer et al. 2007). The CORINE dataset used topographic maps and orthophotos to improve the classification of Landsat-derived Image 2000 data (Büttner and Kosztra 2007). Integration of satellite imagery with field-based data for predictive modeling is well established in ecology (Kerr and Ostrovsky 2003), and is frequently used for spatial extrapolation of known information (e.g., data from previous studies or knowledge of a process) to expand the geographic extent to areas where field-based data are limited or nonexistent (Miller et al. 2004). Similar work in other fields has extrapolated soil properties from local to regional scales using predictive models and existing ancillary spatial data (Lemercier et al. 2012) and also by integrating satellite imagery (Grinand et al. 2008). While these data integration efforts often create valuable new data, the methodological challenges of integration with regard to spatial resolution, temporal offsets and ambiguity in thematic classes are complex and represent a research field in and of itself (e.g., Cruz and Xiao 2008; Hasani et al. 2015). Overall this body of research has shown the benefits of data integration methods for information extraction in the spatial sciences. The higher complexity of data provides additional opportunities for integrating information and constructing valuable knowledge bases. However, little attention has been paid to the use of this encapsulated information in the development of formal methods for spatial and temporal replication using remote sensing imagery. The knowledge base encapsulated in existing spatial databases therefore remains an underused resource.

1.2.3 Dataset Shift as a Major Challenge in Sampling and Active Machine Learning Solutions

Dataset shift (also referred to as domain adaption) is a classification problem where the distributions of the training and test data differ significantly (Quionero-Candela et al. 2009), which affects the robustness of the classifier and prevents the development of general models applicable to different data. In the machine learning literature, dataset shift is described as a violation of the assumption made in many classification algorithms that the training and test data follow the same distribution (Moreno-Torres et al. 2012). Recently, a growing body of research has developed remote sensing classification models for spatial extrapolation or temporal extension of a given training sample that specifically address the dataset shift problem. Working with synthetic experimental data, important theoretical and methodological advancements have been made to better understand the impact of dataset shift on image classification algorithms. For example, a support vector machine (SVM) classifier was applied to Landsat 5 TM imagery and synthetic data to extend a given training sample to a different geographic extent (Bruzzone and Marconcini 2010) and a different point in time (Bruzzone and Marconcini 2009). These studies reported that differences in the distribution of spectral information between the training and test data posed a challenge for the classification algorithm because the training data were not representative for the test data. Other studies confirmed these observations indicating that dataset shift negatively impacts land cover classification results when the underlying imagery for the training sample does not follow the same distribution as the image to be classified (Tuia et al. 2011). These early studies addressed dataset shift by applying an active machine learning framework where the classifier iteratively selects an optimal training sample from a pool of training data. Active learning has been used extensively in computer vision and natural language processing (Settles 2010), and has been successfully applied in remote sensing image classification contexts to address dataset shift (e.g., Tuia et al. 2009).

Extending on early work by Bruzzone and Marconcini (2009, 2010), active learning was used in spatial extrapolation experiments to migrate the training sample iteratively from the training domain to the test domain (Matasci et al. 2012; Persello and Bruzzone 2012). In this iterative procedure, new sample pixels from the test domain were selected and then labeled by the user, while sample pixels from the training domain were gradually removed. This way the distributional differences could be reduced by shifting the distribution of the training sample towards that of the test data. This adaptation approach was tested on high-resolution multispectral imagery and hyperspectral imagery. Results in both applications yielded improved classification accuracy and demonstrated that active learning can effectively address and mitigate dataset shift while requiring a small initial training sample. Related work has applied similar frameworks in the temporal domain as a potential method for updating land cover data (Bahirat et al. 2012; Demir et al. 2013). However, the need to label new training pixels from the test domain throughout the procedure is a significant drawback regarding automation of the general framework. Ideally, the user would not have to intervene in the classification of new training pixels thus avoiding the need of expert knowledge.

Building upon this body of literature, Maclaurin and Leyk (2016a, b) proposed a methodological framework to address the dataset shift problem for spatial and temporal replication of existing land cover data. These studies addressed dataset shift in a different situation where the pool of potential training data was vast (i.e., all pixels in the existing land cover dataset), and therefore the focus was to define an optimal training sample from the underlying imagery that approximated the distribution of the imagery for a different geographic extent or point in time. In the spatial domain, a corrective sampling method was developed to match the training and test data distributions as closely as possible. Dataset shift was mitigated in the temporal domain by applying an orthogonal cross decomposition that jointly transformed the training and test images into a maximally correlated space, and thus improved the performance of the information extraction procedure. Properly mitigating dataset shift allowed for successful replication of the existing land cover data in both domains.

1.2.4 Predictive Machine Learning Models for Information Extraction

Predictive modeling in the spatial sciences has seen recent interest in machine learning algorithms. In particular, support vector machines (SVM) (Drake et al. 2006; Mountrakis et al. 2011) and maximum entropy (MaxEnt) models (Baldwin 2009; Li and Guo 2010) are highlighted in this section because they have been applied broadly in land cover replication and classification. Both of these models belong to a class of general-purpose statistical approaches for modeling incomplete information that are particularly useful when probability based interpretation is desired.

Support vector machines algorithms are non-parametric supervised classifiers that define a hyperplane as the optimal boundary between classes in multivariate space (Mountrakis et al. 2011). In its simplest form, SVM is an optimization problem that maximizes the separation between data points from two classes and the hyperplane. In most practical applications, data points cannot be completely separated and the hyperplane identifies a boundary that optimally classifies the data into classes while minimizing misclassification error based on the training data (Huang et al. 2002). SVM classifiers have been adapted for multi-class applications to create land cover data (e.g., Foody and Mathur 2004), which has broadened their utility. This family of classifiers has been employed extensively in ecological modeling (Drake et al. 2006; Pouteau et al. 2012) and in remote sensing image classification (Pal and Mather 2005; Mountrakis et al. 2011). Recent research on dataset shift in land cover classification (discussed in Sect. 1.2.3) has applied semi-supervised SVM techniques to optimize the training sample and thus improve the final classification (e.g., Persello and Bruzzone 2012; Tuia et al. 2011). In classifying multispectral and hyperspectral remote sensing imagery, SVM methods have shown to handle high-dimensional space effectively and perform well across a broad range of land cover classification applications (Tuia et al. 2009). However, it has also been shown that performance of SVM classifiers is dependent on the quality of the training data (Foody et al. 2006), and that results can vary significantly depending on the input parameters (Li and Guo 2010; Munoz-Mari et al. 2007).

The principle of ***maximum entropy*** is that predictions should be based only on the information that is known while making no assumptions about what is unknown (Jaynes 1957). This means that the model is fit to the training sample while ensuring maximum entropy of the estimated distribution (i.e., probabilities are distributed as evenly as possible while conforming to the information put forth by the training sample). This is done by estimating a probability distribution for class *c* over the finite set *X* (e.g., the underlying remote sensing data) that adheres to a set of specified constraints (determined from the training sample) and has maximum entropy, i.e., the most uniform distribution (Berger et al. 1996). The constraints in a MaxEnt model are implemented as a set of real-valued functions ($f_1,f_2,\ldots,f_n$) on *X*. The expected values of these functions represent what is known from the training sample. Multiple distributions could satisfy the constraints, so based on the principle of maximum entropy the most uniform one that does so is chosen.

MaxEnt has been applied extensively in spatial contexts to model ecological and demographic processes. Phillips et al. (2006) used MaxEnt to model distributions of two species of mammals in South America based on spatial data of climate, elevation and vegetation. Leyk et al. (2013) implemented MaxEnt for dasymetric modeling of census microdata for small area estimation. Wei et al. (2011) estimated the risk of Hantavirus infection based on a MaxEnt ecological niche model of rodent populations in eastern China. The broad application of maximum entropy models in geographic research demonstrates its efficacy for extracting information from spatial data in probabilistic solutions. Lin et al. (2014) used a MaxEnt model to classify urban land across China by combining multispectral satellite imagery with nighttime lights data. Li and Guo (2010) applied MaxEnt for land cover classification from high-resolution aerial imagery, and found it to be more accurate than an SVM classifier. Erkan et al. (2010) assessed the efficacy of MaxEnt for image segmentation and classification using three types of remote sensing data: multispectral, hyperspectral and synthetic aperture radar. They concluded that MaxEnt shows excellent potential and provides a strong alternative to widely used SVM classifiers. Detailed in the following section, Maclaurin and Leyk demonstrated the potential of MaxEnt in a data replication framework for spatial extrapolation (2016a) and temporal extension (2016b) of the NLCD using the underlying Landsat 5 TM imagery. This body of research shows that MaxEnt is well suited for a wide range of sensor types and performs particularly well for land cover classification.

1.3 Is Land Cover Data Replication Feasible?

Two research studies conducted by the authors developed a land cover data replication model for spatial extrapolation and temporal extension based on a MaxEnt active machine learning framework. The model was applied to an existing land cover database – the NLCD – and performed well in extracting class-specific information in both cases, even though the NLCD represents an imperfect data product (Wickham et al. 2010; Wickham et al. 2013). These experiments present important benchmarks showing that typical land cover data products with varying levels of accuracy can be replicated successfully. While the MaxEnt classifier appeared to accommodate imperfect training data in the active learning framework, it showed significant sensitivity to dataset shift. Therefore, separate preprocessing procedures for spatial extrapolation and temporal extension were developed to mitigate dataset shift. A flowchart of this combined framework is shown in Fig. 1.2. In both studies, the model was tested across three study areas to assess performance under different landscape conditions.

In the ***spatial domain,*** the model extracted NLCD class information using Landsat imagery from one geographic extent (i.e., the training area) and then replicated the NLCD from this model for a Landsat 5 TM image covering a different geographic extent (i.e., the test area) (Maclaurin and Leyk 2016a). The framework is efficient and fully automated, relying solely on the NLCD for the training area and Landsat imagery for the training and test areas (Fig. 1.3). The Tasseled Cap (TC) transformation was applied on the Landsat images to reduce the dimensionality to three transformed bands, which has been shown to improve separability of vegetated classes while capturing most of the variance in the imagery (Homer et al. 2004).

A major challenge for the model arose from the inherent covariate shift problem –a special case of the more general dataset shift as described in Sect. 1.2.3, where the conditional distributions are the same between the training and test data, but the distributions themselves differed significantly. This problem was mostly attributed to different proportions of individual NLCD classes between the training and test areas. Other possible causes for observed distributional differences such as illumination and atmospheric or environmental conditions were expected to be minimal since training and test areas were selected from the same Landsat scene. To minimize the effects of covariate shift, a corrective sampling method was implemented to optimally sample the training data such that the distribution approximated the test distribution as closely as possible (Fig. 1.4), and therefore improved performance while increasing the efficiency of the replication algorithm. This spatial extrapolation framework has been shown generalizable, under the assumption that distributional differences due to illumination and atmospheric or environmental variation in the training and test images are minimal. The model achieved similar levels of agreement as the original NLCD when compared against high-resolution reference datasets (Table 1.1).

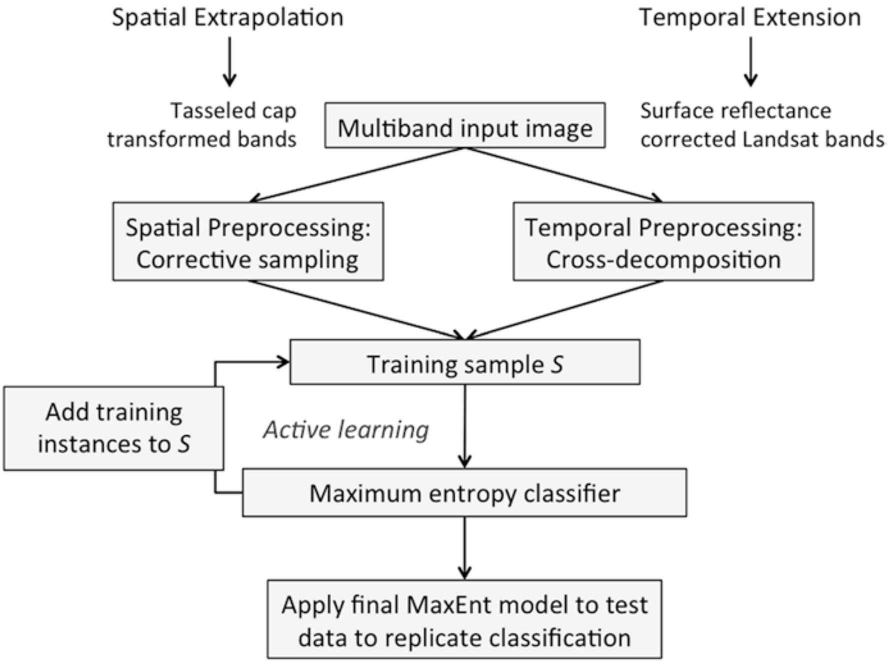

Spatial Extrapolation
Temporal Extension
Tasseled cap transformed bands
Multiband input image
Surface reflectance corrected Landsat bands
Spatial Preprocessing: Corrective sampling
Temporal Preprocessing: Cross-decomposition
Training sample *S*
Add training instances to *S*
Active learning
Maximum entropy classifier
Apply final MaxEnt model to test data to replicate classification

Fig. 1.2 Flow chart of framework for spatial and temporal land cover data replication

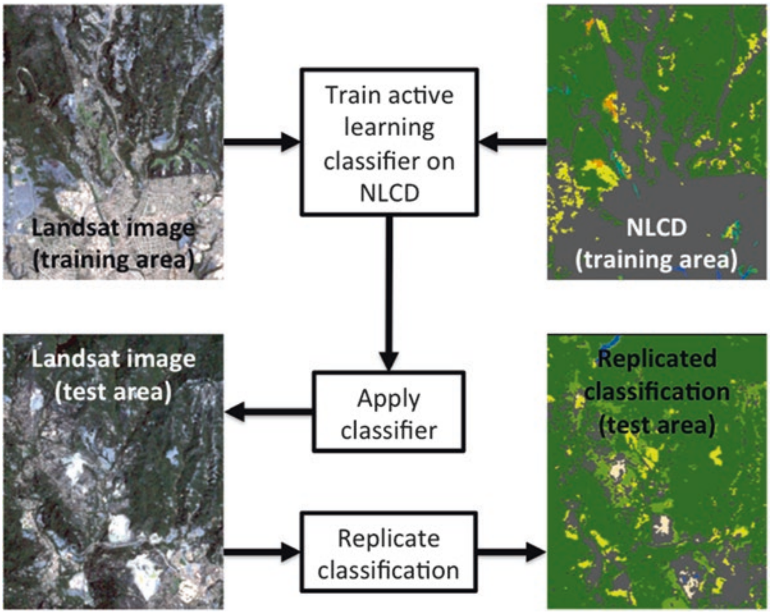

Landsat image (training area)
Train active learning classifier on NLCD
NLCD (training area)
Landsat image (test area)
Apply classifier
Replicated classification (test area)
Replicate classification

Fig. 1.3 Spatial extrapolation of the NLCD

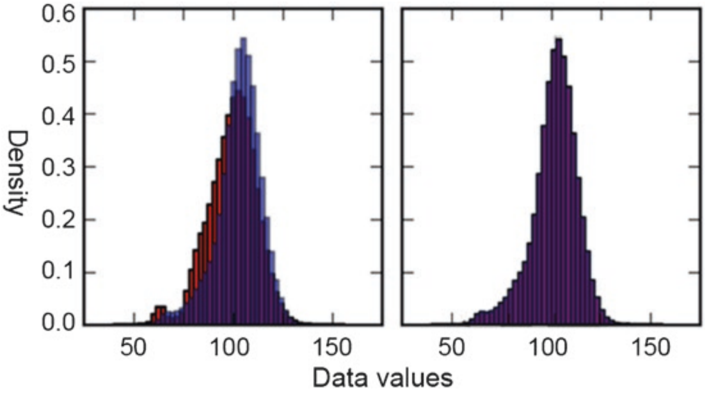


Fig. 1.4 Example of corrective sampling: the distribution of the training data (*red*) is shown before and after covariate shift corrective sampling was applied, and overlain by the test distribution (in *blue* with transparency) resulting in overlapping regions in purple. In order to overcome significant distributional differences between training and test data (*left panel*) a subset sample of training data was generated by the covariate shift corrective sampling method that matches very closely the distributional properties of the test data (*right panel*)

The replication model was modified for the temporal domain using bi-temporal pairs of Landsat imagery corresponding to two NLCD releases (e.g., 2001 and 2011) (Maclaurin and Leyk 2016b). The experimental setup was to extract information from the Landsat imagery for one NLCD release year (e.g., 2001) and then replicate it for a different release year (e.g., 2011) but for the same geographic extent, allowing for direct comparison between the replicated classification and the updated NLCD release. While the use of surface reflectance corrected Landsat image pairs improved bi-temporal extraction from the images, residual differences remained; these differences were reduced by a partial least squares (PLS) cross-decomposition that produced maximally correlated bands between the images (Wegelin 2000). Next, the active learning framework as described above was used to collect the optimal training samples and temporally replicate the NLCD. For each study area, replicated datasets were compared against the NLCD and the high-resolution reference land cover data. The results indicated encouraging replication performance for two of the study sites and moderate performance for the third one (Table 1.1).

When compared against the high-resolution reference land cover datasets, the replication model, in both domains, produced similar levels of overall agreement as the NLCD. Whether the levels of individual class agreement for the replicated classification were surprisingly low or encouragingly high, they were generally very similar to those of the NLCD when both were compared against the reference datasets. For example, the wetland class consistently had the lowest levels of agreement with the reference datasets for the replicated classification. Generally, the NLCD also showed similarly low levels of agreement with the reference dataset for the wetland class. This phenomenon was also observed for classes that performed

Table 1.1 Overall agreement (OA) between the NLCD, the reference dataset, and the replicated classification shown for the spatial extrapolation and temporal extension model results (see Maclaurin and Leyk 2016a, b)

	Spatial extrapolation			Temporal extension		
Study site	NLCD-reference (%)	Replicated-reference (%)	Replicated-NLCD (%)	NLCD-reference (%)	Replicated-reference (%)	Replicated-NLCD (%)
1	77.6	80.2	81.5	77.6	85.8	89.5
2	52.5	55.0	64.1	52.5	51.8	58.5
2	63.7	57.5	56.5	63.7	43.5	47.2

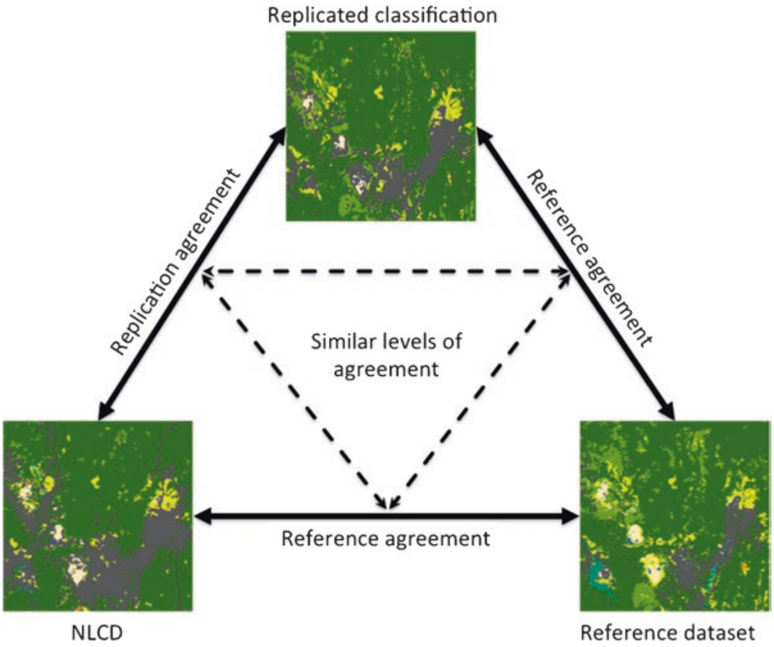


Fig. 1.5 The relationship between the levels of accuracy of the input land cover database (the NLCD), the replicated classification and the reference dataset

very well in both the NLCD and the replicated classification, such as forest or open water. Interestingly, the levels of agreements of the replicated classification compared against the NLCD were similar to the levels of each classification compared against the reference datasets. The relationship between levels of agreement between the three datasets (replicated classification, NLCD and reference dataset) suggests that the quality of the knowledge base extracted from the NLCD for a given class is dependent on the class's level of accuracy (Fig. 1.5).

1.4 Lessons Learned

This section summarizes the most important lessons learned and addresses the remaining limitations based on recent research efforts. We discuss the feasibility of the general idea of reverse engineering land cover data for replication purposes in the spatial and temporal domains, and comment on the generalizability of existing information extraction frameworks. Finally, we address the issue of the imperfectness of existing land cover databases that researchers are trying to replicate. It is the hope of the authors that these key aspects will help others to identify further potential for improvement of existing methods.

1.4.1 Feasibility of Reverse Engineering Frameworks for Land Cover Data Replication

Using remote sensing imagery alone, national scale land cover databases such as the NLCD in the United States, can be effectively replicated in order to create similar-quality data for a different geographic extent or for a different point in time. Thus, there is significant potential to extract the knowledge base encapsulated through the complex classification procedure that often includes rich ancillary datasets. While accuracy levels of large-scale land cover data typically vary regionally and between classes (e.g., Wickham et al. 2010, 2013), the recent research studies presented here have shown that information for most individual classes can be extracted consistently. When applied to different test data in both the spatial and temporal domains these models were able to produce levels of overall and class-specific accuracies similar to those of the original data, whether they were exceptionally high or surprisingly low in the case of the NLCD.

While there are still limitations and further need to test existing methods over larger geographic scales, the reverse engineering frameworks discussed here demonstrate remarkable progress and the potential to create land cover data at fine resolution for geographic extents and points in time with no or limited coverage. Successfully reversing an existing classification procedure with a machine learning classifier (e.g., MaxEnt or SVM) using only remote sensing imagery appears to be a vital solution to automated replication of land cover data.

1.4.2 Lessons Learned from Spatial Extrapolation

The replication of land cover data for different geographic extents benefits from reducing the dimensionality of the imagery (e.g., by using a Tasseled Cap (TC) transformation). The main challenge for spatial extrapolation appears to be addressing the covariate shift problem, which could be related to different compositions of land cover classes (i.e., the frequency of occurrence for each class) between the training and test areas if these are within the same image scene. Active learning frameworks that incorporate corrective sampling steps enable the selection of an optimal training sample for the classifier (e.g., MaxEnt) to match the distribution of the test area. Such corrections have been demonstrated as an effective way to reduce covariate shift resulting in the improvement of the classification result and an increased efficiency of the process (Maclaurin and Leyk 2016a, b; Tuia et al. 2011). While these experiments are encouraging starting points, further research is needed to understand and address covariate shift if the training and test areas are covered by different images as this will add differences in illumination and environmental conditions as causes for covariate shift. This will demonstrate how generalizable these approaches are at larger geographical scales and under different environmental conditions and compositions of land cover classes.

1.4.3 Lessons Learned for Temporal Extension

Recent experiments showed that existing machine learning frameworks for temporal land cover data replication have the potential to work backwards and forwards in time (Bruzzone and Marconcini 2009; Maclaurin and Leyk 2016b). The main challenge in temporal replication is spectral and radiometric differences between the bi-temporal images, which could be seen as a form of the dataset shift problem. Surface reflectance correction models significantly reduce artifacts introduced from illumination and viewing geometry of the sensor (Masek et al. 2006). However, often there are still spectral differences present due to environmental, on-the-ground variation such as temporal differences in productivity of vegetation, states of agricultural land (e.g., different crop schedules or rotation of fallow land) (Lambin et al. 2000), and atmospheric noise (such as thin clouds or haze) (Lunetta et al. 2004). The application of a partial least squares cross-decomposition to the image pairs has been shown as just one way to further reduce these residual differences and improve overall replication (Maclaurin and Leyk 2016b). Nevertheless, the presence of these unexpected differences can still lead to significant misclassification, and thus require further attention before broad, regional scale land cover replication would be feasible.

1.4.4 Generalizability of Replication Methods for Different Landscape Types and Scales

Landscape heterogeneity and strong overlap between spectral signature of classes appear to have some negative impact on the performance of active learning frameworks for land cover classification (Tuia et al. 2011) and for land cover replication

(Maclaurin and Leyk 2016a, b). While the range of landscape types in the studies discussed here was rather limited, results indicate some stability and generalizability across different landscape types and varied compositions of land cover classes. To achieve more confidence, additional confirmatory tests need to be done covering larger areas and greater variability in landscape types and compositions.

Smith et al. (2003) demonstrated that overall land cover classification accuracy tends to increase as landscape homogeneity and average patch size for individual classes increase. Furthermore, examination of scale dependence on accuracy levels in land cover data showed that smaller geographic extents generally had lower levels of overall accuracy (Hollister et al. 2004). These two observations suggest that the impact of landscape composition and environmental variation on both spatial and temporal replication requires further investigation at broader, regional scales.

1.4.5 Imperfect Land Cover Data as Knowledge Bases for Information Extraction

In general, higher prevalence classes are expected to perform better as a direct benefit from a larger sampling pool. An interesting phenomenon that will always pose a challenge in this research is the dichotomy between the concepts of *land cover* and *land use* and how this distinction is blurred in the final data products. *Land cover* describes biophysical properties through direct observation and classification (usually by remote sensing) of the earth's surface, whereas *land use* is defined by how human activities alter, occupy, and manage the physical environment of the earth's surface (Comber 2008; Theobald 2014). One example of this is the developed open space class in the NLCD, which is highly comprised of secondary roads. Theobald (2010) removed secondary roads from the NLCD 2001 (using the same Census TIGER files applied to initially *burn in* the roads), and replaced these pixels with the dominating neighboring class. He found that this reduced the overall percentage of developed land from 5.11 to 2.69% for the conterminous U.S. Two-lane roads in rural areas are less than 15 m wide (FHWA 2014), and are thus particularly difficult to extract from Landsat imagery. Forest canopy can partially obscure these roads from the remote sensor and in non-forested areas roads are usually bordered by herbaceous vegetation (i.e., grassland in the NLCD). The open space developed class in the NLCD is therefore predominately comprised of highly mixed pixels. This can pose serious problems, as a high proportion of sampled pixels for developed land would spectrally represent different or mixed classes.

Another, slightly different example is wetland. Wetland classes are often defined as areas of forest, shrub, or herbaceous vegetation where the soil is periodically saturated or covered with water (Homer et al. 2004). Often the data producer relies on extensive ancillary data for classifying wetlands (Homer et al. 2004, 2007). Soil moisture in vegetated areas is a difficult property to measure with remote sensing imagery due to the small portion of the spectral signature actually coming from the soil (Muller and Decamps 2001). Furthermore, if the soil is not highly saturated at the time the image is captured, the spectral response will not differ significantly from that of other vegetated classes resulting in misclassifications as forest, grassland or shrub. Replication of wetland is further impeded by typically low prevalence, which limits the potential pool of data to train the classifier and thus lower levels of generalizability of the model. This problem is also true for other classes with low occurrence such as bare land.

Importantly, the land cover dataset must be taken as-is when it serves as the knowledge base for a spatial or temporal replication effort. A well-performing data replication approach will recreate each class for a different geographic extent or point in time with approximately the same accuracy. It is unlikely that a replication framework could produce a significantly improved classification, either in space or time, with any consistency. For existing algorithms to be applied as generalizable approaches they must be applicable across different landscapes and be stable for both spatial and temporal replication.

1.5 Conclusions & Outlook

Research studies to date represent an encouraging perspective for extraction of the rich knowledge base encapsulated in existing high-level spatial data using remote sensing imagery for both spatial extrapolation and temporal extension. The extensive spatial and temporal coverage of satellite imagery would enable researchers and industry to scale up such frameworks once they become fully operational. However, studies on real data are only in early phases and future research should continue to test existing and novel approaches on different national or regional land cover databases. Successful temporal replication has great potential for updating, and also for backcasting, existing land cover data. Since updating such databases is generally a highly labor-intensive and expensive operation, procedures for producing updates at higher levels of automation would be highly beneficial and have significant policy and funding implications.

Methodologically, solutions for distributional differences between training and test data, e.g., the covariate shift problem in remote sensing (e.g., Bruzzone and Marconcini 2010; Matasci et al. 2012) or residual differences between two bi-temporal images due to environmental variation, have been a focus in recent studies. With increasing use of machine learning algorithms in remote sensing (Pal and Mather 2005; Rogan et al. 2008), dataset shift has become a more common and recognized problem (Tuia et al. 2011, Persello and Bruzzone 2012) and future research will further develop sampling methods that account for such problems.

Future work in this field should test existing replication frameworks on thematically refined land cover classes which are of greater use in specific application domains such as landscape ecology, land use science or urban planning. Such improvements in the replication procedure could be gained by using multisensor remote sensing imagery and/or multisource data. For example, long wavelength radar has been shown to improve classification of wetland (Rosenqvist et al. 2007), and could help improve the extraction of this problematic class. Another natural future step in this research will be to synthetically improve the accuracy of the database of interest using existing reference datasets. Training samples collected for the extraction can be restricted to pixels that agree with the reference data thus improving the active learning approach for more effective replication.

Information extraction from remotely sensed imagery for spatial data replication is an exciting research frontier that benefits from an ongoing integration of machine learning theories and methods in remote sensing and GIScience research. The development of theoretical frameworks and formal methods for this avenue of research has received little attention, yet shows great potential. Further advancements in the field of automated spatial and temporal replication of existing land cover databases would help overcome spatial and temporal limitations of land cover data for the scientific community.

References

Bahirat K, Bovolo F, Bruzzone L, Chaudhuri S (2012) A novel domain adaptation Bayesian classifier for updating land-cover maps with class differences in source and target domains. Geosci Remote Sens, IEEE Trans 50(7):2810–2826

Baldwin RA (2009) Use of maximum entropy modeling in wildlife research. Entropy 11(4):854–866

Berger AL, Pietra VJD, Pietra SAD (1996) A maximum entropy approach to natural language processing. Comput Linguist 22(1):39–71

Biggs TW, Atkinson E, Powell R, Ojeda-Revah L (2010) Land cover following rapid urbanization on the US–Mexico border: implications for conceptual models of urban watershed processes. Landsc Urban Plan 96(2):78–87

Bruzzone L, Marconcini M (2009) Toward the automatic updating of land-cover maps by a domain-adaptation SVM classifier and a circular validation strategy. Geosci Remote Sens, IEEE Trans 47(4):1108–1122

Bruzzone L, Marconcini M (2010) Domain adaptation problems: a DASVM classification technique and a circular validation strategy. Pattern Anal Mach Int, IEEE Trans 32(5):770–787

Büttner G, Kosztra B (2007) CLC2006 technical guidelines. In: Technical report. European Environment Agency, Copenhagen

Comber AJ (2008) The separation of land cover from land use using data primitives. J Land Use Sci 3(4):215–229

Congalton RG (1991) Remote sensing and geographic information system data integration: error sources and research issues. Photogramm Eng Remote Sens 57(6):677–687

Cruz IF, Xiao H (2008) Data integration for querying geospatial sources. In: Sample J, Shaw K, Tu S, Abdelguerfi M (eds) Geospatial services and applications for the internet. Springer, Boston, pp 110–134

Demir B, Bovolo F, Bruzzone L (2013) Updating land-cover maps by classification of image time series: a novel change-detection-driven transfer learning approach. Geosci Remote Sens, IEEE Trans 51(1):300–312

Drake JM, Randin C, Guisan A (2006) Modelling ecological niches with support vector machines. J Appl Ecol 43(3):424–432

Erkan AN, Camps-Valls G, Altun Y (2010) Semi-supervised remote sensing image classification via maximum entropy. 2010 IEEE Inte Work Mach Learn Sig Process (MLSP):313–318

European Space Agency (ESA) – Data User Element (2010, December 21) GlobCover 2009 (Global Land Cover Map). Retrieved March 16, 2015, from http://due.esrin.esa.int/globcover

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Mag 17(3):37–54

Federal Highway Administration (FHWA) (2014) Lane width. Accessed November 26, 2014 from http://safety.fhwa.dot.gov/geometric/pubs/mitigationstrategies/chapter3/3_lanewidth.cfm

Foody GM, Mathur A (2004) A relative evaluation of multiclass image classification by support vector machines. IEEE Trans Geosci Remote Sens 42(6):1335–1343

Foody GM, Mathur A, Sanchez-Hernandez C, Boyd DS (2006) Training set size requirements for the classification of a specific class. Remote Sens Environ 104(1):1–14

Friedl MA, Sulla-Menashe D, Tan B, Schneider A, Ramankutty N, Sibley A, Huang X (2010) MODIS collection 5 global land cover: algorithm refinements and characterization of new datasets. Remote Sens Environ 114(1):168–182

Geneletti D, Gorte BGH (2003) A method for object-oriented land cover classification combining Landsat TM data and aerial photographs. Int J Remote Sens 24(6):1273–1286

Grinand C, Arrouays D, Laroche B, Martin MP (2008) Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. Geoderma 143(1):180–190

Guo D, Mennis J (2009) Spatial data mining and geographic knowledge discovery—an introduction. Comput Environ Urban Syst 33(6):403–408

Harris PM, Ventura SJ (1995) The integration of geographic data with remotely sensed imagery to improve classification in an urban area. Photogramm Eng Remote Sens 61(8):993–998

Hasani S, Sadeghi-Niaraki A, Jelokhani-Niaraki M (2015) Spatial data integration using ontology-based approach. Inter Arch Photogramm Remote Sens Spat Inf Sci 40(1):293–296

Hollister JW, Gonzalez ML, Paul JF, August PV, Copeland JL (2004) Assessing the accuracy of national land cover dataset area estimates at multiple spatial extents. Photogramm Eng Remote Sens 70(4):405–414

Homer C, Huang C, Yang L, Wylie B, Coan M (2004) Development of a 2001 national land-cover database for the United States. Photogramm Eng Remote Sens 70(7):829–840

Homer C, Dewitz J, Fry J, Coan M, Hossain N, Larson C, Herold N, McKerrow J, VanDriel N, Wickham J (2007) Completion of the 2001 national land cover database for the counterminous United States. Photogramm Eng Remote Sens 73(4):337

Huang C, Yang L, Wylie B, Homer C (2001) A strategy for estimating tree canopy density using Landsat 7 ETM+ and high resolution images over large areas. Third international conference on geospatial information in agriculture and forestry. Denver, Colorado

Huang C, Davis LS, Townshend JRG (2002) An assessment of support vector machines for land cover classification. Int J Remote Sens 23(4):725–749

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106(4):620

Jin S, Yang L, Danielson P, Homer C, Fry J, Xian G (2013) A comprehensive change detection method for updating the national land cover database to circa 2011. Remote Sens Environ 132:159–175

Kerr JT, Ostrovsky M (2003) From space to species: ecological applications for remote sensing. Trends Ecol Evol 18(6):299–305

Lambin EF, Rounsevell MDA, Geist HJ (2000) Are agricultural land-use models able to predict changes in land-use intensity? Agric Ecosystems Environ 82(1):321–331

Latifovic R, Homer C, Ressl R, Pouliot D, Hossain SN, Colditz R, Victoria A (2010) North American land change monitoring system (NALCMS), Remote sensing of land use and land cover: principles and applications. CRC Press, Boca Raton

Lemercier B, Lacoste M, Loum M, Walter C (2012) Extrapolation at regional scale of local soil knowledge using boosted classification trees: a two-step approach. Geoderma 171:75–84

Lenzerini M (2002) Data integration: a theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, pp 233–246

Leyk S, Nagle NN, Buttenfield BP (2013) Maximum entropy dasymetric modeling for demographic small area estimation. Geogr Anal 45(3):285–306

Li W, Guo Q (2010) A maximum entropy approach to one-class classification of remote sensing imagery. Int J Remote Sens 31(8):2227–2235

Lin J, Liu X, Li K, Li X (2014) A maximum entropy method to extract urban land by combining MODIS reflectance, MODIS NDVI, and DMSP-OLS data. Int J Remote Sens 35(18):6708–6727

Liu JY, Zhuang DF, Luo D, Xiao XM (2003) Land-cover classification of China: integrated analysis of AVHRR imagery and geophysical data. Int J Remote Sens 24(12):2485–2500

Lunetta RS, Johnson DM, Lyon JG, Crotwell J (2004) Impacts of imagery temporal frequency on land-cover change detection monitoring. Remote Sens Environ 89(4):444–454

Maclaurin G, Leyk S (2016a) Extending the geographic extent of existing land cover data using active machine learning and covariate shift corrective sampling. Int J Remote Sens 37(21):5213–5233

Maclaurin G, Leyk S (2016b) Temporal replication of the national land cover database using active machine learning. GISci Remote Sens 53(6):759–777

Masek JG, Vermote EF, Saleous NE, Wolfe R, Hall FG, Huemmrich KF, ..., Lim TK (2006) A Landsat surface reflectance dataset for North America, 1990–2000. Geosci Remote Sens Lett IEEE 3(1):68–72

Matasci G, Tuia D, Kanevski M (2012) SVM-based boosting of active learning strategies for efficient domain adaptation. Sel Top Appl Earth Obs Remote Sens IEEE J 5(5):1335–1343

Miller HJ, Han J (2009) Geographic data mining and knowledge discovery: an overview. In: Miller HJ, Han J (eds) Geographic data mining and knowledge discovery. CRC Press, Boca Raton

Miller JR, Turner MG, Smithwick EA, Dent CL, Stanley EH (2004) Spatial extrapolation: the science of predicting ecological patterns and processes. Bioscience 54(4):310–320

Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recogn 45(1):521–530

Morton D, Rowland C, Wood C, Meek L, Marston C, Smith G, Simpson I (2011) Final Report for LCM2007-the new UK land cover map. Countryside Survey Technical Report No 11/07

Mountrakis G, Im J, Ogole C (2011) Support vector machines in remote sensing: a review. ISPRS J Photogramm Remote Sens 66(3):247–259

Muller E, Decamps H (2001) Modeling soil moisture–reflectance. Remote Sens Environ 76(2):173–180

Muñoz-Marí J, Bruzzone L, Camps-Valls G (2007) A support vector domain description approach to supervised classification of remote sensing images. IEEE Trans Geosci Remote Sens 45(8):2683–2692

Norman LM, Feller M, Guertin DP (2009) Forecasting urban growth across the United States–Mexico border. Comput Environ Urban Syst 33(2):150–159

Pal M, Mather PM (2005) Support vector machines for classification in remote sensing. Int J Remote Sens 26(5):1007–1011

Persello C, Bruzzone L (2012) Active learning for domain adaptation in the supervised classification of remote sensing images. Geosci Remote Sens IEEE Trans 50(11):4468–4483

Petit CC, Lambin EF (2001) Integration of multi-source remote sensing data for land cover change detection. Int J Geogr Inf Sci 15(8):785–803

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. Ecol Model 190(3):231–259

Pouteau R, Meyer JY, Taputuarai R, Stoll B (2012) Support vector machines to map rare and endangered native plants in Pacific islands forests. Eco Inform 9:37–46

Quionero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND (2009) Dataset shift in machine learning. The MIT Press

Rogan J, Franklin J, Stow D, Miller J, Woodcock C, Roberts D (2008) Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. Remote Sens Environ 112(5):2272–2283

Rosenqvist AKE, Finlayson CM, Lowry J, Taylor D (2007) The potential of long-wavelength satellite-borne radar to support implementation of the Ramsar Wetlands Convention. Aquat Conserv Mar Freshwat Ecosyst 17(3):229–244

Settles B (2010) Active learning literature survey. Univ Wis Madison 52:55–66. 11

Smith JH, Stehman SV, Wickham JD, Yang L (2003) Effects of landscape characteristics on land-cover class accuracy. Remote Sens Environ 84(3):342–349

Stefanov WL, Ramsey MS, Christensen PR (2001) Monitoring urban land cover change: an expert system approach to land cover classification of semiarid to arid urban centers. Remote Sens Environ 77(2):173–185

Theobald DM (2010) Estimating natural landscape changes from 1992 to 2030 in the conterminous US. Landsc Ecol 25(7):999–1011

Theobald DM (2014) Development and applications of a comprehensive land use classification and map for the US. PLoS One 9(4):e94628

Tuia D, Ratle F, Pacifici F, Kanevski MF, Emery WJ (2009) Active learning methods for remote sensing image classification. Geosci Remote Sens IEEE Trans 47(7):2218–2232

Tuia D, Pasolli E, Emery WJ (2011) Using active learning to adapt remote sensing image classifiers. Remote Sens Environ 115(9):2232–2242

Van den Berg EC, Plarre C, Van den Berg HM, Thompson MW (2008) The South African national land cover 2000. Agricultural Research Council (ARC) and Council for Scientific and Industrial Research (CSIR), Pretoria. Report No. GW/A/2008/86

Verstraete MM, Pinty B, Myneni RB (1996) Potential and limitations of information extraction on the terrestrial biosphere from satellite remote sensing. Remote Sens Environ 58(2):201–214

Wegelin JA (2000) A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. University of Washington, Tech. Rep

Wei L, Qian Q, Wang ZQ, Glass GE, Song S X, Zhang WY, ..., Cao WC (2011) Using geographic information system-based ecologic niche models to forecast the risk of hantavirus infection in Shandong Province, China. Am J Trop Med Hyg 84(3):497–503

Wickham JD, Stehman SV, Fry JA, Smith JH, Homer CG (2010) Thematic accuracy of the NLCD 2001 land cover for the conterminous United States. Remote Sens Environ 114(6):1286–1296

Wickham JD, Stehman SV, Gass L, Dewitz J, Fry JA, Wade TG (2013) Accuracy assessment of NLCD 2006 land cover and impervious surface. Remote Sens Environ 130:294–304

Chapter 2
Geospatial Analysis Requires a Different Way of Thinking: The Problem of Spatial Heterogeneity

Bin Jiang

Abstract Geospatial analysis is very much dominated by a Gaussian way of thinking, which assumes that things in the world can be characterized by a well-defined mean, i.e., things are more or less similar in size. However, this assumption is not always valid. In fact, many things in the world lack a well-defined mean, and therefore there are far more small things than large ones. This paper attempts to argue that geospatial analysis requires a different way of thinking – a Paretian way of thinking that underlies skewed distribution such as power laws, Pareto and lognormal distributions. I review two properties of spatial dependence and spatial heterogeneity, and point out that the notion of spatial heterogeneity in current spatial statistics is only used to characterize local variance of spatial dependence or regression. I subsequently argue for a broad perspective on spatial heterogeneity, and suggest it be formulated as a scaling law. I further discuss the implications of Paretian thinking and the scaling law for better understanding geographic forms and processes, in particular while facing massive amounts of social media data. In the spirit of Paretian thinking, geospatial analysis should seek to simulate geographic events and phenomena from the bottom up rather than correlations as guided by Gaussian thinking.

Keywords Big data • Scaling of geographic space • Head/tail breaks • Power laws • Heavy-tailed distributions

B. Jiang (✉)
Faculty of Engineering and Sustainable Development, Division of GIScience, University of Gävle, SE-801 76 Gävle, Sweden
e-mail: bin.jiang@hig.se

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_2

2.1 Introduction

Geospatial analysis, or spatial statistics in particular, has been dominated by a Gaussian way of thinking, which assumes that things are more or less similar in size, and can be characterized by a well-behaved mean. Based on this assumption, extremes are rare; if extremes do exist, they can be mathematically transformed into normal things (e.g., by taking logarithms or square roots). This Gaussian thinking is widespread, and has dominated the sciences for a very long time. However, Gaussian thinking has been challenged and been accused of misrepresenting our world (Mandelbrot and Hudson 2004; Taleb 2007). Indeed, many things in the world are not well behaved or lack of a well-behaved mean. This can seen from the extreme events such as the September 11 attacks. The extent of devastation of such events was enormous and beyond any predictions and estimations. This is the same for many geographic features, which exhibit a pretty skewed or heavy-tailed distribution such as power laws and lognormal distributions. The heavy-tailed distributions imply that there are far more small geographic features than large ones, namely scaling of geographic space.

A power law distribution is often referred to as scale free, literally meaning a lack of average for characterizing the sizes of things (Barabási and Albert 1999). The power law distribution has been given different formats for it was discovered by different scientists in different disciplines over the past 100 years. Among several alternatives, Zipf's law (1949) and the Pareto distribution (Pareto 1897) are the two formats most frequently referred to in the literature. Zipf's law, with respect to city sizes, implies that there are far more small cities than large ones, while the Pareto distribution indicates that there are far more poor people than rich people, or equivalently far more ordinary people than extraordinary people. The Pareto distribution has been popularized as the 80/20 principle (Koch 1999) or the long tail theory (Anderson 2006) in the popular science and business literature. The heavy-tailed distribution, including power laws, lognormal and others similar, is what underlies the new way of thinking I want to advocate in this paper. The central argument is that geospatial analysis requires a new way of thinking radically different from Gaussian thinking, and spatial heterogeneity should be formulated as a scaling law of geography.

This is an unprecedented time when we face increasing rich geographic data sources based not only on the legacy of traditional cartography and remote sensing imagery, but also emerging from various social media such as Flickr, Twitter, and OpenStreetMap, collectively known as volunteered geographic information (Goodchild 2007). Today, one can amass gigabytes of an entire country's data for geospatial analysis and computing, for both data volumes and computing capacity have increased dramatically. However, our mindsets, subsequently our analysis methods, have been relatively slower to adapt the rapid changes (Mayer-Schonberger and Cukier 2013). For example, we tend to sample data rather than take all data for geospatial analysis; we tend to transform skewed data into "normal" by taking logarithms for example. The sampling and logarithm transformation have distorted the underlying property of the data before the data can yield insights. The old way of thinking, Gaussian thinking, that relies on a well-defined mean to characterize geographic features, is a major barrier to achieving deep and new insights into geographic forms and processes.

Many analytical techniques have been developed, in both standard and spatial statistics, to address outliers, to measure skewness and autocorrelation, and to test significance. However, what I want to argue in this paper is not these techniques per se, but something radical in the way of thinking. Gaussian thinking, based on the assumption of independent things in a simple, static, and equilibrium world, is essentially a typical linear thinking, which implies that small cause small effect, large cause large effect, and the whole is equal to the sum of its parts. This linear thinking is a simple way of thinking guided by the reductionism philosophy, and for understanding a simple world in essence (see Sect. 2.2.2 for more details). The reader may argue that spatial statistics differs from standard statistics in spatial dependence or spatial autocorrelation. It is indeed true, but the notion of spatial dependence or autocorrelation does not help us to go beyond Gaussian thinking assumed by standard statistics, for we tend to characterize things by a well-defined mean with a limited variance. It is well recognized that geographic forms are fractal rather than Euclidean, and geographic processes are nonlinear rather than linear (Batty and Longley 1994; Chen 2009). In other words, a geographic system is a complex nonlinear world, in which there is the butterfly effect, and the whole is greater than the sum of its parts. In this paper, I attempt to argue that the Paretian way of thinking, founded on the assumption of interdependent things in a complex, dynamic, and nonequilibrium world, is more appropriate for geospatial analysis, and for better understanding geographic forms and processes. Geospatial analysis, while facing increasing amounts of social media data, should seek to uncover the underlying mechanisms through simulations from the bottom up rather than simple causality or correlations.

The remainder of this paper is organized as follows. Section 2.2 introduces, in a pedagogic manner, two distinct statistic distributions, namely Gaussian- and Paretian-like distributions, with a particular focus on the underlying ways of thinking. Section 2.3 reviews two unique properties of spatial dependence and spatial heterogeneity, and points out that the notion of spatial heterogeneity in current spatial statistics is only used to characterize local variance of spatial dependence. I therefore argue, in Sect. 2.4, that spatial heterogeneity should be formulated as a scaling law, and suggest some effective ways of detecting and revealing the scaling law and pattern for geographic features. I further discuss, in Sect. 2.5, some deep implications of Paretian thinking and the scaling law before draw a summary in Sect. 2.6.

2.2 Two Distinct Distributions and the Underlying Ways of Thinking

In this section, I first illustrate statistical differences between a homogenous Gaussian-like distribution and a heterogeneous Paretian-like distribution (Note the 'homogenous' is relative to the 'heterogeneous'; see Sect. 2.2.1 for more details), using temperature and population of major US cities, and based respectively on histograms and rank-size plots. The temperature is the annual average maximum during 1981–2010, taken from the site: http://www.prism.oregonstate.edu/products/matrix.phtml, while the population is according to the 2010 US census. I then elaborate on the underlying ways of thinking or world views associated with the two categories of distributions.

2.2.1 *Gaussian- Versus Paretian-Like Distributions*

If we carefully examine two variables – temperature and population – of 720 major U.S. cities with population greater than 50,000 people, we can see that the two variables are very distinct. Although not a normal distribution, the temperature can be well characterized by its mean 20.6 (Fig. 2.1a). One can estimate a city's temperature fairly accurate and precise based on the mean value, since the highest is 31.6, and the lowest is 9.3. In other words, the mean 20.6 is a typical temperature for US cities. The distribution that can be characterized by a well-defined mean is referred to as a Gaussian-like distribution including for example the binomial and Poisson distributions. This temperature distribution can be further assessed from the detailed statistics as shown in Table 2.1 (the temperature column). The range between the highest (31.6) and the lowest (9.3) is not very big (22.3), and the ratio of the highest to the lowest is as little as 3.4. The two measures of central tendency – mean and median – are the same. The standard deviation is

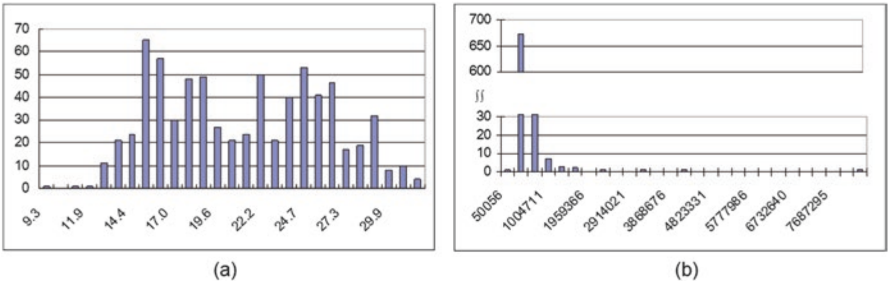


Fig. 2.1 Histograms of (**a**) the temperature, and (**b**) the population of U.S. cities (Note: the two distinct distributions indicate respectively Gaussian-like and Paretian-like distributions)

Table 2.1 Statistics about temperature and population of U.S. cities

Statistics	Temperature	Population
Minimum	9.3	50,056
Maximum	31.6	8,323,732
Range	22.3	8,273,676
Ratio	3.4	166
Mean	20.6	157,467
Median	20.6	82,115
Mode	14.9	62,820
St. Dev.	4.9	393,004

4.9, about one quarter of the range. This statistical picture of the temperature is very distinct from that of the city size or population.

The histogram of the population is extremely right skewed (Fig. 2.1b). This extreme skewness is reflected in several parameters: a wide range (8,273,676), a huge ratio (166), and a large standard deviation (393,004). In such a significantly skewed distribution, the mean of 157,467 make little sense for characterizing the population. In other words, the mean of 157,467 does not represent a typical size of the U.S. cities, since the largest city is as big as 8 millions, while the smallest city is as small as 50,000. The right skewed histogram indicates that there are far more small cities than large ones in the U.S. No wonder that the two measures of central tendency – mean and median – differ from each other significantly; refer to Table 2.1 (the population column) for more details. The standard statistics, or the histogram in particular, is little effective for describing data with a heavy-tailed distribution such as city sizes. Instead, power law based statistics, or rank-size plots in particular, should be adopted for characterizing this kind of data.

Instead of plotting temperature and population on the x-axis (as in the histograms), they are plotted on the y-axis, while the x-axis is the ranking order. This way of plotting is called rank-size plot, or rank-size distribution (Zipf 1949). The largest city (in terms of population) ranks number one, followed by the second largest, and so on. The same arrangement is made for the temperature; the highest temperature city ranks number one, followed by the second highest, and so on. The two distribution lines look very different; the temperature curve drops gradually, and then reaches quickly the minimum, while the population curve drops quickly and then gradually approaches the minimum (Fig. 2.2). Note that the red parts in the figure are those above the averages, called the head, while those below the averages, called the tail, are shown in blue. More specifically, 362 cities (approximately 50%) are above the average temperature 20.6, while only 146 cities (approximately 20%) are above the average city size 157,467. Clearly, a heavy or long tail (80% in the tail) exists for the population distribution, but a short tail (50%) for the temperature distribution. Generally, a heavy-tailed distribution possesses an inbuilt imbalance between the head and the tail (e.g., a 70/30 or 80/20 relationship). This imbalance indicates a nonlinear relationship between the head and the tail. Such an inbuilt

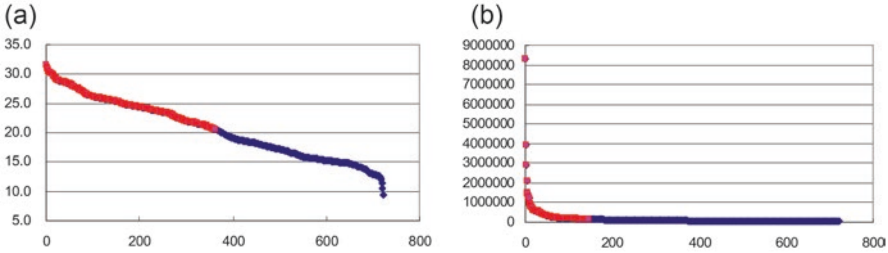


Fig. 2.2 Rank-size plots of (**a**) the temperature, and (**b**) the population of the U.S. cities (Note: values above and below the averages are respectively in *red* and *blue*; clearly, there is a short head and a long tail for the population, forming an unbalanced contrast, while the values above and below the averages are more or less the same for the temperature)

Table 2.2 Comparison between the two ways of thinking

Gaussian thinking	Paretian thinking
With a mean (or scale)	Without a mean (or scale-free)
Static	Dynamic
Simple	Complex
Equilibrium	Non-equilibrium
Linear	Nonlinear
Predictable	Unpredictable

imbalance, or nonlinearity, is clearly missing in a Gaussian-like distribution with a well-balanced relationship between the head and the tail (e.g., 50/50).

2.2.2 *The Underlying Ways of Thinking*

The differences between the two distributions lie fundamentally in different ways of thinking, or different ways of viewing the world, rather than different techniques associated with each distribution. Technically, data with a Paretian-like distribution can be easily transformed into a Gaussian-like distribution, e.g., by taking logarithms. Gaussian thinking implies more or less similar things in a simple, static, and equilibrium world, while Paretian thinking believes in far more small things than large ones in a dynamic, complex, and nonequilibrium world (McKelvey and Andriani 2005; see Table 2.2). Standard statistics teaches us that if the probability of an event is small, then the event occurs rarely. The event can be considered an outlier that is literally distant from the rest of the data. However, in Paretian thinking, an event of small probability, or the highly improbable, has a significant impact (e.g., the September 11 attacks) and thus be ranked highly.

In Gaussian thinking, the world does not change much, and all changes occur around a stable and well-defined mean. Thus, the presumed Gaussian world is static, simple, linear, and predictable. The Newtonian physics is sufficient to understand and deal with the Gaussian world. Why does such a predictable world exist? Such a world reflects a lack of interaction and competition among individual agents; every agent acts independently without influences upon or affects from others. This assumption is fundamental to standard statistics, and of course appropriate for many events in the world like human heights. Spatial statistics differentiate it from standard statistics in spatial dependence, but it does not change fundamentally the underlying way of thinking – Gaussain thinking with a well-defined mean for characterizing things. On the other hand, in Paretian thinking, the world is full of surprises, and changes are often dramatic and unexpected. Thus, there is no stable and well-defined mean for characterizing the surprises and changes. The presumed Paretian world is essentially dynamic, complex, nonlinear, and unpredictable. This unpredictable world is founded on the assumption that everything is related to, or interdependent with, everything else. This interdependence assumption implies that cooperation and competition would eventually lead to unbalanced results characterized by a long-tail distribution (c.f. Sect. 2.3 for more discussions).

Nature is awash with phenomena such as trees, rivers, mountains, clouds, coastlines, and earthquakes that exhibit power laws or heavy-tailed distributions in general (Mandelbrot 1982; Schroeder 1991; Bak 1996). Accordingly, power law has been formulated as a fundamental law in various disciplines such as physics, biology, economics, computer science, and linguistics. People's daily activities are also governed by power laws (Barabási 2010), indicating bursty behaviors of human mobility or activities in general. Power laws are a signature of complex systems that are evolved in nonlinear manners, i.e., small causes often have disproptional large effects. For instance, the top 10% of the most connected streets account for 90% of traffic flows (Jiang 2009). In a 21-block area of Philadelphia, 70% of the marriages occurred between people who lived no more than 30% of that distance apart (Zipf 1949).

The examination of the two ways of thinking suggests that Paretian-like distribution, or Paretian thinking in general, appears more appropriate for understanding geographic forms and processes, for dependence is a key property of spatial statistics (c.f., Sect. 2.3 for more details). In spite of spatial dependence being its key property, spatial statistics is still unfortunately very much dominated by Gaussian thinking. The very notion of spatial heterogeneity refers to local variance of spatial dependence, but from global to local, or from one single correlation coefficient to multiple coefficients (c.f., Sect. 2.3 for more details). In the remainder of this paper, we review two spatial properties of dependence and heterogeneity, and argue that spatial heterogeneity is ubiquitous, and it should be formulated as a scaling law. And we further discuss some deep implications of the scaling law and Paretian thinking for better understanding of geographic forms and processes in the era of big data.

2.3 Spatial Properties of Dependence and Heterogeneity

It is well known that in contrast to the independence assumption of standard statistics, geographic phenomena or events are not random or independent. Geographic events are more likely to occur in some locations than others (spatial heterogeneity), and nearby events are more similar than distant events (spatial dependence). Both spatial heterogeneity and spatial dependence are referred to as spatial properties, indicating respectively that geographic events are related to their locations and to their neighboring events. Spatial dependence is widely known or formulated as the first law of geography: *"Everything is related to everything else, but near things are more related than distant things"* (Tobler 1970). For example, your housing price is likely to be similar (positive correlation) to those of your neighbors. Similarly, the elevations of two locations 10 m apart are likely to be more similar than the elevations of two locations of 100 m apart. Note that "likely" indicates a statistical rather than a deterministic property; one can always find exceptions in statistical trends.

Spatial heterogeneity refers to no average location that can characterize the Earth's surface (Anselin 1989; Goodchild 2004). This is indeed true, while for example referring to the diversity of landscapes and species (animals and plants) on the Earth's surface (Wu and Li 2006; Bonner 2006). This diversity or heterogeneity indicates uneven geographic and statistical distributions involving both landscapes and species – that is, a mix of concentrations of multiple species (biological), terrain formations (geological), environmental characteristics (such as rainfall, temperature, and wind) on the one hand, and various concentrations of various types of species on the other. A variety of habitats such as different topographies, soil types, and climates can accommodate a greater number of species. These are the natural environments of the Earth's surface. Spatial heterogeneity in geography also concerns human-made built environments created by human activities such as industrialization and urbanization, and in particular, for example, the diversity of human settlements or cities in particular. Given the diversity or spatial heterogeneity of the Earth's surface, homogeneous Gaussian-like distribution is unlikely to be the right means to characterize complex geographic features.

Spatial dependence and spatial heterogeneity are properties to spatial data and geospatial analysis (Anselin 1989; Griffith 2003), and probably the two most important principles of geographic information science (GIScience). Goodchild (2004) has been a key advocator for formulating general principles for GIScience. On several occasions, he has made insightful remarks on spatial heterogeneity or spatial properties in general. His definition of spatial heterogeneity as "no average location" is in effect the notion of scale-free used to characterize things that exhibit a power law or heavy-tailed distribution (Barabási and Albert 1999). On the other hand, he stated, with respect to spatial heterogeneity, that all locations are unique, due to which geography might be better considered as an idiographic science, studying the unique properties of places (Goodchild 2004). However, I argue, in contrast to Goodchild, that spatial heterogeneity makes geography a nomothetic science. This is because spatial heterogeneity itself is a law – the scaling law, implying that there are far more small geographic features than large ones. Spatial heterogeneity is a kind of hidden order, which appears disordered on the surface, but possesses a deep order beneath. This kind of hidden order can be characterized by a power law or a heavy-tailed distribution in general.

Current spatial statistics suffers from what I call ‘spatial heterogeneity paradox’. Spatial heterogeneity is defined as no average location, but we tend to use a well-defined mean or average to characterize locations. This paradox implies that our mindsets are still constrained by Gaussian thinking. The current notion of spatial heterogeneity refers to local variance of spatial dependence or regression. This can be seen from the development of local spatial statistics and local statistical models that initially brought spatial heterogeneity into spatial statistics (Anselin 1989). Local spatial statistics concern local variants of spatial autocorrelation or regression, a measure to spatial dependence, including, for example, the local statistical models (Getis and Ord 1992), the LISA techniques (Anselin 1995), and geographically-weighted regression (Fotheringham et al. 2002). The shifting perspective of spatial autocorrelation from global to local brings new insights into spatial dependence, or the heterogeneity of spatial dependence. However, all these techniques and models are essentially based on Gaussian statistics, using a well-defined mean with a limited variance. To paraphrase Mandelbrot (Mandelbrot and Hudson 2004), spatial heterogeneity refers to ‘wild’ variances, but Gaussian-like distribution can only characterize ‘mild’ variances.

Human activities are the major forces behind spatial heterogeneity in the built environments. While carrying out activities, human beings (and their interventions) must respect the spatial heterogeneity of Nature – that is, harmonize with rather than damage the natural environments. Geographic information concerning urban and human geography captures essentially spatial variations of the built environments, which demonstrate ‘wild’ heterogeneity as well. For example, Zipf’s law on city sizes (Zipf 1949) mainly concerns such a spatial variation. Thus, I argue, in contrast to the conventional view, that dependence, or more precisely interdependence, is a first-order effect, while heterogeneity is a second-order effect. Let us do a thought experiment. Imagine that once upon a time, there were no cities, only scattered villages. Over time, large cities gradually emerge through the interactions of villages, so do mega cities through the interactions of cities. The interactions (competition and cooperation) of villages and cities are actually those of people acting individually and/or collectively. These interactions are what we mean by dependence and interdependence. Eventually, there are far more small cities than large ones through for example the mechanism of “the rich get richer.” This observation is the same for the wealth distribution among individuals in a country; far more poor people than rich people, or far more ordinary people than extraordinary people (Epstein and Axtell 1996). The interactions among people and cities reflect the interdependence effect in the formation and evolution of cities and city systems, and the built environments in general.

2.4 Spatial Heterogeneity as a Scaling Law

The subtitle of this paper *'The Problem of Spatial Heterogeneity'* is an homage to the classic work *'The Problem of Spatial Autocorrelation'* (Cliff and Ord 1969), which popularized the concept of spatial dependence. Similarly, spatial heterogeneity under Gaussian thinking is indeed a problem because the Earth's surface cannot be characterized by a well-defined mean. However, in the Paretian way of thinking, spatial heterogeneity is not a problem, but the norm. Spatial heterogeneity should be formulated as a scaling law in geography.

2.4.1 Ubiquity of the Scaling Law in Geography

Geographic features are unevenly or abnormally distributed, so the scaling pattern of far more small things than large ones is widespread in geography (Pumain 2006). The scaling pattern has another name called fractal (Mandelbrot 1982). Fractal-related research in geography has concentrated too much on concepts such as fractal dimension and self-similarity. In fact, the scaling law is fundamental to all of these concepts. In this regard, Salingaros and West (1999) formulated a universal rule for city artifacts; there are far more small city artifacts than large ones, due to which the image of the city can be formed in human minds (Jiang 2013b). With the increasing availability of geographic information, the scaling law has been observed and examined in a wide range of geographic phenomena including, for example, street lengths and connectivity (Carvalho and Penn 2004; Jiang 2009), building heights (Batty et al. 2008), street blocks (Lämmer 2006; Jiang and Liu 2012), population densities (Schaefer and Mahoney 2003; Kyriakidou et al. 2011), and airport sizes and connectivity (Guimerà et al. 2005). Interestingly, the scaling of geographic space has had an enormous effect on human activities; human activities and interactions in geographic space exhibit power law distributions as well (Brockmann et al. 2006; Gonzalez et al. 2008; Jiang et al. 2009). Table 2.3 provides a synoptic view of the ubiquity of power laws in geography, noting that the references listed are non-exhaustive, but for example only.

Despite its ubiquity, ironically the scaling law, or the Paretian way of thinking in general, has not been well received in geospatial analysis as elaborated earlier in the text. Current geospatial analysis adopts a well-defined mean or average to characterize spatial heterogeneity. The two closely related concepts of scale and scaling must be comprehended together, i.e., many different scales, ranging from the smallest to the largest, form a scaling hierarchy. This comprehension should be added, as a fourth one, into the three meanings of scale in geography: cartographic, analysis, and phenomenon (Montello 2001); see more elaborations in this recent paper (Jiang and Brandt 2016). The essence of power laws is the scaling pattern, in which there are far more small scales than large ones. This scaling pattern reflects the true picture of spatial heterogeneity or that of the Earth's surface.

Table 2.3 Power laws in geographic features or phenomena

Geographic phenomena	References (for example)
City sizes	Zipf (1949), Krugman (1996), and Jiang and Jia (2011)
Fractals in cities or geographic space	Goodchild and Mark (1987) and Batty and Longley (1994)
Coast lines and mountains	Mandelbrot (1967) and Bak (1996)
Hydrological networks	Hack (1957), Horton (1945), Maritan et al. (1996), and Pelletier (1999)
Urban and architectural space	Salingaros and West (1999)
Street lengths and connectivity	Carvalho and Penn (2004) and Jiang (2009)
Building heights	Batty et al. (2008)
Street blocks	Lämmer (2006) and Jiang and Liu (2012)
Population density	Schaefer and Mahoney (2003) and Kyriakidou et al. (2011)
Airport sizes and connectivity	Guimerà et al. (2005)
Human mobility	Brockmann et al. (2006), Gonzalez et al. (2008), and Jiang et al. (2009)

2.4.2 Detecting the Scaling Law

What were claimed to be power laws in the literature could be actually lognormal, exponential, or other similar distributions, because the detection of power laws can be very tricky. Given a power law relationship $y = x^{a}$, it can be transformed into the logarithm scales, i.e., $\ln(y) = a \cdot \ln(x)$, indicating that the logarithms of the two variables x and y have a linear relationship. Conventionally, an ordinary least squares (OLS) based method was widely used for the detection. In the fractal literature, the box-counting method is usually used to compute the fractal dimension, which is the de facto power law exponent, and its computation is also based on OLS. There are at least two issues surrounding the power law detection. The first is that the OLS based method is found to be less reliable for detecting a power law, so a maximum likelihood method has been developed (Clauset et al. 2009). It was found that many claims on power laws in the literature are likely to be lognormal or other degenerated formats such as a power law with an exponential cutoff. For the sake of readability, this paper does not cover mathematical details on heavy-tailed distributions and their detection; interested readers can refer to Clauset et al. (2009) and the references therein. The second is that even with the OLS-based method, the definition of fractal dimension is so strict that many geographic features are excluded from being fractal (Jiang and Yin 2014). Given the circumstance, the authors have recently provided a rather relaxed definition of fractals, i.e., a geographic feature is fractal if and only if the scaling pattern of far more small things than large ones recurs multiple times. The number of times plus one is referred to as ht-index (Jiang and Yin 2014), an alternative index of fractal dimension for characterizing complexity of fractals or geographic features in particular.

The idea behind the relaxed definition of fractals, or the ht-index, is pretty simple and straightforward. It is based on the head/tail breaks (Jiang 2013a), a new classification scheme for data with a heavy-tailed distribution. Given a variable whose distribution is right skewed, compute its arithmetic mean, and subsequently split its values into two unbalanced parts: those above the mean in the head, and those below the mean in the tail. The values above the mean are a minority, while the values below are a majority. The ranking and breaking process continues for the head part progressively and iteratively until the values in the head no longer meet the condition of far more small things than large ones. This way both the number of classes and the class intervals are naturally and automatically derived based on the inherent hierarchy of data. Eventually, the number of classes, or equivalently the ht-index, indicates hierarchical levels of the values. One can simply rely on an Excel sheet for the computation of the ht-index. As an example, Fig. 2.3 illustrates the scaling pattern of the US cities, discussed earlier in Sect. 2.2.1, and it has the ht-index of 7.

2.4.3 Revealing the Scaling Pattern

The head/tail breaks can effectively reveal or visualize the scaling pattern if the data itself exhibits a heavy-tailed distribution. This is because the head/tail breaks was developed initially for revealing the inherent scaling hierarchy or the scaling pattern. In this regard, conventional classification methods, mainly guided by Gaussian thinking, failed to reveal the scaling pattern. For example, the most widely used classification natural breaks (Jenks 1967), which is set as a default in ArcGIS, is based on the principle of minimizing within-classes variance, and maximizing between-classes variance. It sounds very natural. In some case like the US cities, the classification result of the natural breaks may look very similar to the one by natural breaks (Fig. 2.3). However, this is just by chance. Essentially, the natural breaks is motivated by Gaussian thinking; each class is characterized by a well-defined mean with a limited or minimized variance. In a contrast, the head/tail breaks is motivated by Paretian thinking, and for data with a Paretian-like or heavy-tailed distribution. The iteratively or recursively defined averages are used as meaningful cutoffs for differentiating hierarchical levels.

The reader probably has got used to the US terrain surface (Fig. 2.4a), which is based on the natural breaks. It is commonly seen in geography and cartography textbooks and atlases. However, I want to challenge this conventional wisdom, arguing that the natural breaks based visualization is little natural. I contend that the head/tail breaks derived visualization is more natural, since it reflects the underlying scaling pattern of far more small things than large ones (Fig. 2.4b). The things here are referred to individual locations, or more specifically, far more low locations than high locations. The left visualization, which distorted the scaling pattern, appears having far more high locations than the visualization to the right, or equivalently far

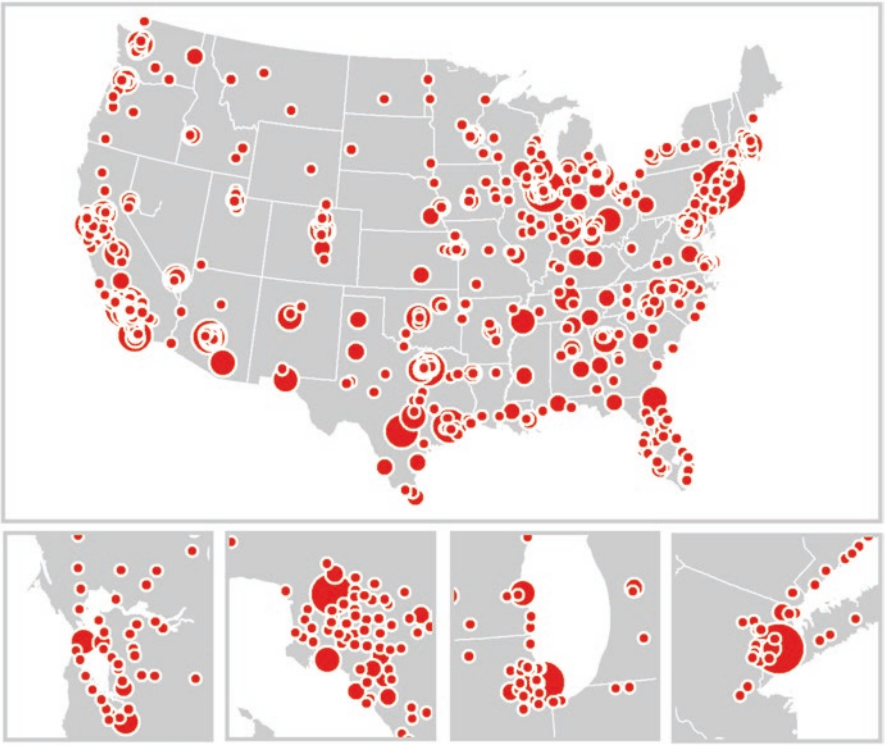


Fig. 2.3 Scaling pattern of US cities with ht-index equal 7 (Note: the four *insets* from the *left* to the *right* provide the enlarged view respectively for San Francisco, Los Angeles, Chicago and New York regions)

more high locations than what it actually has. The visualization to the right reflects well the underling scaling pattern. This can be further seen from the corresponding histograms of the individual classes of the two classifications (Fig. 2.4c, d). What is illustrated by the left histogram is "more low locations than high ones" which is a linear relationship, rather than "far more low locations than high ones", which is a nonlinear relationship. For the left histogram, each pair of the adjacent bars from left to right does not constitute an unbalanced contrast of majority versus minority. For example, the first pair of bars of the left histogram shows a well-balanced contrast of 7–6; in a contrast, the first pair of the right histogram is unbalanced, 14–5. Therefore, the right histogram indicates clearly "far more low locations than high ones". Interestingly, the scaling pattern remains unchanged with respect to different scales of digital elevation models (Lin 2013).

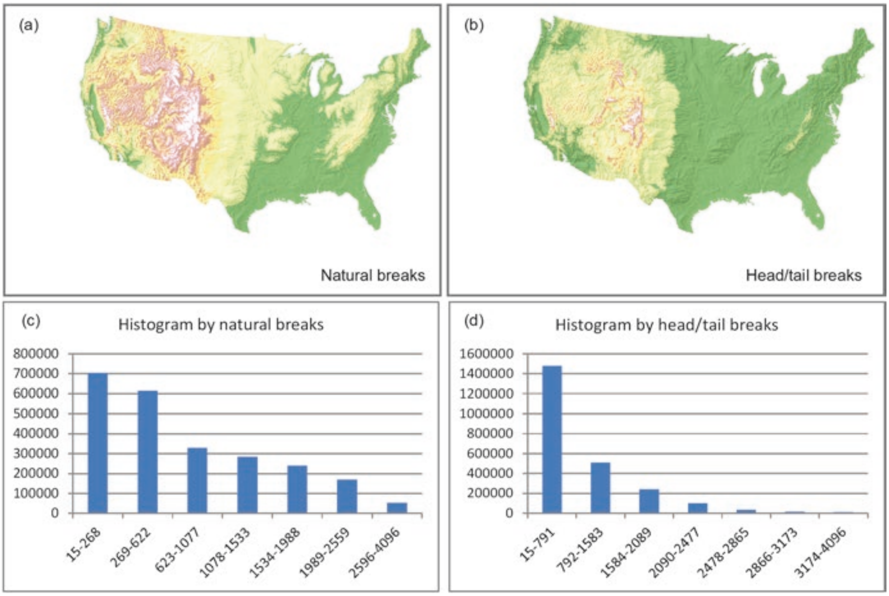


Fig. 2.4 The scaling pattern of US terrain surface is distorted by the natural breaks, but revealed by the head/tail breaks

2.5 Implications of Paretian Thinking and the Scaling Law

Current geospatial analysis concentrates more on geographic forms, but less on why the forms. The forms illustrated are mostly limited to whether they are random, or to what extent they are auto-correlated. As to why the forms, it is usually ended up with simple regressions and causalities. This way of geospatial analysis is much like short-term weather forecasting. Despite its usefulness, the short-term weather forecast adds little to understanding the complex behavior of weather – the long-term weather beyond 2 or 3 weeks. In essence, the long-term weather, or climate change in general, is unpredictable, just like earthquakes and many other events in Nature and society (Bak 1996). If the real world is unpredictable, what can we do as scientists? We can simulate interactions of things from the bottom up in order to understand the underlying mechanisms, which would help improve predictions. In this regard, the emerging social media, in particular location-based social media, provide valuable data for validating the simulation results (Jiang and Miao 2015). The data, unlike traditional statistical or census data that are mainly aggregated, are not only big in size, but are collected at individual levels. The data are not only at individual levels, but linked in time and among individuals. The data can help track the trajectories of individuals and their associations in space and over time. For this kind of social media data, the scaling law and fractals should be the norm.

Current spatial statistics constrained by Gaussian thinking show critical limitations for analyzing or getting insights into big data (Mayer-Schonberger and Cukier

2013). What are illustrated by spatial statistics, either patterns or associations, can be compared to the mental images of the elephant in the minds of the blind men. These images reflect local truths, and are indeed correct partially, but they did not reflect the whole of the elephant. Geospatial analysis should go beyond illustrating spatial autocorrelation, either globally or locally, but towards uncovering the underlying scaling or fractal patterns. Geographic features are essentially and ultimately scaling or fractal. Therefore, any patterns deviating from the scaling pattern or that can be characterized by a well-defined mean are either wrong or biased.

Geographic forms (or phenomena) are not the outcomes of simple processes but the results of complex processes with positive feedbacks. In the built environments, human interventions (interdependence and interactions) of various kinds are the major effects of spatial heterogeneity. As famously stated by Winston Churchill (1874–1965), "*we shape our buildings, and thereafter they shape us*". This statement should be comprehended in a progressive and recursive manner. This comprehension, which underlies Paretian thinking, is essentially a complex system perspective for exploring the underlying processes related to geographic forms (e.g., Benguigui and Czamanski 2004; Blumenfeld-Lieberthal and Portugali 2010). In this regard, complexity science has developed a range of tools such as discrete models, complex networks, scaling hierarchy, fractal geometry, self-organized criticality, and chaos theory (Newman 2011). All these modeling tools attempt to reveal the underlying mechanisms, linking surface complex forms (or complexity) to the underlying mechanisms (or deep simplicity) through simulations from the bottom up, rather than simple descriptions of forms or of geographic forms in particular.

Paretian thinking represents a paradigm shift. Shifting from the sands to the avalanches (Bak 1996), and from the street segments to the natural streets (Jiang 2009), enables us to see something interesting and exciting, i.e., from the things of limited sizes to the things of all sizes. The things of all sizes imply a scaling pattern across all scales. Recognition of the scaling pattern helps us to better understand the underlying universal form of geographic features. This scaling pattern can further be linked to the underlying geographic processes that are dynamic, nonlinear, and bottom-up in nature. This view would position geography in the family of science, since we geographers are interested in not only what things look like (the forms) but also why things look that way (the processes). Spatial heterogeneity is thus not a problem but an underlying scaling law of geography.

2.6 Concluding Summary

This paper argues that geospatial analysis requires a different way of thinking, or world view in general, that underlies the Paretian-like distribution of geographic features. We put the two distinct views in comparison: more or less similar things in a simple, static, and equilibrium world on the one hand, and far more small things than large ones in a complex, dynamic, and non-equilibrium world on the other. Geospatial analysis has been dominated by Gaussian statistics with a well-defined mean for characterizing spatial variation ('mild' variance so to speak). Despite its ubiquity in geography, the Paretian-like heavy-tailed distribution, or the underlying way of thinking in general, has not been well received in geospatial analysis. The current geospatial analysis mainly focuses on how spatial variation deviates from a random pattern, and measuring spatial auto-correlation from global to local (the current spatial heterogeneity), but leaves the underlying processes unexplored. This way of geospatial analysis is inadequate for understanding geographic forms and processes, in particular while facing increasing amounts of social media data.

No average location exists on the Earth's surface. Instead, there are far more small things than large ones in geographic space; small things are a majority while large things are a minority. Importantly, the pattern of far more small things than large ones recurs multiple times (Jiang and Yin 2014). This recurring scaling pattern reflects the true image of spatial heterogeneity that lacks a well-defined mean ('wild' variance so to speak). Spatial heterogeneity is indeed a problem in Gaussian thinking, but it is a law or scaling law in Paretian thinking. In the spirit of Paretian thinking and the scaling law, geospatial analysis should seek to simulate individuals and individual interactions from the bottom up rather than simple correlations and causalities. In this connection, complexity tools such as complex networks, agent-based modeling, and fractal/scaling provide effective means for geospatial analysis of complex geographic phenomena.

Acknowledgments This chapter is a reprint of original paper by Jiang (2015) with the permission of the publisher. The author would like to thank the anonymous referees and the editor Daniel Z. Sui for their valuable comments. However, any shortcoming remains the responsibility of the author.

References

Anderson C (2006) The Long tail: why the future of business is selling less of more. Hyperion, New York

Anselin L (1989) What is special about spatial data: alternative perspectives on spatial data analysis. National Center for Geographic Information and Analysis, Santa Barbara

Anselin L (1995) Local indicators of spatial association – LISA. Geogr Anal 27:93–115

Bak P (1996) How nature works: the science of self-organized criticality. Springer, New York

Barabási A (2010) Bursts: the hidden pattern behind everything we do. Dutton Adult, Boston

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Batty M, Longley P (1994) Fractal cities: a geometry of form and function. Academic, London

Batty M, Carvalho R, Hudson-Smith A, Milton R, Smith D, Steadman P (2008) Scaling and allometry in the building geometries of Greater London. Eur Phys J B 63:303–314

Benguigui L, Czamanski D (2004) Simulation analysis of the fractality of cities. Geogr Anal 36(1):69–84

Blumenfeld-Lieberthal E, Portugali J (2010) Network cities: a complexity-network approach to urban dynamics and development. In: Jiang B, Yao X (eds) Geospatial analysis of urban structure and dynamics. Springer, Berlin, pp 77–90

Bonner JT (2006) Why size matters: from bacteria to blue whales. Princeton University Press, Princeton

Brockmann D, Hufnage L, Geisel T (2006) The scaling laws of human travel. Nature 439:462–465

Carvalho R, Penn A (2004) Scaling and universality in the micro-structure of urban space. Phys A 332:539–547

Chen Y (2009) Spatial interaction creates period-doubling bifurcation and chaos of urbanization. Chaos, Solitons Fractals 42(3):1316–1325

Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51:661–703

Cliff AD, Ord JK (1969) The problem of spatial autocorrelation. In: Scott AJ (ed) London papers in regional science. Pion, London, pp 25–55

Epstein JM, Axtell R (1996) Growing artificial societies: social science from the bottom up. Brookings Institution Press, Washington, DC

Fotheringham AS, Brunsdon C, Charlton M (2002) Geographically weighted regression: the analysis of spatially varying relationships. Wiley, Chichester

Getis A, Ord JK (1992) The analysis of spatial association by distance statistics. Geogr Anal 24:189–206

Gonzalez M, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:779–782

Goodchild M (2004) The validity and usefulness of laws in geographic information science and geography. Ann Assoc Am Geogr 94(2):300–303

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69(4):211–221

Goodchild MF, Mark DM (1987) The fractal nature of geographic phenomena. Ann Assoc Am Geogr 77(2):265–278

Griffith DA (2003) Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin

Guimerà R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci U S A 102(22):7794–7799

Hack J (1957) Studies of longitudinal stream profiles in Virginia and Maryland. US Geol Surv Prof Pap:294-B

Horton RE (1945) Erosional development of streams and their drainage basins: hydrological approach to quantitative morphology. Bull Geogr Soc Am 56:275–370

Jenks GF (1967) The data model concept in statistical mapping. Int Yearb Cartogr 7:186–190

Jiang B (2009) Street hierarchies: a minority of streets account for a majority of traffic flow. Int J Geogr Inf Sci 23(8):1033–1048

Jiang B (2013a) Head/tail breaks: a new classification scheme for data with a heavy-tailed distribution. Prof Geogr 65(3):482–494

Jiang B (2013b) The image of the city out of the underlying scaling of city artifacts or locations. Ann Assoc Am Geogr 103(6):1552–1566

Jiang B (2015) Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity. GeoJournal 80(1):1–13

Jiang B, Brandt A (2016) A fractal perspective on scale in geography. ISPRS Int J Geo-Inf 5(6):95. doi:10.3390/ijgi5060095

Jiang B, Jia T (2011) Zipf's law for all the natural cities in the United States: a geospatial perspective. Int J Geogr Inf Sci 25(8):1269–1281

Jiang B, Liu X (2012) Scaling of geographic space from the perspective of city and field blocks and using volunteered geographic information. Int J Geogr Inf Sci 26(2):215–229

Jiang B, Miao Y (2015) The evolution of natural cities from the perspective of location-based social media. Prof Geogr 67(2):295–306

Jiang B, Yin J (2014) Ht-index for quantifying the fractal or scaling structure of geographic features. Ann Assoc Am Geogr 104(3):530–541

Jiang B, Yin J, Zhao S (2009) Characterizing human mobility patterns in a large street network. Phys Rev E 80:021136

Koch R (1999) The 80/20 principle: the secret to achieving more with less. Crown Business, New York

Krugman P (1996) The self-organizing economy. Blackwell, Cambridge, MA

Kyriakidou V, Michalakelis C, Varoutas D (2011) Applying Zipf's power law over population density and growth as network deployment indicator. J Serv Sci Manag 4:132–140

Lämmer S, Gehlsen B, Helbing D (2006) Scaling laws in the spatial structure of urban road networks. Phys A 363(1):89–95

Lin Y (2013) A comparison study on natural and head/tail breaks involving digital elevation models. Bachelor thesis at University of Gävle, Sweden

Mandelbrot B (1967) How long is the coast of Britain? Statistical self-similarity and fractional dimension. Science 156(3775):636–638

Mandelbrot BB (1982) The fractal geometry of nature. W. H. Freeman, San Francisco

Mandelbrot BB, Hudson RL (2004) The (mis)behavior of markets: a fractal view of risk, ruin and reward. Basic Books, New York

Maritan A, Rinaldo A, Rigon R, Giacometti A, Rodríguez-Iturbe I (1996) Scaling laws for river networks. Phys Rev E E53:1510–1515

Mayer-Schonberger V, Cukier K (2013) Big data: a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt, New York

McKelvey B, Andriani P (2005) Why Gaussian statistics are mostly wrong for strategic organization. Strateg Organ 3(2):219–228

Montello DR (2001) Scale in geography. In: Smelser NJ, Baltes (eds) International encyclopedia of the social & behavioral sciences. Pergamon Press, Oxford, pp 13501–13504

Newman M (2011) Resource letter CS-1: complex systems. Am J Phys 79:800–810

Pareto V (1897) Cours d'économie politique. Rouge, Lausanne

Pelletier JD (1999) Self-organization and scaling relationships of evolving river networks. J Geophys Res 104:7359–7375

Pumain D (2006) Hierarchy in natural and social sciences. Springer, Dordrecht

Salingaros NA, West BJ (1999) A universal rule for the distribution of sizes. Environ Plan B Plan Des 26:909–923

Schaefer JA, Mahoney AP (2003) Spatial and temporal scaling of population density and animal movement: a power law approach. Ecoscience 10(4):496–501

Schroeder M (1991) Chaos, fractals, power laws: minutes from an infinite paradise. Freeman, New York

Taleb NN (2007) The black swan: the impact of the highly improbable. Allen Lane, London

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(2):234–240

Wu J, Li H (2006) Concepts of scale and scaling. In: Wu J, Jones KB, Li H, Loucks OL (eds) Scaling and uncertainty analysis in ecology. Springer, Berlin, pp 3–15

Zipf GK (1949) Human behavior and the principles of least effort. Addison Wesley, Cambridge, MA

Part II
Geographic Data Mining

Chapter 3
A Survey on Spatiotemporal and Semantic Data Mining

Quan Yuan, Chao Zhang, and Jiawei Han

Abstract The wide proliferation of GPS-enabled mobile devices and the rapid development of sensing technology have nurtured explosive growth of semantics-enriched spatiotemporal (SeST) data. Compared to traditional spatiotemporal data like GPS traces and RFID data, SeST data is multidimensional in nature as each SeST object involves location, time, and text. On one hand, mining spatiotemporal knowledge from SeST data brings new opportunities to improving applications like location recommendation, event detection, and urban planning. On the other hand, SeST data also introduces new challenges that have led to the developments of various techniques tailored for mining SeST information. In this survey, we summarize state-of-the-art studies on knowledge discovery from SeST data. Specifically, we first identify the key challenges and data representations for mining SeST data. Then we introduce major mining tasks and how SeST information is leveraged in existing studies. Finally, we provide an overall picture of this research area and an outlook on several future directions of it. We anticipate this survey to provide readers with an overall picture of the state-of-the-art research in this area and to help them generate high-quality work.

Keywords Spatiotemporal data • Semantic data • Data mining techniques

3.1 Introduction

With the wide proliferation of GPS-enabled mobile devices and the rapid advance of sensing technology, recent years are witnessing a massive amount of semantics-rich spatiotemporal data accumulated from various sources. For example, on social media platforms like Twitter, millions of geo-tagged tweets are created every day, where each geo-tagged tweet consists of a timestamp, a location, and short text

Q. Yuan (✉) • C. Zhang • J. Han
Department of Computer Science, University of Illinois at Urbana-Champaign,
201 N Goodwin Ave, Urbana, IL 61801, USA
e-mail: qyuan@illinois.edu; czhang82@illinois.edu; hanj@illinois.edu

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_3

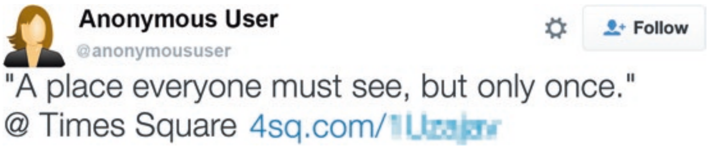

Anonymous User
@anonymoususer
"A place everyone must see, but only once."
@ Times Square 4sq.com/[redacted]

Fig. 3.1 An example of geo-annotated tweet. User name and url link are anonymized for privacy preservation

(Fig. 3.1). For another example, mainstream search engines (e.g., Google, Bing) are continuously collecting queries from GPS-enabled mobile devices. These queries are also associated with timestamps and locations as metadata.

Compared to traditional spatiotemporal data like GPS traces and RFID data, semantics-enrich spatiotemporal (abbreviated as SeST onwards) data is multidimensional in nature. A typical SeST object involves three different data types (location, time, and text) and thus provide a unified where-when-what (three W) view of people's behaviors. As such, the prevalence of SeST data brings new opportunities to spatiotemporal knowledge discovery and opens doors to improving a lot of real-life applications. Consider mobility understanding as an example. While traditional GPS trace data can reveal how an individual moves from one location to another, the SeST data allows us to go beyond that and understand what activities the individual does at various locations. Such semantics-level information is essential in terms of capturing people's mobility patterns, improving applications e.g. location prediction, advertising targeted users, and urban planning.

While SeST data sheds light on improving a wide variety of real-life applications, it is by no means trivial to fully unleash its power. Compared with knowledge discovery in traditional spatiotemporal data, mining SeST data introduces a handful of new challenges:

- **How to integrate diverse data types?** SeST data involves three data types: location, time, and text. Considering the distinct representations of these data types (continuous or discrete) and the complicated correlations among them, it is difficult to effectively integrate them for spatiotemporal knowledge discovery.
- **How to overcome data sparsity?** Unlike intentionally collected tracking data, most SeST data is low-sampling in nature. Take geo-tagged tweets as an example, a user is unlikely to report her activity at every visited location. Such data sparsity makes it challenging to apply classic data mining and machine learning techniques.
- **How to extract useful knowledge from noisy data?** Text in SeST data is usually short and noisy. For example, a geo-tagged tweet contains no more than 140 characters, and most geo-tagged Instagram photos are associated with quite short text descriptions. Moreover, existing studies have revealed that about 40% social media posts are just pointless babbles (Kelly 2009). Still, it is nontrivial to make use of such noisy and incomplete text data to acquire useful knowledge.
- **How to handle large-scale SeST data to build scalable and efficient systems?** Many spatiotemporal applications (e.g., local event detection, location prediction)

requires the back-end system to deal with large-scale SeST data and to respond to users' needs in a timely manner. Since practical SeST data comes in a massive volume, how to develop efficient techniques to handle such big SeST data remains challenging.

Because of the large potential of SeST data in improving various spatiotemporal applications as well as the unique challenges in fully unleashing the power of SeST data, mining SeST data has attracted a lot of research attention from communities like data mining, civil engineering, transportation, and environmental science. SeST data mining has potentially great impact on a variety of fields such as sociology, epidemiology, psychology, public health, etc. We notice several review works on spatial and spatiotemporal data mining (Cheng et al. 2014; Shekhar et al. 2015), but a systematic summarization on state-of-the-art techniques for mining SeST data is still an untouched topic. In this survey, we summarize recent research studies on knowledge discovery from SeST data. Specifically, we introduce data representations, key research problems, methodologies, and future directions. We anticipate this survey to provide an over- all picture of this area, which can help the community better understand the cutting edge and generate quality research results.

The organization of this survey is as follows. In Sect. 3.2, we survey the datasets and the representations of spatial, temporal and semantic information used in existing studies, and introduce the major approaches to SeST data mining. Then, in Sect. 3.3 we review the major tasks that are widely studied in existing SeST data mining works. Important directions for future research are discussed in Sect. 3.4. In the end, Sect. 3.5 summarizes the article.

3.2 Framework of Mining the SeST Data

Mining the SeST data is a general process of acquiring, integrating, analyzing, and mining semantics-enriched spatiotemporal data. In this section, we overview the data sources and representations of SeST data, and then introduce the major approaches to mining SeST knowledge.

3.2.1 Data Sources

Various types of SeST data are used in existing studies. In this section, we list the major data sources and introduce their properties.

- **User generated content.** With the development of social media websites and GPS technology, a great quantity of user generated content has been accumulated, which involves spatial, temporal and semantic information. Examples are social media posts such as Tweets and Facebook statuses, reviews in crowd-sourced review based social networks (e.g., Yelp, Dianping), check-ins in location-based social networks (e.g., Foursquare), events in event-based social networks and travelogues.

- **Survey study data.** Some organizations collect the mobility behaviors of users via survey studies. Representative survey data includes MIT Reality Mining,[1] American Time Use Survey,[2] and Puget Sound Regional Council Household Activity Survey.[3] In the survey data, each visit of an individual involves location, visiting time, and semantics describing the activity (working, shopping, etc.) or visiting purpose. Sometimes, survey data also contains demographic information of individuals, such as age, gender, job, etc. This enables us to study the correlations between user mobility and their demographics.
- **GPS trajectories.** Trajectories, consisting of a series of coordinates-timestamps information, are used to unveil people's mobility. As visits are passively collected, people often need to extract stay points as the locations which a user visited rather than passed by. The stay points are extracted based on other evidence, such as the mobility range in a session and the stay time. Semantic information, such as location categories and descriptions, are often extracted from external data sources, such as Wikipedia, gazetteers, land use around cell towers, etc.
- **Query logs and browsing histories.** As an increasing number of cellphones are 3G/4G enabled, more people search information and browse webpages on the go. As a result, query logs and browsing data are associated with geographic coordinates, representing the current surrounding of the users. The metadata reveal spatial and temporal information, and the text content can be used as the semantic information.
- **News feeds and blogs.** A large amount of news feeds and blogs have location and time information, and thus such data can be also viewed as SeST data. In these datasets, spatial and temporal information can be either collected from metadata, or exacted from the content parsed by some natural language processing (NLP) tools. The text content carries rich and often high-quality semantics.

Among these datasets, the publicly available user generated content are often of large quantity but low quality, in terms of sparsity (some users may only have few posts), noise (users write text in free style), and incompleteness (observations are available only if the users actively submit them). In contrast, survey data is of much higher quality, but they are expensive to get, and the lengths of observations are often short (ranging from 2 to 100 days). Trajectory data often has a reasonable and stable sampling rate, but it is hard to collect and additional steps are needed to extract stay-points and to infer the semantics. Query logs are not publicly available. Only search companies such as Google, Bing, Yahoo! can access such data. News articles have many constraints due to the natures. For example, it cannot be used to analyze user behaviors. In practical data mining tasks, researchers may exploit multiple data source. For example, various data, such as Tweets, News feeds, survey data, and webpages is used to forecast events in a city (Sect. 3.3.4).

[1] http://realitycommons.media.mit.edu/realitymining.html

[2] www.bls.gov/tus/

[3] https://survey.psrc.org/web/pages/home/

3.2.2 Data Representations

In the literature, spatial, temporal, and semantic information can be represented in various forms.

- **Spatial information.** A pair of latitude and longitude is the most representative form of spatial information of a target location. Other representations include lines (e.g., road segments) and polygons (e.g., universities, parks). At the meantime, there are also a lot of studies that index locations by identifiers, such as venues (point-of-interest, POI), cities, grids, etc.
- **Temporal information.** Temporal information can be represented as either continuous or discrete variables. The continuous representation mostly denote time as a real-value offset (e.g., timestamp) with regard to a specific starting time. For the discrete representation, one can choose the appropriate granularity (e.g., hour, day, week, month) depending on the specific tasks. Sometimes temporal information is implicitly modeled as the order of visits (e.g., trajectory mining in Sect. 3.3.6).
- **Semantic information.** Semantic information can be modeled as a categorical variables or plain text, and it can be associated with locations, users, and visits. For example, the semantics of a user could be her jobs, her hobbies extracted from Facebook profiles. For a location, we can get its semantics like category (e.g., hotel, restaurant, airport), and descriptions. We can also get the semantics of a user's visit at a location, which could be a Yelp review or the text content of a geo-annotated tweet. Some studies go one step further beyond text, and extract additional knowledge, such as named entities, emotions, etc., using text mining or NLP techniques. The usage of semantics depends on both specific tasks and data availability.

Figure 3.2 shows the graph representation of SeST data. In this figure, users, POIs and visits are associated with text, and visits and POIs have time and category as metadata, respectively. Based on the time of visits, we can recover the trajectory of a user (e.g., dashed line in Fig. 3.2 for user u_4). In addition, friend links among users may be available in social media data. Although some datasets may have additional objects, e.g., events, the majority of SeST data can be modeled as subgraphs of the figure.

3.2.3 Approaches

Different studies exploit spatial, temporal and semantic information in different ways, which can be categorized into four approaches.

- Some studies use the three types of information independently, and build models for each of them. Then, the results coming out of different models are combined by certain strategies. For example, in order to recommend POIs to users

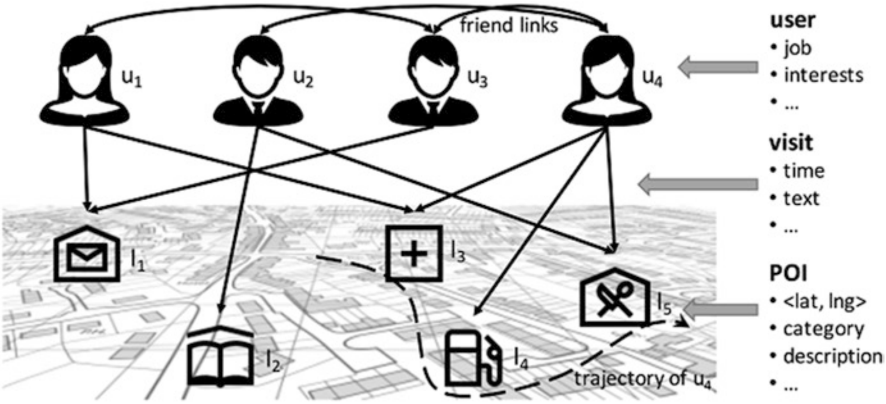


Fig. 3.2 The graph representation of SeST data

- (Section 3.3.2), some researchers build separate models to estimate a target user's spatial, temporal and semantic preference scores of a candidate POI, then these three scores are combined into a final preference score by linear interpolation. Other studies build different classifiers for different types of information and employ co-training strategy to boost the classification performance.
- Some studies extract features from SeST data to train supervised models, such as regression (regression tree, lasso) and classification (support vector machine SVM, Maximum Entropy). This approach is often adopted when a number of features are available, and the target task can be modeled as a supervised or semi-supervised problem, such as quantity prediction (Sect. 3.3.5) and POI typing (Sect. 3.3.8).
- The three types of information are also used as observations for unsupervised models, such as factorization models (tensor factorization, singular value decomposition SVD), graphical models (latent Dirichlet process LDA, hidden Markov model HMM, conditional random field CRF), and graph models (random walk with restart RWR, diffusion model). This approach is often used when the interactions between different information are clear and relatively straightforward to model.
- For some specific tasks, the three types of information are used as optimization constraints or filtering criteria. For example, to plan a trip (Sect. 3.3.7), traveling duration and location category are often modeled as the constraints of an optimization problem. For a second example, many studies on event forecasting do not use the location information in the model. Instead, the location information (e.g., regions, cities) is often used to separate data inside the area of interest from outside.

3.3 Spatiotemporal and Semantic Data Mining Tasks

Many data mining tasks on spatial, temporal, and semantic information have been studied in the literature, the majority of which, however, only exploit at most two dimensions of the three. Recent studies exploit all of the three dimensions, and the integration of space, time, and semantics provides new possibilities of data mining. In this section, we review several most popular SeST data mining tasks, which can be organized as in Fig. 3.3.

3.3.1 Prediction

Prediction aims to infer the candidate dependent variable for a target variable. Under the SeST scenario, the target variable could be user, visit, social media post, etc., while in most studies the dependent variable is location. Representative tasks include next movement prediction for users, home location inference for users, POI inference for geographic coordinates, location estimation for tweets or photos (Hauff and Houben 2012), etc. In most existing studies, the semantic information is the POI category or the text associated with the records. In this section, we take the former two tasks as examples and review existing studies.

Next-place prediction aims at predicting the next place a user is about to visit based on her current location or recent moves, e.g., suppose an office lady just visited a bank branch after work, where is she going to visit next? Next-place prediction is of great importance to user mobility modeling as well as advertisement. Most initial studies only exploit location and time information to construct trajectories, and then predict the next place based off either frequent trajectories of massive people or the user herself. Many recent studies attempt to use semantic information as additional evidence to estimate user mobility preference, where the semantics can be POI categories or text. One thread of works is to extract the frequent trajectory patterns over POI categories. Suppose the pattern *office bank restaurant* is popular in the database, then we can predict that the next place the lady is going to visit is a restaurant. Based on her current location and time, we can select a specific restaurant as the prediction result. Instead of extracting trajectory patterns, we can also infer the transitions between HMM latent states from users' traces for prediction, where each latent state (e.g., office state, home state) defines a semantic topic

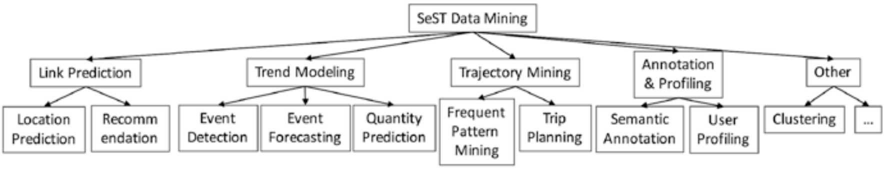


Fig. 3.3 An overview of SeST data mining tasks

(e.g., working), a time range (day-time), and a geographic area (CBD) (Zhang et al. 2016). Another thread is to model the task as a supervised learning problem, in which features like temporal frequencies of categories are extracted to rank candidate POIs by building ranking models.

Home inference is to estimate the home location (e.g., city, state) of a user based on her historical records, such as social media posts, friends, IP address, etc. Home inference is important because users' home information is essential to many tasks such as event detection, personalized recommendation, advertisement, but only few people disclose their home locations on social networks. Pioneering studies exploit social links and user generated text. Suppose a boy has many Facebook friends in New York City, and he has posted a lot about *Net Knicks* and *Bronx*, then he is likely to live in NYC as well. Recent studies use the temporal information to extract the spatiotemporal correlations between text content and locations (Yamaguchi et al. 2014), or extract users' temporal tweeting behavior. For example, if a user posts a tweet right after an earthquake, we can infer that the home of the user should be close to the location of the earthquake. As another example, New Yorkers are more likely to post tweets at 7:00 pm EDT, whereas people who live in Log Angles may tweet less because they are still at work (4:00 pm PDT) in California.

3.3.2 Recommendation

The goal of recommender systems is to suggest new items that a target user might be interested in. While many traditional recommender systems are built on explicit numeric feedbacks (e.g., movie ratings), most recommendation tasks under SeST scenarios deal with implicit binary feedback data, e.g., whether a user will visit a place. Representative tasks include the recommendations of POIs, events, entities (Zhuang et al. 2011), short messages, etc., where the first two tasks are introduced as examples in this section.

POI recommendation aims to suggest unvisited POIs to users based on users' preference. This task has caught a lot of research attention because it can not only help users explore new places but also has great commercial value in advertising. Pioneering studies use spatial and temporal information for recommendation based on the assumptions that users tend to visit their nearby places (Ye et al. 2011b), and a user's preference over POIs is influenced by time (Yuan et al. 2013), e.g., visiting libraries in the morning and bars at night. Some recent works exploit semantic information such as POI categories, check-in text, and reviews to better estimate users' preference implicitly. A straightforward strategy is to recommend the POIs belonging to the categories that the target user visited most. For example, if a user went to many Italian restaurants, then it is safe to continue recommending Italian restaurants to her. We can also infer users' preference transitions over POI categories for recommendation. For example, if a user just visited a restaurant, then we can recommend a theater to her if the pattern *restaurant theater* is frequent in the training set. Rather than using semantics implicitly, we can explicitly take the target user's specific requirements (e.g., *cheesy pizza and spaghetti*) as input, and recommend POIs that best match the target user's semantic profile, mobility behavior and the requirements (Yuan et al. 2015).

Event recommendation aims at recommending local events (e.g., a BBQ party) in event-based social networks (e.g., Meetup[4]) for users to participate in. Initial studies on event recommendation mainly focus on spatial, social and semantic information, based on the assumption that a user tend to participate in events that (1) held close to her, (2) topically attractive to her, and (3) many of her friends also took part in. However, time is also a factor that needs to be taken into consideration because users can only join an event if she is available at that time. To exploit time, several methods (Pham et al. 2015) have been proposed under the frameworks of RWR or ranking models, assuming user tends to attend events held at similar times (e.g., time of a day and day of a week) of the events she attended before.

3.3.3 Event Detection

Event detection is to detect unusual semantic trends that are temporally spiking. Pioneering studies focus on temporal and semantic information to detect global events, e.g., stock market fluctuations and presidential elections, while recent studies exploit the spatial information to detect local events from geo-annotated data, where local event is defined as *something that happens at some specific time and place* (Lee 2012), such as a basketball game or a terrorist attack. Different from the global ones, the local events should be bursty in terms of both location and time. For example, an unlarge number of tweets are talking about *explosion* at Istanbul airport indicate there is an local event terrorist bombing at the airport. The can be detected either by monitoring the changes of spatial and temporal distributions of semantics (Chen and Roy 2009), or comparing the predicted count of tweets generated by regression models with the actual count of tweets for each region (Krumm and Horvitz 2015). Some studies are designed to detect specific types of events such as earthquakes and traffic congestions, in which task-specific evidence is utilized, e.g., the change of massive drivers' routing behavior on road network for traffic anomaly detection (Pan et al. 2013).

3.3.4 Event Forecasting

Event forecasting aims to predict whether an event will happen in the near future. The forecast results make it possible for individuals, organizations and government to prepare for potential crisis in advance. Early studies produce forecasts via either supervised models or time series evolution models that use the temporal and

[4] http://www.meetup.com/

semantic information. The spatial information makes it possible to forecast local events. Existing studies on local event forecasting are domain-specific, i.e., they can detect a specific type of events, such as civil unrest and disease outbreaks. They assume local events can be predicted by monitoring some indicative features, such as keywords counts, tweet cascades, extended vocabulary, etc., extracted from various data sources. For example, if a large portion of tweets in a city are talking about *protest* and *march*, but the portion is small in other cities, there is likely an civil unrest in the city. To forecast whether an event will happen, we can either estimate the development stages (e.g., emerging, uprising, peak, etc.) by monitoring the tweet stream (Zhao et al. 2015), or build regression or classification models on the extracted features.

3.3.5 Quantity Prediction

The availability of SeST information enables us to predict quantity of event or objects based on current observations. Representative tasks include popularity prediction, air quality prediction, traffic volume prediction, etc.

Popularity prediction aims to predict the number of objects adoptions (e.g., hashtags, topics) at specific time in social media, and try to answer the questions like *how many times the hashtag #brexit will be discussed tomorrow in twitter?* Predicting hashtag popularity is important to the identification of commercial and ideological trends. It has been shown that the content of the hashtag (e.g., character length, number of words), the locations mentioned in the tweets, the social network topology (e.g., the number of followers of users who used the hashtag) and the counts of the hashtags in each time intervals are all important features to train a regression model for popularity prediction (Tsur and Rappoport 2012).

Some studies focus on air quality prediction for city regions. This is a challenging task because only a limited number of air quality monitor stations are available in a city, and the air quality depends on various factors, such as meteorology, traffic, land use, etc. This follows our intuition that the air quality in an industry region with high traffic speed is likely to be worse than that in a university. Several supervised or semi supervised models (Zheng et al. 2013) have been developed to predict air-quality based on various spatiotemporal features such as human mobility, traffic speed, the categories of POIs within a region (e.g., factories, parks), etc. Similar strategies are employed to predict the traffic volume in different road segments, where users' activity such as shopping and leisure is an important consideration.

3.3.6 Frequent Pattern Mining

Spatiotemporal frequent pattern mining aims to extract patterns that frequently occur in the given spatiotemporal database. While classic studies on frequent pattern mining in spatiotemporal data can uncover the regularity of people's spatiotemporal movements, the availability of SeST data adds semantics to them and enables us to discover interpretable patterns. Broadly speaking, frequent spatiotemporal patterns in SeST data can be classified into two categories: frequent event patterns, mobility patterns.

Event pattern mining aims to extract frequently co-occurring, cascade, and sequential event patterns from historical SeST data. Co-occurring patterns are events that frequently happen at the same time, e.g., the pattern {*morning, breakfast, at home}* → *{read news}* detected from smartphone context data; cascade patterns are partially ordered subsets of events located together and occurring serially, e.g., the event *bar closing* leads to subsequent event *assault*, and the two together result in *drunk driving* (Mohan et al. 2012); sequential patterns consist of a series events that happen usually in order, e.g., the disease transmission pattern *bird* → *mosquito* → *human being*.

In addition to finding frequent event patterns, researchers have also utilized SeST data to mine frequent movement patterns and tried to understand people's mobility regularity. For example, we can extract sequential patterns over location groups from semantic trajectory databases, where locations in each group are close in distance and consistent in semantic categories (Zhang et al. 2014). For example, a frequent sequential pattern in London could be {House of Parliament, Westminster Abbey} → {*Hyde Park, Regent's Park*}. The former set contains historic sites, while the latter set contains parks. The locations in each set are close to each other.

3.3.7 Trip Planning

The goal of trip planning is to construct a series of locations as travel route for target users. Intuitively, individuals may have limited budget and time for a trip, and different users may have different preferences over places of interests. For example, suppose a girl has only 1 day to travel in London, and she is interested in historic and cultural sites, we should construct a sequence of places within the city, such as *the House of Parliament*, *the British Museum*, *Tower Bridge*, etc., instead of *Wimbledon Tennis Court* or *Windsor Castle* because of they cannot fulfill the girl's interests and the travel time is too long. To incorporate such information, most existing studies (Brilhante et al. 2015) model the trip planning task as an optimization problem by selecting a series of POIs that can meet constraints on time, expense, categories, etc.

3.3.8 Semantic Annotation and Profiling

Semantic annotation aims to infer semantics (categories, descriptions, etc.) for objects (POIs, regions, user visits, trajectories, etc.). Semantic annotation is of great importance to many applications. For example, about 30% of POIs in Foursquare

are lacking any meaningful textual descriptions (Ye et al. 2011a). Annotating these POIs with categories can facilitate both place exploration for users and recommendation services for businesses.

To archive semantic annotation, it is important to exploit spatial, temporal and semantic jointly. To take POI annotation as an example, it is observed that the categories of POIs visited by the same user at the same time are similar. In addition, visitors' demographic information (e.g., age and gender) and the surrounding business are both good indicators of the POI category: a student may stop by a restaurant for lunch at noon because of a break between two classes, and the restaurant is close to other restaurants and grocery stores. Classification models are effective in combining spatiotemporal and semantic features for category estimation. Similarly, the function of a region can be inferred from various SeST evidence such as human mobility and POI categories (Yuan et al. 2012). There are also studies on visit annotation, which uncover the visiting purpose by both static and dynamic features: on the one hand, the visiting purpose is related to the static features such as POI category and region's land use that are invariant to time; on the other hand, the purpose is also influenced by dynamic local events, such as sport games or festival celebrations (Wu et al. 2015). Consider a man who visits Oracle Arena in Oakland on June 19 2016. Then the purpose of his visit might be watching NBA Finals Game 7. Recent studies use static and dynamic features to infer the visiting purpose and achieve satisfactory performance.

Profiling is to characterize entities by spatiotemporal or semantic data. Several papers are published to profile users, POIs and words. Among them, user profiling received the most research interest. Different methods profile users from different aspects, such as individuals' frequent routines (Farrahi and Gatica-Perez 2011), spatiotemporal mobility and topic preference such as movie, hiking (Yuan et al. 2015), and demographic information such as age, gender, marital status (Zhong et al. 2015). It has been shown that the spatial, temporal and semantic information of users' visiting records are all important evidence for profiling.

3.3.9 Clustering

Clustering, which aims to group objects such that objects in the same group are more similar to each other than those in other groups, is a fundamental task in data mining. The availability of spatial, temporal and semantic information enables us to better estimate the relatedness between objects and form clusters. The detected clusters can not only provide a high-level summary of the whole data, but also are important to a number of tasks, such as community detection, frequent pattern mining, recommendation, event detection, next-place prediction, etc. Several methods have been proposed to cluster objects such as tags, hashtags, tweets, users, trajectories, etc. For different objects, the SeST information is used in different ways. For example, to cluster tags or hashtags, the co-occurrence between two objects is an important measure. In other cases, however, the co-occurrence itself is not enough to estimate the relatedness. For example, in Flickr photos, the tags *The Statue of Liberty* and *Times Square* are seldom used together, but the two tags are highly correlated because both of them refer to two famous landmarks in New York City. Thus, in addition to co-occurrence, spatial and temporal features are extracted to cluster tags, based on the assumption that related concepts should have similar spatial and temporal distributions (Zhang et al. 2012). For another example, trajectories can be clustered not only based on locations and visiting orders, but also based on the semantics, e.g., location categories. Now, we can identify groups of objects such as individuals and taxi drivers based on the behaviorally-driven markers of individual and collective movement (Liu et al. 2013).

3.4 Future Directions

Although mining SeST data has been gaining much research attention in recent years, many remaining challenging issues call for new and effective solutions. We outlook some important directions for future research in mining SeST data.

- **Deeper understanding of semantics.** While various techniques have been proposed to incorporate semantics information into the process of mining SeST for useful knowledge, the modeling of the semantics is still built upon simple models e.g., bag of keywords. More accurate methods, e.g., phrase mining (Liu et al. 2015), named entity recognition and typing (Ren et al. 2015), sentiment analysis, etc., are necessary so as to capture the intrinsic semantics more accurately.
- **Managing and integrating multiple data sources.** Current research for mining SeST data mostly consider only one data source. It is interesting and important to integrate the data from different sources (e.g., social media, sensor data) to extract valuable evidence in various aspects.
- **Interactive exploration of SeST data.** In many real-life applications, it is not easy to determine the data mining techniques and model parameters before-hand. Extracting the most useful knowledge from the given SeST data usually involves extensive model and parameter tuning. Therefore, it has become an urgent need to develop techniques that can support interactive exploration of SeST data.
- **Assisting decision making.** How to discover knowledge from SeST data to aid decision making is a promising direction. For example, the semantic enriched mobility of massive people is at great importance to urban planning such as site selection for a new airport. In addition, how to generate interpretable and explorable knowledge is also a critical problem to facilitate decision making processes.

3.5 Summary

With the prevalence of GPS technology and the development of social networks, a sheer amount of SeST data has been accumulated. It involves additional types of information compared with traditional datasets which have up to two dimensions among location, time, and semantics. The multi-dimensional SeST data bring new opportunities along with new challenges to extract knowledge. In this article, we introduced the major challenges, data sources, information representation, and general mining approaches to SeST data mining. We also reviewed nine important tasks and cutting edge studies. Some promising directions for future work were also discussed. To the best of our knowledge, this is the first survey that focuses on summarizing existing techniques for mining SeST data. We hope this article can provide readers with a high-level and systematic overview of the research on SeST data mining.

References

Brilhante IR, Macedo JA, Nardini FM, Perego R, Renso C (2015) On planning sightseeing tours with tripbuilder. Inf Process Manag 51(2):1–15

Chen L, Roy A (2009) Event detection from flickr data through wavelet-based spatial analysis. In: Proceedings of 18th ACM international conference on information and knowledge management, ACM, pp 523–532

Cheng T, Haworth J, Anbaroglu B, Tanaksaranond G, Wang J (2014) Spatiotemporal data mining. In: Fischer MM, Nijkamp P (eds) Handbook of regional science. Springer, Heidelberg, pp 1173–1193

Farrahi K, Gatica-Perez D (2011) Discovering routines from large-scale human locations using probabilistic topic models. ACM Tran on Intell Syst Technol (TIST) 2(1):3

Hauff C, Houben GJ (2012) Placing images on the world map: a microblog-based enrichment approach. In: Proceedings of 35th ACM SIGIR international conference on research and development in information retrieval. ACM, pp 691–700

Kelly R (2009) Twitter study reveals interesting results about usage 40% is “pointless babble”. http://www.pearanalytics.com/blog/2009/twitter-study-reveals-interesting-results- 40- percent-pointless-babble/. Retrieved: January 18, 2012

Krumm J, Horvitz E (2015) Eyewitness: identifying local events via space-time signals in twitter feeds. In: Proceedings of 23rd ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, pp 20:1–20:10

Lee CH (2012) Mining spatio-temporal information on microblogging streams using a density-based online clustering method. Expert Syst Appl 39(10):9623–9641

Liu S, Wang S, Jayarajah K, Misra A, Krishnan R (2013) Todmis: mining communities from trajectories. In: Proceedings of 22nd ACM international conference on information and knowledge management. ACM, pp 2109–2118

Liu J, Shang J, Wang C, Ren X, Han J (2015) Mining quality phrases from massive text corpora. In: Proceedings of 2015 ACM SIGMOD international conference on management of data. ACM, pp 1729–1744

Mohan P, Shekhar S, Shine JA, Rogers JP (2012) Cascading spatio-temporal pattern discovery. IEEE Trans Knowl Data Eng (TKDE) 24(11):1977–1992

Pan B, Zheng Y, Wilkie D, Shahabi C (2013) Crowd sensing of traffic anomalies based on human mobility and social media. In: Proceedings of 21st ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, pp 344–353

Pham TAN, Li X, Cong G, Zhang Z (2015) A general graph-based model for recommendation in event-based social networks. In: Proceedings of 31st IEEE international conference on data engineering (ICDE). IEEE, pp 567–578

Ren X, El-Kishky A, Wang C, Tao F, Voss CR, Han J (2015) Clustype: effective entity recognition and typing by relation phrase-based clustering. In: Proceedings of 21st ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 995–1004

Shekhar S, Jiang Z, Ali RY, Eftelioglu E, Tang X, Gunturi V, Zhou X (2015) Spatiotemporal data mining: a computational perspective. ISPRS Int J Geo-Inf 4(4):2306–2338

Tsur O, Rappoport A (2012) What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In: Proceedings of 5th ACM international conference on web search and data mining. ACM, pp 643–652

Wu F, Li Z, Lee WC, Wang H, Huang Z (2015) Semantic annotaion of mobility data using social media. In: Proceedings of 24th international conference on world wide web. International World Wide Web Conference Steering Committee, pp 1253–1263

Yamaguchi Y, Amagasa T, Kitagawa H, Ikawa Y (2014) Online user location inference exploiting spatiotemporal correlations in social streams. In: Proceedings of 23rd ACM international conference on information and knowledge management. ACM, pp 1139–1148

Ye M, Shou D, Lee WC, Yin P, Janowicz K (2011a) On the semantic annotation of places in location-based social networks. In: Proceedings of 17th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 520–528

Ye M, Yin P, Lee WC, Lee DL (2011b) Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of 34th ACM SIGIR international conference on research and development in information retrieval. ACM, pp 325–334

Yuan J, Zheng Y, Xie X (2012) Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 186–194

Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013) Time-aware point-of-interest recommendation. In: Proceedings of 36th ACM SIGIR international conference on research and development in information retrieval. ACM, pp 363–372

Yuan Q, Cong G, Zhao K, Ma Z, Sun A (2015) Who, where, when, and what: a nonparametric bayesian approach to context-aware recommendation and search for twitter users. ACM Trans Inf Syst (TOIS) 33(1):2

Zhang H, Korayem M, You E, Crandall DJ (2012) Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities. In: Proceedings of 5th ACM international conference on web search and data mining. ACM, pp 33–42

Zhang C, Han J, Shou L, Lu J, La Porta T (2014) Splitter: mining fine-grained sequential patterns in semantic trajectories. Proc VLDB Endowment 7(9):769–780

Zhang C, Zhang K, Yuan Q, Zhang L, Hanratty T, Han J (2016) Gmove: group-level mobility modeling using geo-tagged social media. In: Proceedings of 22ed ACM SIGKDD international conference on knowledge discovery and data mining. ACM

Zhao L, Chen F, Lu CT, Ramakrishnan N (2015) Spatiotemporal event forecasting in social media. In: Proceedings of 15th SIAM international conference on data mining. SIAM, pp 963–971

Zheng Y, Liu F, Hsieh HP (2013) U-air: when urban air quality inference meets big data. In: Proceedings of 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1436–1444

Zhong Y, Yuan NJ, Zhong W, Zhang F, Xie X (2015) You are where you go: inferring demographic attributes from location check-ins. In: Proceedings of 8th ACM international conference on web search and data mining. ACM, pp 295–304

Zhuang J, Mei T, Hoi SC, Xu YQ, Li S (2011) When recommendation meets mobile: contextual and personalized recommendation on the go. In: Proceedings of 13th international conference on ubiquitous computing. ACM, pp 153–162

Chapter 4
Contribution Towards Smart Cities: Exploring Block Level Census Data for the Characterization of Change in Lisbon

Fernando José Ferreira Lucas Bação, Roberto Henriques, and Jorge Antunes

Abstract The interest in using information to improve the quality of living in large urban areas and the efficiency of its governance has been around for decades. Nevertheless, recent developments in information and communications technology have sparked new ideas in academic research, all of which are usually grouped under the umbrella term of Smart Cities. The concept of Smart City can be defined as cities that are lived, managed and developed in an information-saturated environment. However, there are still several significant challenges that need to be tackled before we can realize this vision. In this study we aim at providing a small contribution in this direction, by maximizing the usefulness of the already available information resources. One of the most detailed and geographically relevant information resources available for studying cities is the census, more specifically, the data available at block level. In this study we use self-organizing maps (SOM) to explore the block level data included in the 2001 and 2011 Portuguese censuses for the city of Lisbon. We focus on measuring change, proposing new ways to compare the two time periods, which have two different underlying geographical bases. We proceed with the analysis of the data using different SOM variants, aiming at providing a twofold portrait: showing how Lisbon evolved during the first decade of the twenty-first century and how both the census dataset and the SOMs can be used to produce an informational framework for micro analysis of urban contexts.

Keywords SOM • Clustering • Geo-demographics • Census data • Smart cities

F.J.F.L. Bação (✉) • R. Henriques • J. Antunes
Nova Information Management School, Universidade Nova de Lisboa,
Campus de Campolide, 1070-312 Lisbon, Portugal
e-mail: bacao@novaims.unl.pt; roberto@novaims.unl.pt

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_4

4.1 Introduction

The challenges posed by cities and the exploration of urban trends have both received significant attention with regards to sustainable development (Braulio-Gonzalo et al. 2015; Huang et al. 2016), namely socio-economic development, environmental management and urban governance (United Nations 2013). In order to attain these goals, it is necessary to tackle three major issues: the increasing economic division between rich and poor, climate change and the efficient management of public goods (Birch and Wachter 2011).

In this challenging environment, the concept of Smart Cities emerges referring to the opportunity of developing new strategies to tackle the problems that the urban society faces every day (Roche 2014). Although the cities' very complexity has led to a non-agreement as to what exactly defines a "Smart City" (Lombardi et al. 2012), it is possible to identify a major set of activities focused on the implementation of technology and strategies aimed to improve the city itself. Current research present essentially two main streams of thought: a) the solutions based in information technologies that fully rely on IoT (Internet of Things) and IoS (Internet of Services) as enablers of smart cities that use an unified ICT (Information and Communications Technology) platform (Hernández-Muñoz et al. 2011) and b) more systemic solutions that rely on management, organization, technology, governance, policy, the people and the communities, economy, infrastructure and the natural environment. The last one sees the city as organic and complex systems with multiple and diverse stakeholders, high levels of interdependence, competing objectives and values and social and political complexity (Chourabi et al. 2012).

The urban fabric can be very complex and hard to understand. The inclusion of attributes such as population, infrastructures, social-economic environment, among others, arriving in constant streams, reveal the inherit complexity of this task (Lee and Rinner 2014). The increased sophistication of the Geographic Information Systems (GIS) can contribute to mitigate some of the aforementioned difficulties. The georeferencing of non-spatial data, particularly high-dimension data, allows us to visualize the underlying context. These links can manifest as a cartographic representation of multivariable groups depending on the information one desires to attain. The nongeographic information is processed by computational tools and the results are expressed in maps, where it is easier to interpret the output (Koua and Kraak 2004; Penn 2005; Skupin and Hagelman 2005; Skupin 2002).

To a human being, the visualization of several dimensions of data at once represents a problem, since it is not possible to apprehend a large number of dimensions in an interpretable way (Bação et al. 2004). It is therefore necessary to implement techniques that improve the perception of high dimensional data. Dimension reduction aims to decrease the data parameters to the bare minimum needed to explain the data properties, also known as intrinsic dimensionality (Fukunaga 1990). As a result, it facilitates classification, visualization and compression of high-dimensional data, among others (van der Maaten et al. 2009).

Lisbon makes for an excellent case study due to the intense modifications that occurred during the last decades and particularly throughout the most recent years (Silva and Syrett 2006; Veiga 2014).

The goal in this work is to characterize the population and the residential infrastructure through the use of clustering techniques. In order to support the future implementation of the Smart City, the city portrait includes the targeting of the most affluent areas, where the use of the most recent IT tools and services have the highest probability of acceptance and diffusion. Thus, it is also possible to identify more deprived areas where people and infrastructures are unprepared to deal with future uses of the Internet.

In order to improve the desired portrait, the modifiable areal unit problem must be mitigated during the process. It is imperative to successfully tackle the land use by reallocating the population from a statistical unit to what is the true urban tissue if we want to characterize the area at block level.

Looking for a reliable and efficient solution in the present study, Self-Organizing Maps were selected to produce an informational framework, where the inclusion of other sets of attributes won't be a challenge, but instead an opportunity to evolve towards more specific or broad subjects.

At the same time, spatiotemporal analysis between 2001 and 2011 was performed once pattern and trends identification enabled the understanding of environmental phenomena and a better insight of socio-economic behaviors.

4.2 Context of Research

The concentration of people, investment and resources turn cities into hubs of economic development, innovation and social interaction (Longo et al. 2005; Polèse 2010). This environment brought forth the concept of "Smart Cities", bringing with itself a whole new perspective. The term "Smart" means having or showing a quick-witted intelligence, which is why in some literature it is possible to find the term "intelligent city", or even "digital city", the latter being a more concrete way to define how it operates.

There are several definitions of "Smart Cities" (Nam and Pardo 2011), but generally speaking, they all refer to the opportunity to develop strategies to face the challenges that urban society encounters every day (Roche 2014). The main components of a Smart City were concatenated by Nam and Pardo (2011) in technology, institutional and human factors, fetching a large set of attributes, that can be used to transform a city into a smart, sustainable city.

The use of spatial features attached to non-spatial attributes such as economic, social and demographic data brings new opportunities to develop methods to understand environmental phenomena and socio-economic behaviors (Bação et al. 2005c).

Spatiotemporal data poses serious challenges to analysts. The number of distinct places can be too large, the time period under analysis can be too long and/or the attributes depending on space and time might be too numerous. Therefore, human analysts require proper support from computational methods able to deal with large, multidimensional data (Andrienko et al. 2008, 2010; Andrienko and Andrienko 2006).

In that way, it's necessary to have an observational data source based on valid, reliable, timely, useful, accessible and cost-effective information criteria (Howard et al. 2011).

Looking for data sources that enable the development of methods to understand environmental phenomena and socio-economic behaviors and work as a means of construct an informational framework, the censuses have the required features. The attributes applied (infrastructure, demography, morphology and economic), as well as the granularity (Skupin and Agarwal 2008), cover every need, from national to block level, which enables every sort of analysis (Skupin and Hagelman 2005). The census enumeration units assemble the collected information from the inhabitants from predefined areas. Similar to zone design, the n areal units are aggregated into k zones (Bação et al. 2005a).

For legal and confidentiality reasons, the censuses' specific information cannot be publicly released (Nelson and Brewer 2015), and thus the data must be aggregated in a way that is tractable enough to provide the necessary outputs (Openshaw 1984). Knowing that the resolutions taken affect the reliability of the conclusions reached (Fiedler et al. 2006), it is necessary to avoid inferences that could lead to ecological fallacies once the data we are dealing with is a really aggregated (Nelson and Brewer 2015; Root 2012).

The last sentence refers to the data aggregation that is scale-dependent and obtained from smaller areas (scale effect) (Jelinski and Wu 1996). The use of the land cover data sources enables the mitigation of the previous stated limitation in order to partially tackle the aforementioned MAUP. For that, it is necessary to cross-reference both data sets and a lot of research has been performed particularly on Areal Interpolation, which is formally defined as the process of transferring spatial data from one set of units to another (Bloom et al. 1996; Fisher and Langford 1996). Taking into account the biased aggregated data that was presented, an effort was done to solve the MAUP, albeit unsuccessfully (Manley 2014).

The census tracts are a well-known data source, which includes all types of land use. However, the variables relevant to the present study are the ones that are located in the urban tissue. The cross-referencing of variables, particularly population characteristics, infrastructure and land use, has been explored to examine different subjects such as urban planning, public health and transportation (Wier et al. 2009; Waddell 2007).

The use of location grids enables the identification of different land uses at lower levels, but no major cross-referencing has been performed with the census tracts to reallocate the population at block level. The method that was employed that most closely resembles that cross-referencing, which is also the one used in the present study, is the one performed by Eicher and Brewer (2001), the Polygon Binary Method, where the data was distributed between inhabitable and uninhabitable areas at the county scale.

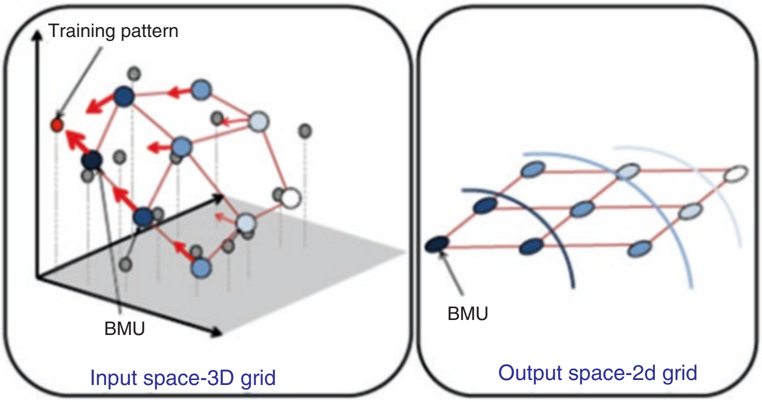


Fig. 4.1 SOM training phase. A training pattern is presented to the network and the closest unit is selected (Best Matching Unit – *BMU*). Depending on the leaning rate, this unit moves towards the input pattern. Based on the BMU and on the neighborhood function, neighbors are selected on the output space. Neighbors are also updated towards the input pattern (Henriques et al. 2012)

Looking for the best way to explore data, the clustering techniques show several advantages, namely the ability to group similar objects in one cluster and dissimilar ones in another. This enables correct labeling and permits acting accordingly in relation to their features and characteristics. The objects are automatically assigned to a cluster, meaning that every item can be described by its representative. These techniques are often scalable and easy to handle, even with different attribute types.

This path takes us to a method that used in intensive computation and applied to large data sets, the Kohonen Map or Self-Organizing Map (SOM) (Kohonen 2013), which performs vector quantization and projection (Kohonen 2013). The output is a two-dimensional grid, easy to visualize and understand. The use of a two-dimensional surface mitigates the human inability to comprehend the visualization of multidimensional data (Bação et al. 2005c; Koua 2003).

Using this technique, it is possible to perform a clustering operation by aggregating the items or objects to the nearest neuron or unit as their representative (Jain et al. 1999; Mennis and Guo 2009) (Fig. 4.1).

4.2.1 Self-Organizing Map

The SOM is an unsupervised Artificial Neural Network, a clustering technique based on the classical vector quantization. The idea is to simulate the brain maps where it is possible to ascertain that certain single neural cells in the brain respond selectively to some specific sensory stimuli (Kohonen 2013). SOM displays high-dimensional data in generally two dimensions, preserving the relations already existent in the input source, meaning that the components that are the most similar will appear near to each other and far away from the ones where it is identified a bigger dissimilitude. The SOM algorithm performs a number of iterations in order to represent the input patterns by reference vectors as best as possible. The movements can be visualized as exemplified in Fig. 4.3:

This way, we can ascertain the Best Matching Unit as well as its closest neighbors. The SOM algorithm is described in Kohonen (2013).

4.2.2 SOM Views

The SOM's visualization is possible through a series of methods like component planes, distortion patterns, clustering and U-matrix (Skupin and Agarwal 2008). A major set of representations can also be found and used (Vesanto 1999).

The input and output space perspectives are directly connected once the output space tries to preserve the topology of the input space (Gorricha and Lobo 2012; Kohonen 2013; Bação et al. 2005b).

As referred previously, there are other methods to visualize the SOM. The U-Matrix is a solution that shows the clusters and it is the most used method to perform the SOM visualization (Ultsch and Siemon 1990). The U-matrix uses colors in order to identify the neighbor's distance. Closest units use lighter colors while darker tones are used for units farther away (Kohonen 1995).

The other option is to use component planes to view each variable individually. Using the same locations units as in the U-matrix, it is possible to analyze the distribution of each variable based on the clustering result (Hajek et al. 2014).

The results were attained through a combination of this technique with geovisualization (Henriques et al. 2012; Koua 2003).

4.3 Data and Software

"The population and housing census represents one of the pillars for the data collection on the number and characteristics of the population of a country" (United Nations Economic Commission for Europe 2006). Based on the quality of this data source, this study focuses on the information extracted from the 2001 and 2011 censuses. The data was collected from Instituto Nacional de Estatística (INE), Portugal's national statistics institute (censos.ine.pt).

As referred previously, to georeference the inquiries' datasets, cartography is essential to determinate the global position. In order to fulfill this requirement, the authors used the Portuguese "Based Geographical Referencing of Information" (BGRI) (mapas.ine.pt). The statistical territorial unit used is the Subsection Stats, which represent the smallest area unit, increasing the available study resolution.

The main attributes used in this study are Building, Family Accommodation, Classic Family and Resident Individual. From these, the censuses provided us with an enormous subset of variables, obtained directly from the census or by processing.

The dataset refers to the Lisbon municipality and includes 3623 Enumeration Districts and 122 original variables in 2011 that are able to characterize social, demographic and economic stats in the applied areas. Related to 2001, there are 4390 Enumeration Districts and 99 original variables. As can be perceived, the basis of the census has changed as a result of spatiotemporal mutations over the selected areas. In order to accomplish consistent results, the influence of different population and housing sizes was reduced, and a ratio that increases analysis effectiveness and reduces description clutter was also included. Implementing relations between variables is a part of good scientific practice (Fink 2009), and so is the ability to avoid the high dimensionality curse (Donoho 2000). The vector dimension was reduced to the bare minimum of attributes that could be used to explain the results and to decrease the occurrence of spurious relations.

In order to better understand the city, we must be able to properly map the data extracted from the census. The land cover information is extremely useful in these cases. There are two main available datasets, the CORINE Land Cover and the Portuguese COS2007. The source used in the present study was the COS2007 level 2, since it has a smaller mapping unit (1 ha vs 25 ha) making for a more precise reading of the maps**.**

In order to better allocate the population and the buildings, the target areas are the ones, which represent urban tissue. Based on recent years, the Lisbon urban tissue didn't suffer significant changes from 2007 to 2011, when the census was performed. However, the present research process always depicts the relation between the BGRI2011 and COS2007.

We used the GeoSOM suite tool, which allows for a number of operations such as training self-organizing maps that apply the standard SOM algorithm and produce several representations of the input and output data. This tool is implemented in Matlab and uses the public domain SOM toolbox (Henriques et al. 2012).

4.4 Methodology

As referred previously, the study area is the Lisbon municipality data, which was extracted from public institutional websites.

As explained, the census subsections have changed from 2001 to 2011, in order to improve the quality of the information obtained from the inquiry. One of the goals of this study is to address the mutations that occurred Lisbon municipality (whose borders haven't changed) during that decade.

First, we must merge the COS2007 with the BGRI to obtain a real distribution of the population and buildings (urban fabric), excluding all the other areas. The process of getting a final dataset to be used as input in the SOM algorithm is obtained through a sequential process of intersection among BGRI and COS2007 on a N2 and a 50×50 m grid resulting in what can be seen in Fig. 4.2:

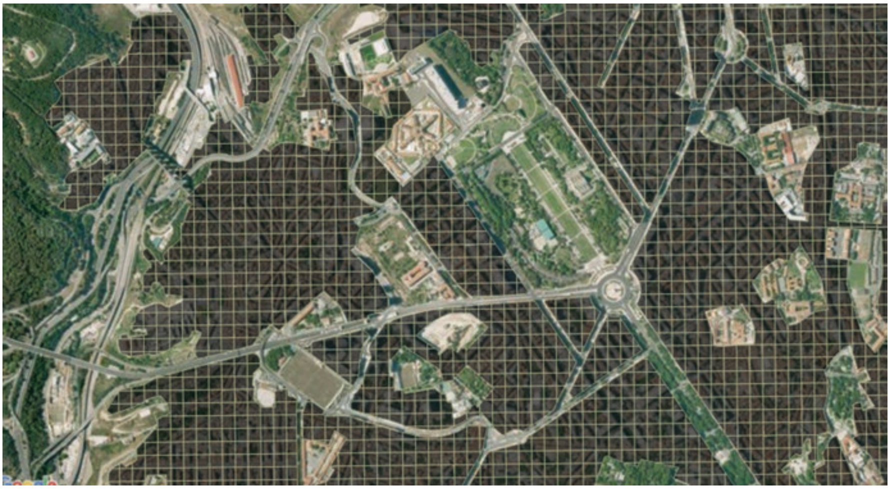


Fig. 4.2 Grid result after intersection process

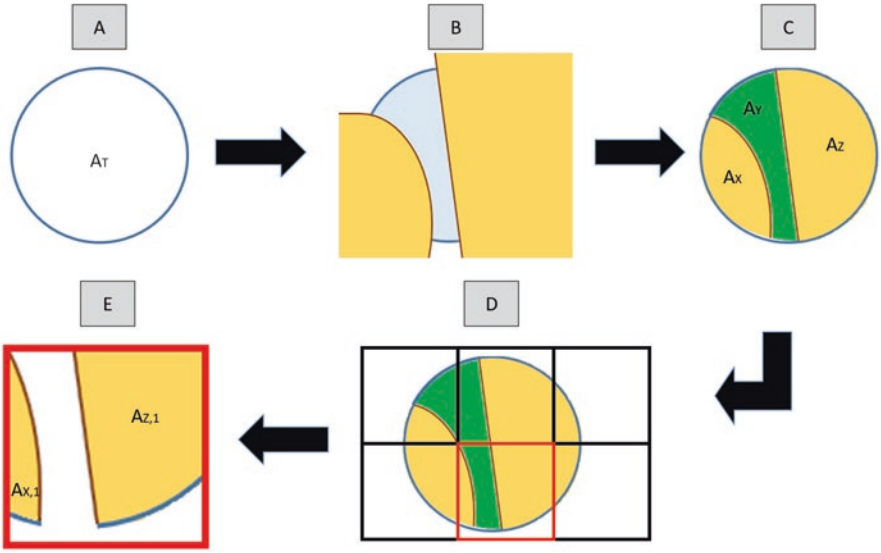


Fig. 4.3 Intersection methodology process

The distribution of the attribute values among the generated pixels was performed recurring to proportional areas. Below, it is possible to see how this process is enacted.

Figure 4.3 shows the process involved in fitting the BGRI relative areas to a single pixel. For example, imagine we use the individual residents to create a new distribution on the final grid. Figure 4.5a shows a subsection BGRI polygon (SBP) with the Area (A_T) and number of Individuals (I_T). Crossing the SBP and the urban fabric, as shown in Fig. 4.5b (from COS2007N2), results in what is presented in Fig. 4.5c. A green area (A_Y), representing a park (no population allocated), is created along with two new populated areas (A_X and A_Z).

The areas' percentages are as follow:

$$A_X = 25\%;\ A_Y = 25\%;\ A_Z = 50\%$$

This means that the new considered area is:

$$\text{Area}_{\text{new}} = A_X + A_Z = A_T \times 0.25 + A_T \times 0.50$$

Since the **EDs** from 2001 and 2011 don't have the same spatial delimitation, we created a grid over the urban fabric to map the changes occurred in that period (Fig. 4.5d).

Now, we have a new Area that will be distributed by the pixels. Each pixel will include the areas of one or more SBPs. In this case, the selected Pixel includes two Areas ($A_{X,1}$ and $A_{Z,1}$) that represent the proportion inside the new SBP pixel.

Let's assume that the new SBP area's percentages are as follows:

- $A_{X,1} = 5\% \Rightarrow 0.05 \times \text{Area}_{\text{new}}$
- $A_{Z,1} = 20\% \Rightarrow 0.2 \times \text{Area}_{\text{new}}$

Then the population allocated to this red Pixel is the following:

$$\text{Ind}_{\text{pixel}} = \sum_{\text{BGRI}=j}^{m} \sum_{\text{Area}=i}^{n} I_T \times \frac{\text{Area}_i}{\text{Area}_{\text{newSBP}}}$$

The issues that rise from this data transformation are already known as MAUP. Using the example provided again, the population would be better allocated in 75% of the total area than in 100%, although the final distribution by each pixel assumes that the attribute values spreading is uniform.

4.4.1 SOM

The resulting shapefile, with the attributes previously described in the Data and Software section, was imported to the Geo-SOM Suite, where the configuration algorithm processing take place. The parameters were chosen with the intent to use the best characteristics and features of the Kohonen map that fit the object of our study. Deciding on the map size was ultimately based on the results obtained in the study performed by Bação et al. (2005a). Selecting a wider map allows for a better distribution of units per neurons and increases the interpretability of the grid distances in order to identify the clusters present. Nevertheless, an extremely large map is not without its drawbacks, since what we gain in reliability we lose in interpretability.

Based on these premises, a grid of 20 × 15 was used, representing around 10% of the total analyzed population **(of subsections).** The final network architecture that run on SOM was hexagonal based on the advantages referred by Kohonen (1995). The input data was standardized and the learning process occurred under a Gaussian neighborhood function. The resulting U-Matrix has a large number of neurons, and therefore, makes it extremely difficult to understand which ones are the general characteristics applied to each cluster. The interpretability problem was solved by the use of hierarchical clustering (HC), a known clustering technique, in the process flow, in order to define the final clusters.

The hierarchical clustering is depicted in trees or dendrograms, nesting, but not deterministic partitions, once no random initial conditions are given, except for the method itself. The hierarchical clustering method selected was Ward. This method measures the distance between two clusters (or units), searching for the increasing sum of squares, also known as The Merging Cost. The clustering technique applied had been studied by the academia for a while (Jain et al. 1999; Steinbach et al. 2000). The demonstrated advantages fit our purpose, resulting in a more reliable way to aggregate the units obtained from the neural network. Finally, through an average of all the corresponding input grid elements, we are given the resulting cluster, which is then mapped and used to perform spatial-temporal comparisons.

In order to perform an accurate comparison between the two datasets (2001 and 2011), we've classified the 2001 elements based on the resulting hierarchical clusters obtained from the 2011 study process.

The classification was performed through an Euclidean distance calculation among the 2001 elements and the obtained units (neurons) from the 2011 dataset. In a way, it is possible to say that the 2001 element looks for the Best Matching Unit without changing the network. The remaining parameters like location update and learning rate are part of the initial conditions set by the analyst that were tuned during the iterative process.

4.5 Results

The resulting U-Matrix (20 × 15) (Fig. 4.5b) reveals the existence of some clusters with different sizes. It was also possible to locate the Best Matching Units with the elements present in the Lisbon municipality urban selection map. For example, after the selection of six units (red hexagons Fig. 4.4a), they also appear selected in red on the map (Fig. 4.4b).

However, the 300 units obtained with the SOM represent a wide variety of clusters. Hierarchical clustering was performed as merging technique to obtain the higher cluster level. The decision to cut at the "height = 11" parameter was based on the ability to perform enough clustering to get the most interpretability without losing the "special" differentiating characteristics each cluster should have. The result-

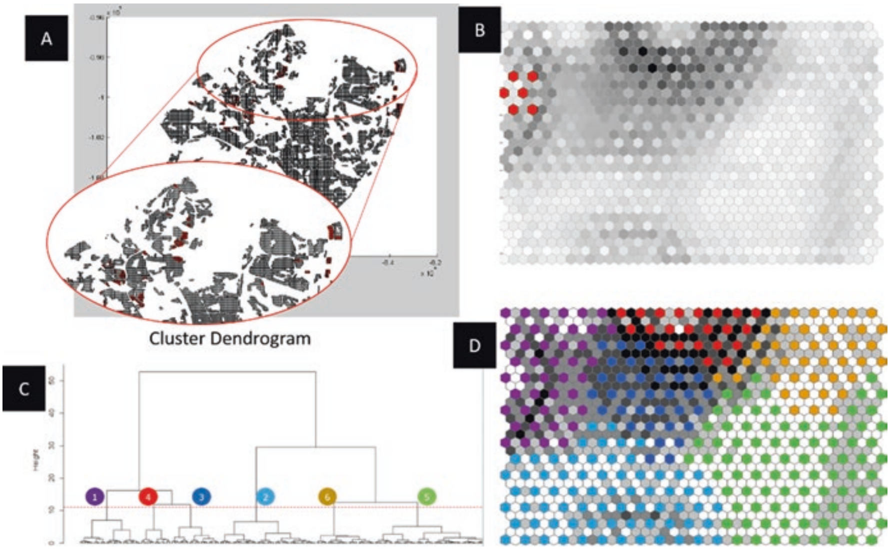


Fig. 4.4 (**a**) Map output view; (**b**) SOM classical U-matrix; (**c**) cluster dendrogram; (**d**) SOM HC U-matrix

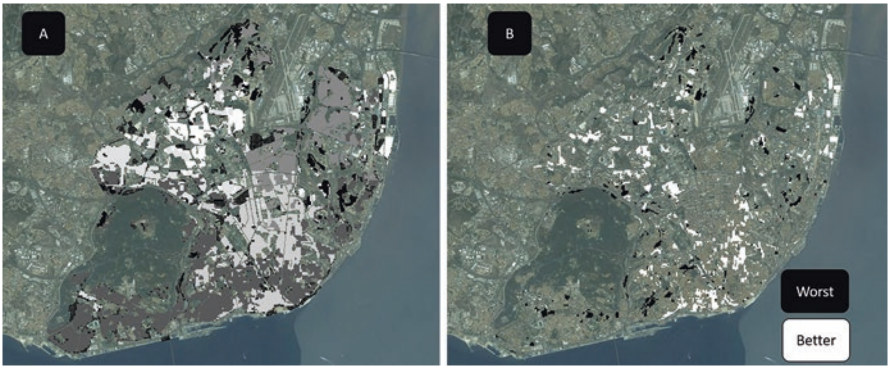


Fig. 4.5 (**a**) SOM hierarchical mapping; (**b**) comparison of 2001 and 2011 choropleths

ing U-Matrix (Fig. 4.4d), with the new six grouped clusters, is presented with the same matching colors/numbers in both Fig. 4.4c, d.

The centroid was calculated for every HC in order to understand which cluster belongs to which element (Table 4.1).

The choropleth map (Fig. 4.5a) shows the 2011 hierarchical clustering distribution display how the population is organized and structured in the Lisbon municipality. The reason why we apply hierarchical clustering over the BMUs instead of the items themselves is because HC doesn't produce satisfactory results in big datasets and the final dendrogram isn't easy to understand due to the huge amount of height links (Guha et al. 1998). The clusters geolocation is heterogeneous, meaning

Table 4.1 SOM hierarchical cluster description

Cluster (name/number)	Description
Upper class TB – 1	New, tall buildings
	Largely occupied by the owners themselves (i.e. not rented)
	Highly educated population with a low unemployment rate
	Population usually composed of young, active people
Upper class SB – 3	Older Buildings, mainly occupied by the owners
	Families tend to be larger (more than 4 members)
	Highly educated population
Middle class TB – 2	Older, taller Buildings
	Families tend to be smaller
	Population tends to be highly educated, but also older
Middle class SB – 5	Older, smaller Buildings
	Families tend to be smaller
	Population tends to have an average education
	Older population (low density)
Middle lower class SB – 6	Older, smaller Buildings, with a high rent rate
	Population tend to have a lower education level with a high unemployment rate
	Average population age (low density)
Lower class – 4	High density of population and **accommodations** in newer buildings, with the highest rent rate
	Large families
	Lower education level with the highest unemployment rate
	Youngest population of all

The cluster mapping was performed for both the 2001 and 2011 datasets
SB Short Buildings (less than four floors), *TB* Tall Buildings (five or more floors)

that urban fabric doesn't follow a precise pattern when growing, especially when compared with the non-geographic features.

Figure 4.5b aims to show how the city mutated over the decade (2001–2011) and shows that in fact, the historic city center is getting clear improvement spots, represented in white.

4.6 Conclusions

In this paper, we focus our efforts in analyzing an available data resource, which possesses high quality standards, the census. It was possible to link the data mining techniques with the Smart Cities concept.

The census data revealed itself an excellent source of information to assess the environmental phenomena and socio-economic behaviors that depict the city. We used exploratory spatial data analysis and clustering techniques that successfully gave a deeper insight into the Lisbon municipality.

The MAUP problem was addressed and a new methodology that mitigates errors in population and building assignment was presented. This new methodology crosses the land cover use data with the census tracts to increase the reliability of the areal aggregated data.

The use of Unsupervised Neural Networks and SOM allowed for a broader analysis of the change of the spatial situation over time by implementing a new geographic base that tackled the statistical territorial unit mutation.

The spatiotemporal change (2001–2011) was pictured through SOM and a new informational framework, one able to receive additional data from different sources, was created. This portrait of the population and residence infrastructure was achieved through map grids. However, to tackle the human inability to deal with large amounts of data, we executed a hierarchical clustering and generalized the applied areas effectively.

Looking to particular areas in the paper, the historic city center and Parque das Nações, it was possible to identify patterns and trends. Parque das Nações assumes its position as an upper class area. On the other hand, the historic city center shows an increased gentrification, since a change from lower level clusters to higher ones is clearly identifiable.

References

Andrienko NV, Andrienko G (2006) In: Springer (ed) Exploratory analysis of spatial and temporal data: a systematic approach. Science & Business Media, London

Andrienko G et al (2008) Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. Inf Vis 7(3–4):173–180

Andrienko G et al (2010) Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. Comput Graph Forum 29(3):913–922

Bação F, Lobo V Painho M (2004) Geo-self-organizing map (Geo-SOM) for building and exploring homogeneous regions. Geogr Inf Sci

Bação F, Lobo V, Painho M (2005a) Applying genetic algorithms to zone design. Soft Comput 9(5):341–348

Bação F, Lobo V, Painho M (2005b) Self-organizing maps as substitutes for k-means clustering. Comput Sci–ICCS 2005(3516):476–483

Bação F, Lobo V, Painho M (2005c) The self-organizing map, the Geo-SOM, and relevant variants for geosciences. Comput Geosci 31(2):155–163

Birch EL, Wachter SM (2011) Global Urbanization. J Reg Sci 51(5):1026–1028

Bloom LM, Pedler PJ, Wragg GE (1996) Implementation of enhanced areal interpolation using MapInfo. Comput Geosci 22(5):459–466

Braulio-Gonzalo M, Bovea MD, Ruá MJ (2015) Sustainability on the urban scale: proposal of a structure of indicators for the Spanish context. Environ Impact Assess Rev 53:16–30

Chourabi H et al (2012) Understanding smart cities: an integrative framework. In: 2012 45th Hawaii international conference on system sciences, pp 2289–2297

Donoho D (2000) High-dimensional data analysis: the curses and blessings of dimensionality. In: AMS math challenges lecture, pp 1–33

Eicher CL, Brewer CA (2001) Dasymetric mapping and areal interpolation: implementation and evaluation. Cartogr Geogr Inf Sci 28(2):125–138

Fiedler R, Schuurman N, Hyndman J (2006) Improving census-based socioeconomic GIS for public policy: recent immigrants, spatially concentrated poverty and housing need in Vancouver. ACME 4(1):145–169

Fink EL (2009) The FAQs on data transformation. Commun Monogr 76(January 2015):379–397

Fisher PF, Langford M (1996) Modeling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping. Prof Geogr 48(3):299–309

Fukunaga K (1990) Statistical pattern stas-tical pattern recognition. Pattern Recogn 22(7):833–834

Gorricha J, Lobo V (2012) Improvements on the visualization of clusters in geo-referenced data using Self-Organizing Maps. Comput Geosci 43:177–186

Guha S, Rastogi R, Shim K (1998) Cure. ACM SIGMOD Rec 27(2):73–84

Hajek P, Henriques R, Hajkova V (2014) Visualising components of regional innovation systems using self-organizing maps-Evidence from European regions. Technol Forecast Soc Chang 84:197–214

Henriques R, Bacao F, Lobo V (2012) Exploratory geospatial data analysis using the GeoSOM suite. Comput Environ Urban Syst 36(3):218–232

Hernández-Muñoz JM et al (2011) Smart cities at the forefront of the future internet. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6656, pp 447–462

Howard G, Lubbe S, Klopper R (2011) The impact of information quality on information research. Manag Inf Res Des 288

Huang L, Yan L, Wu J (2016) Assessing urban sustainability of Chinese megacities: 35 years after the economic reform and open-door policy. Landsc Urban Plan 145:57–70

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Jelinski DE, Wu J (1996) The modifiable areal unit problem and implications for landscape ecology. Landsc Ecol 11(3):129–140

Kohonen T (1995) Self organizing maps. Springer series in information sciences, 30. Springer, Berlin, p 521

Kohonen T (2013) Essentials of the self-organizing map. Neural Netw 37:52–65. Available at: http://dx.doi.org/10.1016/j.neunet.2012.09.018

Koua EL (2003) Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets. In: Proceedings of 21st international cartographic renaissance (ICC), pp 1694–1702

Koua EL, Kraak M (2004) Alternative visualization of large geospatial datasets. Cartogr J 41(3):217–228

Lee ACD, Rinner C (2014) Visualizing urban social change with self-organizing maps: Toronto neighbourhoods, 1996–2006. Habitat Int 45:92–98

Lombardi P et al (2012) Modelling the smart city performance. Innov: Eur J Soc Sci Res 25(2):137–149

Longo G, Gerometta J, Haussermann H (2005) Social innovation and civil society in urban governance: strategies for an inclusive city. Urban Stud 42(11):2007–2021

Manley D (2014) Scale, aggregation, and the modifiable areal unit problem. In: Handbook of regional science, pp 1157–1171

Mennis J, Guo D (2009) Spatial data mining and geographic knowledge discovery—an introduction. Comput Environ Urban Syst 33(6):403–408

Nam T, Pardo TA (2011) Conceptualizing smart city with dimensions of technology, people, and institutions. In: Proceedings of the 12th annual international digital government research conference on digital government innovation in challenging times – dg.o ‘11, p 282

Nelson JK, Brewer CA (2015) Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem. Cartogr Geogr Inf Sci 0406(October):1–16

Openshaw S (1984) Ecological fallacies and the analysis of areal census data. Environ Plan A 16(1):17–31

Penn BS (2005) Using self-organizing maps to visualize high-dimensional data. Comput Geosci 31(5):531–544

Polèse M (2010) The resilient city: on the determinants of successful urban economies. INRS, Montreal, p 32

Roche S (2014) Geographic information science I: why does a smart city need to be spatially enabled? Prog Hum Geogr 38(5):0309132513517365

Root ED (2012) Moving neighborhoods and health research forward: using geographic methods to examine the role of spatial scale in neighborhood effects on health. Ann Assoc Am Geogr 102(5):986–995

Silva CN, Syrett S (2006) Governing Lisbon: evolving forms of city governance. Int J Urban Reg Res 30(March):98–119

Skupin A (2002) A cartographic approach to visualizing conference abstracts. Ieee Comput Graph Appl 22(February):50–58

Skupin A, Agarwal P (2008) Introduction: what is a self-organizing map? Self-organising maps: applications in geographic information science. Wiley, Chichester

Skupin A, Hagelman R (2005) Visualizing demographic trajectories with self-organizing maps. GeoInformatica 9(2):159–179

Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. In: KDD workshop on text mining, pp 1–2. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4721382

Ultsch A, Siemon HP (1990) Kohonen's self organizing feature maps for exploratory data analysis. In: Proceedings of the International Neural Network Conference (INNC-90), pp 305–308

United Nations (2013) World economic and social survey 2013. Available at: http://esa.un.org/wpp/documentation/pdf/WPP2012_ KEY FINDINGS.pdf

United Nations Economic Commission for Europe (2006). Conference of European statisticians recommendations for the 2010 censuses of. In New York and Geneva

van der Maaten L, Postma E, van den Herik J (2009) Dimensionality reduction: a comparative review. J Mach Learn Res 10(February):1–41

Veiga L (2014) Economic crisis and the image of Portugal as a tourist destination: the hospitality perspective. Worldw Hosp Tour Themes 6(5):475–479

Vesanto J (1999) SOM−based data visualization methods. Intell Data Anal 3(2):111–126

Waddell P (2007) UrbanSim: modeling urban development for land use, transportation, and environmental planning. J Am Plan Assoc 68(3):297–314

Wier M et al (2009) An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accid Anal Prev 41(1):137–145

Chapter 5
The Application of the SPAWNN Toolkit to the Socioeconomic Analysis of Chicago, Illinois

Julian Hagenauer and Marco Helbich

Abstract The SPAWNN toolbox is an innovative toolkit for spatial analysis with self-organizing neural networks. It implements several self-organizing neural networks and so-called spatial context models which can be combined with the networks to incorporate spatial dependence. The SPAWNN toolkit interactively links the networks and data visualizations in an intuitive manner to support a better understanding of data and implements clustering algorithms for identifying clusters in the trained networks. These properties make it particularly useful for analyzing large amounts of complex and high-dimensional data. This chapter investigates the application of the SPAWNN toolkit to the socioeconomic analysis of the city of Chicago, Illinois. For this purpose, 2010 Census data, consisting of numerous indicators that describe the socioeconomic status of the US population in detail, is used. The results highlight the features of the toolkit and reveal important insights into the socioeconomic characteristics of the US.

Keywords Self-organizing neural networks • Spatial clustering • Spatial analysis

J. Hagenauer (✉)
Leibniz Institute of Ecological Urban and Regional Development,
Weberplatz 1, 01217 Dresden, Germany
e-mail: j.hagenauer@ioer.de

M. Helbich
Department of Human Geography and Planning, Utrecht University,
Heidelberglaan 2, 3584 CS Utrecht, The Netherlands
e-mail: m.helbich@uu.nl

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_5

5.1 Introduction

Technological advances have facilitated acquiring, sharing, processing, and storing spatial information. As a result, we are confronted with a massive increase of spatial data (Miller and Goodchild 2014). This data typically contains hidden and unexpected information, which can hardly be explored using traditional methods from the field of statistics, since these typically require hypothesis testing and are not amenable to handle large amounts of data (Miller and Han 2009). To address these issues, spatial data mining emerged as a new field, borrowing mainly methods from the fields of artificial intelligence, machine learning, and spatial database systems, to extract information and to ultimately transform it into new and potentially useful knowledge (Yuan et al. 2004).

One of the most important methods of spatial data mining is clustering. It organizes observations into clusters such that the similarity within a cluster is maximized while the similarity between different clusters is minimized (Jain 2010). Such an organization represents structural organization of the data, which alleviates data exploration of the data. These tasks are often done by a human analyst. Since the humans' ability to perceive and understand visual patterns exceeds the capabilities of computational algorithms (Keim 2002; Ware 2012), it is common practice to combine clustering methods with appropriate visualizations and interactive means in a combined toolkit.

Spatial clustering is different from common clustering in that it takes spatial dependence into account. Spatial dependence means that observations that are spatially close to each other also tend to have similar characteristics (Sui 2004). Without spatial dependence, the variation of phenomena would be independent of location, and thus, the notion of a region would be less meaningful (Goodchild 1986). Mostly it is necessary to take spatial dependence explicitly into account when clustering spatial data, because the available data is not sufficient to accurately model the spatial varying phenomena. Consequently, neglecting spatial dependence has a high risk of resulting in incorrect clusters, leading to a limited understanding of the spatial patterns (Openshaw 1999).

Because of the importance of clustering for data analysis, many different clustering algorithms for spatial and non-spatial data have been developed in the past (see, e.g., Guo 2008; Jain 2010; Parimala et al. 2011). However, only very few neural network-based clustering approaches that explicitly take spatial dependence into account have been developed. The GeoSOM (Bação et al. 2005) and contextual neural gas (CNG) (Hagenauer and Helbich 2013) represent to notable approaches. Both are adaptations of basic self-organizing network algorithms which utilize the spatial configuration of the neurons to account for spatial dependence. However, both approaches are purely computational; a human analyst is still inevitable to interpret the resulting clusters, taking domain-specific knowledge into consideration, and to adjust the parameter settings if needed. To facilitate these tasks, it is necessary to integrate different self-organizing neural network-based clustering methods, where each comes with its unique advantages, in an interactive toolkit with other computational, visual, and geographic methods. Such toolkit should be intuitive and easy to use so that its usage is promoted across different spatial disciplines.

To address the lack of such toolkit, the SPAWNN-toolkit (Hagenauer and Helbich 2016) has been developed. It implements the self-organizing map (SOM) (Kohonen 1982, 2001) and neural gas (NG) (Martinetz and Schulten 1991; Martinetz et al. 1993) and allows these self-organizing networks to be combined with either the CNG or the GeoSOM approach, or with alternative spatial context models, in order to account for spatial dependence. Furthermore, the toolkit provides different visualizations and links between the neurons and a geographic map, which permits the analyst to interactively select neurons or observations and to visually inspect the mapping between them in order to explore the results of the trained networks in detail, and implements a set of powerful clustering algorithms for post-processing the network models. This permits the analyst to interactively select neurons or observations and to visually inspect the mapping between them for exploring the results of the trained networks in detail.

These properties make it particularly useful for analyzing large amounts of complex and high-dimensional data. This chapter investigates the application of the SPAWNN toolkit to the socioeconomic analysis of the city of Chicago, Illinois, using US census 2010 data. This data consists of a large amount of observations and numerous complex indicators that describe the socioeconomic status of the US population in detail. The results of the present analysis bring out the useful features of the toolkit and also give important insights into the socioeconomic characteristics of the city of Chicago.

The paper is structured as follows: Sect. 5.2 presents the methods that are used in this study and which are all part of the SPAWNN toolkit. In Sect. 5.3 the application of these methods to the socioeconomic analysis of Chicago, Illinois, is investigated. Finally, Sect. 5.4 concludes the paper and discusses future work.

5.2 Methods

This chapter utilizes the SPAWNN toolkit for analysis. The toolkit consists of numerous components, which can be useful for different applications (for a detailed description of the toolkit, see Hagenauer and Helbich (2016)). This section describes the main methods and algorithms that were used for the analysis of the present chapter.

5.2.1 Self-Organizing Networks

Self-organizing neural networks represent a class of artificial neural networks (ANNs) that are trained in an unsupervised manner. The SPAWNN toolkit implements two basic self-organizing networks, which are both used in this chapter because of their complementary useful properties (see Hagenauer and Helbich 2016). The first network is the SOM (Kohonen 1982, 2001). The SOM consists of an arbitrary number of neurons that are connected to adjacent neurons by a neighborhood relation that defines the topology of the map. While in principle the dimension of a SOM is arbitrary, two-dimensional SOMs are more commonly used in practice because of their ease of visualization. Each of the SOM's neurons is associated with a prototype vector that is of the same dimension as the input space. During the training, input vectors are presented to the SOM, and the neuron with the smallest distance to the input vector, referred to as the best matching unit (BMU), is identified. The prototype vector of the BMU and the prototype vectors within a certain neighborhood on the map are then moved in the direction of the input vector. The magnitude of the movements depends on the distance of the neurons to the BMU on the map and on the actual learning rate. Both the size of the neighborhood and the learning rate are monotonically decreased during the learning. After the training, the SOM represents a low-dimensional map of the input space. Each neuron of the SOM represents some portion of the input space and distance relationships of the input space are mostly preserved.

The second network is NG (Martinetz and Schulten 1991). While NG is inspired by the SOM, it has some significant differences. Similar to the SOM, it consists of an arbitrary number of neurons. However, in contrast to the SOM, the NG's neurons are not subjected to any topological restrictions, which typically results in a superior quantitative performance compared to the SOM (Martinetz et al. 1993; Cottrell et al. 2006). Associated with each of the NG's neurons is a prototype vector of the same dimension as the input space. During the training, input vectors are presented to the NG and each neuron is moved in the input vector's direction. The magnitude of the movement depends on the neurons' ranking order with respect to the distance to the input vector, the learning rate, and the neighborhood range. The neighborhood range and learning rate are typically set to decrease with training time. After a sufficient number of training steps, the prototype vectors typically approximate the probability density function of the input space with near-minimum quantization error.

In contrast to the SOM, NG does not have a predefined topology, which reflect the similarity relationships between the neurons (Martinetz and Schulten 1991). A topology is particularly useful because it can reveal valuable information about the underlying data. In order to learn a topology, competitive Hebbian learning (Martinetz and Schulten 1991; Martinetz 1993) can be applied to NG in a post-processing step as follows: For each input vector, the two closest neurons are identified and a connection between these two neurons is added to the total set of connections, whereas closeness is usually measured by Euclidean distance.

When all input vectors have been processed, the resulting set of connections represents the topology. The number of connections that have been added between two neurons can be used to indicate the strength of their relationship (Hagenauer 2014).

5.2.2 Spatial Context Models

Hagenauer and Helbich (2016) introduced the concept of the spatial context model. A spatial context model describes the relationships between spatial observations and the neurons of a self-organizing neural network during the training or when applying the trained network to data. They have previously been considered as a integral part of a self-organizing network (see, e.g., Bação et al. 2005; Hagenauer and Helbich 2013). However, it is reasonable to distinguish between self-organizing networks and spatial context for the following reasons: First, such a distinction maintains the modularity of the toolkit. This is desired because it facilitates reuse of existing code, the implementation of new features, and its further extension. Second, and more important, it allows to combine different self-organizing networks with different spatial context models and thus increases the analytical capabilities of the toolkit (Hagenauer and Helbich 2016).

While the SPAWNN toolkit implements a variety of different spatial context models, this study uses the GeoSOM (Bação et al. 2005) and Contextual Neural Gas (CNG) (Hagenauer and Helbich 2013). A particular advantage of both is that they can produce accurate mappings and do not require to scale the spatial dimensions of the input data to match the feature space dimensions (Hagenauer and Helbich 2016).

The GeoSOM (Bação et al. 2005) is a variant of the SOM algorithm that adapts the idea of Kangas (1992) for quantizing, clustering, and visualizing spatial data. The main difference with the basic SOM is that the GeoSOM uses a two-step procedure to determine the BMU. In the first step, the neuron that is spatially closest to the input vector is identified. In the second step, the closest neuron to the input vector, but within a fixed radius of this neuron (in terms of map distance), is identified. This neuron is then designated as the final BMU. The size of the radius affects the strength of spatial dependence that is incorporated into the learning process. The smaller the radius, the more the final ordering of the map is determined by spatial closeness.

CNG (Hagenauer and Helbich 2013) is a vector quantization and clustering algorithm that combines the concepts of the GeoSOM with the NG algorithm. Analogous to the GeoSOM, CNG enforces spatial proximity between the observations and neurons by utilizing the spatial arrangement of the neurons. However, since its neurons are not topologically ordered in a map, CNG applies a two-step procedure for determining a rank ordering: In the first step, the neurons are ordered with respect to spatial closeness. In the second step, the first k neurons of the resulting spatial ordering are reordered within their ranks according to input vector similarity. The parameter k controls the degree of spatial dependence that is incorporated in the adaptation process. The smaller the parameter *k*, the more is the adaptation of neurons determined by spatial closeness.

5.2.3 Network Clustering

The SPAWNN toolkit provides several powerful clustering algorithms as well as means for manually outlining and visualizing clusters in the trained networks. The advantage of using clustering algorithms is that the results depend less on the subjective decisions of an analyst and are more convenient to obtain for large or complex networks. Because contiguity constrained clustering has already been shown to be effective for clustering self-organizing neural networks (see e.g. Murtagh 1995), this algorithm is also used in this chapter. The algorithm works as follows: At the beginning each neuron represents one cluster. Than, two clusters are determined which have at least one neighboring neurons and maximize a certain optimization criterion. These clusters are then merged and the procedure is repeated until the number of desired clusters is reached. While many different criterion for hierarchical clustering exist, this study uses Ward's criterion (Ward 1963). This criterion evaluates the smallest increase of sum of squares that results from a merge of two clusters.

5.3 Case Study

This sections investigates the application of the SPAWNN toolkit to the analysis of the socio economic characteristics of the city of Chicago, Illinois. The analysis consists of several steps, namely an outlier analysis, a correlation analysis, and a cluster analysis.

Chicago is situated in the Midwestern United states, in the northeast of Illinois on the southwestern shores of the Lake Michigan and consists of an area of approximately 606 km^2. The city is currently the third largest city in the United States with an estimated population of 2.7 million people in 2012. It is also the principal city and cultural center of the Chicago Metropolitan Area with an estimated population of 9.9 million people. Two thirds of the cities population are members of minority groups and segregation is on an exceptional high level (Kaufman 1998). Consequently, it is expected that the analysis will reveal the socioeconomic characteristics of the city strongly vary across space and that the segregation of minority groups will emerge as distinct clusters.

The case study uses freely available tract-level data extracted from the 2010 US Census about ethnicity, age, housing, and households in Chicago. While census tracts are mostly homogeneous with respect to population characteristics, economic status, and living conditions, there are census tracts which exhibit pronounced ethnic heterogeneity below the census tract level. These tracts do not affect the applicability of the toolkit, but they must be considered when interpreting the results. The following eight variables are used: percentage of white population, percentage of African Americans, percentage of Asians, percentage of Hispanics, percentage of renter-occupied houses, percentage of population younger than 25 years old, percentage of population older than 64 years and the average size of households. Tracts without population are removed from the data set beforehand, and all attributes are standardized to zero mean and unit variance to make them comparable. The study site consists of 797 census tracts in total. While the SPAWNN toolkit can be applied to data sets of arbitrary size, the rather small number of census tracts in this case study facilitates the visualization of the results.

In the following, a GeoSOM and CNG are trained with the following settings. The GeoSOM consists of 12 × 8 neurons and the CNG consists of 96 neurons. Preliminary tests have shown that these numbers represent a fair compromise between computational effort for training the networks and quantization performance. Both networks are trained for 500,000 iterations. Moreover, the neighborhood of the GeoSOM is chosen Gaussian and its radius is at the beginning set to 10 and approaches 1 at the end of the training. The learning rate of the GeoSOM decreases linear from 1 to 0.01. The neighborhood range and learning rate of CNG are chosen to decrease linear with training time. Starting with 48, the neighborhood range reaches 0.01 at the end of the training. The learning rate starts with 0.5 and also ends with 0.01.

A critical role plays the choice of the parameters that determine the strength of spatial autocorrelation that is incorporated in the mapping, the radius of the GeoSOM and k of NG. The radius of the GeoSOM is set to 4, while k of the CNG is set to 35. These settings provide a fair compromise between quantization performance of the networks and incorporation of spatial dependence. Moreover, the ratio of quantization performance and spatial coherence, measured as the quantization error of the spatial coordinates, is for both settings approximately the same.

5.3.1 Outlier Analysis

First, a GeoSOM and CNG are trained for outlier detection. The identification of outliers is a crucial task, because outliers distort the distribution of the data and thus can significantly affect the results of subsequent analysis.

In the distance matrix representation of the resulting GeoSOM (Fig. 5.1), outliers can be identified by neurons that have high median distance to neighboring values and are, given that the size of the map is sufficiently large, located at the border of the matrix (Muñoz and Muruzábal 1998). In total, the GeoSOM identified 25 census tracts as outliers. In the distance-based representation of the resulting CNG (Fig. 5.2), outliers can be identified by having also a high median distance to neighboring neurons but also by being sparsely connected to other neurons (Hagenauer and Helbich 2016). In total, the CNG identified three outliers, one in the north and

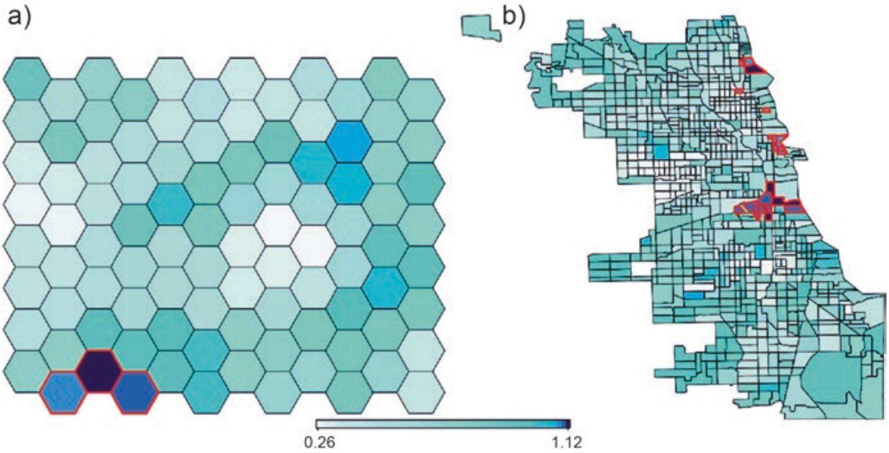


Fig. 5.1 Distance matrix (**a**) and cartographic map (**b**) of the GeoSOM. Identified outliers are outlined in *red*

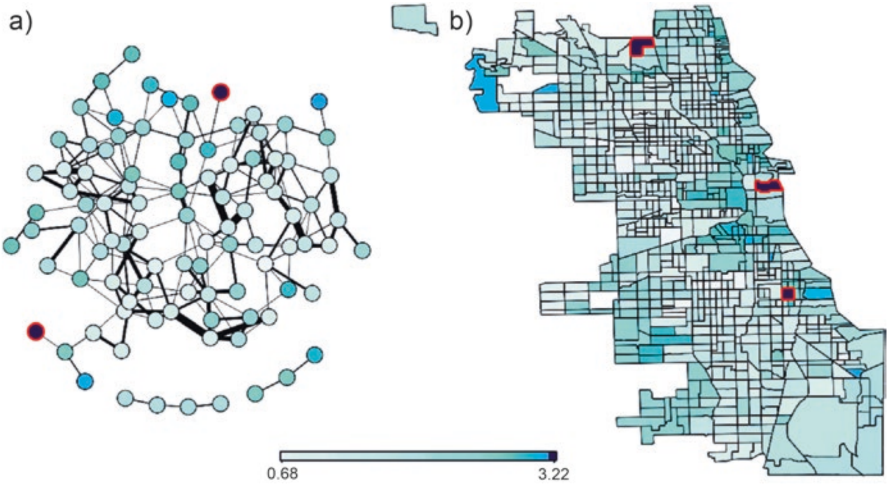


Fig. 5.2 Distance-based representation of the CNG (**a**) and the geographic map (**b**). Identified outliers are outlined in *red*

two close to the shore in the west of the city. The northern outlier is mapped by a different neuron than the other two neurons.

Comparing the identified outliers (outlined in red) in both representations shows that they mostly do not correspond. Only one census tract in the west has been identified by both networks as an outlier. A reason for the differences might be that selecting a median distance threshold is a rather subjective task. This matter is typically less crucial for CNG, because the learned topology provides an additional guidance for the identification of outliers (Hagenauer and Helbich 2016). Moreover, the generally better quantization performance of CNG (see Hagenauer and Helbich 2013) permits a more accurate representation of the data. Indeed, close inspection of the outliers identified by the GeoSOM does not reveal a common pattern that helps to understand why the GeoSOM identified these particular tracts as outliers. Many of the outliers of the GeoSOM share similarities and can barely considered outliers. Inspecting the outliers of the CNG in detail reveals that the northern tract is considered an outlier because of its very high rate of old population. None of the tracts in its neighborhood has comparable age characteristics. Similarly, the two tracts in the west are considered outliers by CNG, because their rates of young and white population exceed the ones of every other tract in their neighborhoods. Thus, it can be concluded from this section that the results of the CNG are more consistent with the actual data than the ones of the GeoSOM and that the CNG is more appropriate for identifying outliers than the GeoSOM.

5.3.2 Correlation Analysis

As a second analysis step, correlation analysis is performed which identifies and evaluates the associations between different attributes of the data. For this purpose, the identified outliers are removed from the data set first and a GeoSOM and a CNG are trained.

A common approach for identifying correlations in the data is to compare component planes (e.g., Vesanto and Ahola 1999; Barreto and Pérez-Uribe 2007). Correlations become apparent by similar (positive correlation) or complementary (negative correlation) patterns in identical areas of the network.

This approach for identifying correlations has several advantages over standard correlation analyses: First, the SOM as well as the NG provide a nonlinear map of the data which allows the identification of nonlinear correlations. Second, by comparing multiple component planes multivariate correlations become apparent. Third, local correlations can be identified by partially matching patterns.

Figure 5.3 exemplarily depicts the GeoSOM component planes for the rates of African Americans and white population. The component planes reveal complementary patterns, indicating a strong negative correlation and therefore high segregation between the African American and white populations. Furthermore, it can be seen that the patterns for both rates span the whole map. Since the GeoSOM is spatially explicit, this indicates that the segregation between the African American and white population is not restricted to local neighborhoods, but rather is present for most parts of the city.

Similar, Fig. 5.4 shows the neurons of the CNG, which are also colored according to the percentage of African Americans (a) and the percentage of white population (b). Even though the complementary patterns are also present, these patterns are more difficult to perceive due to the seemingly unordered arrangement of the

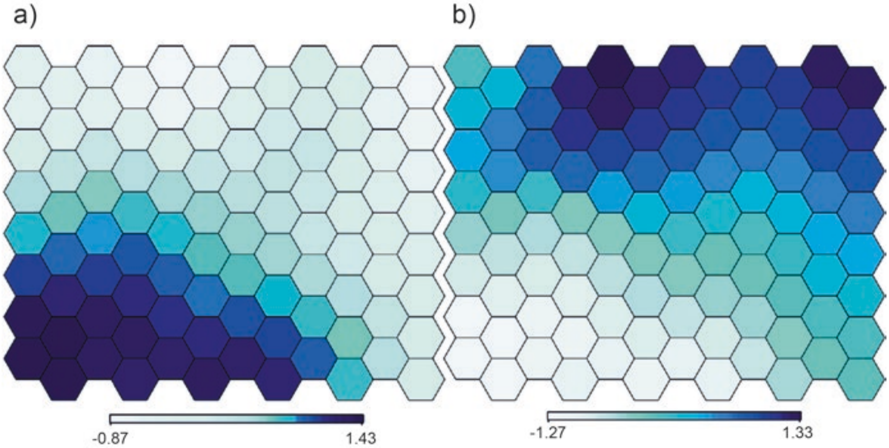


Fig. 5.3 GeoSOM component planes for the rates of African Americans (**a**) and white population (**b**)

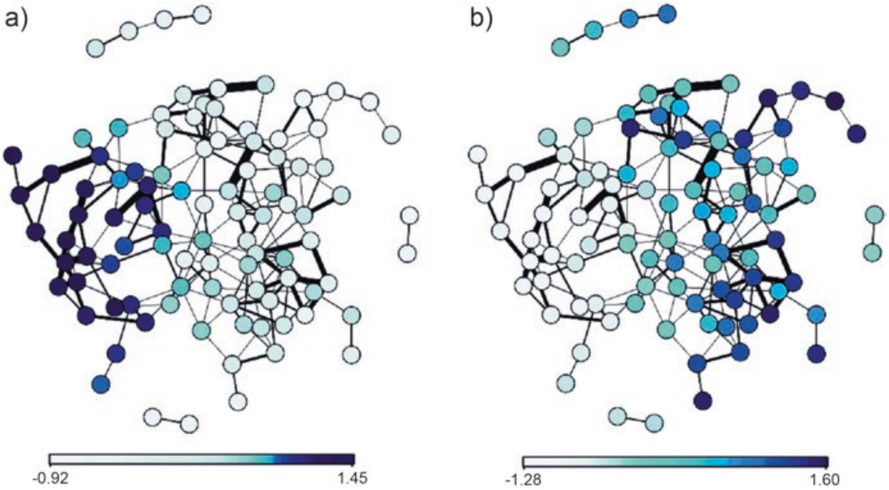


Fig. 5.4 Neurons of the CNG, colored according to the rates of African Americans (**a**) and white population (**b**)

neurons and overlapping topology of the network. In fact, it is hardly feasible to arrange the CNG's neurons on a two-dimensional plane while preserving the neurons' topological relationships. This problem typically becomes even more severe as the dimension of the input space increases. Thus, it is hard to perceive the distance relationships of the neurons from the network topology alone.

Figure 5.3 exemplarily depicts the GeoSOM component planes for the rates of Hispanics and average household size. The component planes reveal more complex relationships between these variables. The similar dark coloring in the upper left

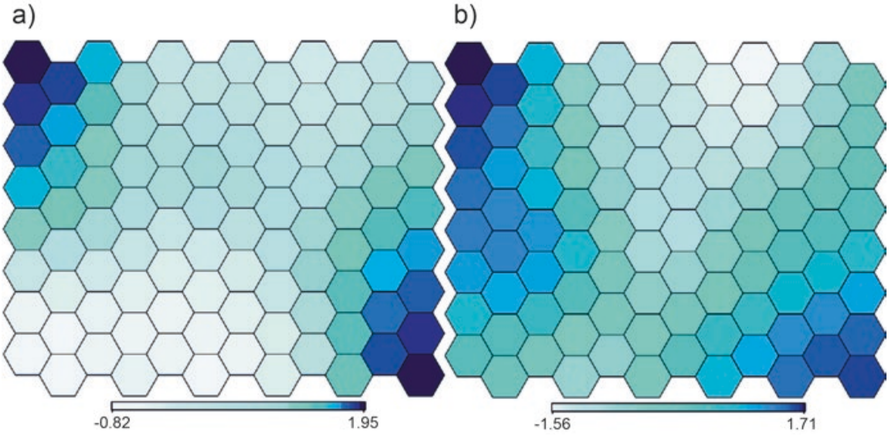


Fig. 5.5 GeoSOM component planes for the rates of Hispanics (**a**) and average household size (**b**)

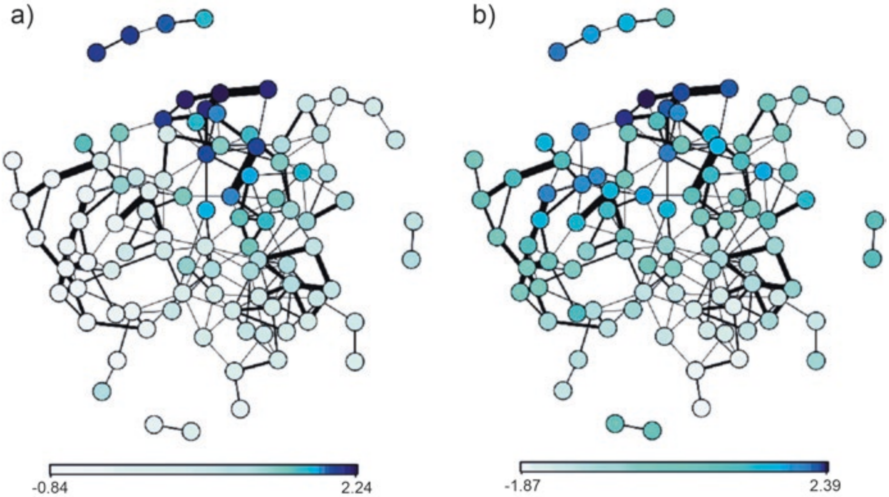


Fig. 5.6 Neurons of the CNG, colored according to the rates of Hispanics (**a**) and average household size (**b**)

and lower right of both component lanes indicates clearly that for some areas of the city high rates of Hispanic population correlate strongly with large average household sizes. However, the coloring of the planes is rather different in the middle left of the component planes. This indicates, that large average household sizes are not exclusively related to high rates of Hispanic population.

Analogously, Fig. 5.4 shows the neurons of the CNG, which are also colored according to the percentage of Hispanics (a) and the average household size (b). It can be clearly seen that in this representation high rates of Hispanic population correlate strongly with large household sizes. Besides that, further patterns that reveal further insights into the relationships of these variables are again hardly perceivable.

To conclude, both networks show a high segregation of Chicago and the correlation between high rates of Hispanic population and large average household sizes. However, the GeoSOM provides a more clear representation of the relationships of the variables than the CNG and is thus more appropriate for visually analyzing correlations.

5.3.3 *Cluster Analysis*

As a last step showing the application of the SPAWNN toolkit to the socioeconomic analysis of Chicago, the trained GeoSOM and CNG from the preceding section are used to detect spatially contiguous clusters within the study area. The applied clustering algorithm is contiguity-constrained hierarchical clustering using Ward's criterion. Figure 5.7 maps the results for the CNG, while the results for the GeoSOM are shown in Fig. 5.8. In addition, to facilitate interpretation of the results boxplots are depicted in Fig. 5.9 for CNG and Fig. 5.10 for the GeoSOM.

Both algorithms detected very different clusters, even though there are also some notable similarities. For example, cluster 3 of the GeoSOM matches cluster 3 and 7 of CNG, while the outline of cluster 4 of CNG is similar to cluster 4 and 6 of the GeoSOM. However, no clusters of GeoSOM and CNG are perfectly identical.

In general, the high segregation present throughout the city facilitates the interpretation of the clustering results. Cluster 8 of the GeoSOM, for instance, outlines accurately the census tracts with the highest rates of Asian population. However, the GeoSOM does not distinguish between tracts where Asians represent a majority and

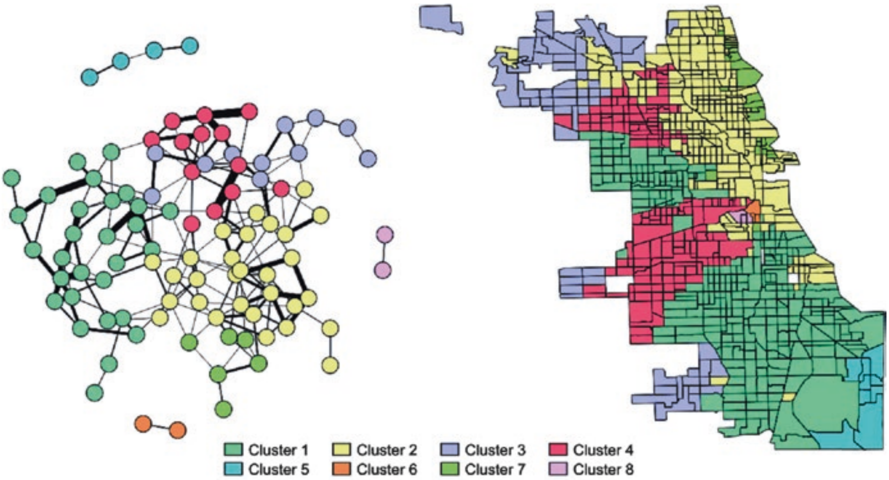


Fig. 5.7 Clustering results for CNG

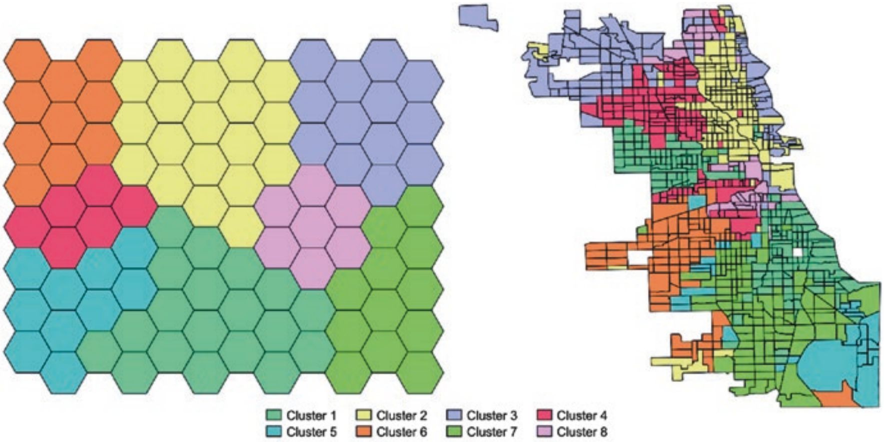


Fig. 5.8 Clustering results for the GeoSOM

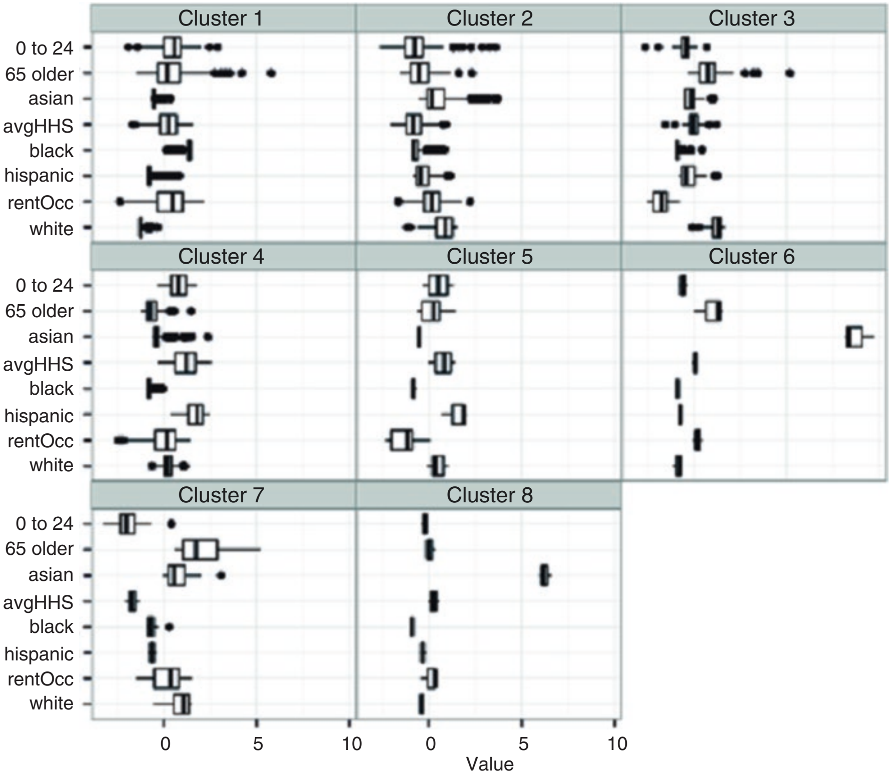


Fig. 5.9 Boxplots of the clustering results for CNG

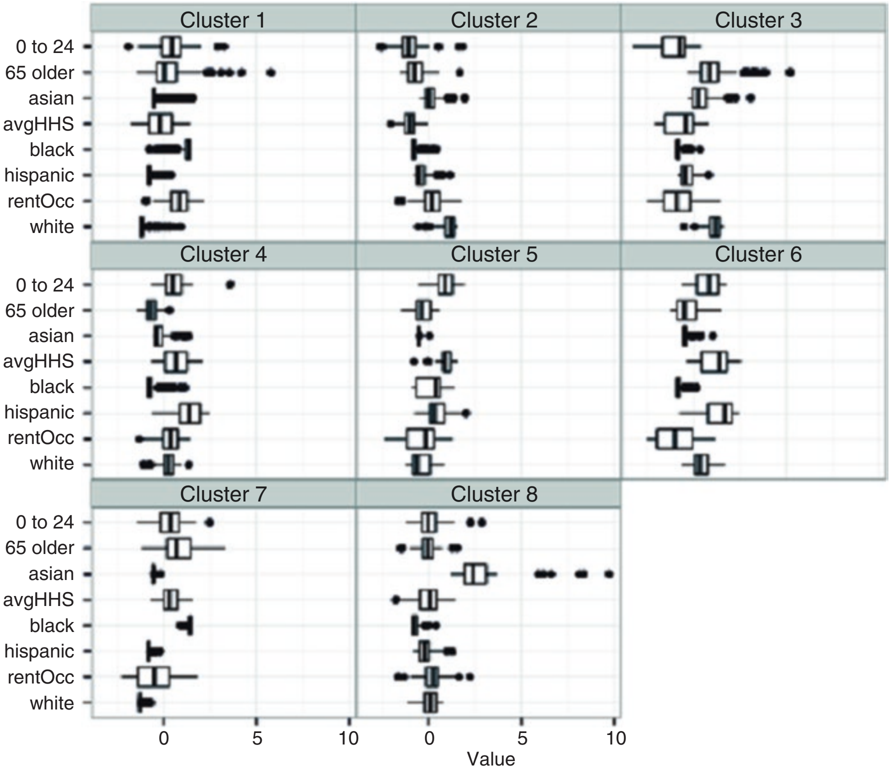


Fig. 5.10 Boxplots of the clustering results for the GeoSOM

minority of the population. By contrast, CNG does not outline all census tracts with high rates of Asian population, in particular not the ones in the north of the city, but it clearly identifies the census tracts where Asians represent the majority of the population (cluster 6 and 8). Even though Cluster 6 and 8 are located close to each other, CNG distinguished between them. The reason for this is that the rate of Asian population and the proportion of people older than 65 is higher for cluster 6. In fact, cluster 6 mainly covers the Chinatown neighborhood of Chicago, which is the oldest persisting settlement of Chinese in the city (Santos et al. 2008).

African Americans represent the largest minority in Chicago and its segregation is particularly strong (Kaufman 1998). This characteristic is clearly visible for the clustering of CNG, where cluster 1 almost exclusively consists of census tracts where African Americans represent the majority of population. In the clustering of the GeoSOM, predominantly African American census tracts are mainly represented by two different clusters, cluster 5 and 7 in the south of the city. Such a distinction between clusters can be meaningful, because the clusters represent different regions of the city, which might have undergone different social and economic developments in the past. In addition, cluster 5 also consists of some tracts with significant rates of African Population. Thus, the high segregation of the African American population is more apparent in the clustering of the CNG.

The second largest minority of the city represent Hispanics. Similar to the rates of African Americans, CNG outlines the predominantly Hispanic census tracts by a single cluster (cluster 4), while the GeoSOM basically represents these tracts by two separate clusters (cluster 4 and 5). However, some tracts of cluster 5 of the GeoSOM, particularly in the southwest of the city, have low rates of Hispanic population. Thus, the Hispanic population is more faithfully represented by the clustering of the CNG.

Cluster 3 of the GeoSOM is characterized by a high proportion of white and old population. It is very similar to cluster 3 of CNG, but the latter assigns the census tracts on the east coast to a separate cluster (cluster 7). The reason for this is that the rate of renter occupied houses is much higher for these tracts than for the rest of the tracts of cluster 3. This reflects the actual geography of the city: While the northeast of the city is characterized by middle class family homes, the tracts of cluster 3 of CNG are characterized by many large apartment buildings. Therefore, the clustering of the CNG is more reasonable than the on of the GeoSOM.

In conclusion, while the GeoSOM is particularly useful for relating clusters to component planes in order to inspect data relationships, the clustering of the CNG is geographically more accurate

5.4 Conclusion

This chapter presented the application of the SPAWNN toolkit, a new and powerful exploratory toolkit for spatial analysis and clustering, for the analysis of the socioeconomic characteristics of the city of Chicago, Illinois, using US census data. The results showed the complementary advantages of CNG and GeoSOM and how these networks can be used to get a better understanding of the data. In addition, they pointed out that Chicago is faced with high segregation across the cityscape and challenged by socioeconomic diversity.

References

Bação F, Lobo V, Painho M (2005) The self-organizing map, the Geo-SOM, and relevant variants for geosciences. Comput Geosci 31:155–163

Barreto SMA, Pérez-Uribe A (2007) Improving the correlation hunting in a large quantity of SOM component planes. In: de Sá JM, Alexandre L, Duch W, Mandic D (eds) Artificial neural networks – ICANN 2007, vol 4669 of Lecture Notes in Computer Science. Springer, Berlin/Heidelberg, pp 379–388

Cottrell M, Hammer B, Hasenfuß A, Villmann T (2006) Batch and median neural gas. Neural Netw 19(6):762–771

Goodchild MF (1986) Spatial autocorrelation. CATMOG. Geo Books, Norwich

Guo D (2008) Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). Int J Geogr Inf Sci 22(7):801–823

Hagenauer J (2014) Clustering contextual neural gas: a new approach for spatial planning and analysis tasks. In: Helbich M, Jokar Arsanjani, J, Leitner M (eds), Computational approaches for urban environments. Springer

Hagenauer J, Helbich M (2013) Contextual neural gas for spatial clustering and analysis. Int J Geogr Inf Sci 27(2):251–266

Hagenauer J, Helbich M (2016) Spawnn: a toolkit for spatial analysis with self-organizing neural networks. Transactions in GIS

Jain AK (2010) Data clustering: 50 years beyond k-means. Pattern Recogn Lett 31(8):651–666

Kangas J (1992) Temporal knowledge in locations of activations in a self-organizing map. In: Aleksander I, Taylor J (eds) Artificial neural networks, 2, vol 1. North-Holland, Amsterdam, pp 117–120

Kaufman JL (1998). Chicago: segregation and the new urban poverty. In: Urban segregation and the welfare state: inequality and exclusion in Western Cities. Routledge, pp 45–63

Keim DA (2002) Information visualization and visual data mining. Visual Comput Graph IEEE Trans 8(1):1–8

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

Kohonen T (2001) Self-organizing maps. Springer, New York

Martinetz T (1993) Competitive hebbian learning rule forms perfectly topology preserving maps. In: Gielen S, Kappen B (eds) ICANN ‘93. Springer, London, pp 427–434

Martinetz T, Schulten K (1991) A “neural-gas” network learns topologies. Artif Neural Netw 1:397–402

Martinetz T, Berkovich S, Schulten K (1993) “neural-gas” network for vector quantization and its application to time-series prediction. IEEE Trans Neural Netw 4(4):558–569

Miller HJ, Goodchild MF (2014) Data-driven geography. GeoJournal 1–13

Miller HJ, Han J (2009) Geographic data mining and knowledge discovery. CRC Press, Boca Raton

Muñoz A, Muruzábal J (1998) Self-organizing maps for outlier detection. Neurocomputing 18(1):33–60

Murtagh F (1995) Interpreting the kohonen self-organizing feature map using contiguity-constrained clustering. Pattern Recogn Lett 16(4):399–408

Openshaw S (1999) Geographical data mining: key design issues. In: 4th international conference on geocomputation. Mary Washington College, GeoComputation CD-ROM, Fredericksburg

Parimala M, Lopez D, Senthilkumar N (2011) A survey on density based clustering algorithms for mining large spatial databases. Int J Adv Sci Technol 31(1):59–66

Santos CA, Belhassen Y, Caton K (2008) Reimagining chinatown: an analysis of tourism discourse. Tour Manag 29(5):1002–1012

Sui DZ (2004) Tobler’s first law of geography: a big idea for a small world? Ann Assoc Am Geogr 94(2):269–277

Vesanto J, Ahola J (1999) Hunting for correlations in data using the self-organizing map. In: Proceeding of the international ICSC congress on Computational Intelligence Methods and Applications (CIMA ‘99), pp 279–285

Ward JH (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244

Ware C (2012) Information visualization: perception for design. Elsevier, Amsterdam

Yuan M, Buttenfield B, Gahegan M, Miller H (2004) Geospatial data mining and knowledge discovery. In: McMaster RB, Usery EL (eds) A research agenda for geographic information science. CRC Press, Boca Raton

Part III
Spatial Modelling, System Dynamics and Geosimulation

Chapter 6
The Evolution of the Land Development Industry: An Agent-Based Simulation Model

Jonatan Almagor, Itzhak Benenson, and Daniel Czamanski

Abstract Urban spatial structure is shaped by decisions of land developers that both react to and influence urban plans. The paper presents an agent-based model of the evolution of the land development industry in a city regulated by a land-use plan that is modified from time to time by the planner. At the heart of the model are investment decisions of developers that generate profits and accumulated assets, which in turn affect investment decisions. In the model, the economic state of the developers is initially equal. Over time, certain developers accumulate wealth that enables them to make larger investments and take higher risks by investing in low priced lands that are not zoned for urban development. These risky investments are motivated by the prospect of obtaining land-use variance. We demonstrate that when the land market favors large developers who are more likely to obtain construction permits from the planner, a positive feedback effect is created, which leads to an oligopolistic market, controlled by a few large developers. We also demonstrate that the interaction between risk-taking developers and a flexible planner who approves incremental amendments and periodic updates to the land-use plan may result in bifurcations of the city structure, which leads to a polycentric city.

Keywords Agent-based model • Land developers • Urban development • Land-use plan • Spatial structure

J. Almagor (✉)
The Porter School of Environmental Studies, Tel Aviv University, Tel Aviv, Israel
e-mail: yonialmagor@hotmail.com

I. Benenson
Department of Geography and Human Environment, Tel Aviv University, Tel Aviv, Israel
e-mail: bennya@post.tau.ac.il

D. Czamanski
Faculty of Architecture and Town Planning, Technion – Israel Institute of Technology, Haifa, Israel
e-mail: ardaniel@tx.technion.ac.il

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_6

6.1 Introduction

There is limited literature concerned with the land development industry. The existing literature suggests that there is a large variance in size distribution of developers. Somerville (1999) finds a rich variation in the market structure of developers across metropolitan areas. He demonstrates that the mean size of developer firms is larger in the more active housing markets, where more undeveloped land is available and where the probability of carrying out land assembly is lower. He concludes that the systematic variation in developer firm size is consistent with treating homebuilding as an imperfectly competitive industry. Our study is motivated by this stylized regularity. We seek to understand the repercussions of investment behavior of land developers on the industrial organization of the industry and its impact on urban spatial structure. We examine the claim that under particular land-use plan dynamics and market conditions, rent seeking by developers leads to an imperfectly competitive industrial structure and specific patterns of spatial evolution of cities.

Regulation is an important factor affecting the land development industry. Land-use plans restrict land supply, and therefore owners of residential land sites acquire a degree of monopoly power (Ball 2003). This may lead to the emergence of larger developers who exploit the opportunity to control supply and reduce competition in both land and housing markets. Buzzelli (2001) analyzes the evolving firm size structure of developers in North America based on census data. He finds no long-term trend toward rising market concentration that is characterized by a few large development firms. Rather the industry passes through cycles in the levels of concentration, and even when concentration peaks, it never approaches the degree of centralization common in other industries. In contrast, Coiacetto (2009) points to evidence that supports the tendency of the land development industry to concentrate. Such evidence includes the rise of large development firms that dominate the industry in the UK and are increasingly common in Australia. The use of product branding by development firms, which is associated with oligopolistic strategy and locally oriented development, lead to spatial monopoly. Coiacetto (2009) suggests that variation in the development industry structure depends on local factors and sectors, where instances of high oligopoly can exist and some degree of monopoly can be achieved.

The study of economic behavior of developers demands explicit description of their behavior. Agent-based modeling (Benenson and Torrens 2004) makes it possible to incorporate microeconomic fundamentals into the agents' decision-making rules. Parker and Filatova (2008) outline a conceptual agent-based model of land market dynamics with the agents representing buyer households, relocating seller households and developers. Seller and buyers negotiate and bid prices evolve in respect to the ratio of active buyers and sellers at the market. Ettema (2011) develops this idea further by examining the dynamics of housing market where prices are established according to the agent's perception of the availability of housing. Magliocca et al. (2011) expand the scheme proposed by Parker and Filatova (2008) and link developers' rent expectations and bidding at the land market. Magliocca et al. (2015) investigate how changes in landscape characteristics and heterogeneity in consumer preference impact land prices, timing of land sales and location and density of development.

A fundamental assertion at the backdrop of this paper is that to understand the evolution of the land development industry, it is imperative to frame land developers' behavior in the spatial context of cities and the land-use plans that regulate them. The behavior of land developers reflects parsimoniously all the relevant information concerning urban land markets and aims at profit maximization. However, the outcomes of developers' decisions are always uncertain and bear risk. One of the significant risks developers face is associated with obtaining construction permits (Sevelka 2004). It may be high in areas not zoned for construction, where the likelihood of obtaining the necessary approvals and permits is uncertain; Whereas, in areas with an approved development plan, the permitting risk is reduced and the regulatory process required for obtaining construction permits is shortened. The delay in obtaining construction permits creates additional costs for the developer and delays revenue generation from the development. The duration of the development process, from the time of land purchase up to the sale of the developed land or real-estate products, constitutes the critical variable in developers' decision-making processes (Czamanski and Roth 2011). Recent analysis of empirical data is consistent with the theory that regulatory delays reduce the probability of subdivision development on a parcel (Wrenn and Irwin 2015).

In this paper, we investigate the emergence of a land development industry as the outcome of individual developers who make land investment decisions within a city constrained by a land-use plan. By means of an agent-based model (ABM), we investigate the consequences of developers' actions following two basic assumptions grounded in theory and empirical evidence. First, the perception of risk is dependent on the financial state of the land developer. We assert that large developers are less risk-averse and more accepting of investments in land not zoned for development than small developers. This is because they are more likely to obtain land-use variances from the planning authorities by using political influence and negotiating capabilities (Molotch 1976; Stone 1993; Dalton et al. 1989; Ruming 2010).

Second, land-use plans are not static. Planning authorities periodically adjust plans in response to development pressures (Booth 1995; Booth 2003; Janssen-Jansen and Woltjer 2010; Alfasi et al. 2012; Abrantes et al. 2016). This implies the possibility of qualitative changes in entire sections of the city's spatial pattern, following the accumulation of local land-use variances. These deviations of development from the land-use plan are later included by the planning agencies in the next comprehensive land-use plan (Alfasi et al. 2012). By this reason, we deviate from the popular modeling assumption that urban development plans do not change throughout the simulation (Huang et al. 2013a, b) and, in the ABM presented in this paper, incorporate the possibility of modifications in the land-use plan as the simulation progresses. The ABM enables explicit consideration of multiple decision-making of developers operating under changing planning conditions and the study of the consequences on the land development industry and urban patterns.

The model starts with developer-agents who are homogeneous in terms of wealth and simulates the economic repercussions of their investments in land. Agent heterogeneity emerges as some receive abnormal profits from construction. The accumulation of wealth by larger developer-agents makes it possible for them to make even larger investments and assume greater risks. As a result, and under different conditions of competition and regulation, we obtain various size-distributions of developers and various urban development patterns.

The remainder of this paper includes four sections. In Sect. 2, we discuss two main planning systems and lay out the context of our model. In Sects. 3 and 4 we describe the structure and dynamics of the ABM. In Sect. 5 we present the results of our simulations, which we discuss in Sect. 6.

6.2 Regulatory Versus Discretionary Planning Systems

Developers operate under the restrictions of planning systems. Planning legislation has two basic forms. The first is based on land-use regulation. The second is a general planning policy with discretionary decision-making. In this paper, we assume that developers operate within a regulatory planning system that is widely used in the majority of European countries, Australia, and most of the United States. Long-term comprehensive outline plans form the pillars of this system. The outline plan includes land-use scripts and ordinances that assign specific land-uses to different zones and determine both spatial location and the extent of land-use development. The regulatory planning system assumes that future land development can be predicted and directed and provides an excess supply of land for each activity by which the environment is developed and organized (Alfasi and Portugali 2004; Alfasi and Portugali 2007). As such, regulatory planning is mainly passive. While assigning the location and extent of development, it does not initiate actual development. The implementation of the plan is in the hands of landowners, developers and other public bodies (Dalton 1989).

The discretionary planning approach consists of a general guiding policy. This planning approach is implemented in the United Kingdom, where the central government supervises and controls planning policies and publishes national and regional guidelines that constitute instructions for planning decisions for local authorities on numerous policy topics. There are no national or regional spatial plans, such as comprehensive land-use plans. Local authorities grant planning permissions and, thus, regulate development. The permission to develop land commonly involves extensive negotiation processes between private developers and local planning authorities (Janssen-Jansen and Woltjer 2010; Buitelaar et al. 2011; Booth 2003).

On the one hand, discretionary planning provides the planning authorities with flexibility in dealing with the uncertain nature of reality, including unexpected development, local initiatives, innovations and changing public needs that were not predicted by the plan. On the other hand, a regulative planning system ensures certainty for an extended period and minimizes risks for owners of development rights (residents, landowners and developers) (Booth 1995). However, Alfasi (2006) argues that in regulative planning systems there is a growing gap between the official planning system and its conduct in actuality. This gap is the result of the conflict between the need to create a firm picture of future development and the need to leave room for discretionary actions to incorporate unexpected needs and initiatives. In actuality, local authorities develop a leeway for discretion in decision-making by using exemption procedures that amend the operative land-use plans that were not foreseen by legislators (Janssen-Jansen and Woltjer 2010).

Local pressure from developers, combined with the interest of local municipalities to attract new building projects that enlarge the city's revenues and prestige, push the local planning commission to approve local changes to the plan (Booth 2002). The wide use of spot zoning – the procedure for approving local amendments to the plan – characterizes Israel's planning institutions. It results in substantial deviations from the comprehensive plan policy. As local plans deviate from the comprehensive policy set by higher order plans, the regulatory planning system hierarchy is breached. The scale of the deviations of actual development from the comprehensive plan was examined recently in the context of the implementation of Israel's central district plan – DOP/3 (Alfasi et al. 2012). These deviations are frequent, ranging from single constructions to groups of buildings and extended urban areas which, over the years, are incorporated into the comprehensive land-use plan prepared for the next period. These deviations confirm an explanation for the gap between the regulatory planning and its actual implementation brought by Dalton (1989) who asserts that the dependence on regulation as the primary form of plan implementation "reinforces a decision making process that emphasizes bargaining with applicants, permitting piecemeal adjustments to the plan over time" (p. 162). The planning agency is "captured" by the development industry because of the close interaction between the project review planner and the developers.

In what follows, we incorporate into the model the aforementioned discretionary mechanisms that take place in actuality under regulative planning systems. We assume that the statutory land-use plan is not static but constantly modified. Developers submit applications for a land-use modification and the planning authorities (represented in the model by a developer-agent), based on policy rules, decide whether to accept them. In this way, unplanned development becomes possible.

6.3 The Evolution of the Land Development Industry: An Agent-Based Model

6.3.1 Concise Introduction to the Model

The goal of the model is to simulate the process of city formation and wealth distribution by developers under varies assumptions regarding the competition in the land market and land-use regulation. Two types of agents operate and interact in the modeled city: developers and a planner. The planner-agent aims at preserving the city compactness by limiting development to the urban zone established by a land-use plan. Developer-agents purchase lands and construct housing with the goal to maximize their profits. All developer-agents start with an equal endowment of wealth, which they invest in land purchase and residential development. The developer-agents' financial state is based on the land they own and their profits. The term large/small is used when referring to the developer-agent's financial state. The model focuses on the possible advantage that large developers have over smaller ones in their interaction with the planner and during the competition for lands, and studies the consequences of this advantage.

The Landscape The landscape of the model is constructed from square cells representing land parcels. The center of the city is situated in the middle of the landscape. A land-use plan divides the landscape into two zones: *urban* and *non-urban*. Initially, the *urban zone* covers the area around the city center, beyond that is the *non-urban* zone. Within the *urban zone* housing construction is allowed in accordance with development rights as specified by the plan for each parcel. Over this virtual landscape the developer-agents operate and their main activity is land purchase and housing construction. The land parcels are initially non-built; as a result of developers' actions they change state (Fig. 6.1). After the parcels are purchased by developers they are constructed and the housing is sold and populated. As the simulation progresses the city develops.

Land-Use Plan The land-use plan is established by a planner for several years ahead and defines an *urban zone* where construction is permitted during the planned period. Each parcel in the urban zone is assigned with building right that determines the number of floors allowed for development in the parcel. Construction outside the urban zone, which is called the *non-urban zone,* is prohibited. Every few years, the planner extends the urban zone by including a part of the non-urban zone in order to supply residential space for the growing population.

Developers The model simulates a group of *land developer-agents* operating in a city. Developer-agents purchase lands, construct buildings and sell them to residents. Developer-agents aim to maximize their profits. Their decisions regarding land investment at a certain year is influenced by their current financial state, and the outcomes of their investment decisions influence their financial state and decision-making in the future. The model follows their accumulation of wealth.

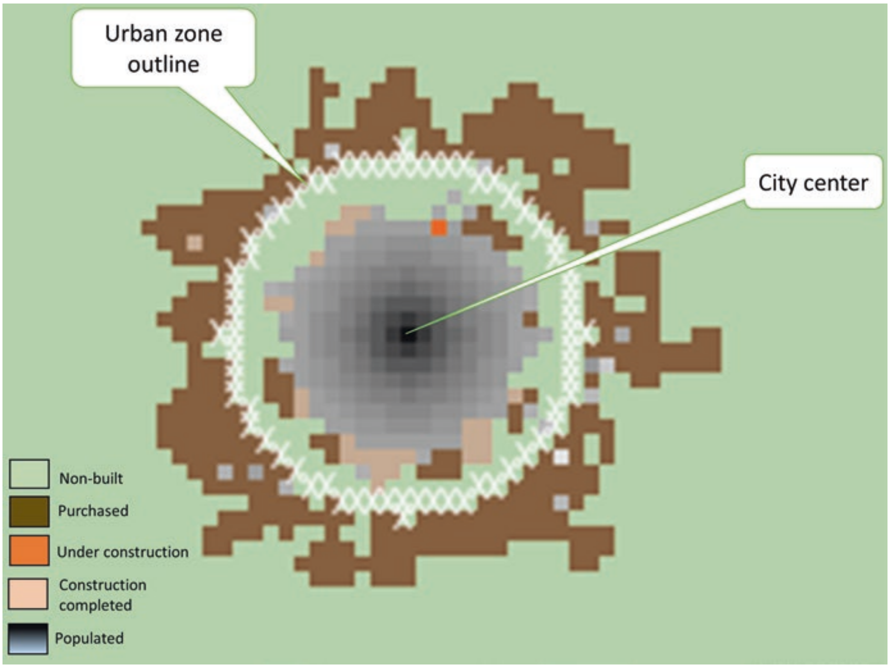


Fig. 6.1 The landscape of the ABM. Cells represent land parcels that change state as a result of developers' actions. The *gray scale* represents the height of the building, where the highest building is *black* and height declines towards *light gray*

Planner The *planner-agent* represents a municipal planning commission that establishes and manages the land-use plan. The policy of the planner-agent is to constrain outward sprawl while providing the necessary floor space for the city's growing population.

Developer-Planner Interactions The land-use plan determines development rights and, therefore, affects land values. Urban boundary (i.e. the boundary of the urban zone), is assigned by the plan and results in a decline in land price. However, following developer-agent's pressure on the planner-agent, the plan may be occasionally amended by the planner-agent using assignment of special construction permits for lands in the non-urban zone. These occasional amendments are considered by developer-agents as a source of wealth, due to the potential appreciation of excessively high land values for lands located in the non-urban zone after their development (Christensen 2014).

Developers' Decision-Making Developer-agents compare expected profits from alternate land parcels. Profit expectations vary among developer-agents and are dependent on their size (accumulated wealth). The value of parcels within the urban zone depends on their development potential, and therefore their value is high. These parcels can be developed immediately after they are acquired. The value of parcels in the non-urban zone is low and capital realization for these parcels is highly uncertain. The developer-agents will assume the risk and purchase non-urban parcels only when their expected profit, after accounting for the time and expenses spent on lobbying and obtaining construction permits, is higher than those in the urban zone.

Planner's Decision-Making The planner-agent occasionally grants a limited number of special construction permits to developer-agents who own lands located in the non-urban zone. This is assumed to be the outcome of negotiations and pressures by the developer-agents. The planner-agent's decision to grant permits is the result of the resolution of two opposing forces: the political pressure from the developer-agent to allow construction of land zoned as non-urban; and the policy guidelines to restrict development activities and prevent urban sprawl.

Size Advantage We investigate the profit accumulation dynamics in imperfect land markets. In such markets, large developers hold more knowledge on available land supply, have beneficial government regulations and are more likely to secure loans from banks. We call these conditions *size advantage* and investigate the city's dynamics depending on the potency of this advantage.

6.3.2 Comprehensive Model Description

6.3.2.1 Population Growth and Residential Demand

The dynamics of the model are driven by the growing population that results in the demand for housing. The city's population grows at a rate:

$$\mathrm{Pop}\left(t+1\right)=\mathrm{Pop}\left(t\right)^{*}\left(1+R\right)$$

Where P(t) denote population numbers in year t and R annual growth rate, normally distributed with average $R_{avg} = 0.02$ and standard deviation $R_{std} = 0.025$.

The demand for floor space D(t), at time t is:

$$D\left(t\right)=\mathrm{Pop}\left(t+1\right)-C\left(t\right)$$

Where C(t) is an amount of constructions available at the beginning of year t.

In case the completed constructions exceed the demand, completed constructions are sold in a random order until demand is satisfied, and the remaining constructions are left available for the next year.

6.3.2.2 The Land-Use Plan and the Planner

The Land Use Plan

The land-use plan divides the urban space into urban-zones and non-urban zones. Within the urban zone, parcel development is allowed in accordance with the building rights assigned by the land-use plan. These rights specify the maximal number of floors allowed for construction. It is assumed that the maximal height assigned by the plan for a parcel monotonously decreases with the increase in the distance between a parcel and the nearest center of the city (Fig. 6.2, Appendix 1).

Expansion of the Urban Zone

Once in T years, the planner-agent estimates the floor space required to accommodate expected population growth in the next T years and expands the urban zone, by including part of the non-urban area (Appendix 2). We assume that the planner-agent aims at minimizing urban sprawl and maintaining continuity of the build-up area. To implement this policy, the planner-agent seeks to include the densest build-up areas that were developed outside the urban zone into the extended urban zone (Appendix 3).

Granting Special Construction Permits

Developer-agents who possess lands in the non-urban zone exert pressure on the planner-agent to obtain special construction permissions. The planner approves some of these requests and each year issues a few special construction permits. The

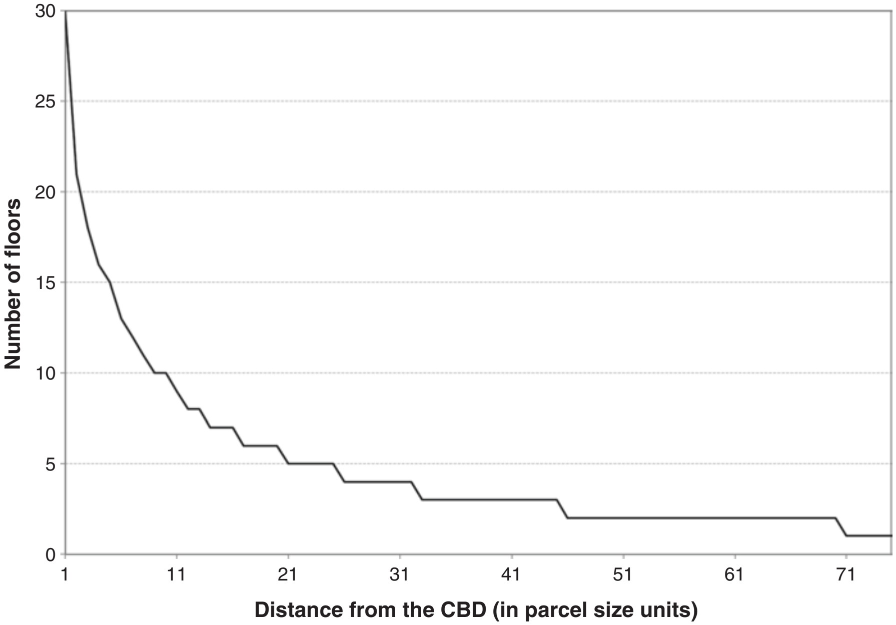


Fig. 6.2 Permitted number of floors as dependent on the distance between the parcel and the nearest center of the city

planner-agent's decision to grant a construction permit outside the urban zone is affected by two main factors:

(a) The planning policy that aims to reduce urban sprawl by permitting development adjacent to already built-up areas.
(b) The political power of the developer owning the parcel to negotiate with the planning commission.

These factors are translated into probabilities to issue special permits for each of the purchased parcels located in the non-urban zone (Appendix 4). The total amount of special construction permits granted by the planner-agent constitutes a fraction n_{sp} of the planner's estimated demand for floor space M(t) (Appendix 2).

6.3.2.3 Land Market Regulations

The land market is regulated according to the following principles:

- The total number of parcels transactions at a year t is limited by the the planner's estimated annual demand M(t) (Appendix 2).
- To reflect an anti-trust law, the planner-agent restricts annual purchases by one developer-agent to m*M(t), where 0 < m < = 1.

6.3.2.4 Behavior Procedures of the Developer-Agents

Each year t, the developer-agent chooses potentially profitable parcels in the urban zone and in the non-urban zone and competes with other developer-agents for the right to purchase them. After purchasing an urban parcel, the developer-agent begins construction immediately. For parcels located in the non-urban zone, the developer-agent must wait for a construction permit from the planner-agent. Once completing a construction, the developer-agent attempts to sell the housing.

Formally, each year t, the developer-agent implements several procedures in an order presented below (Fig. 6.3):

- Evaluates expected profit from constructing on non-built parcels within the urban and non-urban zones and chooses parcels for purchasing according to the value of her liquid assets $D_L(t)$.
- Competes for the chosen parcels with the other developer-agents and purchases them if she wins the competition.
- Starts construction on parcels she owns, for which she has construction permits. Construction lasts 1–3 years (the length of construction is equally probable).
- Sells completed buildings. Completed buildings are randomly selected for selling until the demand for floor space at time t is satisfied.
- Updates the state of her assets based on purchases and profits.

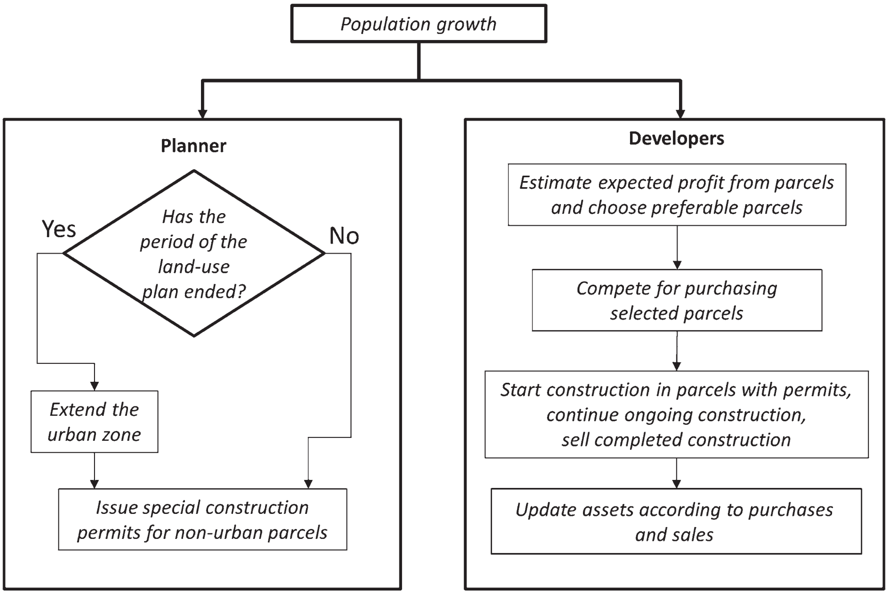

Population growth
Planner
Has the period of the land-use plan ended?
Yes
No
Extend the urban zone
Issue special construction permits for non-urban parcels
Developers
Estimate expected profit from parcels and choose preferable parcels
Compete for purchasing selected parcels
Start construction in parcels with permits, continue ongoing construction, sell completed construction
Update assets according to purchases and sales

Fig. 6.3 Planner-agent and developer-agents' decisions taken each year in reaction to the varying demand

6.3.2.5 Choice of Parcels for Purchasing

Developer-agent D is characterized by Liquid assets $D_L(t)$ and Total assets $D_T(t)$. The latter is comprised of $D_L(t)$ plus the value of parcels owned by the developer-agent at t. Developer-agent D initially allocates her investments between urban and non-urban parcels. A constant share s_{urban} of her liquid assets is invested in urban parcels, the rest, $1 - s_{urban}$ in non-urban parcels. A developer-agent invests in a non-urban parcel only if it is potentially more profitable than any parcel within the urban zone. The developer-agent's estimate of parcel's potential profitability includes an estimate of the time to realization of building permits, which is uncertain. Larger size developer-agents tend to evaluate shorter time to realization (Appendix 7). When developer-agents do not invest in non-urban parcels, all liquid assets are used to purchase parcels within the urban zone.

Developer-Agents as Satisficers We assume that the rationality of the developer-agents is limited and regard them to be satisficers who consider alternatives for which payoffs are above a certain value (Simon 1955, 1959; Daniels 1998; Mohamed 2006, 2009). Interpreting this view, we assume that the developer-agent can only roughly asses the most profitable parcel. Therefore, she chooses one of the ten most profitable parcels available, not necessarily the most profitable. Once that parcel is found, the developer-agent concentrates the rest of her purchases around it, in order to form a *continuous area*. A continuous area of parcels aims at reducing development costs in the future. *Parcels are chosen based on return on investment:* The developer-agent first chooses one parcel in the urban and one parcel in the non-urban zone. To choose she searches for one of the ten most profitable patches in each of the zones. After locating these parcels, she chooses parcels adjacent to them until their total cost reaches her investment budget (Fig. 6.4).

The profitability of choosing a parcel P is estimated by the developer-agent D based on the return on investment $\mathrm{ROI_D(P)}$:

$$\mathrm{ROI_D}(\mathrm{P}) = \mathrm{E_D}(\mathrm{P}) / (\mathrm{V}(\mathrm{P}) + \mathrm{N}(\mathrm{P}))$$

Where V(P) is the price of parcel P (Appendix 5), N(P) is the construction cost on parcel P, and $\mathrm{E_D(P)}$ is the expected profit of developer-agent D from construction on parcel P (Appendix 6). $\mathrm{E_D(P)}$ is defined by the zone, where P is located and the distance to a city center and D's estimate of the delay in obtaining construction permit. We assume that developer-agent's estimate of the delay is considered inversely proportional to the developer-agent's total assets $\mathrm{D_T(t)}$ (Appendix 7). That is, larger developer-agents are ready to take higher risks, as they are more

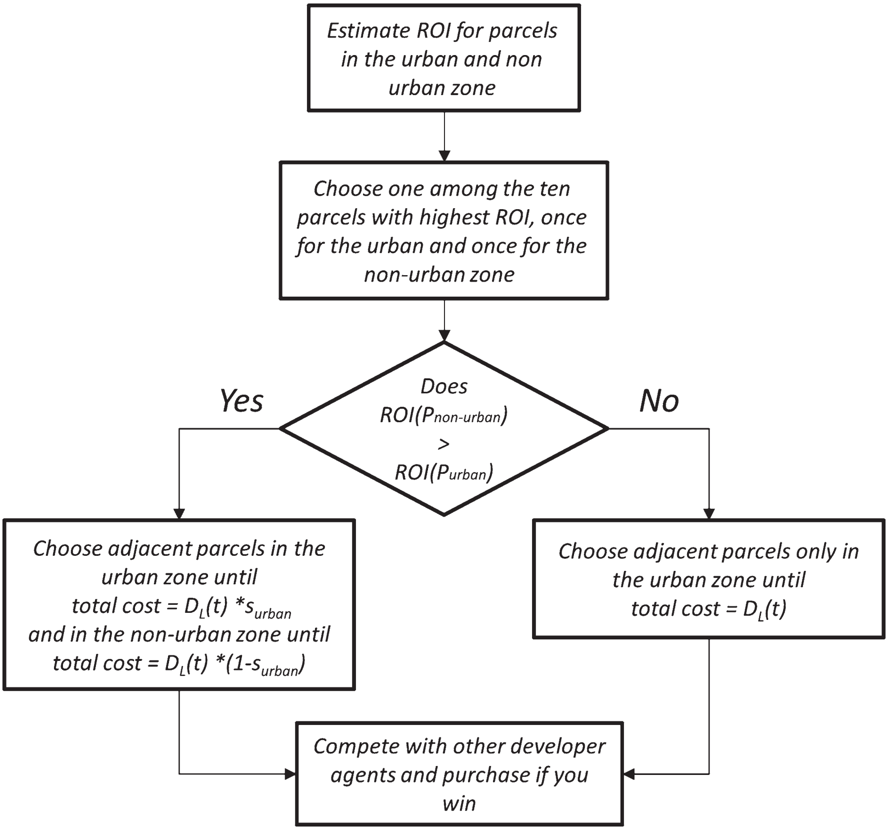

Estimate ROI for parcels in the urban and non urban zone
Choose one among the ten parcels with highest ROI, once for the urban and once for the non-urban zone
Does ROI($P_{non-urban}$) > ROI(P_{urban})
Yes
No
Choose adjacent parcels in the urban zone until total cost = $D_L(t)$ *s_{urban} and in the non-urban zone until total cost = $D_L(t)$ *(1-s_{urban})
Choose adjacent parcels only in the urban zone until total cost = $D_L(t)$
Compete with other developer agents and purchase if you win

Fig. 6.4 The decision-making process for choosing parcels by developer-agents

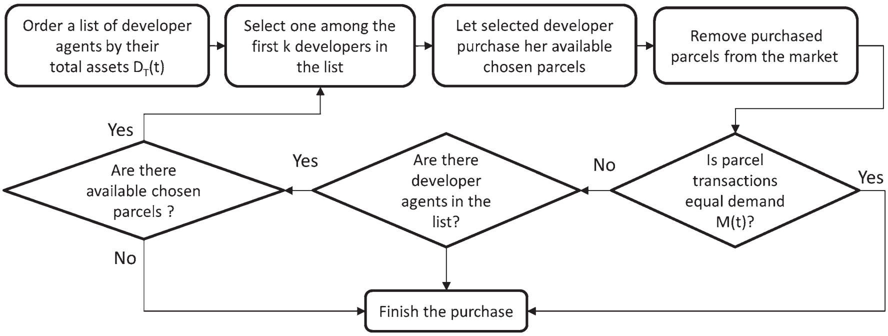

Order a list of developer agents by their total assets $D_T(t)$
Select one among the first k developers in the list
Let selected developer purchase her available chosen parcels
Remove purchased parcels from the market
Is parcel transactions equal demand M(t)?
No
Yes
Are there developer agents in the list?
Yes
Are there available chosen parcels ?
Yes
No
Finish the purchase

Fig. 6.5 Procedures for competition over land purchases

experienced in negotiating with planning committees and have a greater influence on the decisions of the planner-agents.

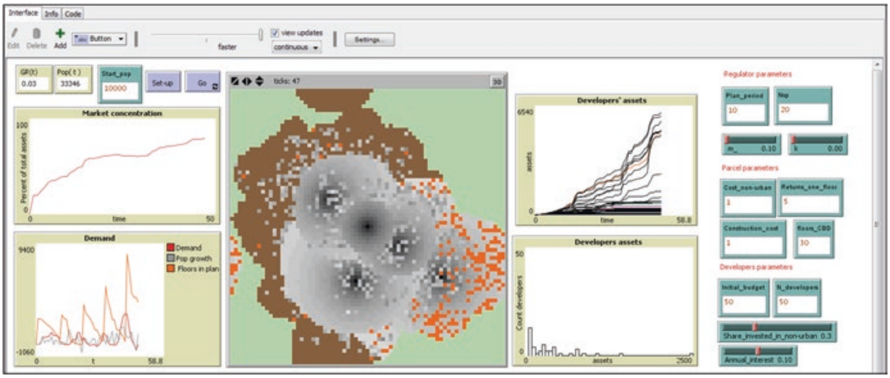
6.3.2.6 Competing for Land Purchase

After developer-agents have selected their preferable parcels, they compete for purchasing them. A developer-agent's success in competing for parcels is determined by her total assets $D_T(t)$.

At the beginning of the competition developer-agents are ordered according to their total assets $D_T(t)$ (Fig. 6.5). Parameter k reflects the nature of completion in the market. When k = 1, the competition in the land market is completely imperfect and the purchase order is completely dependent of the size of the developers. While with the increase in k the market becomes more competitive (Appendix 8). The competition starts as one among the first (k) largest developer-agents is selected, equally probable, and can purchase her chosen parcels as long as market regulation limits are not reached. The purchased parcels are removed from the market, and one of the (k) largest among the remaining developer-agents is selected to purchase parcels.

6.3.2.7 Framework of Implementation

The model is implemented with Netlogo, a multi-agent simulation environment (Wilensky 1999; Sklar 2007). The user interface includes a parameter setup, a map of parcels state, temporal charts representing the value of assets by developer-agents and the distribution of developer-agents by total assets. Several aggregate indicators are also plotted, such as market concentration expressed by the share of assets owned by the ten largest developer-agents; population growth; unsatisfied demand; and available floor space in the urban zone (Fig. 6.6).



GP(t) 0.03
Pop(t) 33346
Start_pop 10000

Regulator parameters
Plan_period 10
Nsp 20
m_ 0.10
k 0.00

Parcel parameters
Cost_non-urban 1
Returns_one_floor 5
Construction_cost 1
Room_CBD 30

Developers parameters
Initial_budget 50
N_developers 50
Share_invested_in_non-urban 0.3
Annual_interest 0.10

Fig. 6.6 User interface (Netlogo)

6.4 Method of Analysis

The ABM investigates the evolution of assets accumulation by the land development industry and the formation of city structure under various conditions of regulation and competition in the land market. Three types of conditions are investigated in the simulation scenarios:

Flexibility of the planner: Represented by a parameter that controls the number of special construction permits approved by the planner each year for development of land that is not assigned for development by the original land-use plan.

Regulation of the land market: Represented by a parameter that restricts the amount of land permitted for purchasing by a single developer each year.

Competition in the land market: Represented by a parameter that controls the level of advantage that larger (wealthier) developers have in the competition for purchasing land.

All model runs for a certain set of parameters are repeated 100 times and the results are averaged. The simulation is run for a period of 50 years in each scenario. The parameters' values used in the model are presented in Table 6.1. The values presented are kept constant in all simulation scenarios. Parameters values vary according to the scenario being investigated and are specified in the context of each scenario.

The industry structure that emerges under different scenarios is presented in a rank-size form. Developer-agents are sorted in descending order of their assets and presented as a function of their rank.

Our benchmark, or "null hypothesis", represents a perfectly competitive market. In such a market there are no abnormal profits and accumulation of wealth. To illustrate this we sampled 50 observations from uniform distribution, on a [50, 1500] interval, Fig. 6.7. The density of such a rank-size distribution is a linear decreasing function on the [1, 50] interval.

Table 6.1 Parameter values common for all simulation scenarios

Parameter	Value	Description
P(0)	10,000	Initial city population
N	50	Number of developer-agents
$D_L(0)$	50	Developer-agent's initial liquid assets, equal for all developer-agents
$s_{non\text{-}urban}$	30%	The share of developer-agent's liquid assets used for purchasing non-urban parcels
α	0.1	Annual interest rate for developer-agent's alternative investment
T	10 years	The length of the planned period, in years
R	2%	Annual population growth rate
g_{CBD}	5	Returns for selling one floor in the city center
P_{CBD}	30	Number of floors permitted for construction in the city center
$c_{non\text{-}urban}$	1	Price of a parcel assigned by the plan as non-urban
n_{sp}	Varies	Percentage of current demand for housing that is issued with special construction permits every year
m	Varies	Maximal share of current demand for floor space that one developer-agent can purchase in the urban and non-urban zones every year
k	Varies	Number of the largest developer-agents that have identical advantage when purchasing parcels
Grid	70*70	Size of the model space (4900 parcels)

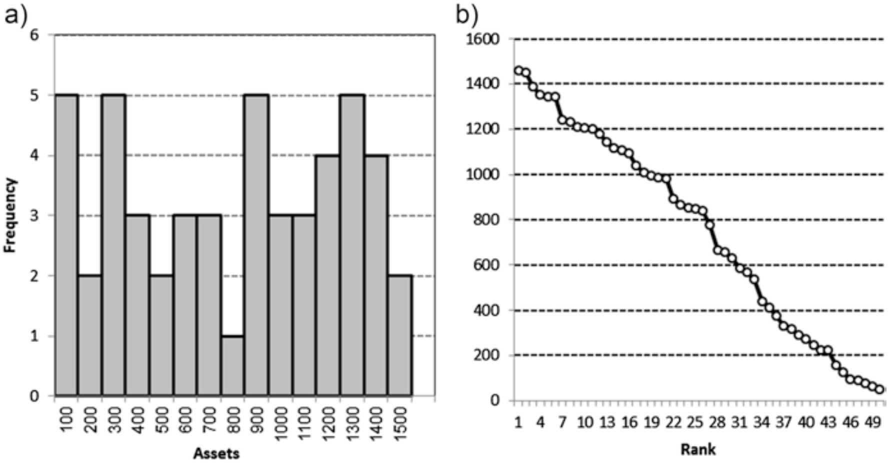


Fig. 6.7 A sample of 50 observations from a uniform distribution on [50, 1500], represented in a standard form (**a**) rank-size form (**b**)

For measuring the difference between the rank-size distribution of developer-agents derived by the different simulation scenarios and the distribution presented above, we employ the following U-measure:

$$\mathrm{U}_{\mathrm{k}} = \sum_{r=1}^{n} \left(\left| D_r\left(scenario\right) - D_r\left(uniform\right) \right| \right) / N$$

Where D_r*(scenario)* is the size of a developer-agent of rank r in the simulation scenario and D_r*(uniform)* is the size of a developer-agent of rank r in the uniform distribution.

6.5 Results

6.5.1 The Effects of Size Advantage when Purchasing Parcels

To begin with, there are two polar scenarios of market conditions. In the first, size does not influence the developer-agent's chance to purchase attractive parcels. Formally (see Sect. 3.2.6), this is the scenario of k = N. The second scenario represents the opposite case of market conditions that provides absolute *advantage of size*, k = 1. We employ the values of n_{sp} = 20% and m = 0.1 in both.

In the first scenario, developer-agents' assets at t = 50 vary between 0.3% and 4.6% of the total assets of all developer-agents and developer-agents' size distribution remains uniform. This distribution represents an absolute competitive market.

In the second scenario, the distribution of assets at t = 50 is bimodal (Fig. 6.8, Table 6.2) and the ratio of the amount of assets of the largest to that of the median developer-agent is 23.4, which is essentially higher than the value 2.1, characteristic of the uniform distribution. The development industry in this scenario is controlled by ten developer-agents who hold 77.6% of the total assets, compared to 30.3% obtained in the first scenario (Fig. 6.8, Table 6.2).

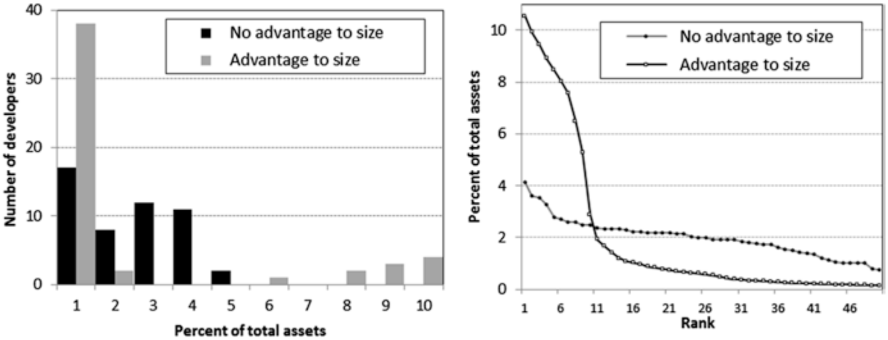


Fig. 6.8 Developer-agents' rank-size distribution at t = 50, based on absolute advantage to size (k = 1) and for no advantage to size (k = 50)

Table 6.2 Characteristics of the rank-size distributions of developer-agents by their assets at t = 50 for the cases of no advantage to size (k = 1), and of absolute advantage to size (k = 50)

Index	No advantage of size (k = 50)	Absolute advantage of size (k = 1)
Percent of assets owned by 10 largest developer-agents	30.3%	77.6%
Ratio of the largest/median	2.1	23.4
Comparison to the uniform distribution, χ^{50} and p	$\chi^{50} = 9.1$ (p ~ 0.17)	$\chi^{50} = 131.4$ ($p < 0.0001$)

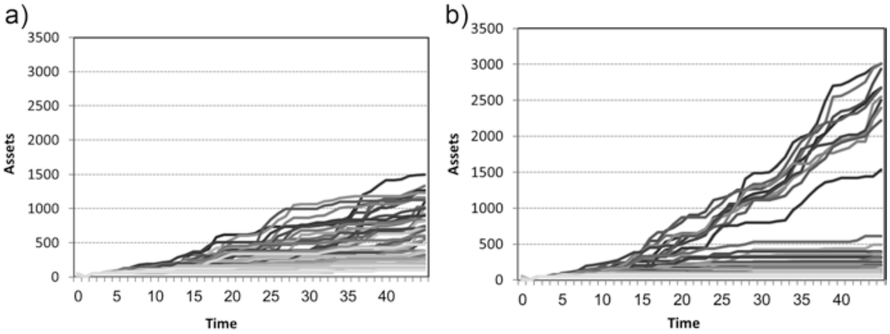


Fig. 6.9 Dynamics of asset accumulation, for each one of 50 developer-agents (each line represents a developer-agent).no advantage of size (k = 50) (**a**) advantage of size (k = 1) (**b**)

6.5.2 Dynamics of Assets Accumulation

Over time, in the case of market conditions with *no advantage of size* (k = N), developer-agents' size distribution remains uniform all the time (Fig. 6.9a). This is not so in the scenario of absolute advantage of size (k = 1). Time-evolution of developer-agents' size in this scenario exhibits two phases (Fig. 6.9b). During phase 1 (up to the year 10), developer-agents remain similar in size and their size distribution remains close to uniform. Phase 2 starts when one or more developer-agents, by chance, become significantly larger than the rest. From that time on, these developer-agents have the advantage in purchasing land and, as a result, increase their assets faster than the smaller developer-agents. This positive feedback results in the separation between the large and the small developer-agents towards t ~ 20. The full control of large developer-agents over the land market is eventually reached towards t = 50.

6.5.3 Model Sensitivity to Competition Over Land

In what follows, developer-agents' size distribution is simulated under different market conditions that vary in the competition level (k) of having an *advantage of size* when purchasing land. Figure 6.10 and Table 6.3 present developer-agents'

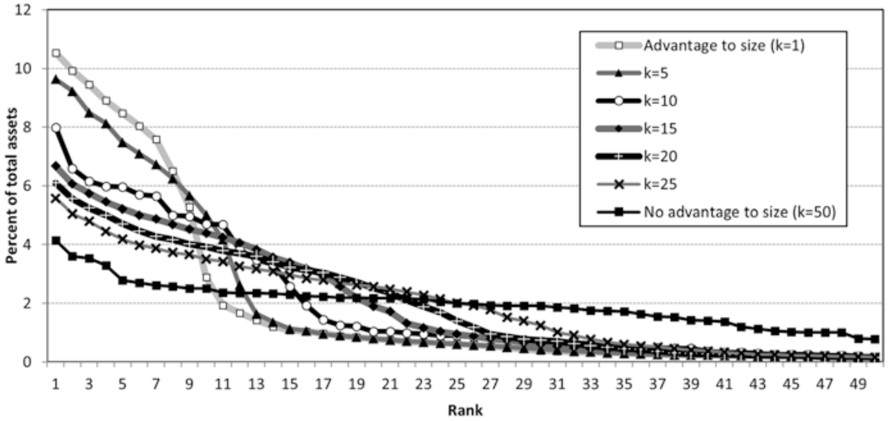


Fig. 6.10 Developer -agents' rank-size distribution at t = 50 as dependent on the uncertainty in purchasing as expressed by parameter k

Table 6.3 Characteristics of the developer-agents' rank-size distributions for different levels of uncertainty of having size advantage (k) when purchasing parcels

Index	k = 1	k = 5	k = 10	k = 15	k = 20	k = 25	k = 50
Number of large developer-agents	10	12	15	18	21	25	–
Total share of large developer-agents	77.6%	80.5%	76.6%	80.5%	81.9%	83.5%	–
Largest/median developer-agent ratio	17.5	16.3	9.1	7.3	4.7	2.8	2.1
Assets owned by 10 largest developer-agents	77.6%	73.7%	58.7%	52.6%	47.5%	42.8%	30.3%
U_k measure	1.89**	1.82**	1.39*	1.27	1.07	0.8	0.43

Mann-Whitney test, comparison to the uniform distribution, significance levels: ** – 0.0001, * – 0.01

rank-size distributions and their aggregate characteristics as dependent on k preserving the values of n_{sp} = 20% and m = 0.1.

We classify as "large" the developer-agents who are larger than the developer-agents of the same rank in the case of the uniform distribution, when k = 50 (see Fig. 6.10). The influence of uncertainty is summarized in Table 6.3.

As can be seen from Table 6.3, with the increase in the level of uncertainty the distribution of developer-agents' size converges to uniform distribution. However, this convergence is non-linear, the difference between the developer -agents' distribution and the uniform remains high up to k = 5 and then drops and decreases with k linearly. Note that the total share of large developer-agents rises by 6% only between k = 1 and k = 25, while the number of large developer-agents more than doubles.

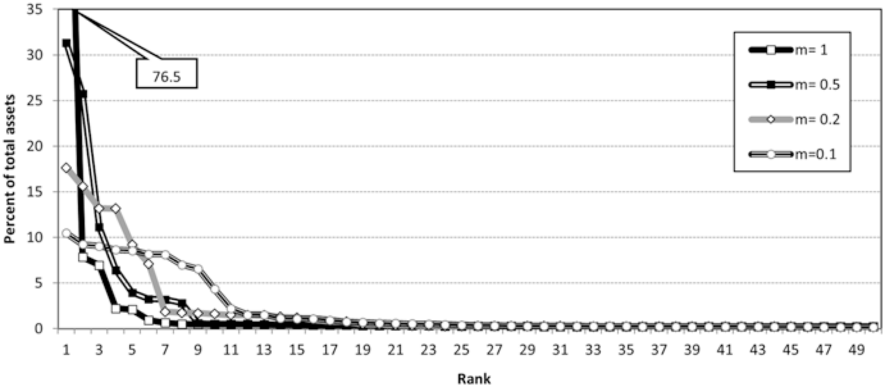


Fig. 6.11 Assets distribution at t = 50 as dependent on maximal market share m

6.5.4 Restrictions on Land Purchases

Restrictions on land purchases evidently affect the potential of a developer-agent to accumulate assets. Such restrictions are represented in the model by m –the maximal share of total demand for floor space that one developer-agent is allowed to purchase per year. To explore the influence of m on the developer-agents' rank-size distribution, we compared model outcomes for the scenario of absolute advantage of size (k = 1) for four values of m – 0.1, 0.2, 0.5 and 1 and n_{sp} is kept 20%. The resulting rank-size distributions of the developer-agents' assets are presented in Fig. 6.11.

According to the Fig. 6.11, the higher m, the lower the number of large developer-agents and the "wealthier" the largest developer-agent. This phenomenon can be expected: Given m, the entire unsatisfied demand is divided every year between 1/m largest developer-agents. That is, the higher m, the higher the advantages of the large developer-agents and the lower the chances of small developer-agents for growth. Generally, the value of 1/m defines essential model invariants. For example, the total assets of the 1/m large developer-agents at t = 50 is close to the 75% of the total assets in the city (Table 6.4).

6.5.5 Size Distribution and the Flexibility of the Planner

The flexibility of the planner-agent is represented in the model by the number of special permits she approves annually- n_{sp}. The more special construction permits issued by the planner-agent, the more certain the investment in purchasing parcels located in the non-urban zone. To investigate dependence of developer-agents' distribution by size on the number of special permits, we compare scenarios in which the annual number of special construction permits differs. The number of special

Table 6.4 Percent of total assets owned by the 1/m largest developer-agents, for different values of m

Maximal market share that can be purchased by a single developer – m	100%	50%	20%	10%
Percent of assets owned by 1/m developer-agents	76.5%	73%	73.4%	77.6%

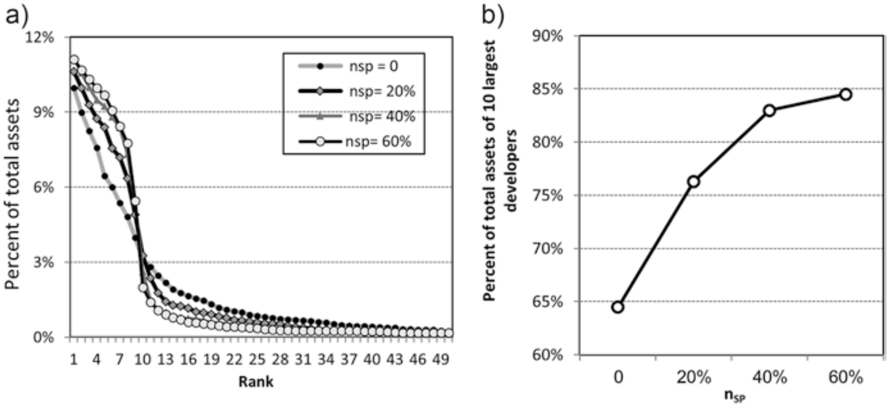


Fig. 6.12 Rank-size distribution at t = 50 (**a**) and the percent of total assets at t = 50 held by 10 largest developer-agents (**b**) as dependent on number of special construction permits n_{sp}

construction permits is measured as the percentage n_{sp} of the current demand for floor space. Four scenarios of n_{sp} = 0%, 20%, 40% and 60% are investigated for the case of absolute advantage of size (k = 1) and m = 0.1.

As seen in Fig. 6.12a, the higher the share of special construction permits, the sharper the distinction between the groups of large and small developer-agents, and the higher is the share of the total assets held by the ten largest developer-agents. This share depends on n_{sp} non-linearly and the overall share of total assets owned by ten largest developer-agents seems to stabilize at a level of 85% as far as the amount of special permits reaches 60% (Fig. 6.12b).

6.5.6 *The Dynamics of the City Pattern*

We illustrate here the impact of two planning policies on the spatial urban pattern. The first policy establishes a land-use plan with a planning horizon of 10 years, which is modified at the end of that period with respect to the population forecast for the next decade. The second policy has a long-term planning horizon and establishes a land-use plan for the whole simulation period based on a population forecast. We model the city pattern for a period of 50 years for both policies and present the results of three simulation runs for each of the scenarios. In both scenarios the advantage of size is absolute (k = 1), the maximal market share that can be

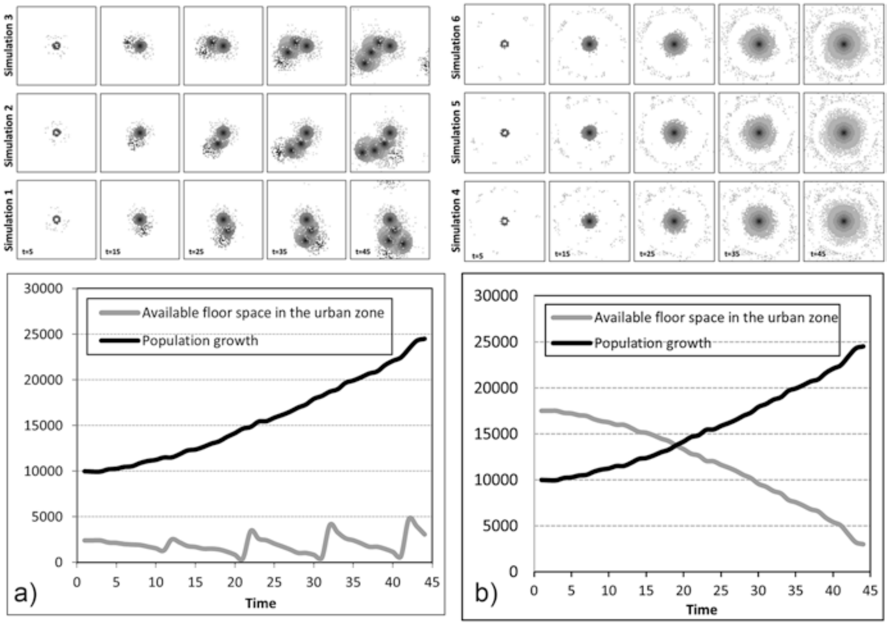


Fig. 6.13 Dynamics of the city pattern, total population and planned floor space in the urban zone in case the plan is modified every 10 years (**a**) and a long-term plan (**b**). *Darker color* marks higher buildings in the parcel

purchased by one developer (m = 0.1) and the fraction of special permission (n_{sp} = 20%).

As seen in Fig. 6.13a, plan extension every 10 years essentially influences urban pattern dynamics. The location of new urban centers reflects the history of land purchases in the non-urban zone and every simulation produces a different urban pattern. In contrast, in all three simulations, the long-term plan produces the same mono-centric city (Fig. 6.13b). Development takes place from the center out towards to the fringes, with scattered development outside the urban zone, which is the outcome of Special construction permits.

6.5.7 *Qualitative Comparison to Real Data*

Our model is a stylized theoretical exercise. However, the model's distributions of developer-agents' assets (Fig. 6.14a) obtained in two scenarios of absolute advantage of size (k = 1), and a high maximal share of parcels available for purchasing by one developer-agent, m = 1 and m = 0.5, strongly resemble the size distribution of Israeli land development firms (Fig. 6.14b). Different from reality, the model predicts a larger share for the largest developer(s). A possible reason for this difference is that, in reality, large development firms are involved in other fields of business

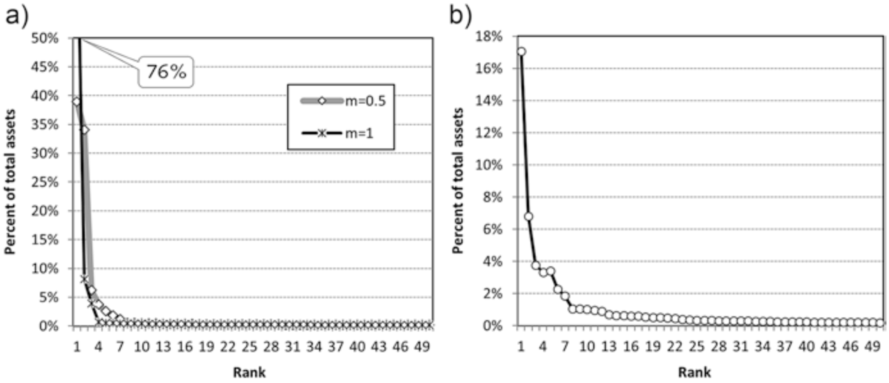


Fig. 6.14 Rank-size distribution of developer-agents' assets at t = 50 obtained in scenarios of k = 1, n_{sp} = 20% and m = 1 and 0.5 (**a**) versus rank-size distribution of development firms in Israel in 2013 (Ranking of Construction and Development 2013) (**b**)

and divert essential share of their profits from land development into other activities. With that, the qualitative resemblance of the results indicates that market conditions in Israel provides advantage of size that result in imperfect market.

6.6 Conclusions

By means of an agent-based model, we study the evolution of the land development industry in a city regulated by a land-use plan, under different levels of competition for purchasing land. Starting with developers endowed with equal assets, the differences in accumulation of profits result in increasingly non-uniform distribution of developers' size. The model allows larger developers to assume greater risks and amass even greater assets. The risks in the model are associated with the nature of urban zones where land development requires exertion of political pressure and negotiation that at times yields building permits and abnormal profits. This spatial context of profit accumulation contributes to the emergence of an oligopolistic industry structure.

The model includes a dynamic land-use plan, managed by a planner-agent, aiming to maintain excess supply of planned land for residential development. On the developer's side, investment within the planned urban zone is certain but when compared to lands in the non-urban zone, in the long run, it can be less profitable. However, the investment in non-urban zone is less certain and demands plan modification, either by the expansion of the planned urban zone or by special construction permits. Pressures imposed by developer-agents on the planner-agent can lead to modifications of the land-use plan by obtaining construction permits outside the planned urban zone. In time, these advantages result in abnormal profits for some developers and influence the evolution of the land development industry and formation of city patterns.

Like any other, our model is a simplification of reality, which intends to explore certain aspects of a phenomenon based on stated assumptions (Benenson and Torrens 2004). Our ABM uses a simplified space comprised of square parcels of identical size, which are differentiated based on distance from the urban centers. Real cities are much more complicated and include many factors that characterize parcels and influence their demand and price. With that said, we intend to reveal the major processes that affect the land development industry. A simple representation of space enables the isolation of these processes without the "noise" created by random variations of other factors that may change from city to city.

Our model keeps land and housing prices stable throughout the simulation process and ignores the boom and slump cycles that may occur over long periods in the market (Gillen and Fisher 2002). We are not certain to what extent these cycles affect the system dynamics and we intend to incorporate them in future research.

Our model demonstrates the evolution of a monopoly and oligopoly in an imperfect market. In imperfect markets, where larger developers both have an advantage in competing for land and are more likely to obtain special construction permits from the planning authorities, a positive feedback is generated. This feedback causes increasing returns for some developers and, over the long run, a divergence of the development industry into two distinct groups of large and small developers. This divergence does not emerge in markets where land purchase is competitive and do not entail advantage of size to some developers. In that case, the outcome is a uniform distribution of developers' assets that is preserved over time.

Even when developers are initially equal in size, small initial differences in developers' profits are growing over time and lead to a situation in which some of them become larger than others. In imperfect markets, the initial divergence in size activates positive feedback and larger developers begin to exploit their size advantage by accumulating more assets and by further investing in land purchase and construction. Consequently, the level of market concentration steadily grows and eventually a few ever-growing large developers control the development industry. The effect of the level of competition in the land purchasing market is non-linear, and in order to achieve a competitive land development industry, competitiveness in the land purchasing market should be significantly high. That is, market conditions should not favor larger developers over smaller developers when competing for land.

Planners' approval of special permits for construction in the non-urban zone strengthen the positive feedback and increase market concentration. Special construction permits are more beneficial to large developers and therefore accelerates the convergence of the industry into being controlled by a few large developers who control larger parts of the market.

Recent studies on agent-based models demonstrate the influence of agents' heterogeneity on the formation of urban spatial pattern (Ligmann-Zielinska 2009; Irwin 2010; Hatna and Benenson 2012; Broitman and Czamanski 2012; Huang et al. 2013b). Our study specifies the process of emergence of agent heterogeneity and the resulting pattern of city development. Agent heterogeneity emerges and increases over time, as developers receive profits from construction and their economic state differentiates. We also demonstrate the importance of the planner – developer interactions when considering urban patterns. The urban development plan aims at governing developers' decisions regarding land purchases and construction. However, planning policy requires flexibility, and plan's modifications entail a co-adaptation of the planner and the developers. As we demonstrate, these processes are history-dependent and thus, hardly predictable. These processes give rise to qualitative consequences on the city structure. The interaction between risk taking developers and a flexible planner who approves incremental amendments and periodic updates of the land-use plan may result in bifurcations of city structure, which leads to a polycentric city.

Finally, our model suggests three policy implications. First, the efforts of planners to prevent sprawl by issuing land-use plans can be counterproductive when land can be rezoned within a reasonable time frame. Land-use plans create discontinuity in land prices, making land not zoned for development, such as agriculture, nature and open space, attractive to developers who foresee the potential for rezoning. Second, when rezoning is widely used by planners, it creates windfall profits for developers willing to take risks, and therefore reinforces the establishment of large development firms. Third, the land development industry has a high tendency to produce oligopoly and monopoly without any regulations. Therefore, in order to prevent such market failure, strong intervention by regulation in the land purchasing market is required.

Appendix

1. The amount of construction permitted on a parcel P is expressed by the number of floors F(P). Within the urban zone, F(P) monotonously decays with the increase in the distance Dist(P, P_{CBD}) between parcel P and the nearest central parcel P_{CBD} as:

   $$F(P) = \max\{1, \mathrm{INT}[F(P_{CBD})/(1 + \delta^{*}\mathrm{Dist}(P, P_{CBD}))]\}$$

   Where $F(P_{\mathbf{CBD}})$ is the number of floors in the nearest central parcel and INT(X) is a closest integer to X. In what follows, the value of $F(P_{\mathbf{CBD}})$ is set, for all central parcels, equal to 30 floors. To guarantee that the minimal possible height of 1 floor will be reached far away from the center, the value of δ is chosen equal to 0.2, providing the distance from the city center equal to 70 parcels.
2. To evaluate expected demand (ED) for floor space for the next planning period of T years, the planner-agent estimates population growth and assesses the currently available floor:

   $$ED(t + T) = Pop(t)^{*}((1 + R_{avg})^{T}) + A(t) - C(t)$$

   Where C(t) is the amount of construction available at year t, and A(t) is the potential floor space, yet developed, in the current urban zone at t.

   The demand for floor space M(t) is estimated by the planner based on the yearly average demand in the next 3 years:

   $$M(t) = ((Pop(t)^{*}((1 + R_{avg})^{3}) - C(t) - L(t))/3$$

   Where L(t) is the floor space that is currently under construction.

3. The density of built area in the non-urban zone is estimated as a moving average of the buildings' height within a circle of a 6-unit radius around a parcel P. The location, with the highest density is considered as a *new central parcel*. The planner-agent aims at including central parcels into the plan when extending the plan. To ensure plan's contiguity, the extension includes, together with the central parcel, all parcels within a buffer zone of a shortest path between the central parcel and the existing urban zone. The width of the buffer zone is increased, until the accumulated amount of floor space is sufficient to accommodate the expected demand ED(t + T).

   After the plan is expanded, the *central parcel* is assigned with a permitted number of floors for construction equal to $F(P_{CBD})$ and becomes an anchor for calculating the permitted floors for construction for the parcels in its vicinity (as in formula 1). If the highest built density is equal in a few parcels, the one closer to the urban zone is selected.
4. The probability to obtain a special permit is based on the density of buildings within a 6-unit neighborhood around the parcel and the relative wealth of the developer-agent D who owns the parcel. The higher is the density of construction around a parcel and the higher is the wealth rank of the parcel's owner the higher is the relative weight that the parcel will be granted a permit for construction. A Special construction permit assigns to the non-urban parcel amount of floors for construction F(P) according to a normal distribution, with mean equal to $F(P_{CBD})/3$ and STD is equal to 2.
5. To determine parcel price in the model, we follow the Ricardian rent theory and consider, land prices as residuals of the housing prices relative to construction costs (Ball 1983).

   The market price $V(P_{urban})$ of an urban parcel P_{urban} is defined as a fraction r of expected revenue G (P_{urban}) from selling the construction. In what follows, we assume that G (P_{urban}) is proportional to the number of floors in the construction and inversely proportional to the distance between P and the city center:

$$G\left(P_{urban}\right) = g_{CBD}{}^{*}\mathrm{Dist}\left(P_{urban}, P_{CBD}\right)^{-\beta *} F\left(P_{urban}\right)$$
$$V\left(P_{urban}\right) = r^{*}G\left(P_{urban}\right)$$

   where $\mathbf{g_{CBD}}$ is the return per one floor in the center of the city, $F(P_{urban})$ is given by (1), the power β = −0.1 is a decay of the price away from the nearest central parcel and r is uniformly distributed on [0.3, 0.5].

   The market price of all non-urban parcels $P_{non\text{-}urban}$ is constant:

$$V(P_{non\ -\ urban}) = v_{non\ -\ urban}$$

   where $v_{non\text{-}urban}$ is a value which is below the price of any parcel within the urban zone.
6. For a parcel P_{urban} in the urban zone, expected profit $E_D(P_{urban})$ is estimated as:

$$E_D(P_{urban}) = G(P_{urban}) - V(P_{urban}) - N(P_{urban})$$

   where $G(P_{\mathbf{urban}})$ is the expected revenue from selling the construction, $V(P_{\mathbf{urban}})$ is the price of the parcel and $N(P_{urban})$ is construction cost at P_{urban}.

In what follows, we assume that the construction cost of one floor is the same in all parcels, where it is equal to 20% of the returns from selling one floor $\mathbf{g_{CBD}}$ in the center of the city, see (appendix 5)

$$N(P) = 0.2^{*}g_{CBD}{}^{*}F(P)$$

The developer-agent's D estimate of profit from a non-urban parcel $\mathbf{P_{non\text{-}urban}}$ accounts for the possible delay in capital realization and therefore accounts for opportunity cost:

$E_D(P_{non-urban}) = G(P_{non-urban}) - V(P_{non-urban})^{*}(1+\alpha)^{\tau D(t)} - N(P_{non-urban})$ where α is the annual interest rate from an alternative investment, and $\tau_D(t)$ is developer-agent's D estimate of the length of the delay necessary to obtain a special construction permit at $\mathbf{P_{non\text{-}urban}}$.

7. The delay in obtaining construction permit $\tau_D(t)$ is calculated as:

$$\tau_D(t) = \mathrm{Int}(2^{*}r_D(t)^{0.5})$$

where $r_D(t)$ is a rank of developer-agent D at t according to her total assets $D_T(t)$.

8. The higher k is, the higher the uncertainty of the outcome of the purchasing process. For k = 1 the developer-agent with the largest total assets $D_T(t)$ will always be first to purchase, then the second largest will enter the bid to purchase, etc. For k ~ N/2 any of the developer-agents whose assets are higher than the median assets, can be the first. For the case k = N, there is no advantage of size.

References

Abrantes P, Fontes I, Gomes E, Rocha J (2016) Compliance of land cover changes with municipal land use planning: evidence from the lisbon metropolitan region (1990–2007). Land Use Policy 51:120–134

Alfasi N (2006) Planning policy? Between long-term planning and zoning amendments in the israeli planning system. Environ Plan A 38(3):553

Alfasi N, Portugali J (2004) Planning just-in-time versus planning just-in-case. Cities 21(1):29–39

Alfasi N, Portugali J (2007) Planning rules for a self-planned city. Plan Theory 6(2):164–182

Alfasi N, Almagor J, Benenson I (2012) The actual impact of comprehensive land-use plans: insights from high resolution observations. Land Use Policy 29(4):862–877. doi:http://dx.doi.org.rproxy.tau.ac.il/10.1016/j.landusepol.2012.01.003

Ball M (1983) Housing policy and economic power the political economy of owner occupation. Methuen

Ball M (2003) Markets and the structure of the housebuilding industry: an international perspective. Urban Stud 40(5–6):897–916. doi:10.1080/0042098032000074236

Benenson I, Torrens PM (2004) Geosimulation: automata-based modeling of urban phenomena. Wiley, Chichester

Booth P (1995) Zoning or discretionary action: certainty and responsiveness in implementing planning policy. J Plan Educ Res 14(2):103–112. doi:10.1177/0739456X9501400203

Booth P (2003) Promoting radical change: the loi relative a 'la solidarite' et au renouvellement urbains in france. Eur Plan Stud 11(8):949–963. doi:10.1080/0965431032000146141

Broitman D, Czamanski D (2012) Polycentric urban dynamics – heterogeneous developers under certain planning restrictions. J Urban Reg Inf Syst Assoc 24(1)

Buitelaar E, Galle M, Sorel N (2011) Plan-led planning systems in development-led practices: an empirical analysis into the (lack of) institutionalisation of planning law. Environ Plan A 43(4):928

Buzzelli M (2001) Firm size structure in north american housebuilding: persistent deconcentration, 1945–98. Environ Plan A 33(3):533–550

Christensen FK (2014) Understanding value changes in the urban development process and the impact of municipal planning. Land Use Policy 36(0):113–121. doi:http://dx.doi.org.rproxy.tau.ac.il/10.1016/j.landusepol.2013.07.005

Coiacetto E (2009) Industry structure in real estate development: is city building competitive? Urban Policy Res 27(2):117–135

Czamanski D, Roth R (2011) Characteristic time, developers' behavior and leapfrogging dynamics of high-rise buildings. Ann Reg Sci 46(1):101–118

Dalton LC, Conover M, Rudholm G, Tsuda R, Baer WC (1989) The limits of regulation evidence from local plan implementation in california. J Am Plan Assoc 55(2):151–168. doi:10.1080/01944368908976015

Daniels T (1998) When city and country collide: managing growth in the metropolitan fringe. Island Press, Covelo

Ettema D (2011) A multi-agent model of urban processes: modelling relocation processes and price setting in housing markets. Comput Environ Urban Syst 35:1–11

Gillen M, Fisher P (2002) Residential developer behaviour in land price determination. J Prop Res 19(1):39–59

Hatna E, Benenson I (2012) The schelling model of ethnic residential dynamics: beyond the integrated – segregated dichotomy of patterns. J Artif Soc Soc Simul 15(1):6

Huang Q, Parker DC, Filatova T, Sun S (2013a) A review of urban residential choice models using agent-based modeling. Environ Plan B Plan Des 40

Huang Q, Parker DC, Sun S, Filatova T (2013b) Effects of agent heterogeneity in the presence of a land-market: a systematic test in an agent-based laboratory. Comput Environ Urban Syst 41(0):188–203. doi:http://dx.doi.org.rproxy.tau.ac.il/10.1016/j.compenvurbsys.2013.06.004

Irwin EG (2010) New directions for urban economic models of land use change: incorporating spatial dynamics and heterogeneity*. J Reg Sci 50(1):65–91

Janssen-Jansen LB, Woltjer J (2010) British discretion in dutch planning: establishing a comparative perspective for regional planning and local development in the Netherlands and the United Kingdom. Land Use Policy 27(3):906–916

Ligmann-Zielinska A (2009) The impact of risk-taking attitudes on a land use pattern: an agent-based model of residential development. J Land Use Sci 4(4):215–232

Magliocca N, Safirova E, McConnell V, Wall M (2011) An economic agent-based model of coupled housing and land markets (CHALMS). Comput Environ Urban Syst 35:183–191

Magliocca N, McConnell V, Walls M (2015) Exploring sprawl: results from an economic agent-based model of land and housing markets. Ecol Econ 113:114–125

Mohamed R (2006) The psychology of residential developers: lessons from behavioral economics and additional explanations for satisficing. J Plan Educ Res 26(1):28–37. doi:10.1177/0739456X05282352

Mohamed R (2009) Why do residential developers prefer large exurban lots? Infrastructure costs and exurban development. Environ PlanB Plan Design 36(1):12

Molotch H (1976) The city as a growth machine: toward a political economy of place. Am J Sociol 82(2):309–332. doi:10.2307/2777096

Parker DC, Filatova T (2008) A conceptual design for a bilateral agent-based land market with heterogeneous economic agents. Comput Environ Urban Syst 32:454–463

Ranking of Construction & Development (2013) http:www.duns100.dundb.co.il.rproxy.tau.ac.il/ts.cgi?tsscript=/2013h/e40a61, Accessed 24 Apr 2014

Ruming KJ (2010) Developer typologies in urban renewal in sydney: recognising the role of informal associations between developers and local government. Urban Policy Res 28(1):65–83

Sevelka T (2004) Subdivision development: risk, profit, and developer surveys. Apprais J 72(3)

Simon HA (1955) A behavioral model of rational choice. Q J Econ 69(1):99–118. doi:10.2307/1884852

Simon HA (1959) Theories of decision-making in economics and behavioral science. Am Econ Rev 49(3):253–283

Sklar E (2007) NetLogo, a multi-agent simulation environment. Art&Life 13(3):303–311

Somerville CT (1999) The industrial organization of housing supply: market activity, land supply and the size of homebuilder firms. Real Estate Econ 27(4):669–694

Stone CN (1993) Urban regimes and the capacity to govern: a political economy approach. J Urban Aff 15(1):1–28. doi:10.1111/j.1467-9906.1993.tb00300.x

Wilensky U (1999) Netlogo. http://cclnorthwesternedu/netlogo/indexshtml. Accessed 8 June 2014

Wrenn DH, Irwin EG (2015) Time is money: an empirical examination of the effects of regulatory delay on residential subdivision development. Reg Sci Urban Econ (0) doi:http://dx.doi.org.rproxy.tau.ac.il/10.1016/j.regsciurbeco.2014.12.004

Chapter 7
Dynamic Relationships Between Human Decision Making and Socio-natural Systems

Andreas Koch

Abstract The following article presents an attempt to model and simulate processes of urban socio-spatial segregation by focusing on both the local scale of households' decision-making processes and the macro scale of institutional and market determinants. In so doing, the theoretical domain of the model's purpose is to highlight the mutual relationships between individual acting conditions and structural ordering conditions, embedded in the context of intra-urban moves. The methodological domain of the model's purpose is to model residential agents as truly *individual* units. One aim of this paper is to alter some of the agents' premises and neighborhood rules of evaluation and movement in order to strictly individualize the defined entity of households. The city of Salzburg, Austria, serves as a test bed for this approach.

Keywords Residential mobility • Individual autonomy • Institutional meaning

7.1 Introduction

The distribution of households in urban space, its underlying mechanisms, processes and spatiotemporal structures, is a complex phenomenon (see, for instance, Iltanen (2012, p. 75f) and Mitchell (2009, pp. 3ff and 145ff) as a reference for complexity science). The social and spatial patterns which arise and – subtly and dynamically – change over time, such as segregation, residential up- and down-grading, gentrification, places of inclusion and exclusion, are significantly influenced and determined by numerous rules, markets, political attitudes, and behavior of different stakeholders. In order to adequately analyze and evaluate how social and spatial forces are mutually linked attempts to model issues of households' location-allocation patterns

A. Koch (✉)
Department of Geography and Geology, University of Salzburg,
Hellbrunnerstr. 34, 5020 Salzburg, Austria
e-mail: andreas.koch@sbg.ac.at

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_7

and to simulate their processes have to explicitly take into consideration the three crucial domains of ‘scales’, ‘entities’, and ‘interactions’.

The following article presents an attempt at modeling and simulating processes of urban socio-spatial segregation by focusing on both the local scale of households’ decision-making processes and the macro scale of institutional and market determinants. In so doing, the theoretical domain of the model’s purpose is to highlight the mutual relationships between individual acting conditions and structural ordering conditions, embedded in the context of intra-urban moves. An actor-centered modeling perspective appears to be important because it makes the scope and constraints explicit and visible from the bottom up. This perspective, however, needs to be complemented by macro determinants, such as housing markets, the capitalization of these markets, estate agencies, urban planning strategies, as well as social norms and cultural attitudes towards lifestyle and neighborhood building, which affect individual decisions in a top-down manner. Putting emphasis on intra-urban moves is justified since they make up a large proportion of all residential mobility; for European cities, for instance, they vary between approx. 20 movers per 1000 inhabitants in Irish cities and up to 121 in Finnish cities, per year (Knox and Pinch 2006, p. 252; comparative data is from 1980). This remarkable range can be used as one indicator for the mutual relationship between the local-individual and global-social scale.

The methodological domain of the model’s purpose is to model residential agents as truly *individual* units. The Schelling-style segregation model serves thereby as a starting point and benchmark. One of the great benefits of this model type lies in its emphasis on emerging spatial patterns at the macro level which cannot be thoroughly explained by investigating the motives and interaction patterns at the micro level. One aim of this paper is to alter some of the agents’ premises and neighborhood rules of evaluation and movement in order to strictly individualize the defined entity of households. The problem of many Schelling-style models is that they do not consistently account for the individual but refer only to collective actions: from the initialization of the model up to the agents’ evaluation and decision to move, a homogenized procedure to set the rules has been implemented.

In the remaining part of this chapter we first take a look at some of the macro-social regularities which determine the agents’ power to act. These regularities have conceptually been put into model practice by inserting a model component that aims to represent the institutional functioning of urban planning and real estate agencies through the implementation of large housing units. This component represents the socio-natural system approach. This is followed by a discussion about the need to modify Schelling-style segregation models. An individualized model approach is then presented, describing first the agents’ characteristics and then illustrating the effects by presenting some of the relevant model’s results. Thereby the human decision process domain is incorporated.

7.2 The Macro Scale of Segregation

The “Nature of Cities”, published by Harris and Ullman in 1945, has changed economically, socially, politically, culturally, and geographically in structure, shape, meaning and function. With respect to the function of housing in general and the different socio-spatial patterns of segregation in particular several theoretical approaches have been developed, putting the macro scale forces to the fore. One early but, to date, influential approach to investigate gentrification as a specific kind of segregation (which we are interested in here) is the rent gap theory by Smith (1979, 1996). The core concern of this theory is a profit-driven economic explanation of the processes of social, cultural, and architectural upgrading accompanied by social homogenization of gentrified neighborhoods and the displacement of less affluent households. The emergence of a rent gap in city areas has been linked to suburbanization of urban agglomerations. “Inner-city decline and suburban expansion has therefore led to a **rent gap** – a disparity between the potential rents that could be commanded by inner-city properties and the actual rents they are commanding” (Knox and Pinch 2006, p. 145). Rent-gap driven socio-spatial revaluation remains an important aspect in urban redevelopment, mainly as a consequence of the contemporary financial and economic crises. This in turn means that classical models, such as the Chicago model of urban structure, are “[…] now largely redundant” (Fyfe and Kenny 2005, p. 128). On the other hand, an exclusive rent-gap perspective “[…] leaves little room for human agency or consumer preferences. Thus, by itself, the rent-gap theory cannot explain which cities, and which areas within cities, are most likely to be regenerated” (Knox and Pinch 2006, p. 146).

A different but closely related macro-scale approach, highlighting the processes of residential mobility and neighborhood change, is the concept of invasion-succession cycles (Clay 1979). It describes the invasion of low-income households, referred to as pioneers (e.g. students, artists, people with an alternative lifestyle), into obsolescent city areas. During this phase a cultural revaluation, by establishing new and alternative stores (e.g. ethnic food restaurants) and services (e.g. galleries), and a modest architectural revaluation, the so called incumbent upgrading, takes place. A second cycle is initiated by the invasion of gentrifies who prefer the cultural atmosphere but dislike the built environment. Due to large-scale and costly architectural revaluation, often realized by large property development companies, housing becomes significantly more expensive, resulting in extensive displacement of less affluent households, including the pioneers of the first phase. Although this theory represents a socio-spatial generalization of city development which is often triggered by urban planning strategies and political desires of urban uniqueness under conditions of global competitiveness, it is simultaneously understood as a bottom-up approach to explaining individual preferences. It is, however, not a strict and pervasive individual perspective, but one that takes into account communities (milieus, lifestyles) as more or less homogenous entities (see, e.g. Baumgärtner 2009, p. 66).

A further determinant influencing individual decision making and agency is given with institutional stakeholders. Dangschat (1997, p. 643; 2007), among others, proposes an integrative approach for a theory of segregation by including a theory of social inequality, a theory of spatial inequality, and a theory of allocation of living space and city districts to households and social aggregates. Apart from builders, developers, mortgage lenders, and government agencies it is the estate agents who do play a crucial role within the network of mediators between sellers and buyers or landlords and tenants. Residential property management, insurance, data collection and analysis, and financing are only a few of the tasks they undertake and contribute to housing market mechanisms. "They are not simply passive brokers in these transactions, however; they influence the social production of the built environment in several ways. In addition to the bias introduced in their role as mediators of information, some estate agents introduce a *deliberate* bias by *steering* households into, or away from, a specific neighborhood in order to maintain what they regard as optimal market conditions. […] Thus the safest response for realtors is to keep like with like and to deter persons from moving to areas occupied by persons 'unlike' themselves" (Knox and Pinch 2006, p. 143). Though real estate agents complexify segregation processes one should keep in mind that simplified decentralized bilateral trading mechanisms, too, can exhibit remarkable complexity at the individual level, as is shown by Filatova et al. (2009).

There are undoubtedly several more approaches which deal with macro implications of segregation. One may think of the theory of fragmenting development which links global trends of economic and political developments to locally fragmented but homogenous processes of residential structuring (Scholz 2002); other determinants are social housing policies and the public housing sector. The creation of large housing complexes that give rise to influences on the individual scale of households appears also to be an appropriate theoretical representative for top-down driven socio-spatial changes. Though all these criteria are relevant to understanding segregation in a more comprehensive way, it is usually difficult to represent them all in an adequate manner. Apart from data availability at high resolution and the necessity to suitably translate them into procedural code, it is sometimes hard to connect the (potential) knowledge to a valid synoptic unity (e.g. how do you weight all the information properly?). The following model incorporates the above mentioned determinants in a preliminary and approximate way by (i) inserting two large and socially different housing estates into a model representation of the city of Salzburg, Austria, and (ii) using rent price as a cumulative indicator which has been disaggregated by statistical techniques. This appears to be justified by the model's purpose, because it has its thematic priority in a comprehensive individualization of agents' agency. Before presenting some results, the theoretical counterpart of the model about incorporating actual individual agents will be discussed.

7.3 With Schelling Beyond Schelling: Individualization of Agents

Urban segregation appears to be an amalgamation of smaller-scale intra-urban moves and larger-scale in- and out-migration. Knox and Pinch (2006, p. 254ff) refer to empirical results of intra-urban mobility which allow for some generalization in the search for regularities of social homogenization at the neighborhood scale. We refer to these regularities as a coarse framing of agents' social and spatial evaluation and decision-making behavior, in addition to the macro-social influential forces mentioned above. "The most significant regularities in intra-urban movement patterns […] relate to the relative *socioeconomic status* of origin and destination areas. The vast majority of moves […] take place within census tracts of similar socioeconomic characteristics" (ibid., p. 254). This result correlates closely with the distance of moves which have been found to be comparatively short, however, varying by income, tenure, ethnic identity and suchlike. Another determinant is the distinction between voluntary and involuntary moves, which are not recognized as a sharp dichotomy in the following model (and voluntary moves are in themselves different, which causes ongoing discourses on the evaluation of 'voluntary' moves; see, e.g. Dangschat and Alisch 2012, p. 36).

The search for a new home is commonly biased, too. Criteria such as living-space, tenure, dwelling preferences, built environment, and social neighborhood all need to be considered. In addition, the search space is strongly correlated with local knowledge about urban districts, influenced by spatial activity patterns and information sources (e.g. media, friends). "It follows that different subgroups of households, with distinctive activity spaces and mental maps, will tend to exhibit an equally distinctive spatial bias in their search behaviour" (Knox and Pinch 2006, p. 262). For the following model it is assumed that agents are flexible in their search space, being able to get information about living costs (the spatial rent domain) and the social status of neighbors (the social attitude domain) (Ioannides 2013, p. 101).

Against this background the benefit of Schelling-style segregation models lies in its explicit focus on local circumstances as reasons for processes of socio-spatial homogenization. The model's purpose is directed towards the micro-macro link of individual agency (Schelling 2006, O'Sullivan and Perry 2013, p. 83). One of the most important results of Schelling's model approach is the phenomenon that, if agents act according to their subjective aspirations of residing in close proximity to other agents with equal social characteristics, a global pattern arises which cannot be derived from these aspirations automatically. The detection of emerging segregation to a much stronger extent than individually anticipated and intended is, on the other hand, a result of a standardization of the individual agent.

For the development of a conceptual segregation model the difficulty is to combine methodologically true individual entities with socially similar characteristics and behavior. This challenge is framed by complex empirical knowledge. Just to give an example about the intention of contributing to socially homogenous neighborhoods: Squazzoni (2012, p. 89), by highlighting a continuum of nearness and distance as a determinant for social relations, argues: “Therefore, for extensive co-residing as in modern cities, individuals have also developed subtle, complex and sometimes partially unconscious and unintended ways of dissociating from and discriminating against others”. Contrary to this ‘unconscious’ and ‘unintended’ behavior effect, Sennett (1970, p. 48) stresses an explicit desire “[…] for coherence, for structured exclusion and internal sameness […]”, which in turn seems to be in contradiction with contemporary aims of living in vivid mutual supportive communities: “Innate to the process of forming a coherent image of community is the desire to avoid actual participation. Feeling common bonds without common experience occurs in the first place because men are afraid of participation, afraid of the dangers and the challenges of it, afraid of its pain”. (ibid., p. 42) Both aspects can be observed empirically, the latter, for instance, in gated communities and (partly) gentrified neighborhoods, while the former is more common in organically grown, multicultural urban neighborhoods.

A reliable conflation of individual agency (subjective opportunities for freedom with respect to aspirations, preferences, but also constraints) and social influences (power relations, cultural bias, or recognition of capabilities) in segregation modeling is justified by empirical experience and progressive debates on modifications and alterations of Schelling-like models (see, e.g. Bruch and Mare 2006; Fossett 2006; Fossett and Waren 2005; Pancs and Vriend 2007; Zhang 2009; for a review of these studies see O’Sullivan and Perry (2013, pp. 83–87) and Squazzoni (2012, pp. 88–97). Squazzoni (2012, p. 96) draws the conclusion “[…] that if individual preferences and perceived differences between groups refer just to one characteristic, such as ethnicity, religion or political position and decision is binary, segregation is unavoidable and social integration is impossible”. From a modeling perspective the unrealistic determination is due to the univariate reference and the binary decision scheme. Moreover, relaxing the causal link between the decision-threshold of (dis-)satisfaction and corresponding action should surmount the seemingly inevitability of segregation.

Furthermore, the individual-community link complexifies the discussion about segregation when taking majority-minority relationships into consideration. In this case processes of inclusion and social cohesion can conflate with displacement, and political claims of integration and neighborhood diversity converge or diverge scale-dependently by quite similar transformations. From a model perspective it is very hard (or even impossible) to reasonably disentangle the bundle of interwoven processes and relations. “The language of ‘preference’ and ‘tolerance’ surrounding the Schelling model can give the appearance that such claims are being made. It is important to keep in mind that we could just as easily interpret the movement of minority households into friendly neighborhoods as arising from an inability to access neighborhoods with a high presence of the majority, a reading that would make the driving mechanism not preference but discrimination. Such debates are not about the model’s outcomes but about its interpretation and what can be inferred from it” (O’Sullivan and Perry 2013, p. 85).

To sum up: the consequent disaggregation of agents’ characteristics and behavior towards the individual level is an attempt to avoid homogenous community building that derives deterministically from an *a priori* standardized setting of attributes which leaves agents indistinguishable. The inherent sameness of agents at the group level in the original model contains segregation within itself as a predictable outcome to some degree. Truly distinguishable agents do recognize social conditions, nonetheless, but they do it differently. We put the emphasis on similar agents acting similarly with regard to similar decision making.

7.4 The Simulation Model

The current version of the segregation model is a conceptual simulation model which serves predominantly as an instance to verify the model's purpose. It has been built to simulate intra-urban residential mobility in the city of Salzburg, Austria, by highlighting both emerging and downward-causing patterns of socio-spatial cohesion. A quantity of census data from 2001 to 2011 has been used to carry out a factor analysis followed by a cluster analysis. The data is applied to initialize spatial raster cell characteristics on an approximate empirical basis. Future model development is intended to transform the pre-given model raster cell resolution to a 250 by 250 m scale in accordance with the officially available data provided by the Austrian Statistics Authority (Statistik Austria 2014). A validation of segregation processes over the period of one decade will then be possible.

7.4.1 Agents' Properties

The entire agent population is subdivided into four different subgroups, representing cluster characteristics which have been derived from a factor analysis with 14 socio-demographic variables (e.g. education, age cohorts, religion, household size, nationality). Clusters represent demographic characteristics of the city of Salzburg at district level and have been disaggregated randomly at cell level.

The realization of creating individual agents refers to agents' characteristics and decisions as well as executed actions. With respect to characteristics, 'income' is used as a prototypical randomized variable using a normal distribution function with a small standard deviation to individualize an agent's economic situation. In addition to the variation of income within the four subgroups, a variation has been applied between them. This is to vary the economic wealth of agents at the collective scale, representing social status.

Due to the normal distribution of income values there is no sharp distinction between agents as economically defined entities. Because of this, we included a second variable, 'attitude', which represents agents' social preferences (or disaffirmations) in a qualitatively generalized way; an agent's 'attitude' is represented by color. This is according to the idea mentioned above. The fluent transition of economic property (income) correlates with a clear distinction of the social characteristic (attitude) and allows for a reasonable diversity of agents; for instance, two socially similar agents share the same attitude but differ in income, and two economically similar agents may have similar income, but differ in their attitudes.

The reference variable for every agent to derive its initial income is given with the 'rent' value of the cell (patch) in which an agent is situated initially. Thus, a valid relationship of agent-location interaction is achieved.

The evaluation procedure of an agent's neighborhood embraces the affordability of the current location and the (dis-)satisfaction with its neighborhood. Whilst in the first case agents must move if their income is less than the costs for housing, in the latter they are equipped with some flexibility when assessing their neighbors' income and attitude. The common approach in agent-based segregation modeling is a single unified threshold value which will be applied to all agents. In contrast to this approach we first use two thresholds (for income and attitude, separately) and second vary the values of income by applying a range of values around the mean income (the qualitative indicator of 'attitude' remains as a binary variable). The income threshold which, as mentioned earlier, represents the approximate socio-economic status of an agent, is then transformed by a normal distribution function in order to individualize the decision-making process for residential moves.

In addition, and different from traditional segregation modeling, we have inserted two more modifications. First, not all agents move, even if they fulfill the condition of being dissatisfied. The reason for this can be justified with empirical observations: residential relocation is a complex fact, involving lots of criteria which must be pondered deliberately and diligently (which is represented here quite inaccurate by just two dimensions). The desire of retaining social ties developed over a long period or the established familiarity of everyday activities may represent reasons which imply some inertial behavior though dissatisfaction is a significant counter force (Knox and Pinch 2006, p. 253). Furthermore, even if one feels dissatisfied with one's current social and spatial neighborhood situation, relocation is not the one and only obligatory response to it. One may think of political activism or community engagement in order to improve local social well-being.

Secondly, and contrary to the first modification, agents may wish to move even though they might be satisfied with their current neighborhood. Reasons for such a decision might be the inheritance of a house or apartment, change of work place, family situation, lifestyle changes, or simply the search for the perfect home.

Agents who are not able to find an affordable dwelling in the city of Salzburg within a certain period of time because of economic reasons are forced to move to the suburban region. In turn, agents with sufficient income can either return to the city or immigrate (and might return again). With this procedure we have included migration in addition to intra-urban mobility.

7.4.2 Spatial Entities' Properties

Spatial entities (patches) represent housing costs as scaled values derived from cluster data. The scaling of values is based initially on a normal distribution and then adapted to the cluster characteristics. The data used represents a coarse approximation of the socio-demographic situation at census district level and is disaggregated statistically but verified as a proven approximation by experts from city authorities.

7.4.3 Scale Relevant Model Issues

The model has been implemented in NetLogo 5.2.0 (Wilensky 1999). The parameters 'income' of agents and 'housing costs' of spatial entities increase marginally per time step, i.e. there is an interaction pattern not only among agents, but also between them (the social domain) and the cells (the spatial domain). This interaction pattern thus represents the local scale of dynamic relationships between human decision making and socio-natural systems. The "between-agent" type of interaction is currently implemented in an abstract way because only a global trend of housing costs is included due to restrictions in data availability. There is, however, no local modification of this global trend at the moment. One may think, for instance, of a higher/lower dynamic in areas of high/low status neighborhoods. What has been implemented, however, is an accidental above-average increase and decrease, respectively, of housing costs, whereby 5% of all patches are then affected by an increase and 3% by a decrease. The extent can be altered interactively; for the subsequent model results the increase is set to 20% and the decrease to 10%.

In order to incorporate the institutional and market domains, i.e. the macro-structural determinants of the dynamic relationships between human decision making and socio-natural systems, two large housing estates have been implemented into the simulation model. They are located at different city districts and vary in social policy: "The socially homogeneous housing estate with a high proportion of social housing has been realised in a less affluent district (Itzling) in order to support affordable living conditions for the local communities. This housing estate represents an actually existing urban planning realisation. The second estate, following a "social mix" housing concept, has been realised in a wealthy city district of Salzburg (Aigen) and represents an actual realisation of future urban planning. In order to analyse the question of how large housing estates influence individual migration decisions under the pre-conditions of individual affordability and preferences, the size of the two housing estates has been varied by three steps. In addition, migration dynamics have not only been investigated within the housing estates, but also in the immediate local surroundings" (Koch 2016, np).

7.5 Selected Model Results

The resolution of the city is set to approx. 15,000 spatial entities inhabited by 150,000 citizens. A number of 6000 agents is selected as potential intra-urban movers which is an estimated 4% of the total population. Initially, agents of all four subgroups are randomly distributed over the urban space, according to the price per patch, i.e. initially, every agent can afford the dwelling she/he is living in. Census districts are colored according to the cluster they belong to, representing housing costs which vary from cheap to expensive in the following sequence: brown-orange-blue-pink-green-yellow.

The standard model (thereafter *sm*) has the following settings: the proportion of each subgroup is the same (25%), the preferences of similarity for each group is 25% for 'income similarity' and 20% for 'attitude similarity'. Income and housing price growth rates are set equal to 0.5% per time step. Seventy percent of dissatisfied agents actually move, but also 5% of satisfied agents do so. Two remarkable results are noteworthy: (1) Compared to a Schelling-type model, segregation is no longer a common phenomenon, being distributed evenly over the urban space (Fig. 7.1).

Instead, segregation is concentrated in affordable districts (colored brown and orange), and its spatial manifestation is given at a small-scale level. In high-price districts (green and yellow) socio-spatial community building of similar agents is much harder to achieve, even for the most affluent (red agents). (2) Segregation takes place, notwithstanding. Ultimately different degrees of neighborhood evaluation, of decision-making processes, and agents' as well as patches' characteristics do not avoid clustering of similar agents. The exceptional fact is the declining degree of segregation, most obvious for the least affluent (blue agents) and only very limited for the

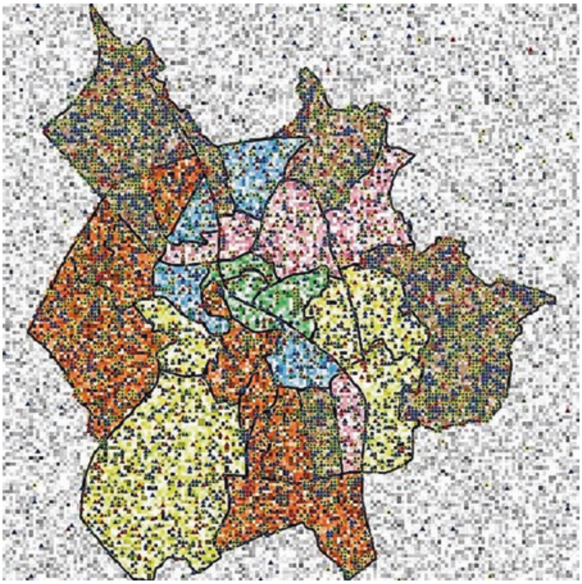


Fig. 7.1 Spatial representation of segregation with the standard simulation model

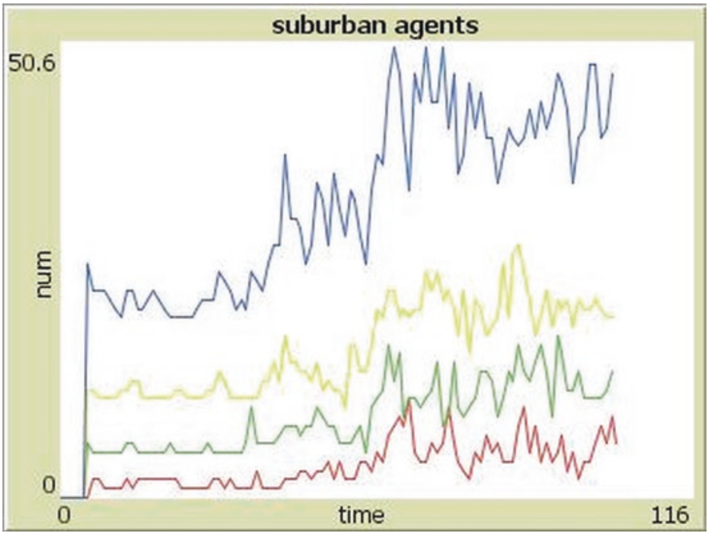


Fig. 7.2 Number of agents forced to the city (sm)

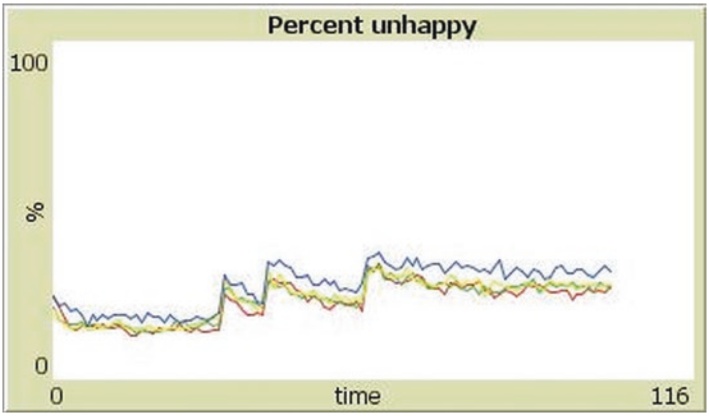


Fig. 7.3 Agents dissatisfied (%) and move from sudden increases of rent prices (sm)

most affluent. As an attempt to interpret one might be tempted here to draw a distinction between involuntary moves, caused by displacement, and more opportunities for freedom with respect to affordability, income and attitude preferences. For the current model there is only a qualitative statement of experts' empirical experiences, verifying the segregation clusters in the north of Salzburg but of less scope in the southern district. Other interesting results refer to the extent of outmigration to the suburban region, which significantly depends on the agent's income and the city's housing cost situation and which is most problematic for the least affluent subgroup (Fig. 7.2). Sudden dramatic rises of rents do affect all tenants in more or less the same way (Fig. 7.3).

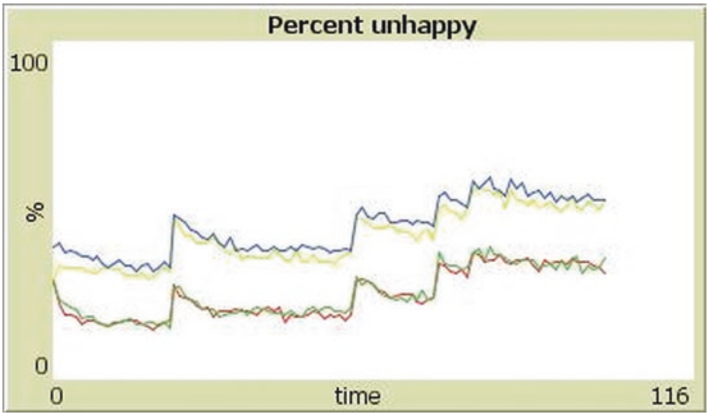


Fig. 7.4 Agents dissatisfied (%) and sudden increases of rent prices

The standard segregation simulation model is provided with eleven parameters to modify and alter the behavior space of agents according to the theoretical requirements mentioned above. The behavior space, therefore, offers a wide range of opportunities for if-then-analyses and scenario building. In what follows, four model variations which appear to be relevant for the (potential) emergence of segregation will be discussed briefly.

The first modification refers to the variation of the (individualized) thresholds of 'income' and 'attitude'. One hypothesis here is that the more affluent subgroups (red and green agents) appreciate higher degrees of income-similarity while the less affluent prefer higher degrees of attitude-similarity. Thus, income-similarity of red and green agents is set to 40%, and attitude-similarity to 15%. Yellow and blue agents' preferences are set to 25% for income-similarity and 45% for attitude-similarity. As Fig. 7.4 illustrates (with four sudden significant leaps of housing costs), a differentiation in the quality of preferences leads to two different levels of realized homogenous neighborhoods of similar agents. The trajectories of either two subgroups remain, however, relatively similar. It turns out that the emergence of neighborhoods with higher aspirations of income-similarity is less difficult to achieve than it is the case for attitude-similarity. This can be explained with the continuous variance of the income variable which makes the arrangements of local co-residing much easier than for the dichotomous attitude variable, even in the case of individualized agents. Segregation patterns are in part similar to the standard model – higher proportions of small-scale segregation have been evolved in the cheaper districts, but are also different from that model, because red and green agents have now been more successful in the creation of homogenous neighborhoods. Finally, the effect of the cost jumps is different for the four subgroups (Fig. 7.5); while for the least affluent agent population (blue) displacement is a cumulative force from the very beginning of the simulation, it is of only marginal relevance for the more and most affluent during the first half of simulation time and remains less influential in the second half.

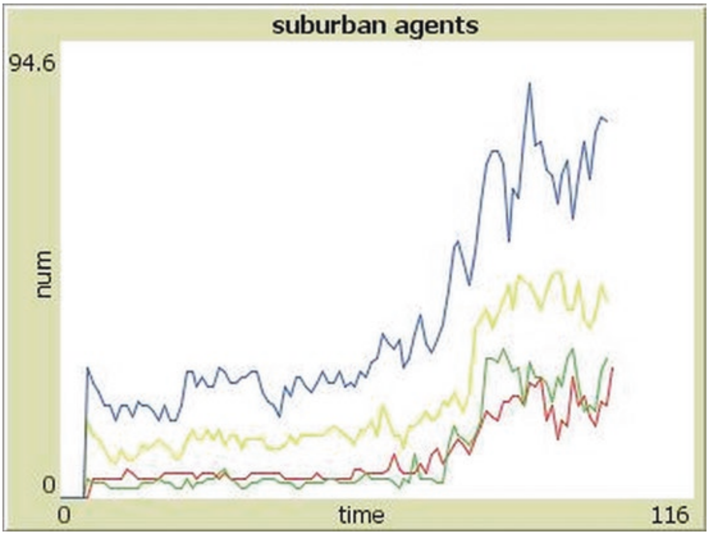
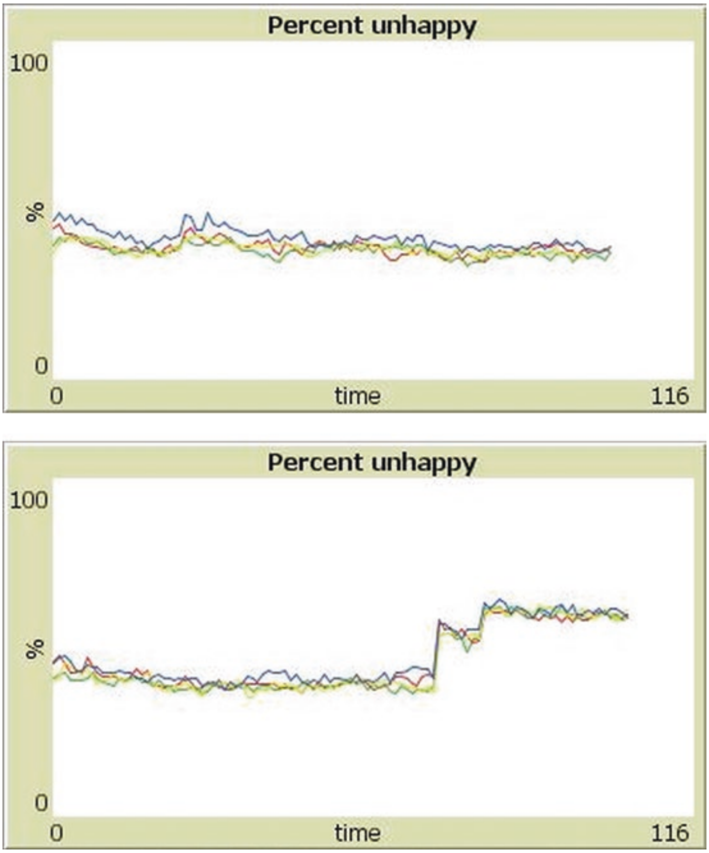


Fig. 7.5 Number of agents forced to move from the city (Notice: Figures 7.4 and 7.5 represent the change-of-preference model)

The parameters that determine the proportion of unhappy agents who actually move and happy agents who move notwithstanding, are relatively stable. For the first case we varied the proportion between 60% and 90% without significant changes in the model output. The latter has been varied between 3% and 10%, again without significant changes. These results hold true if the preferences for income-similarity and/or attitude-similarity are being varied (between 20% and 40% for both kinds across agents' subgroups).

A third variation of the standard simulation model considers the independent variables of 'increase of income' and 'increase of housing costs' as influential for neighborhood composition. Income-similarity and attitude-similarity is set to 35% for every subgroup. If income development is substantially higher than the development of housing costs (the subsequent model used an earnings growth of 0.7% per time step and a growth of housing costs of 0.3% per time step) then the, more or less, expected result is that all subgroups are able to live in the city. The differences between groups are marginal and even sudden leaps of rising prices do not affect agents' residential behavior significantly (Fig. 7.6).

In fact, the percentage of dissatisfied agents is slightly decreasing. Furthermore, even small clusters of homogenous neighborhoods of medium- and high-income households in more expensive districts have evolved. The situation changes with a reverse relationship but less significant than expected. Remarkably, the four subgroups do not differ in their capability to create homogenous neighborhoods (Fig. 7.7). Sudden changes of housing costs do, however, influence this capability explicitly. In the simulation illustrated in Fig. 7.7 the percentage of unhappy agents increases from approx. 37% before up to 63% after the jump in prices.



Figs. 7.6 and 7.7 Agents dissatisfied (%) and sudden increase of rent prices (change-of-income-cost-relationship model), with 0.3/0.7% (*above*) and 0.7/0.3% (*below*) cost-income change

The last variation takes the range of action of minorities into account. A first modification refers to the situation of having one minority in the city, starting with the least affluent subgroup (blue agents). They represent 10% of the urban population while the other three subgroups make up 30% each. The poor minority seeks to live in a neighborhood with at least 50% of agents sharing a similar attitude; its aspiration towards income-similarity is comparatively low (20%). The remaining subgroups all have a relationship of 35% income-similarity and attitude-similarity, respectively. Surprisingly, the poor minority does not have completely different troubles in dealing with its preferences of co-residing (Fig. 7.8) though there are fewer opportunities because of the small size of this group. The size of the group might be a suitable explanation for this result since majorities – primarily the socially adjacent group of yellow agents with slightly higher income and less restrictive preferences – are more powerful competitors in the housing market. Rent gaps and public housing allocation policy, as mentioned above, may contribute to

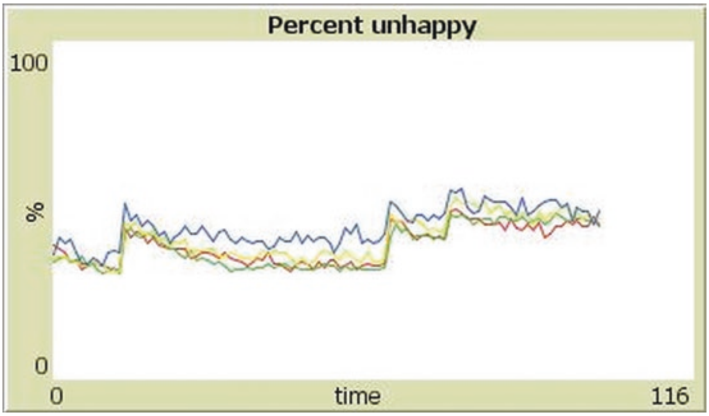


Fig. 7.8 Agents dissatisfied (%) and move from the city

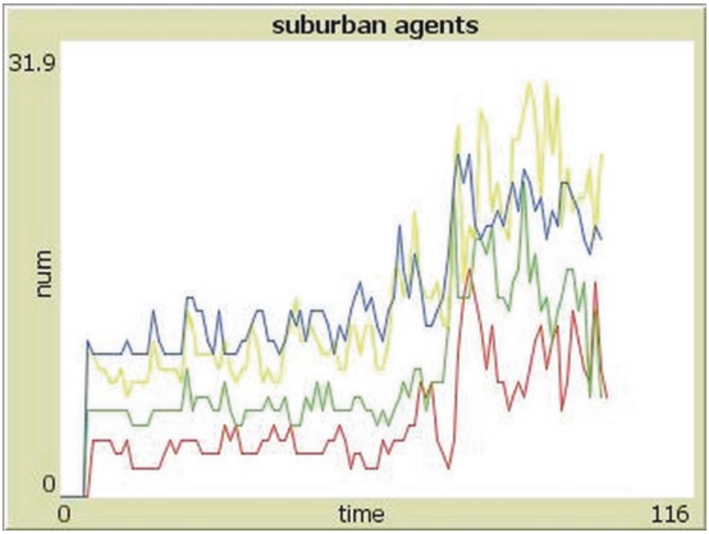


Fig. 7.9 Number of agents forced to sudden increases of rent prices (Notice: Figures 7.8 and 7.9 represent the poor-minority model)

amplify or mitigate this competitive change in socio-spatial distribution. Figure 7.9 confirms this thesis in part: the competition among the three relative majorities outperforms the competition between them with the minority.

If the most affluent subgroup in the city is in a minority situation – and their aspirations are more directed towards income-similarity (50%) and less towards attitude-similarity (20%) – then the underlying principle does change visibly: the affluent agents do much more to achieve their preferences, and the percentage of unhappy fellows is significantly larger than it was for the poor agents (Fig. 7.10). Simultaneously, the least affluent agent group exhibits a contradictory fact: on the

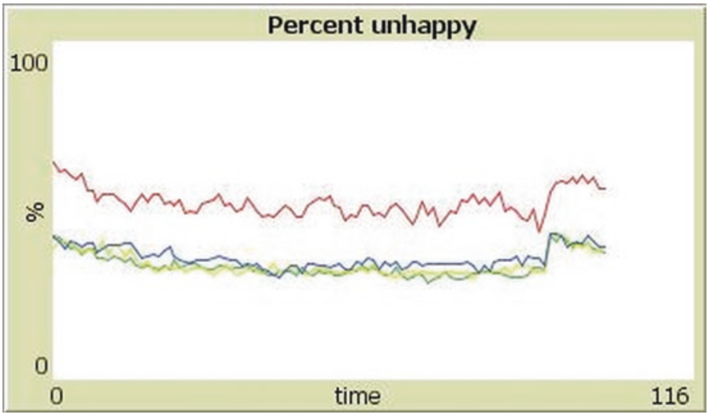


Fig. 7.10 Agents dissatisfied (%) and move from the city

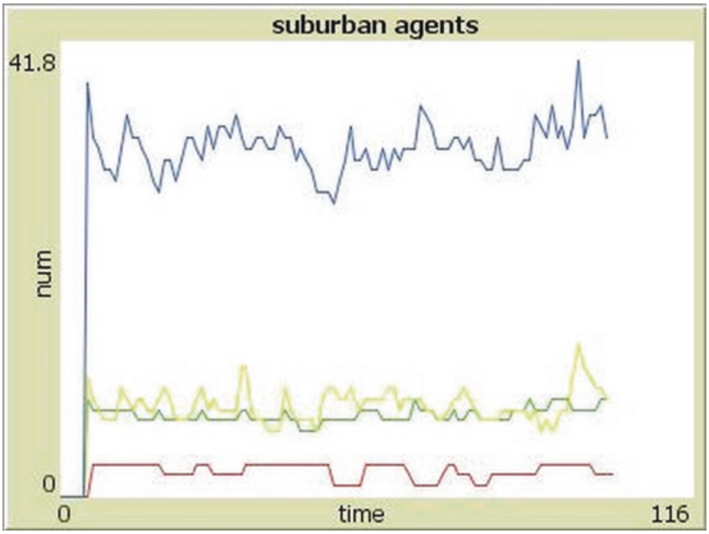


Fig. 7.11 Number of agents forced to sudden increases of rent prices (Notice: Figures 7.10 and 7.11 represent the rich-minority model)

one hand it is much easier for them to segregate themselves in the affordable districts; on the other hand the number of agents who were forced to move is much higher than it is for the other three groups (Fig. 7.11). One explanation might again lie in greater competition between socially and economically similar communities of the same population size.

The second modification of majority-minority relationship investigates the constellation of two minorities. In the first scenario the most and least affluent agent groups find themselves in a minority position. While the richest agents prefer

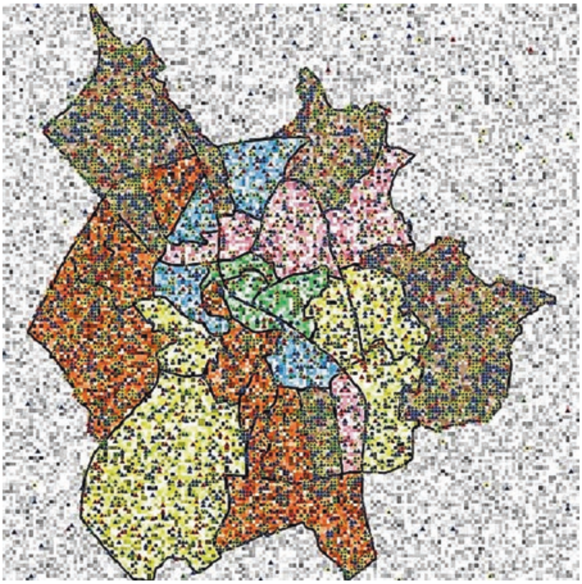


Fig. 7.12 Spatial representation of segregation with the two minorities' simulation model

income-similarity (50% compared with 20% attitude-similarity) the poorest like it the other way round. In addition, housing costs growth is higher than income growth. Now, both minorities have significantly greater difficulties in segregating themselves. They live in some kind of diaspora while the two middle-class majority populations are able to create small-scale but widespread homogenous neighborhoods. They, however, struggle most against displacement.

On the assumption that social polarization took place in the city, with high proportions of most and least affluent agent groups (80%), and a strong minority of middle-classes (20%) which prefer attitude-similarity to a higher degree (50% compared with 20% of income-similarity) then, again, it becomes obvious that displacement of the poor to the affordable districts (and to the suburban region) does play a crucial role in intra-urban residential mobility (Fig. 7.12).

On the other hand, no large-scale gentrification of the richest agents in the most expensive districts takes place. This modeling outcome is contrary to the empirical reality and thus confirms the necessity to take macro social conditions more seriously into account.

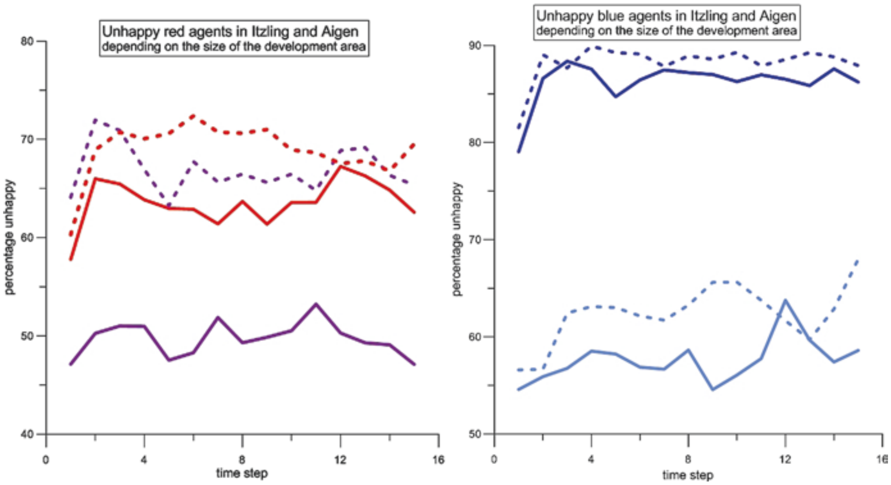
The consideration of these conditions has been incorporated by two large housing estates, as mentioned earlier. The following results are based on an extended model and a first and preliminary discussion is published in Koch (2016) where we refer to this article. With the simulation model it is possible to vary the spatial area of investigation, i.e. the neighborhood area that influences potential segregation outcomes. Interestingly, the two housing block areas do not exert a significant influence at the *large-scale city level*. Indeed, both the most and the least affluent agents appear to be easily able to establish socially similar neighborhoods. The least affluent agent group is, however, less often able to do so, likely because they are restricted to cheaper places in the city which hardly exist.

More interesting are the *small-scale dynamics at the district level* (within and around the newly created large housing estates). The case study area of Itzling with a high proportion of social housing highlights the fact that all four agent populations are able to achieve quite similar levels of satisfaction, which are relatively independent from the size of the neighbourhood area. However, the larger the area is the less easily are rich agents capable of realizing their preferences and needs. One explanation might be that they are not as easily able to achieve their goals in terms of status satisfaction, which is obvious from an empirical perspective, as social housing is dedicated to less affluent households. The case of Aigen with a mix of social housing and private property is somewhat different. The poor agents have the relatively highest percentage of dissatisfaction, irrespective of the size of the neighbourhood area; living in a high-price place is most difficult for them, although urban planning has developed strategies of inclusion. Surprisingly, unhappy rich agents are ranked second, but maybe this is again due to the social policy of socially mixed areas, "[…] which makes it more difficult to achieve the goal of homogeneous social neighbourhoods in sustainable ways" (Koch 2016, np).

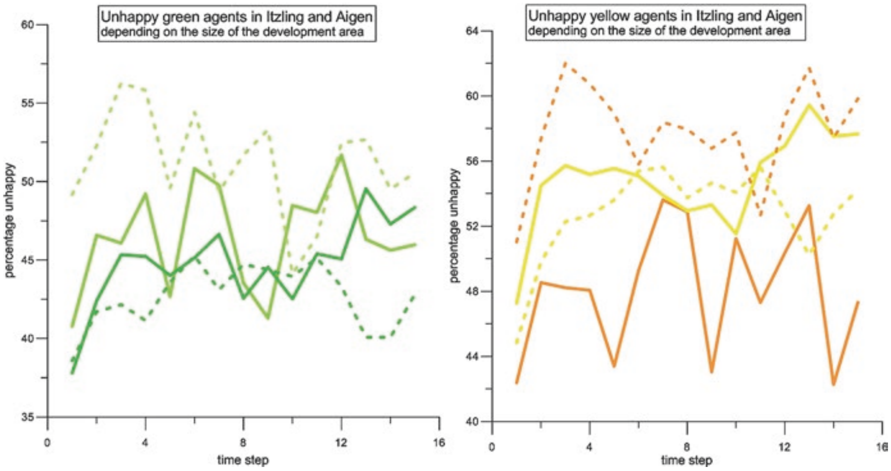
The study of relationships between individual motives and collective behavior reveals another interesting insight when comparing patterns of within-agent population decision processes: the different dynamics of how agents struggle with their social and spatial environment (labelled as "unhappy agents"). In this case the residential mobility behavior of all four agent populations over the two districts that inhabit the housing blocks is compared. As Figs. 7.13, 7.14, 7.15, and 7.16 illustrate, this dynamic is less pronounced for the most affluent (red) and least affluent (blue) agents, as is the case for the middle classes (green and yellow).

As has been stated by Koch (2016, np): "While red agents feel most comfortable in a small housing estate in Itzling and much less comfortable in larger units in both districts, the blue agent population exhibits a sharp distinction between Itzling (smaller degrees of unhappiness) and Aigen. The common competition for affordable living space provides the red agents with an advantage due to a higher degree of economic power, which in turn provides them with more opportunities to live in close proximity – even in areas that are spatially less attractive. The blue agents, on the other hand, are displaced to locations with a high degree of social housing, preventing social interactions with other social groups."

The middle-income agent populations, represented as green and yellow graphs in Figs. 7.15 and 7.16, display a different migration behavior which is much more volatile but with lower levels of dissatisfaction at the same time. In addition, a distinct segregation pattern cannot be detected, "[…] neither in terms of the size of the housing estate nor of the city district" (Koch 2016, np). Consequently, the social planning strategy (social housing or social mix) does not induce a sharp distinction.



Figs. 7.13 and 7.14 Dissatisfied rich agents (*left*; *red* graphs = Aigen, *purple* graphs = Itzling) and dissatisfied poor agents (*right*; *dark blue* graphs = Aigen, *light blue* graphs = Itzling). *Solid lines* represent small housing estates, *dotted lines* large housing estates according to the variability given with the simulation model



Figs. 7.15 and 7.16 Dissatisfied *upper* middle-class agents (*left*; *dark green* graphs = Aigen, *light green* graphs = Itzling) and lower middle-class agents (*right*; *yellow* graphs = Aigen, *light orange* graphs = Itzling). *Solid lines* represent small housing estates, *dotted lines* large housing estates according to the variability given with the simulation model

7.6 Conclusion

The theoretical aim of the paper was to develop a segregation simulation model which implements a coherent and reliable representation of the mutual influences und dependencies between individual agents and institutional structures. While the notion of the "individual" comprises the versatile characteristics and decision patterns of acting units (households in our case) it is the notion of "structure" that – with its rules, norms, attitudes, and other forms of collectivization – acts as a counterpart. The model's purpose therefore was to highlight both the bottom-up processes which may help to understand better the emerging patterns of urban socio-spatial homogenization and the top-down processes of downward-causation which are an independent force in altering or maintaining human decisions.

The observation of Ioannides (2013, p. 123), "despite the elegance of Schelling's model, empirics show neighborhoods are overall quite mixed", inspired us to use the basic ideas of this model-type in order to create true individual agents consistently. The benefit of Schelling's model approach, apart from it being the most influential approach in computational segregation research, is that it integrates a locational model and a "bounded-neighborhood model" (ibid., p. 115). From this starting point the results of the extended and modified simulation model presented here have demonstrated that segregation is a strong, though small-scaled process when viewed from the bottom up. Even though: (1) agents individually vary in attitudes, decision making, actions, and characteristics; (2) a universal threshold has been avoided; (3) macro social determinants have approximatively been included with urban planning realizations of large and socially diverse housing estates and housing costs, *segregation took place*.

Some empirical validation is given: while affluent households voluntarily tend to exclude themselves from the rest of the society, poorer households are mostly forced to segregate, although they would prefer to live in a socially mixed environment. Segregation is a controversial topic – in science, spatial planning, and politics. Among many others, one aspect has become increasingly crucial in debates on segregation: the knowledge transfer between people of different social statuses has been more and more interrupted, because of the creation of tangible and intangible borders. These borders tend to be used to make exclusion, injustice, and poverty invisible. A computational approach thus remains an important technique and provides a scientific contribution to detect the hidden mechanisms of these processes.

References

Baumgärtner E (2009) Lokalität und kulturelle Homogenität. transcript Verlag, Bielefeld

Bruch E, Mare RD (2006) Neighborhood choice and neighborhood change. Am J Sociol 112:667–709

Clay PL (1979) Neighborhood renewal. Middle-class resettlement and incumbent upgrading in American neighborhoods. Lexington Books, Lanham

Dangschat J (1997) Sag' mir wo du wohnst, und ich sag' dir wer Du bist! Zum aktuellen Stand der deutschen Segregationsforschung. In: PROKLA, Zeitschrift für kritische Sozialwissenschaft, Jg. 27, pp 619–647

Dangschat J (2007) Segregation. In: Häußermann H (ed) Großstadt. Soziologische Stichworte. 3. Auflage. VS Verlag, Wiesbaden

Dangschat JS, Alisch M (2012) Perspektiven der soziologischeen Segregationsforschung. In: May M, Alisch M (eds) Formen sozialräumlicher Segregation. Verlag Barbara Budrich, Opladen/Berlin/Toronto

Filatova T, Parker D, van der Veen A (2009) Agent-based urban land markets: agent's pricing behavior, land prices and urban land use change. J Artif Soc Soc Simul 12(1 3). Available at: http://jasss.soc.surrey.ac.uk/12/1/3.html. (2014-03-08)

Fossett M (2006) Ethnic preferences, social distance dynamics, and residential segregation: theoretical explorations using simulation analysis. J Math Sociol 30(3–4):185–273

Fossett M, Waren A (2005) Overlooked implications of ethnic preferences for residential segregation in agent-based models. Urban Stud 42(11):1893–1917

Fyfe NR, Kenny JT (2005) The urban geography reader. Routledge, London

Harris CD, Ullman EL (1945) The nature of cities. In: The annals of the American academy of political and social science. 11/1945, 242, pp 7–17

Iltanen S (2012) Cellular automata in urban spatial modelling. In: Heppenstall AJ, Crooks AT, See LM, Batty M (eds) Agent-based models of geographical systems. Springer, Dordrecht/Heidelberg/London/New York, pp 69–84

Ioannides Y (2013) From neighborhoods to nations. The economics of social interactions. Princeton University Press, Princeton/Oxford

Knox P, Pinch S (2006) Urban social geography, 5th edn. Pearson Education Limited, Harlow

Koch A (2016) The impact of macro-scale determinants on individual residential mobility behaviour. In: Jager W (ed) Proceedings of SSC 2015, Groningen. Springer, Dordrecht/Heidelberg/London/New York. (in press)

Mitchell M (2009) Complexity. Oxford University Press, Oxford

O'Sullivan D, Perry GLW (2013) Spatial simulation. Exploring pattern and process. Wiley, Chichester

Pancs R, Vriend NJ (2007) Schelling's spatial proximity model of segregation revisited. J Public Econ 92(1–2):1–24

Schelling TC (2006) Micromotives and macrobehavior, rev edn. Norton & Co, New York

Scholz F (2002) Die Theorie der 'fragmentierenden Entwicklung'. In: Geographische Rundschau, Jg. 54, Nummer 10, pp 6–11

Sennett R (1970) The uses of disorder. Personal identity and city life. Norton & Co, New York

Smith N (1979) Toward a theory of gentrification. A back to the city movement by capital not people. J Am Plan Assoc 45(4):538–548

Smith N (1996) The new urban frontier: gentrification and the revanchist city. Routledge, London

Squazzoni F (2012) Agent-based computational sociology. Wiley, Chichester

Statistik Austria (2014) Raster data, available at: http://www.statistik.at/web_de/klassifikationen/regionale_gliederungen/regionalstatistische_rastereinheiten/index.html (2014-03-12)

Wilensky U (1999) NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston

Zhang J (2009) Tipping and residential segregation: a unified Schelling model. J Reg Sci 51:167–193

Chapter 8
Lessons and Challenges in Land Change Modeling Derived from Synthesis of Cross-Case Comparisons

Robert Gilmore Pontius Jr., Jean-Christophe Castella, Ton de Nijs, Zengqiang Duan, Eric Fotsing, Noah Goldstein, Kasper Kok, Eric Koomen, Christopher D. Lippitt, William McConnell, Alias Mohd Sood, Bryan Pijanowski, Peter Verburg, and A. Tom Veldkamp

Abstract This chapter presents the lessons and challenges in land change modeling that emerged from years of reflection and numerous panel discussions at scientific conferences concerning a collaborative cross-case comparison in which the authors have participated. We summarize the lessons as nine challenges grouped under three themes: mapping, modeling, and learning. The mapping challenges are: to prepare data appropriately, to select relevant resolutions, and to differentiate types of land change. The modeling challenges are: to separate calibration from validation, to predict small amounts of change, and to interpret the influence of quantity error. The learning challenges are: to use appropriate map comparison measurements, to learn about land change processes, and to collaborate openly. To quantify the pattern validation of predictions of change, we recommend that modelers report as a percentage

R.G. Pontius Jr. (✉)
Clark University, Worcester, MA, USA

Michigan State University, East Lansing, MI, USA

Universiti Putra Malaysia, Serdang, Malaysia

Purdue University, West Lafayette, IN, USA

University of Twente, Enschede, The Netherlands
e-mail: rpontius@clarku.edu

J.-C. Castella
Institut de Recherche pour le Développement, Montpellier, France
e-mail: j.castella@ird.fr

T. de Nijs
National Institute for Public Health and the Environment, De Bilt, The Netherlands
e-mail: Ton.de.Nijs@rivm.nl

Z. Duan
China Agricultural University, Beijing, China
e-mail: Duanzengqiang@aliyun.com

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_8

of the spatial extent the following measurements: misses, hits, wrong hits and false alarms. The chapter explains why the lessons and challenges are essential for the future research agenda concerning land change modeling.

Keywords CLUE • CLUE-S • Environment Explorer • Geomod • Land Transformation Model • Land change • Land Use Scanner • LUCC • Map • Model • Prediction • SAMBA • SLEUTH • Validation

8.1 Introduction

The first author of this chapter extended an open invitation to the community of land change modelers to participate in a cross-case comparison of spatially explicit land change modeling applications. The focus was the assessment of pattern validation

E. Fotsing
Computer Science, University of Dschang, Bandjoun, Cameroon
e-mail: fosting@gmail.com

N. Goldstein
Lawrence Livermore National Laboratory, Livermore, CA, USA
e-mail: goldstein@navigant.com

K. Kok
Soil Geography and Landscape, Wageningen University, Wageningen, The Netherlands
e-mail: Kasper.Kok@wur.nl

E. Koomen • P. Verburg
Spatial Analysis & Modelling, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Vrije Universiteit Amsterdam, Institute for Environmental Studies, Amsterdam, The Netherlands
e-mail: ekoomen@feweb.vu.nl; Peter.Verburg@ivm.vu.nl

C.D. Lippitt
Geography & Environmental Studies, The University of New Mexico, Albuquerque, NM, USA
e-mail: clippitt@unm.edu

W. McConnell
Center for Systems Integration and Sustainability, Michigan State University, East Lansing, MI, USA
e-mail: mcconn64@msu.edu

A. Mohd Sood
Universiti Putra Malaysia, Selangor, Malaysia
e-mail: ms_alias@putra.upm.edu.my

B. Pijanowski
Forestry and Natural Resources, Purdue University, West Lafayette, IN, USA
e-mail: bpijanow@purdue.edu

A.T. Veldkamp
Faculty of Geo-Information Science and Earth Observation, University of Twente, Enschede, The Netherlands
e-mail: a.veldkamp@twente.nl

of the mapped output of such models, so the invitation requested that participants submit for any case study three maps of land categories: (1) a reference map of an initial time 1 that a land change model used for calibration, (2) a reference map of a subsequent time 2 that could be used for validation, and (3) a prediction map of same time 2 that the land change model produced. Ultimately, we compiled 13 cases from nine countries, which were submitted from seven different laboratories. Pontius et al. (2008) derived and applied metrics to compare those various cases. We presented our work at several scientific conferences. Pontius et al. (2008) has been cited more than 396 times as of September 2017 according to scholar.google.com, thus has had a substantial influence on the constantly growing field of land change modeling (Paegelow et al. 2013). A frequent initial reaction that audiences have when they first hear about our exercise is to ask "Which model is best?" However, the exercise never intended to rank the models. The audience's unintended reaction has been one of the inspirations for this follow-up chapter. The popularity of the question indicates that we must be careful to interpret the results properly, because the purpose of the exercise can be easily misinterpreted. We have found that the exercise's methods and results inspire quite disparate conclusions from various scientists. The purpose of the exercise was to gain insight into the scientific process of modeling, in order to learn the most from our modeling efforts. Therefore, this chapter shares the lessons that survived after years of reflection on both participation in the cross-case comparison and interactions with colleagues.

Figure 8.1 shows how we think of the lessons in terms of the flows and feedbacks of information among the various components of modeling. The figure begins with the landscape in the upper left corner. Scientists create data to summarize the landscape. There is a tremendous amount of information that scientists can derive from simply analyzing the maps from two or more time points (Aldwaik and Pontius 2013; Runfola and Pontius 2013). Scientists anticipate that they can learn even more by engaging in a modeling procedure that produces a dynamic simulation of land change. Scientists usually use a conceptual understanding of landscape dynamics to guide the selection or production of algorithms that express those dynamics. This chapter uses the word "model" to refer to such a set of algorithms, and the word "case" to refer to an application of the model to a particular study site. One way to assess a case is to examine the output that the model produces. Ultimately, a major purpose of the analysis is for scientists to learn from the measurements of the data and the outputs from the model. Scientists can use this learning to revise the mapping, the modeling and/or the measurements of the data and the model's output. The components of Fig. 8.1 reflect the structure of this chapter in that this chapter's Methods section summarizes the techniques to measure both the data and the model's output, while the subsequent Results and Discussion section presents the most important lessons, organized under the themes of mapping, modeling, and learning.

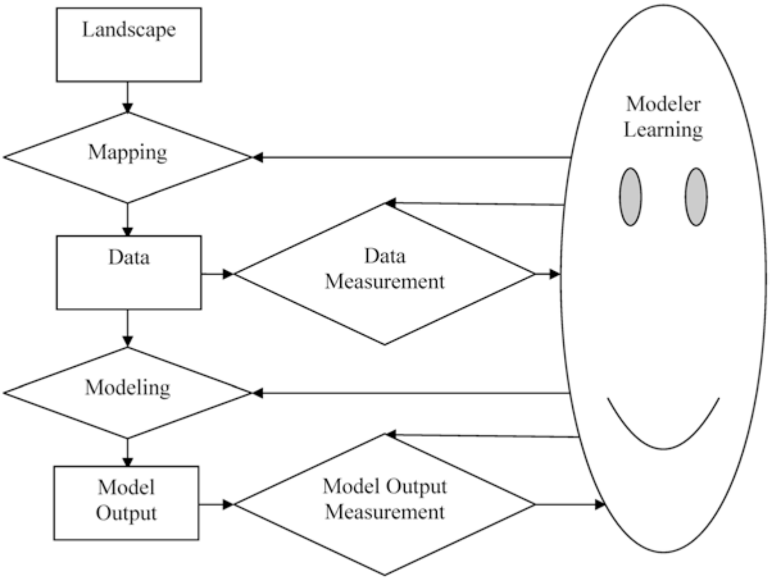


Fig. 8.1 Conceputal diagram to illustrate flows and feedbacks of information among components and procedures for a systematic analysis. *Rectangles* are components of the research system; *diamonds* are procedures; the *oval* is the modeler whose learning can inform methods of mapping, modeling and measuring

8.2 Methods

All of the models have been published in peer-reviewed journals and books. Raster maps have been submitted by scientists from the laboratories that developed the models. Collectively, the sample of models and their applications cover a range of some of the most common modeling techniques such as statistical regression, cellular automata, and machine learning. SAMBA is the single agent-based model in the collection. Table 8.1 offers specific characteristics of the nine models used for the 13 cases. These cases offer illustrations of these models that have been applied with various objectives, extents and resolutions. The model characteristics in Table 8.1 are necessary for proper interpretation. Geomod, Logistic Regression, and Land Transformation Model (LTM) use maps for which each pixel shows the land as either undeveloped or developed. These three models predict a single transition from the undeveloped category to the developed category. The other six models use maps of more than two categories to predict multiple transitions. For seven of the models, the user can set exogenously the quantity of each land cover category for the predicted map, and then the model predicts the spatial allocation of the land categories. SLEUTH and SAMBA do not have this characteristic. The cases that derive from LTM, CLUE-S, and CLUE use the quantity of each category in the reference map of time 2 as input to the model. For these cases, the model is assured to simulate the correct quantity of each category at time 2, thus the purpose of the

Table 8.1 Characteristics of the nine models as applied in the 13 cases

Model	Predicted transitions	Exogenous quantity	Used year 2 quantity	Pure pixels	Case	Literature
Geomod	Single	Yes	No	Yes	Worcester	Pontius et al. (2001), Pontius and Malanson (2005), Pontius and Neeti (2010)
SLEUTH	Multiple	No	No	Yes	Santa Barbara	Dietzel and Clarke (2004), Goldstein (2004), Silva and Clarke (2002)
Land Use Scanner	Multiple	Yes	No	No	Holland(8)	Hilferink and Rietveld (1999), Hoymann, (2010), Koomen and Borsboom-van Beurden (2011), Kuhlman et al. (2013)
Environment Explorer	Multiple	Optional	No	Optional	Holland(15)	de Nijs et al. (2004), Engelen et al. (2003), Verburg et al. (2004)
Logistic regression	Single	Yes	No	Yes	Perinet	McConnell et al. (2004)
SAMBA	Multiple	No	No	Yes	Cho Don	Boissau and Castella (2003), Castella et al. (2005a, b)
Land Transformation Model	Single	Yes	Yes	Yes	Twin Cities, Detroit	Pijanowski et al. (2000, 2002, 2005)
CLUE-S	Multiple	Yes	Yes	Yes	Kuala Lumpur, Haidian, Maroua	Duan et al. (2004), Fotsing et al. (2013), Tan et al. (2005), Verburg and Veldkamp (2004), Verburg et al. (2002)
CLUE	Multiple	Yes	Yes	No	Costa Rica, Honduras	de Koning et al. (1999), Kok and Veldkamp (2001), Kok et al. (2001), Veldkamp and Fresco (1996), Verburg et al. (1999)

modeling application is to predict the spatial allocation of change. Most of the models are designed to use pixels that are categorized as exactly one category, while Land Use Scanner, Environment Explorer and CLUE can use heterogeneous mixed pixels for both input and output.

Both Land Use Scanner and Environment Explorer are applied to the entire country of The Netherlands. One substantial difference between these two cases is that the number of categories in the output map for the application of Land Use Scanner is eight, while the number of categories for the application of Environment Explorer is 15. LTM, CLUE-S, and CLUE are applied to more than one study area, which allows us to see variation in how a single model can behave in various case studies. Our sample does not include cases of how a single model can produce various outputs for a single extent depending on how the model is parameterized. The possible variation due to parameterization of a single model is one reason why we do not rank the performance of the models.

Figure 8.2 shows the mapped results for each of the 13 cases. Each map in Fig. 8.2 derives from an overlay of the three maps that a modeler submitted. The first 11 of the 13 cases share the same legend, while Costa Rica and Honduras share a different legend because those two cases have mixed pixels. We encourage the profession to use the following short names for the categories in the legend of Fig. 8.2 (Brown et al. 2013). Misses are erroneous pixels due to observed change predicted as persistence. Hits are correct pixels due to observed change predicted as change. Wrong hits are erroneous pixels due to observed change predicted as change to the wrong gaining category. False alarms are erroneous pixels due to observed persistence predicted as change. Correct rejections are correct pixels due to observed persistence predicted as persistence.

Figure 8.3 summarizes the results where a segmented bar quantifies each case in terms of the legend of Fig. 8.2. Each bar is a Venn diagram where one set is the observed change and the other set is the predicted change, as the brackets illustrate for the case of Perinet. The “figure of merit” is a summary measurement that is a ratio, where the numerator is the number of hits and the denominator is the sum of hits, wrong hits, misses and false alarms (Pontius et al. 2007, 2011). If the model’s prediction were perfect, then there would be perfect intersection between the observed change and the predicted change, in which case the figure of merit would be 100%. If there were no intersection between the observed change and the predicted change, then the figure of merit would be zero. Figure 8.3 orders the cases in terms of the figure of merit, which is expressed as a percent at the right of each bar. It is also helpful to consider a null model for each case. A null model is a prediction of complete persistence, i.e. no change between time 1 and time 2 (Pontius et al. 2004a). Consequently the accuracy of the null model is 100% minus the percent of observed change. Figure 8.3 shows that the accuracy of the land change model exceeds the accuracy of its corresponding null model for 7 of the 13 cases at the resolution of the raw data.

Figure 8.4 plots for each case the figure of merit versus the percentage of observed change. Figure 8.4 reveals two clusters. The tight cluster near the origin shows that all of the cases that have a figure of merit less than 15% also have an observed change less than 10%. We analyzed many factors that we suspected might explain the predictive power for the 13 cases and found that the percentage of change observed in the reference maps had the strongest relationship with predictive accuracy.

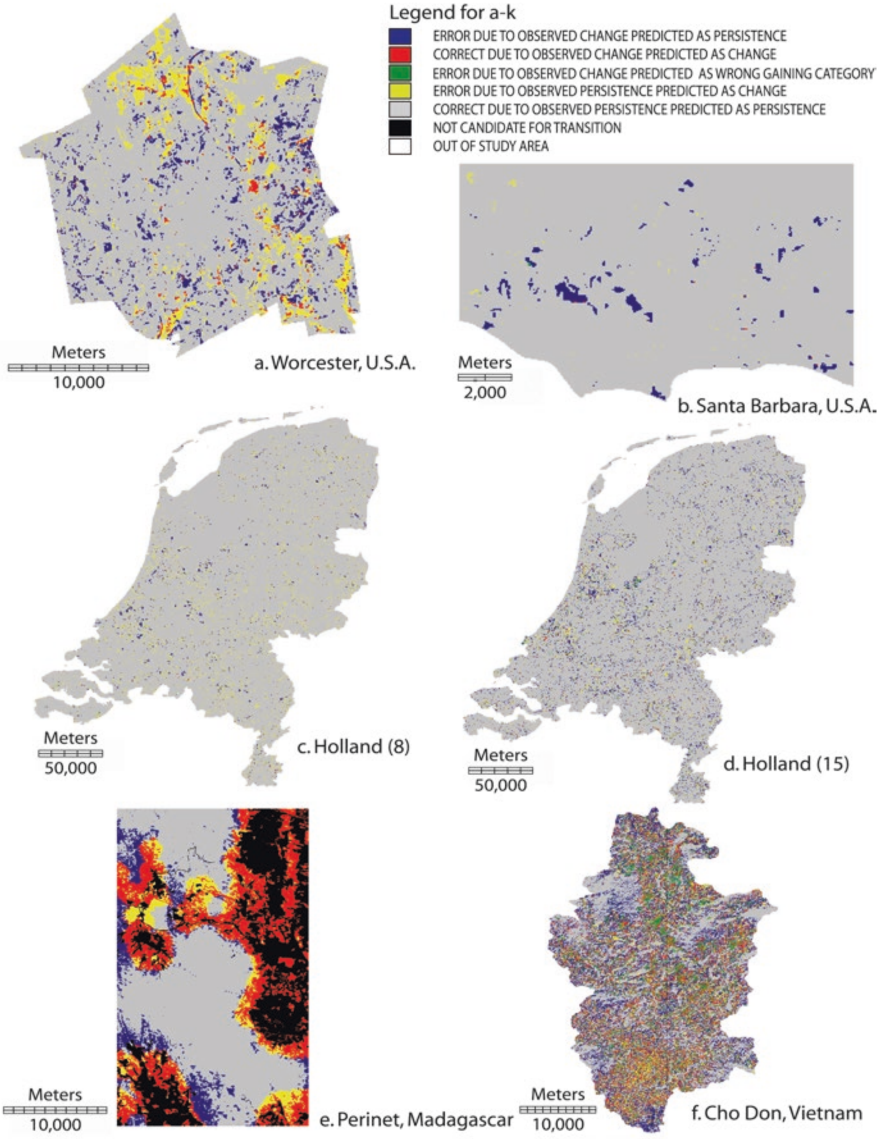

Legend for a-k
ERROR DUE TO OBSERVED CHANGE PREDICTED AS PERSISTENCE
CORRECT DUE TO OBSERVED CHANGE PREDICTED AS CHANGE
ERROR DUE TO OBSERVED CHANGE PREDICTED AS WRONG GAINING CATEGORY
ERROR DUE TO OBSERVED PERSISTENCE PREDICTED AS CHANGE
CORRECT DUE TO OBSERVED PERSISTENCE PREDICTED AS PERSISTENCE
NOT CANDIDATE FOR TRANSITION
OUT OF STUDY AREA
Meters
10,000
a. Worcester, U.S.A.
Meters
2,000
b. Santa Barbara, U.S.A.
Meters
50,000
c. Holland (8)
Meters
50,000
d. Holland (15)
Meters
10,000
e. Perinet, Madagascar
Meters
10,000
f. Cho Don, Vietnam

Fig. 8.2 Maps of misses, hits, wrong hits, false alarms and correct rejections

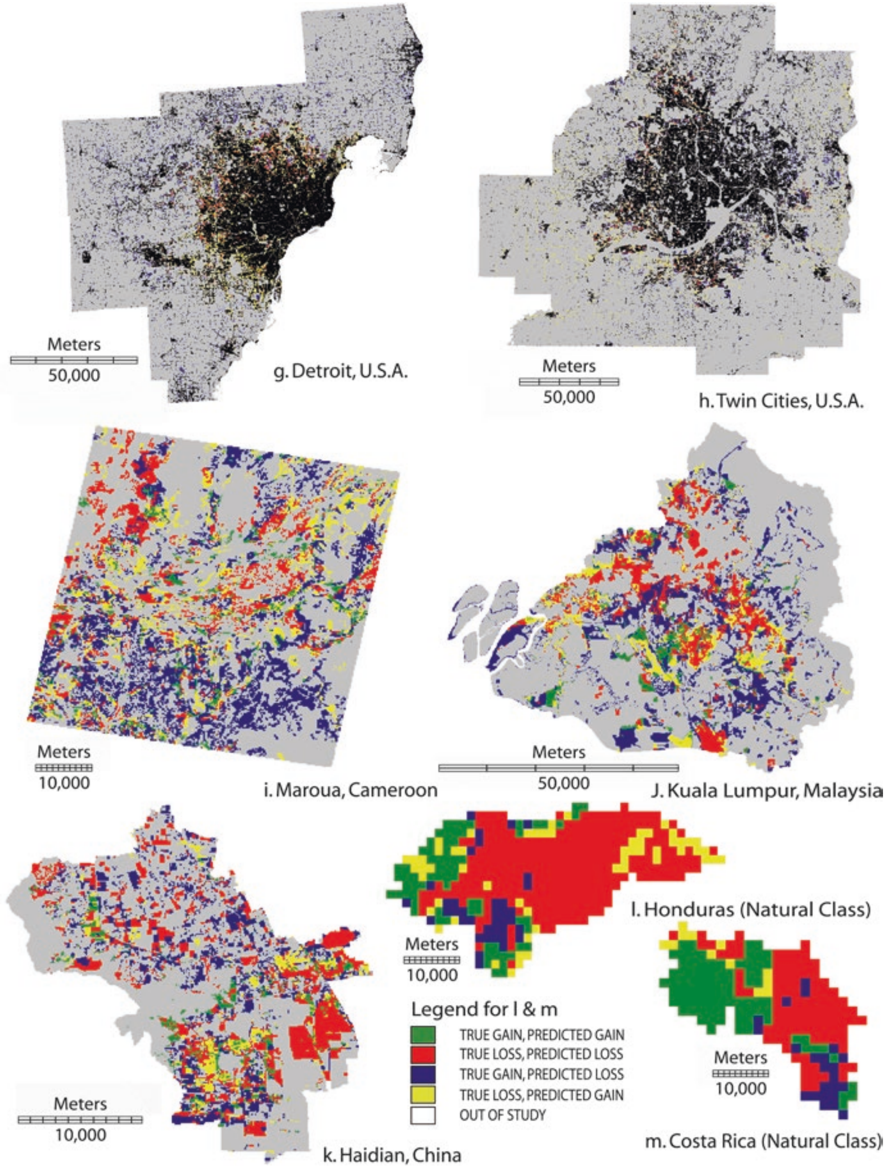


Fig. 8.2 (continued)

We have been soliciting feedback on our exercise since the initial invitation to participate in 2004. We have presented our work at five international scientific conferences: the 2004 Workshop on the Integrated Assessment of the Land System in Amsterdam The Netherlands, the 2005 Open Meeting of the Human Dimensions of Global Environmental Change Research Community in Bonn Germany, the 2006 Meeting of the Association of American Geographers in Chicago USA, the 2007 World Congress of the International Association for Landscape Ecology in

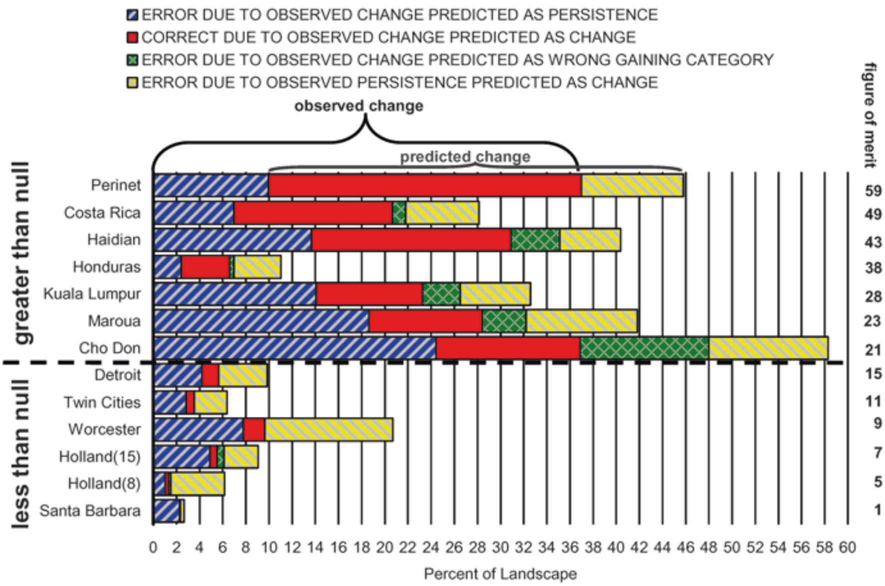
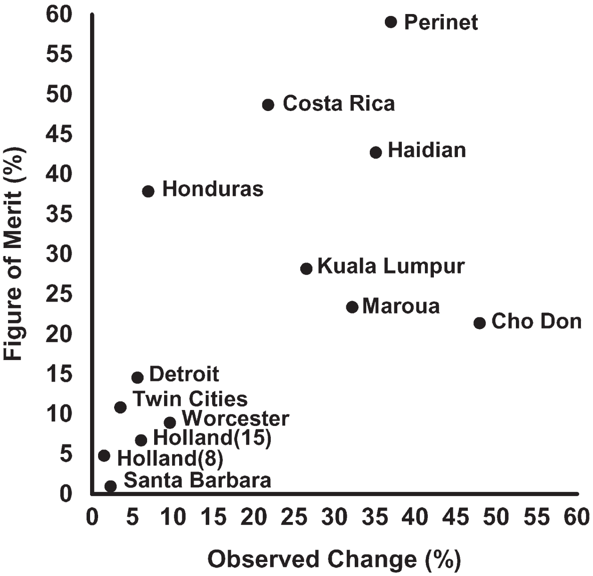


Fig. 8.3 Misses, hits, wrong hits, and false alarms for pattern validation of 13 cases. Correct rejections are 100% minus the length of the entire segmented bar. Each *bar* is a Venn diagram where the union of hits and wrong hits is the intersection of observed change and predicted change

Fig. 8.4 Relationship between predictive accuracy and observed change



Wageningen The Netherlands, and the 2007 Transatlantic Land Use Conference in Washington DC USA. There were panel discussions in Amsterdam, Chicago and Wageningen, where authors shared their experiences and audience members shared their reactions. The next section of this chapter synthesizes the lessons that have withstood more than a decade of examination of this cross-case comparison.

8.3 Results and Discussion

This section offers nine lessons. Each lesson has implications concerning the agenda for future research; therefore each lesson corresponds to a sub-section that articulates a challenge for future modeling efforts. The lessons are grouped under three themes: mapping, modeling, and learning. These groupings emerged as the authors reflected on the various types of lessons. The first theme demonstrates that the selection of the spatial extent and the production of the data have a substantial influence on the results, so scientists must pay as much attention to the mapping procedure as they do to the modeling procedure. This message reinforces known fundamental concepts in mapping, which scientists must keep at the front of their minds. The second theme concerns the modeling process. The challenges under this second theme derive from insights that have emerged specifically as a result of this cross-case exercise. They have implications for how scientists design and assess modeling procedures. The third theme focuses on learning, thus it emphasizes careful reflection on mapping and modeling. If mapping and modeling are not interpreted properly, then modelers can exert a tremendous amount of time and energy without learning efficiently. This third theme contains ideas for how modelers can maximize learning from mapping and modeling.

8.3.1 Mapping Challenges

8.3.1.1 To Prepare Data Appropriately

The decisions concerning how to format the data are some of the most influential decisions that scientists make. In some cases, scientists adopt the existing format of the available data, while in other cases scientists purposely format the data for the particular research project. Scientists must think carefully about the purpose of the modeling exercise when determining the format of the data. Formatting decisions concern the spatial, temporal and categorical scales in terms of both extent and resolution. The apparent complexity of a landscape is a function of how scientists choose to envision it, which is reflected in their mapping procedures. If scientists choose a great level of detail, then any landscape can appear to be greatly complex; while if scientists choose less detail, then the same landscape can appear simpler than what the more detailed data portray. For example, the Dutch landscape is not inherently more complex than the Perinet landscape. However the data for Perinet were formatted to show a one-way transition from forest to non-forest while the data for Holland(15) were formatted to show multiple transitions among 15 categories based on the data formatting decisions of the modelers. One could have attempted to analyze the Dutch landscape as two categories of built versus non-built, and could have attempted to analyze the Perinet data as numerous categories of various types of uses and covers. For example, Laney (2002) chose to analyze land change in Madagascar at a much finer level of detail and deeper level of complexity than McConnell et al. (2004). Anyone can choose a great level of detail for the data that will overwhelm the computational and predictive ability of any particular model. More detail does not necessarily lead to a more appropriate case study, just as less detail does not necessarily lead to a more appropriate case study. Scientists face the challenge to select a spatial resolution, spatial extent, temporal resolution, temporal extent, and set of categories for which a model can illuminate issues that are relevant for the particular purpose of the inquiry.

Decisions concerning the format and detail of the data are fundamental for understanding and evaluating the performance of the model (Dietzel and Clarke 2004). The Holland(8) case demonstrates this clearly as it relates to the reformatting from maps that describe many heterogeneous categories within each pixel to maps that describe the single dominant category within each pixel. The Land Use Scanner model was run for heterogeneous pixels of 36 categories, and then the output was reformatted to homogenous pixels of eight categories for the three-map comparison presented in Fig. 8.2. This reformatting is common to facilitate the visualization of such mixed pixel data. A major drawback of this reformatting is that it can introduce substantial overrepresentation of categories that tend to cover less than the entire pixel but more than any other category within the pixel (Loonen and Koomen 2009). Consequently, the reformatting can also introduce substantial underrepresentation of minority categories. These artifacts due to reformatting can generate more differences between the maps than the differences that the model generates by its predicted change. Such biases substantially influenced the analysis of the Holland(8) case and caused the apparent error of quantity for the predicted change to be larger than the error of quantity for the null model.

Decisions concerning how to format the data are influential, but scientists lack clear guidelines concerning how to make such decisions. It makes sense to simplify the data to the level that the calibration procedure and validation procedure can detect a meaningful signal of land change. It also makes sense to simplify the data so that the computer algorithms focus on only the important transitions among categories, where importance is related to the practical purpose of the modeling exercise. Scientists who attempt to analyze all transitions among a large number of categories face substantial challenges. For the Santa Barbara, Holland(8), and Holland(15) cases, each particular transition from one category to another category in the reference maps occurs on less than 1% of the spatial extent. Each of these individual transitions would need to have an extremely strong relationship with the independent variables in order for a model to predict them accurately. Scientists can alleviate the challenge by aggregation from a set of numerous detailed categories to a set of fewer coarser categories. Aldwaik et al. (2014) offer an algorithm for how to aggregate categories while maintaining the signals of land change.

Decisions concerning the data are related closely to decisions concerning the level of complexity of the models. Models that simulate only a one-way transition from one category to one other category can be simpler than models that simulate all possible transitions among multiple categories. If scientists choose to analyze very detailed data, then they may be tempted or forced to use very complex models. It is not clear whether it is worthwhile to include great detail in the data and/or in the models, because it is not clear whether more detail leads to better information or to more error.

Modelers should consider the certainty of the data, because much of the apparent land change between two time points could be due to error in the reference maps at the two time points (Enaruvbe and Pontius 2015; Pontius and Lippitt 2006; Pontius and Petrova 2010). Participating scientists suspect that error accounts for a substantial amount of the observed difference between the two reference maps for Maroua, Kuala Lumpur, and Holland(15). Scientists should use data for which there is more variation over time due to the dynamics of the landscape than due to map error. This can be quite a challenge in situations where map producers are satisfied with 85% accuracy, which implies up to 15% error, while many data sets show less than 15% land change.

8.3.1.2 To Select Relevant Spatial Resolutions

Spatial resolution is a component of data format that warrants special attention because: (1) spatial resolution can have a particularly strong influence on results, (2) spatial resolution is something that modelers usually can influence, and (3) it is not obvious how to select an appropriate spatial resolution. The spatial resolution at which landscapes are modeled is often determined by data availability and computational capacity. For example, if a satellite image dictates the resolution and extent, as it did in the Maroua case (Fotsing et al. 2013), then the boundaries of the study area and the apparent unit of analysis are determined in part by the satellite imaging system, not necessarily by the theoretical or policy imperatives of the modeling exercise. Kok et al. (2001) argue that the selection of resolution should take into consideration the purpose of the modeling application and the scales of the land change processes. For example, the Worcester case uses 30-m resolution data, but we know of no stakeholders in Worcester who need a prediction of land change to be accurate to within 30 m. Some stakeholders would like to know generally what an extrapolation of recent trends would imply over the next decade to within a few kilometers, which is a resolution at which Geomod predicts better than a null model as revealed by a multiple-resolution analysis of the model's output. Therefore, it is helpful from the standpoint of model performance to measure the accuracy of the prediction at resolutions coarser than the resolution of the raw data. Pontius et al. (2008) show that 12 of the 13 case studies have more error than correctly predicted change at the fine resolution of the raw data. However, for 7 of the 13 cases, most of the errors are due to inaccurate spatial allocation over relatively small distances. Multiple-resolution analysis shows that the errors shrink when the results are assessed at a resolution of 64 times the length of the side of the original pixels.

Errors of spatial allocation shrink as resolution becomes coarser, but errors of quantity are independent of resolution when assessed using an appropriate multiple-resolution method of map comparison (Pontius et al. 2004a).

If there is more allocation error than correctly predicted change at the resolution of the raw data, then it means that the data have a resolution that is finer than the ability of the model to predict allocation correctly. This can be a desirable characteristic because it means that the modeling exercise is not limited by the coarseness of the spatial resolution of the data. If there is more correctly predicted change than allocation error than at the resolution of the raw data, then it might be an undesirable characteristic because it might mean that the modeling exercise is limited by the coarseness of the spatial resolution of the data. The size of the error is larger than the size of correctly predicted change for 12 of 13 of our case studies at the spatial resolution of the raw data. Some scientists might conclude that the models are not accurate, while it may be more appropriate to conclude the data are more detatiled than necessary.

Advances in mapping technology have made it increasingly easy to find data that have a resolution finer than is necessary to address various research questions. If data are available at the meter resolution, then it does not imply that scientists are obligated to simulate changes accurately to within a meter. It might be desirable to run the model at a fine resolution, but to analyze the output at coarser resolutions in order to find a spatial resolution for which the model predicts sufficiently given the goals of the modeling exercise.

8.3.1.3 To Differentiate Types of Land Change

Scientists should select the types of land change that are of interest before deciding which model to use, because some types of land change present particular challenges for models. It is useful to think of two major types of change: quantity difference and allocation difference. Quantity difference refers to the difference in the size of the categories in the reference maps of time 1 and time 2, while allocation difference refers to the difference in the spatial allocation of the categories given the quantity difference (Pontius et al. 2004b; Pontius and Millones 2011; Pontius and Santacruz 2014). Allocation difference exists when a category experiences loss at some places and gain at other places during a time interval. The reference maps for Holland(15), Cho Don, Haidian, Honduras and Costa Rica demonstrate more allocation than quantity difference. In particular, Costa Rica demonstrates about ten times more allocation than quantity difference. When there is substantial allocation difference in the observed data, the model is faced with the challenge to predict simultaneous gains in some pixels and losses in other pixels for a single category in order to predict the change accurately. This can be much more challenging than to predict a one-way transition from one category to one other category. For example, the Worcester, Perinet, Detroit, and Twin Cities cases use models that are designed to simulate only the gross gain of only one category, while all the other cases use models that are designed to allow for simultaneous transitions among several categories. It is particularly challenging to write an algorithm for situations where more than one category competes to gain at a particular pixel.

8.3.2 Modeling Challenges

8.3.2.1 To Separate Calibration from Validation

Calibration is the procedure to set the parameters of a model, based on information at or before time 1. Validation is the procedure to assess how the predicted change compares to the reference change from time 1 to time 2. Proper validation of temporal prediction requires that calibration must be separate from validation though time. However, most of the cases used for calibration some information subsequent to time 1 in order to predict the change between time 1 and time 2. In 7 of the 13 cases, the model's calibration procedure used information directly from the reference map of time 2 concerning the quantity of each category. Other cases used influential variables, such as protected areas, that derive from contemporary time points subsequent to time 1. In these situations, it is impossible to determine whether the model's apparent accuracy indicates its predictive power through time. If a model uses information from both time 1 and time 2 for calibration, then the model's so-called prediction map of time 2 could be a match with the reference map of time 2 because the model parameters might be over fit to the data. The apparent accuracy would reflect a level of agreement higher than the level of agreement attributable to the model's predictive power into an unknown future.

There are some practical reasons why modelers use information subsequent to time 1 to predict the change between time 1 and time 2. Some reasons relate to the purpose of the model; other reasons relate to data availability.

The cases that applied LTM, CLUE-S and CLUE used information directly from the reference map of time 2 concerning the quantity of each category, because the priority for those applications was to predict the spatial allocation of land change. The user can specify the quantity of each category independently from the spatial allocation for these models, which can be an advantage in allowing them to be used with tabular data and other types of models that generate non-spatial information concerning only the quantity of each land category. For example, CLUE-S and CLUE can set the quantity of each category by using case-study-specific and scale-specific methods ranging from trend extrapolations to complex sectoral models of world trade.

Some models such as SAMBA require information that is available only for years after time 1. SAMBA is an agent-based modeling framework that uses information from interviews with farmers concerning their land practices. For the Cho Don case, these interviews were conducted subsequent to time 2. Furthermore, the purpose of the SAMBA model is to explore scenarios with local stakeholders, not to predict the precise allocation of land transitions. The SAMBA team has been developing other methods for process validation of various aspects of their model (Castella et al. 2005b; Castella and Verburg 2007).

There are costs associated with separating calibration from validation information, because strict separation prohibits the use of some variables that are known to influence land change but are available only for time points beyond the calibration time interval. The Worcester case accomplished separation between calibration information and validation information by restricting the use of independent variables. For example, maps of contemporary roads and protected areas are available in digital form, but those maps contain some post-1971 information. The scientists for the Worcester application refrained from using these variables that are commonly associated with land change. Consequently, the Worcester case uses only slope and surficial geology as independent variables. Nevertheless, Pontius and Malanson (2005) show that there would not have been much increase in hits by using the map of protected areas, because such a map shows the places where change is prohibited, not the few places where change is likely to occur.

8.3.2.2 To Predict Small Amounts of Change

All 13 of the cases have less than 50% observed change, seven of the cases show less than 10% observed change, while the Holland(8), Santa Barbara, and Twin Cities have less than 4% observed change. Land change during a short time interval is usually a rare event, and rare events tend to be difficult to predict accurately. Figure 8.4 gives evidence that smaller amounts of change in the reference maps are associated with lower levels of predictive accuracy.

The challenge to detect and to predict change is made even more difficult by insisting upon rigorous separation of calibration data from validation data, especially in situations where data are scarce. For example, many models such as Environment Explorer are designed to examine change during a calibration interval from time 0 to time 1, and then to predict the change during a validation interval from time 1 to time 2. The Holland(15) case separates calibration information from validation information using this technique, where the calibration interval is only 7 years and the validation interval is only 4 years. In such situations, models may have difficulty in detecting a strong relationship between land change and the independent variables during the calibration interval, and the validation measurements may fail to find a strong relationship between the predicted land change and the observed land change during the validation interval. One solution would be for scientists to invest the necessary effort to digitize maps of historic land cover, so scientists can have a longer temporal extent and finer temporal resolution during which to calibrate and validate.

8.3.2.3 To Interpret the Influence of Quantity Error

Models that do not use the correct quantity of each category for time 2 must somehow predict the quantity for each category for time 2. Modelers need to be aware of how error in the prediction of quantity influences other parts of the validation process. Models typically fail to predict the correct allocation precisely; so models that predict more change are likely to produce more false alarms than models that predict less change, when assessed at fine spatial resolutions. For example, the Worcester case predicts more than the observed amount of change, which leads to false alarms. If the model were to predict less than the observed amount of change, then its output would have fewer false alarms and more correct rejections. In contrast, SLEUTH predicts less than half of the amount of observed change for the

Santa Barbara case, thus its error is close to that of a null model. It does not make sense to use criteria that reward systematic underestimates or overestimates of the quantity of each category. This is a weakness of using the percentage correct and the null model as benchmarks for predictive accuracy, and is a reason why Pontius et al. (2008) used the figure of merit as a criterion.

It is difficult to evaluate a model's prediction of spatial allocation when there is large error in quantity, especially when the model predicts less than the amount of observed change in the reference maps. We can assess the model's ability to predict spatial allocation somewhat when the model predicts the correct quantity, which is one reason modelers sometimes use the correct quantity at time 2 for simulation. Nevertheless, if we use only one potential realization of the model's output map, then the model's specification of spatial allocation is confounded with its single specification of quantity. The Total Operating Characteristic (TOC) is a quantitative procedure that can be used to measure a model's ability to specify the spatial allocation of land change in a manner that allows the modeler to consider various specifications of quantity (Pontius and Si 2014). Scientists can compute the TOC for cases where the model generates a map of relative priority for the gain of a particular category, which many models do in their intermediate steps. The TOC allows scientists to measure a model's ability to predict the few locations that change and a model's ability to predict the majority of locations that persist. The TOC is a recent advancement inspired by the Relative Operating Characteristic (Swets 1988; Pontius and Parmentier 2014).

8.3.3 Learning Challenges

8.3.3.1 To Use Appropriate Map Comparison Measurements

Scientists have invested a tremendous amount of effort to create elaborate algorithms to model landscape change. We are now at a point in our development as a scientific community to begin to answer the next type of question, specifically, "How well do these models perform and how do we communicate model performance to peers and others?" Therefore, we need useful measurements of map comparison and model performance. Pontius et al. (2008) derived a set of metrics to compare maps in a manner that we hope is both intellectually accessible and scientifically revealing, because analysis using rigorous and clear measurements is an effective way to learn. The initial invitation to participants asked them to submit their recommended criteria for map comparison. Few participants submitted any criteria, and those who did typically recommended the percentage of pixels in agreement between the reference map of time 2 and the prediction map of time 2.

This percentage correct criterion is one that many modelers consider initially. However, percentage correct can be extremely misleading, especially for cross-case comparisons. Percentage correct fails to consider the landscape dynamics, because percentage correct fails to include the reference map of time 1. For example, the Santa Barbara case has by far the largest percentage correct, 97%, simply because there is very little observed change on the landscape and the model predicts less than the amount of observed change. On the other hand, the Cho Don case has the smallest percentage correct, 54%, primarily because the Cho Don case has more observed change than any other case. The Perinet case has the largest figure of merit, while its percentage correct of 81% ranks just below the median of the 13 cases. Producer's Accuracy, User's Accuracy, and Kappa are other indices of agreement that are extremely common in GIS and can be quite misleading in assessing the accuracy of land change models (Pontius and Millones 2011). The figure of merit has properties that are more desirable than metrics that are frequently used for pattern validation of land change models (Pontius et al. 2007, 2011). We recommend the figure of merit for situations when it is necessary to rank numerous model runs with a single measurement. However, a single measurement offers only one bit of information thus fails to convey various important aspects of a pattern validation. For example, the figure of merit fails to convey the size of the reference change relative to the size of the predicted change.

We recommend much more highly that modelers report the sizes of misses, hits, wrong hits and false alarms, which are the components of the figure of merit. That combination of four measures is helpful in a variety of respects. For example, the false alarms are fewer than the misses when the model predicts less change than the reference change; and the false alarms are more than the misses when the model predicts more change than the reference change. If there exist false alarms at some locations and misses at other locations, then there exists allocation error. It is helpful to distinguish allocation error from quantity error, because the two types of error can have different implications for practical interpretation depending on the model's purpose. For example, if the purpose of the model is to simulate total carbon dioxide emissions due to deforestation, then allocation error is less important than quantity error for spatial extents where forest biomass is homogeneous (Gutierrez-Velez and Pontius 2012).

We need to continue to invest effort to improve methods of map comparison. The Map Comparison Kit includes a variety of new tools (Visser and de Nijs 2006). Modules in the GIS software TerrSet allow scientists to compare maps where the pixels have simultaneous partial membership to several categories, which is essential for multiple resolution comparison (Pontius and Connors 2009). The free software R contains packages that land change scientist will find helpful. The TOC package computes the Total Operating Characteristic (Pontius and Si 2014). The diffeR package gives components of difference at multiple spatial resolutions for two maps that show a single variable, such as maps from times 1 and 2 (Pontius and Santacruz 2014). Moulds et al. (2015) created in R the lulcc package, which performs a variety of operations, including the multiple resolution calculation of misses, hits, wrong hits, false alarms and correct rejections as derived by Pontius et al. (2011).

8.3.3.2 To Learn About Land Change Processes

During the panel discussions, participants agreed that a main purpose of modeling land use and cover change (LUCC) is to increase understanding of processes of LUCC, and that scientists should design a research agenda in order to maximize learning concerning such processes, not merely to increase predictive accuracy. Therefore, scientists should strive to glean from a validation exercise useful lessons about the processes of land change and about the next steps in the research agenda.

Some attendees at the panel discussions expressed concern that this chapter's validation exercises focus too much on prediction to the exclusion of increasing our understanding of the underlying processes of LUCC. Many scientists profess to seek explanation, not necessarily prediction. Some scientists think that a model can predict accurately for the wrong reasons; in addition these scientists think a model can capture the general LUCC processes, but not necessarily predict accurately due to inherent unpredictability of the processes. These participants reminded the audience that pattern validation examines the output maps from the simulation models but does not examine whether the structure of the algorithm matches theory concerning the processes of change. Process validation is required to validate the structure of the algorithm for process based models, especially when path dependence plays a role (Brown et al. 2005).

Other scientists see pattern validation as a means to distinguish better explanations from poorer explanations concerning the LUCC processes. For these other scientists, pattern validation allows a modeler to gain insight concerning the degree to which the simulated change is similar to the observed change. Furthermore, scientists must test the degree to which the past is useful to predict the future because this allows scientists to measure the scales at which LUCC processes are stable over time. A model's failure to predict accurately may indicate that the process of land change is non-stationary in time and/or space, in which case pattern validation can reveal information that is helpful to learn about LUCC processes (Chen and Pontius 2010; Pontius and Neeti 2010). Thus there is need for new methods, such as Intensity Analysis, that test for stationarity at various levels, even before any predictive model is run (Aldwaik and Pontius 2013; Runfola and Pontius 2013). If scientists interpret the validation procedure in an intelligent manner, then they can perhaps learn more from inaccurate predictions than from accurate ones. Consequently, inaccurate predictions do not mean that the model is a failure, because validation can lead to learning regardless of the revealed level of accuracy.

This difference in views might explain the variation in the LUCC modeling community concerning how best to proceed. One group thinks that models are too simple so that future work should consider more variables and develop more complex algorithms so the models can generate a multitude of possible outcomes. A second group insists that such an approach would only exacerbate an existing problem that models are already too complicated to allow for clear communication, even among experts. From this second perspective, contemporary models lack aspects of scientific rigor that would not be corrected by making the models more complex. For example, many models fail to separate calibration information from validation information, fail to apply useful methods of map comparison, and fail to measure how scale influences the analysis. For this second group of scientists, it would be folly to make more complicated algorithms and to include more variables before we tackle basic issues, because we will not be able to measure whether more complex models actually facilitate learning about LUCC processes until we develop and use helpful measures of model performance. This apparent tension could be resolved if the scientists who develop more complex models collaborate with the scientists who develop clearer methods of model assessment.

8.3.3.3 To Collaborate Openly

Participants at the panel sessions found the discussions particularly helpful because the sessions facilitated open and frank cross-laboratory communication. Many conference participants expressed gratitude to the co-authors who submitted their maps in a spirit of openness for the rest of the community to analyze in ways that were not specified a priori. The design of the exercise encouraged participation and open collaboration because it was clear to the participants that the analysis was not attempting to answer the question "Which model is best?"

Some participants in the conference discussions reported that they have felt professional pressure to claim that their models performed well in order for their manuscripts to be accepted for publication in peer-reviewed journals. We hope that this chapter opens the door for honest and helpful reporting about modeling results. In particular, we hope that editors and reviewers will learn as much from this study as the conference participants did, so that future literature includes useful information about model assessment. The criterion for acceptance of manuscripts should be rigor of method and clarity of presentation, not results concerning predictive accuracy, and certainly not vacuous claims of success.

There is clearly a desire to continue this productive collaboration because it greatly increases learning. One particularly constructive suggestion is to build a LUCC data digital library so that scientists would have access to each others' data, models, and modeling results. The data would be peer-reviewed and have metadata sufficient so that anyone could perform cross-model comparison with any of the entries in the library. In order for this to be successful, scientists need sufficient motivation to participate, which requires funding and professional recognition for participation.

8.4 Conclusions

The collective experience of the co-authors supports the statement that all models are wrong but some are useful (Box 1979). All 13 of the models are wrong in the respect that the outputs have errors. Errors in pattern validation mean that the patterns extrapolated from the calibration time interval were not stationary with the patterns observed during the validation time interval. These errors are a reflection of the landscape as much as they are a reflection of the model. If the scientists interpret the results in a useful manner, then scientists can learn; and if scientists learn from a model, then the model was successful at advancing science. It is essential to use measurements that can be interpreted with respect to a model's intended purposes in order to facilitate learning. Clarity and rigor are necessary to establish procedures and measurements for informative judgments concerning model performance. This chapter illuminates common pitfalls and offers guidance for ways to overcome the pitfalls. Specifically, we recommend modelers report the sizes of misses, hits, wrong hits, and false alarms. Those four measurements are based on the mathematical ideas concerning the intersection of sets, which are regularly taught to elementary school students. If scientists meet the challenges specified in this chapter, then we are likely to learn efficiently, because meeting these challenges can help scientists prioritize a research agenda for land change science. To facilitate open collaboration, we have made the raster maps used in this cross-case comparison available for free at www.clarku.edu/~rpontius.

Acknowledgments The C.T. DeWit Graduate School for Production Ecology & Resource Conservation of Wageningen University sponsored the first author's sabbatical, during which he led the collaborative exercise that is the basis for this chapter. The National Science Foundation of the USA supported this work via the grant "Infrastructure to Develop a Human-Environment Regional Observatory (HERO) Network" (Award ID 9978052). Clark Labs (www.clarklabs.org) produces the software TerrSet®, which we used for the GIS analysis. Our colleagues shared valuable insights during several discussion sessions at professional conferences.

References

Aldwaik SZ, Pontius RG Jr (2013) Map errors that could account for deviations from a uniform intensity of land change. Int J Geogr Inf Sci. doi:10.1080/13658816.2013.787618

Aldwaik SZ, Onsted JA, Pontius RG Jr (2014) Behavior-based aggregation of land categories for temporal change analysis. Int J Appl Earth Obs Geoinf 35:229–238

Boissau S, Castella J-C (2003) Constructing a common representation of local institutions and land use systems through simulation-gaming and multi-agent modeling in rural areas of Northern Vietnam: the SAMBA-Week methodology. Simul Gaming 34(3):342–347

Box GEP (1979) Robustness in the strategy of scientific model building. In: Launer RL, Wilkinson GN (eds) Robustness in statistics. Academic, New York, pp 201–236

Brown DG, Page S, Riolo R, Zellner M, Rand W (2005) Path dependence and the validation of agent-based spatial models of land use. Int J Geogr Inf Sci 19(1):153–174

Brown, D.G., Band, L.E., Green, K.O., Irwin, E.G., Jain, A., Lambin, E.F., Pontius Jr, R.G., Seto, K.C., Turner II, B.L., Verburg, P.H. (2013). Advancing land change modeling: opportunities and research requirements. The National Academies Press: Washington, DC. 145. http://www.nap.edu/catalog.php?record_id=18385

Castella J-C, Verburg PH (2007) Combination of process-oriented and pattern-oriented models of land-use change in a mountain area of Vietnam. Ecol Model 10(1):410–420

Castella J-C, Boissau S, Trung TN, Quang DD (2005a) Agrarian transition and lowland-upland interactions in mountain areas in northern Vietnam: application of a multi-agent simulation model. Agric Syst 86(3):312–332

Castella J-C, Trung TN, Boissau S (2005b) Participatory simulation of land use changes in the Northern Mountains of Vietnam: the combined use of an agent-based model, a role-playing game, and a geographic information system. Ecol Soc 10(1):27

Chen H, Pontius RG Jr (2010) Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. Landsc Ecol 25:1319–1331

de Koning GHJ, Verburg PH, Veldkamp TA, Fresco LO (1999) Multi-scale modelling of land use change dynamics in Ecuador. Agric Syst 61:77–93 de Nijs TCM, de Niet R, Crommentuijn L (2004) Constructing land-use maps of the Netherlands in 2030. J Environ Manag 72(1–2):35–42

Dietzel CK, Clarke KC (2004) Spatial differences in multi-resolution urban automata modeling. Trans GIS 8:479–492

Duan Z, Verburg PH, Fengrong Z, Zhengrong Y (2004) Construction of a land-use change simulation model and its application in Haidian District, Beijing. Acta Geograph Sin 59(6):1037–1046. (in Chinese)

Enaruvbe G, Pontius RG Jr (2015) Influence of classification errors on intensity analysis of land changes in southern Nigeria. Int J Remote Sens 31(1):244–261

Engelen G, White R, de Nijs T (2003) The Environment Explorer: spatial support system for integrated assessment of socio-economic and environmental policies in the Netherlands. Integr Assess 4(2):97–105

Fotsing E, Verburg PH, De Groot WT, Cheylan J-P, Tchuenté M (2013) Un modèle intégré pour explorer les trajectoires d'utilisation de l'espace. ARIMA J 16:1–28

Goldstein NC (2004) Brains vs. Brawn – comparative strategies for the calibration of a cellular automata-based urban growth model. In: Atkinson P, Foody G, Darby S, Wu F (eds) GeoDynamics. CRC Press, Boca Raton, pp 249–272

Gutierrez-Velez V, Pontius RG Jr (2012) Influence of carbon mapping and land change modelling on the prediction of carbon emissions from deforestation. Environ Conserv 39(4):325–336

Hilferink M, Rietveld P (1999) Land use scanner: an integrated GIS based model for long term projections of land use in urban and rural areas. J Geogr Syst 1(2):155–177

Hoymann J (2010) Spatial allocation of future residential land use in the Elbe River Basin. Environ Plan B: Plan Des 37(5):911–928

Kok K, Veldkamp TA (2001) Evaluating impact of spatial scales on land use pattern analysis in Central America. Agric Ecosyst Environ 85(1–3):205–221

Kok K, Farrow A, Veldkamp TA, Verburg PH (2001) A method and application of multi-scale validation in spatial land use models. Agric Ecosyst Environ 85(1–3):223–238

Koomen, E, Borsboom-van Beurden, J. (2011). Land-use modelling in planning practice. GeoJ Libr 101, Dordrecht: Springer

Kuhlman T, Diogo V, Koomen E (2013) Exploring the potential of reed as a bioenergy crop in the Netherlands. Biomass Bioenergy 55:41–52

Laney RM (2002) Disaggregating induced intensification for land-change analysis: a case study from Madagascar. Ann Assoc Am Geogr 92(4):702–726

Loonen W, Koomen E (2009) Calibration and validation of the land use scanner allocation algorithms, PBL publication number 550026002. Netherlands Environmental Assessment Agency (PBL), Bilthoven

McConnell W, Sweeney SP, Mulley B (2004) Physical and social access to land: spatio-temporal patterns of agricultural expansion in Madagascar. Agric Ecosyst Environ 101(2–3):171–184

Moulds S, Buytaert W, Mijic A (2015) An open and extensible framework for spatially explicit land use change modelling: the lulcc R package. Geosci Model Dev 8:3215–3229

Paegelow M, Camacho Olmedo MT, Houet T, Mas J-F, Pontius RG Jr (2013) Land change modeling: moving beyond projections. Int J Geogr Inf Sci 27(9):1691–1695

Pijanowski BC, Gage SH, Long DT (2000) A land transformation model: integrating policy, socioeconomics and environmental drivers using a geographic information system. In: Harris L, Sanderson J (eds) Landscape ecology: a top down approach. CRC Press, Boca Raton, pp 183–198

Pijanowski BC, Brown DG, Manik G, Shellito B (2002) Using neural nets and GIS to forecast land use changes: a land transformation model. Comput Environ Urban Syst 26(6):553–575

Pijanowski BC, Pithadia S, Sellito BA, Alexandridis K (2005) Calibrating a neural network-based urban change model for two metropolitan areas of the Upper Midwest of the United States. Int J Geogr Inf Sci 19(2):197–215

Pontius RG Jr, Connors J (2009) Range of categorical associations for comparison of maps with mixed pixels. Photogramm Eng Remote Sens 75(8):963–969

Pontius RG Jr, Lippitt CD (2006) Can error explain map differences over time? Cartogr Geogr Inf Sci 33(2):159–171

Pontius RG Jr, Malanson J (2005) Comparison of the structure and accuracy of two land change models. Int J Geogr Inf Sci 19(2):243–265

Pontius RG Jr, Millones M (2011) Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int J Remote Sens 32(15):4407–4429

Pontius RG Jr, Neeti N (2010) Uncertainty in the difference between maps of future land change scenarios. Sustain Sci 5:39–50

Pontius RG Jr, Parmentier B (2014) Recommendations for using the Relative Operating Characteristic (ROC). Landsc Ecol 29(3):367–382

Pontius RG Jr, Petrova S (2010) Assessing a predictive model of land change using uncertain data. Environ Model Softw 25(3):299–309

Pontius RG Jr, Santacruz A (2014) Quantity, exchange and shift components of differences in a square contingency table. Int J Remote Sens 35(21):7543–7554

Pontius RG Jr, Si K (2014) The total operating characteristic to measure diagnostic ability for multiple thresholds. Int J Geogr Inf Sci 28(3):570–583

Pontius RG Jr, Cornell J, Hall C (2001) Modeling the spatial pattern of land-use change with GEOMOD2: application and validation for Costa Rica. Agric Ecosyst Environ 85(1–3):191–203

Pontius RG Jr, Huffaker D, Denman K (2004a) Useful techniques of validation for spatially explicit land-change models. Ecol Model 179(4):445–461

Pontius RG Jr, Shusas E, McEachern M (2004b) Detecting important categorical land changes while accounting for persistence. Agric Ecosyst Environ 101(2–3):251–268

Pontius RG Jr, Walker R, Yao-Kumah R, Arima E, Aldrich S, Caldas M, Vergara D (2007) Accuracy assessment for a simulation model of Amazonian deforestation. Ann Assoc Am Geogr 97(4):677–695

Pontius RG Jr, Boersma W, Castella J-C, Clarke K, de Nijs T, Dietzel C, Duan Z, Fotsing E, Goldstein N, Kok K, Koomen E, Lippitt CD, McConnell W, Mohd Sood A, Pijanowski B, Pithadia S, Sweeney S, Trung TN, Veldkamp AT, Verburg PH (2008) Comparing the input, output, and validation maps for several models of land change. Ann Reg Sci 42(1):11–47

Pontius RG Jr, Peethambaram S, Castella J-C (2011) Comparison of three maps at multiple resolutions: a case study of land change simulation in Cho Don District, Vietnam. Ann Assoc Am Geogr 101(1):45–62

Runfola D, Pontius RG Jr (2013) Measuring the temporal instability of land change using the flow matrix. Int J Geogr Inf Sci. doi:10.1080/13658816.2013.792344

Silva EA, Clarke KC (2002) Calibration of the SLEUTH urban growth model for Lisbon and Porto, Portugal. Comput Environ Urban Syst 26:525–552

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

Tan M, Li X, Xie H, Lu C (2005) Urban land expansion and arable land loss in China – a case study of Beijing-Tianjin-Hebei region. Land Use Policy 22(3):187–196

Veldkamp AT, Fresco L (1996) CLUE-CR: an integrated multi-scale model to simulate land use change scenarios in Costa Rica. Ecol Model 91:231–248

Verburg PH, Veldkamp TA (2004) Projecting land use transitions at forest fringes in the Philippines at two spatial scales. Landsc Ecol 19:77–98

Verburg PH, de Koning GHJ, Kok K, Veldkamp A, Bouma J (1999) A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. Ecol Model 116(1):45–61

Verburg PH, Soepboer S, Veldkamp TA, Limpiada R, Espaldon V, Sharifah Mastura SA (2002) Modeling the spatial dynamics of regional land use: the CLUE-S model. Environ Manag 30(3):391–405

Verburg PH, de Nijs TCM, van Eck JR, Visser H, de Jong K (2004) A method to analyse neighbourhood characteristics of land use patterns. Comput Environ Urban Syst 28(6):667–690

Visser H, de Nijs T (2006) The map comparison kit. Environ Model Softw 21(3):346–358

Part IV
Multi-scale Representation and Analysis

Chapter 9
Applications of 3D City Models for a Better Understanding of the Built Environment

Bruno Willenborg, Maximilian Sindram, and Thomas H. Kolbe

Abstract The administration of modern cities is a complex task involving various disciplines. To satisfy their specific needs regarding planning and decision making, all of them require a virtual representation of the city. Semantic 3D city models offer a reliable and increasingly available virtual representation of real world objects in an urban context. They serve as an integration platform for information and applications around the city system, because data from different domains can be linked to the same objects representing real world urban objects. This work gives an overview on the current state of applications based on semantic 3D city models and how they can be categorized. Three use cases are explained in detail. Based on city models according to the CityGML standard, first a tool for estimating the solar irradiation on roofs and facades is introduced. By the combination of a transition model, sun position calculation, and an approximation of the hemisphere the direct, diffuse and global irradiation as well as the SkyViewFactor are computed. Second, an application for the simulation of detonations in urban space is presented. The city model is converted to a field-based representation for running a Computational Fluid Dynamics (CFD) simulation. By storing logical links between the object and the field-based representation of the city model, information exchange between the simulation tool and the city models is realized. The third application demonstrates the estimation of the energy demand of buildings based on official statistical data and the simulation of refurbishment measures. All three applications use a cloud-based 3D web client for visualization of the city model and the application results including interactive analysis capabilities.

Keywords CityGML • Semantic 3D city model • Solar potential analysis • Façade • Detonation simulation • Energy demand estimation

B. Willenborg (✉) • M. Sindram • T.H. Kolbe
Department of Civil, Geo and Environmental Engineering, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany
e-mail: b.willenborg@tum.de1; maximilian.sindram@tum.de1; thomas.kolbe@tum.de1

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_9

9.1 Introduction

Virtual 3D city models have been used for many years to capture and explore the view of a city. Visualization and visibility analysis have been (and still are) key applications. The requirements on a 3D city model for these type of applications are rather low. Basically, a Digital Surface Model (DSM) is used to describe the geometry of the Earth's surface – including the shape of the natural and built environment like trees or buildings. In addition, photographic images are mapped onto the DSM providing color and appearance information. Due to major progress in photogrammetry and remote sensing technology and methodology over the last ten years these 3D models can be generated from airborne or terrestrial mapping campaigns in a fully automated way. Good examples are the 3D models provided in Google Earth or in Apple's map application.

While the visual aspects of the built environment are well covered, the before mentioned models do not carry any knowledge of what they are representing. Visualization models in principle just consist of geometric elements like 3D polygons, volumes, or meshes with additional appearance information. The interpretation of the rendered 3D model happens completely by the (human) viewer relying on his capability to recognize and discriminate the individual urban objects like buildings, bridges, roads, trees etc.

The administration and development of modern cities is a complex task involving many disciplines, each of them with their own requirements. To satisfy their specific needs regarding planning and decision making, all of them require a virtual representation of the cityscape, that allows for much more than mere visualization. For instance, to determine the total roof surface area of a city quarter, information on which surfaces represent roofs and a geometric representation allowing area calculation are indispensible.

Semantic 3D city models not only represent the shape and graphical appearance of urban objects but contain semantic information describing their thematic proper ties, taxonomies, aggregations and interrelations. As depicted in Fig. 9.1, the visual quality of semantic 3D city models (right image) may be lower than in visualizations models (left image), but it is possible for machines/algorithms to distinguish urban object like buildings (see highlighted building in the right image) and use their rich thematic and geometric information for queries, statistical computation, simulation, and visualization. Driven by the growing availability of semantic 3D city models and the expanding number of thematic classes for different object types (e.g. roads, vegetation, bridges, tunnels, etc.) new applications in the context of urban planning arise. In the following, we introduce the main modelling concepts and show selected application scenarios that demonstrate the added value of semantic 3D city models coping with current social, ecological, and economical challenges.

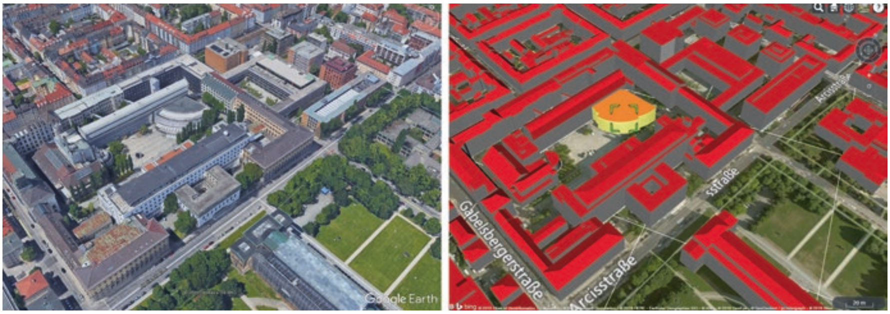


Fig. 9.1 Comparison of visualization only (*left image*, source: Google Earth) and semantic 3D city models (*right image*, source: State mapping agency of Bavaria (LDBV)): while the visual quality is higher in the visualization model, individual objects, as for instance the *highlighted building*, can be discriminated computationally within the semantic model

9.2 Semantic 3D City Models and CityGML

The international standard City Geography Markup Language (CityGML) is an open data model and encoding format that has been developed for the representation and exchange of *semantic virtual 3D city and landscape models*. CityGML comprises information on the geometry, appearance, semantics and topology of objects in an urban context. The city objects are decomposed following logical criteria which can be observed in the real world according to the ISO 19109 definition of geographic objects (ISO 19109:2005(E) 2005). The exchange format defined by CityGML is based on the Extensible Markup Language (XML) and the ISO 19100 standards family, for instance the ISO 19107 standard (Herring 2001). The standard is an application schema of the Geography Markup Language version 3.1.1 (GML3). Its latest issue, version 2.0.0, was released in 2012 as an official standard of the Open Geospatial Consortium (OGC) (Kolbe 2009).

The CityGML standard was designed to serve as a universal topographic information model independent of specific subject areas. It defines a common understanding of the segmentation of the most relevant features classes of a city and their attributes. Hence, the standard serves as an information model for a broad range of applications like urban planning, civil engineering, environmental simulations or tourism. Figure 9.2 gives an overview on the modular structure of CityGML. Based on a core module 10 thematic modules for e.g. buildings, transportation systems or vegetation are defined which can be freely combined according to the given application context (Kolbe 2009).

However, in practical applications it is frequently required to store and exchange additional information, which is not covered by the predefined classes mentioned above. Therefore, CityGML supports two extension mechanisms, *generics* and *Application Domain Extension (ADE)* as shown in Fig. 9.2. All city objects can carry an arbitrary number of *generic attributes*, which are defined by a name, data type and value. Moreover, *generic city objects* with arbitrary geometries and generic

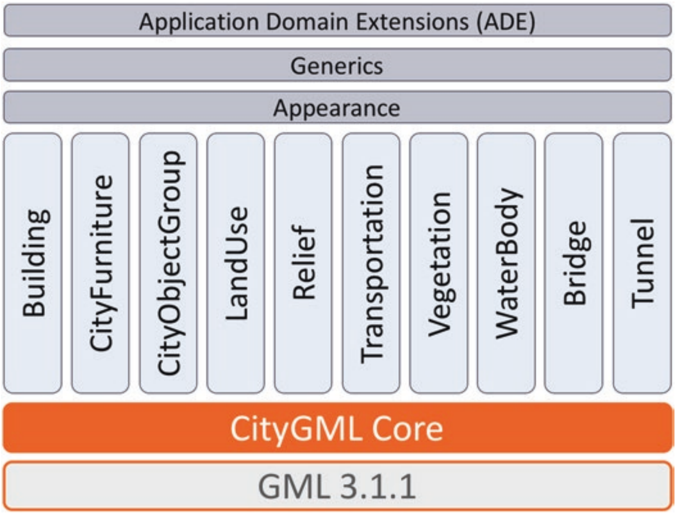


Fig. 9.2 Modularization of CityGML. *Vertical* modules contain the semantic modeling for different thematic domains. The *horizontal* modules contain core functionality and mechanisms for different kinds of graphical appearance of city objects and for extending the predefined thematic modules

attributes can be defined. The second extension concept are so called ADEs. ADEs allow the extension of existing thematic modules and the creation of new feature classes. In contrast to generics, ADEs are defined in a separate XML schema definition file with their own namespace. Hence, they are formally specified and instance files can be validated against the schema of the ADE (Kolbe 2009). An example ADE for modelling traffic noise emissions (Noise ADE (Czerwinski et al. 2006)) is provided within the CityGML 2.0 specification (Gröger et al. 2012). Other popular ADE examples are the Energy ADE extending the CityGML model with features for building heat demand (Nouvel et al. 2015) and the currently developed Utility Networks ADE enabling the modelling of supply and disposal networks for analysing the urban supply situation (Becker et al. 2011, 2013; Kutzner and Kolbe 2016).

On the subject of variable resolution requirements of different applications, CityGML supports a multi-scale representation of objects with five consecutive Level of Details (LoDs). Objects become more detailed both geometrically and thematically with increasing LoD. Each object can be stored in different LoDs simultaneously, allowing its analysis and visualization according to the degree of detail, the given application context requires. Level of Detail 0 (LoD0) is a coarse representation of the earth's surface, Level of Detail 1 (LoD1) is the well know blocks model, where all 3D objects are created by vertical extrusions of footprints. Level of Detail 2 (LoD2) offers distinctive roof structures for buildings, while Level of Detail 3 (LoD3) denotes architectural models with detailed wall and roof surfaces, windows and doors. Level of Detail 4 (LoD4) adds building interiors like rooms, stairs and furniture. The LoD concept applies to all other CityGML features types as well (Kolbe 2009).

Semantic 3D city models are predominantly created and provided by public mapping agencies, which ensures their sustainable maintenance and updating. They are derived in fully or semiautomated workflows from official 2D cadastral data and elevation information from airborne laser scanning or aerial images. However, the automated creation of CityGML models based on open data is feasible as well. Kolbe et al. (2015) created a model of New York City based on 26 different data sets from the New York City Open Data Portal, comprising all buildings, land parcels, roads, parks, the digital terrain model, and water bodies – all with 3D geometries and between 10 and 80 thematic attributes. At the national level the Working Committee of the Surveying Authorities of the States in Germany (AdV) is prescribing a uniform and nationwide dissemination of building models in Germany by the mapping agencies. In Germany, almost all of the existing buildings are currently available in LoD1. This comprises more than 50 million single building objects. As of December 31st, 2016 models in LoD2 for the total building stock are available for the German states North Rhine-Westphalia, Rhineland-Palatinate, Saarland, Saxony, and Saxony-Anhalt with most other german states to be completed by 2018 (Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV): Produktblatt – 3D-Gebäudemodelle LoD2 2016). At the European level, a unified and standards-based availability of building models is determined by the Infrastructure for Spatial Information in the European Community (INSPIRE) directive (INSPIRE: EU Directive 2007). As one of 34 themes, the theme *Buildings* is covering building specific data for different use cases. Gröger et al. (2013) are proposing a CityGML-based encoding for the INSPIRE Data Specification on buildings allowing their use as CityGML buildings and thus bridging the gap between political requirements and data availability for semantic 3D city models.

The upper part of Fig. 9.3 shows some examples of the large number of international cities such as Singapore, Paris, Zurich, Vienna, London, New York, Vancouver, Montreal, and Helsinki that provide and maintain a semantic 3D city model. The federal German state North Rhine-Westphalia and the city of Berlin are even distributing their city models at no charge for both commercial and non-commercial use to foster the usage of the data and accelerate the development of new applications. These applications cover, amongst others, energy demand and production estimation, noise immission simulation and mapping, real estate and urban facility management and vulnerability analysis and disaster management. To ensure, that these tools can be applied in cities all over the globe, a common understanding of the most important urban features and standardization of the underlying data model and exchange format is required. This enables software developers to design tools for a broader audience and facilitates data exchange between software components of different domains and development teams based on the objects of the city model.

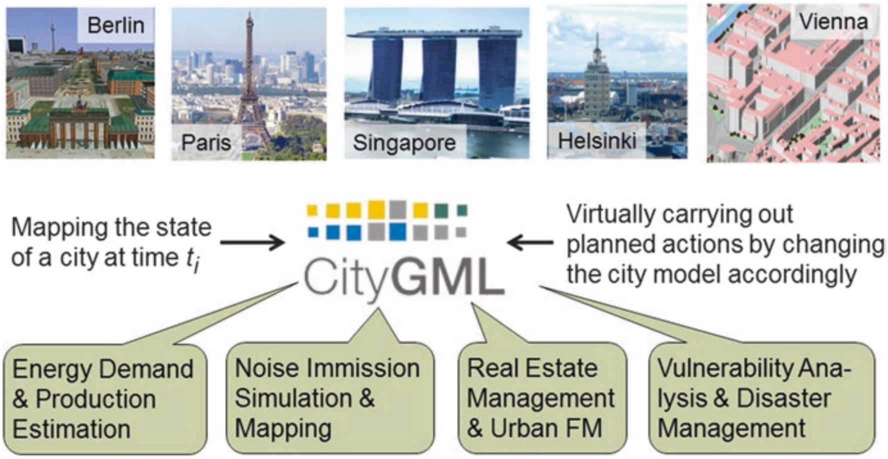

Berlin
Paris
Singapore
Helsinki
Vienna
Mapping the state of a city at time t_i
CityGML
Virtually carrying out planned actions by changing the city model accordingly
Energy Demand & Production Estimation
Noise Immission Simulation & Mapping
Real Estate Management & Urban FM
Vulnerability Analysis & Disaster Management

Fig. 9.3 Semantic 3D city models serve as information hub for different application scenarios from various disciplines. The CityGML standard harmonizes access to the most common urban features. Thus, applications that are based on the standard are guaranteed to work among different cities

9.3 Applications of 3D City Models: An Overview

Today, 3D city models are used in a wide range of applications covering diverse use cases and application domains. The work of Biljecki et al. (2015) gives an extensive overview on the current utilization of 3D city models and introduces a hierarchical terminology for their segmentation, which is briefly recapped in the following.

According to Biljecki et al., the biggest issue, avoiding a straight forward inventory of applications of 3D city models, are many undefined terms in the context of 3D spatial information like use case, application, or operation. Even the definition of 3D city models is not consistently used. Hence, a well-defined categorization of 3D city model applications based on *application domains*, *use cases* and *spatial operations* is not feasible as these terms are overlapping.

Therefore, the authors decided to focus on the listing of use cases. Additionally, applications for a better understanding of the individual use case are collected. When trying to find a taxonomy for use cases, that is both mutually exclusive (a use case can only be part of one category) and collectively exhaustive (all categories cover all use cases) the only valid criteria that could be identified is the visualization aspect. Hence, use cases are categorized into the two following groups.

One the one hand, *non-visualization* use cases are described, which require visualization neither of the 3D city model nor the results of the spatial operations the use case comprises. For instance, the solar potential analysis discussed in Sect. 9.4, an application of the use case of estimating solar irradiation, falls into that category. The simulation results can be visualized, but this is not essential to achieve the purpose of the use case. The information the simulation generates is meant to be used for the identification of suitable areas for solar energy generation. As the results are written to a database, this task can be performed using a query without visualization.

On the other hand, the authors delineate visualization-based use cases. They include cases, where visualization is very important but not essentially required. An example for this is navigation, which works fine with state-of-the-art text to speech software, but greatly benefits from visualization. Second, visualization-only use cases, like virtual reality or communication of urban information are covered by that category. This categorization is consistent with the separation into visual 3D models and semantic 3D city models as described in Sect. 9.1.

In an extensive literature review, the authors identified more than 29 use cases including more than 100 applications, which are arranged into these categories. The non-visualization use cases comprise the estimation of solar irradiation, energy demand estimation, aiding positioning, determination of floorspace and classification of building types and is much smaller than the visualization-based use cases category, that includes more than 20 entries. A complete list of the use cases and a brief description of the included applications can be found in the original work of Biljecki et al. (2015). A brief summary of the most important use cases grouped in four topics is given in the following.

The topic *energy* comprises the use cases of estimating solar irradiation and building energy demands. To compute the insolation on a building the geometric information of the city model building surfaces like the inclination, orientation and area is taken as input for solar empirical models to evaluate its suitability for solar energy generation (photovoltaics (PV) or solar thermal collector (ST)). Please find a detailed application example in Sect. 9.4. For the estimation of the energy demand of a building both geometric and thematic properties of the city model are taken into account. The combination of a buildings' volume, shared wall surface areas and its construction year allows the estimation of its heating energy demand, as discussed in the detailed example given in Sect. 9.6.

The second topic is *homeland security and vulnerability*. One of its central use cases is visibility analysis, where the Line of Sight (LoS) between two points is computed based on the geometries of the city model. For instance, this information is used for optimizing the placement of security cameras (Yaagoubi et al. 2015) or evaluating the hazards of sniper terrorism (Vanhorn and Mosurinjohn 2010). Another relevant use case in this context is emergency response. 3D city models contribute valuable information for the preparation for emergency situations and quick response scenarios like building entry points (doors and windows) or even detailed indoor models for improving evacuation planning or fire fighter ladder positioning (Chen et al. 2014; Kwan and Lee 2005; Tashakkori et al. 2015). Becker et al. (2011) use 3D city models including utility networks (Kutzner and Kolbe 2016) for estimating cascading effects of critical infrastructure failure in cases of disasters or emergency situations. Moreover, an application example applying a Computational Fluid Dynamics (CFD) simulation for the assessment of blast effects in an urban context based on the thematic and geometric information of 3D city model buildings is discussed in Sect. 9.5.

The most relevant use cases for the third category, *traffic and mobility*, are visualization for navigation and routing. 3D city model objects like buildings are of- ten familiar landmarks that help users with orientation in navigation applications. The 3D geometry representation of city models is more realistic than the symbolic representation provided by 2D maps and contains more navigation cues (Oulasvirta et al. 2009; Schilling et al. 2005). Moreover, semantic 3D city models allow for optimizing the 3D view based on the thematic information they provide (Mao et al. 2015; Nedkov 2012). 3D city models have gained interest for routing purpose as 3D navigation techniques become available (Hildebrandt and Timm 2014) and they contain objects that, are not available in 2D maps like steps and ramps, that, for instance, influence the navigable space for pedestrians (Slingsby and Raper 2008). If the 3D city model contains information on the interior of buildings, this information can be use for way finding and accessibility applications (Khan Aftab and Kolbe Thomas 2013; Khan and Kolbe 2012; Khan et al. 2015; Liu and Zlatanova 2011; Thill et al. 2011).

Climate and environment is the fourth topic. Its most prominent use cases are the estimation of noise propagation and CFD simulations for various phenomena including flooding. The estimation of noise propagation benefits from 3D geometries, as the noise level varies for different height levels due to refraction (Kubiak and Ławniczak 2015). Seman tic information's can be used to obtain noise propagation simulation parameters like traffic density, as the work of Czerwinski et al. (2006), (2007) shows 3D city models are a common basis for CFD simulation. Most applications are found in field of microclimate analysis for e.g. evaluation of air quality and pollutant dispersion (Ujang et al. 2013), wind comfort (Janssen et al. 2013) or the urban thermal environment (Maragkogiannis et al. 2014). Estimating the extend and impact of flood events can be enhanced compared to 2D methods using 3D city models as well (Schulte and Coors 2008). The multi-resolution flood simulation approach developed by Varduhn et al. (2015) utilizes the drainage system of a City Geography Markup Language (CityGML) city model to include pipe network interactions an allow predictions for individual buildings. As discussed in Sect. 9.5, the exchange of semantic information between simulation system and city model can be beneficial for both sides.

9.4 Estimation of Solar Irradiation Using Semantic 3D City Models

Solar irradiation is a clean, silent, secure and abundantly available energy source. Due to decreasing costs and improvements in technology and acceptance, Photovoltaics (PV) and solar thermal collectors (STs) are going to play a key role in the future energy production, especially in urban areas where a significant portion of the energy is consumed. In 2010, the EU Directive 2010/31/EU introduced the Nearly Zero Energy Buildings concept, requiring that the local energy production of all new buildings after the year 2020 covers their local energy demand. PV and ST systems foster this concept of decentralized energy production due to their high modularity. Furthermore, transmission and distribution losses are avoided, as the energy is produced at its point of use (Brito et al. 2012; Redweik et al. 2013).

To meet the requirements of EU legislation, in the future much larger areas for PV will be required. As facades are much larger than roofs in modern cities and are mostly devoid of building installations and infrastructure like chimneys, dormers, air conditioning units or elevator engines and usually present better maintenance conditions than PV panels on roofs, as vertical surfaces do not accumulate so much dust and are usually free of snow in the winter, they increasingly gain interest for deployment of PV in residential areas (Redweik et al. 2013). Moreover, the combination of energy production with other building functions like heat insulation, cladding or illumination with semi–transparent photovoltaic modules may offer interesting benefits (Catita et al. 2014). For the successful deployment of PV and ST systems in the urban area, the local potential of roofs and facades needs to be investigated, taking influencing variables like the local meteorological and climate conditions and shadowing effects of the surrounding topographic features into account. Semantic 3D city models are an ideal data source for such assessment as they combine a detailed representation of the cityscape with visualization and analytic capabilities.

9.4.1 Estimation of Urban Solar Energy Potential for Facades and Roofs

The method for estimating solar irradiation in urban areas introduced in this section is based on the Master's Thesis of Wolfgang Zahn (2015). His work has been revised and implemented in Java, as a plugin for the 3DCityDB Importer/Exporter, the standard database management utility for the 3DCityDB for CityGML (3DCityDB). The application has been enhanced for increased performance and functionality with the main objective to develop a user-friendly tool that enables non-expert users to perform and evaluate solar potential analysis for city models of arbitrary size.

The model computes the direct and diffuse solar irradiation and the SkyViewFactor (SVF) for roofs and facades considering shadowing effects of buildings and a Digital Terrain Model (DTM) while ignoring the influence of reflected radiation. Ground features are not being processed, as those areas are usually not available for solar energy generation in cities and would needlessly increase runtime. The only input data required is a 3D city model in Level of Detail 2 (LoD2) according to the City Geography Markup Language (CityGML) standard, where building roofs and facades are modeled as thematic surfaces (Gröger et al. 2012). Optionally, a DTM can be integrated as well.

The direct solar irradiation is modeled using a combination of the transition model developed by Fu and Rich (1999) and an algorithm for computing the position of the sun from the work of Grena (2012). First, the sun positions for 1 year are computed for a freely selectable observation point with geographic coordinates given in Latitude (LAT)/Longitude (LON) and height above sea level. Usually, the center of the city model is being used. Time intervals between the sun positions can be configured in steps of hours and days where typically an 1-h-interval is selected, considering performance and quality aspects. The sun positions are described by two angles, one for orientation (azimuth α) and the other for height (zenith θ) relative to the observation point applying an algorithm introduced by Grena (2012) providing a maximum error of 0.19°. The resulting sun positions are stored as point features in a radius of 100,000 km around the observation point in the Coordinate Reference System (CRS) of the city model in the 3DCityDB with their radiation power as attribute. Second, the radiation power [kWhm−2] of each sun point is calculated using a simplified transition model based on Fu and Rich (1999) considering the transmissivity (τ) of the atmosphere depending on the height of the sun point and the relative optical path length m (θ).

For robustness against regional atmospheric differences the transmissivity and the fraction of the diffuse irradiation of the global irradiation are calibrated using freely available data from the NASA Atmospheric Science Data Center. Implementing an iterative approach, both parameters are adjusted until they match 22 year mean values of the NASA Surface meteorology and Solar Energy (SSE) mission (Langley Research Center 2016), which are queried online by LAT/LON coordinates for each simulation run, allowing a worldwide application of the tool with sufficient result quality.

The diffuse irradiation and the SVF are computed using a simplified approximation of the sky dome with points, where each point represents a spherical segment. The azimut (orientation) and zenit (height) angle distance between the points and the azimut angle offset between individual zenit angle layers can be configured enabling the creation of a hemisphere, where each point represents the same fraction of the area, which produces the best results according to Zahn (2015). Moreover, hemispheres with variable point density can be created to adapt to performance and quality requirements of the given use case. To compute the radiation power of the hemisphere points the Standard Overcast Sky (SOC) model according to Fu and Rich (1999) is used. Analog to the sun points, the hemisphere points are created in a 100,000 km radius around the observation point as point features in the 3DCityDB with their radiation power as an attribute.

For the creation of a computational basis on roofs and facades a regular point grid is placed on the building surfaces. Each point represents the same fraction of the area of the building surface which is determined by averaging. The points are used as reference points for the estimation of the solar irradiation and are stored in the 3DCityDB with the inclination and orientation of the surface they belong to as attributes. Considering performance and quality criteria the density of the point grid can be configured. To prevent the points from intersecting the surface they are placed on during the ray tracing, the point grid is created with a small offset (5–20 cm) in direction of the surface normal. As this slightly increases the field of vision, rays with a incidence angle smaller than 0° according to Eq. 9.1 have to be

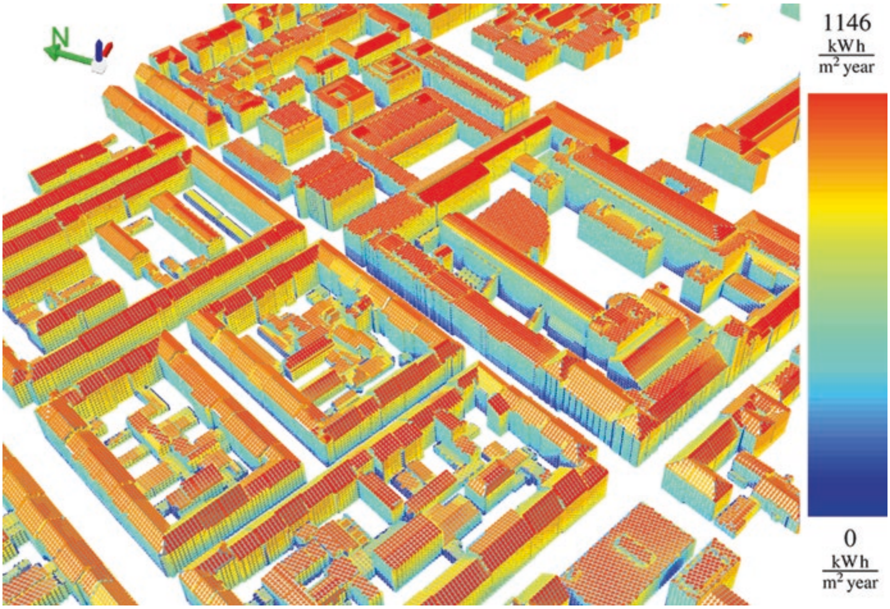


Fig. 9.4 Yearly irradiation sum (kWh/m²/year) for building points at the TUM campus and surrounding buildings

filtered, depending on the position of the sun and the inclination and orientation of the surface.

$$AngIn_{\theta,\alpha} = \arccos\left(\cos(\theta) \bullet \cos(G_z) + \sin(\theta) \bullet \sin(G_z) \bullet \cos(\alpha - G_a)\right) \quad (9.1)$$

The shadows cast by the surrounding constructions are considered by performing a visibility analysis applying a ray tracing approach. For each building point rays to all sun and hemisphere points are created which are tested for intersection with building geometries and the DTM using a ray/triangle intersection test according to Möller and Trumbore (2005). Therefore, all building geometries are triangulated using the Java3D library (Oracle: Java 3D Project Website 2016) in advance. The resulting triangles are stored in a bounding volume octree index structure significantly decreasing the number of expensive intersection tests. Both ray creation and intersection test are implemented using a thread pool allowing to process several surfaces in parallel to increase scalability. The influence of the visibility analysis can be observed in Fig. 9.4. Points in narrow corners and close to the ground receive less irradiation than points at roof tops or at unobstructed walls.

For the processing of large models (e.g. whole cities) a tiling strategy has been implemented. The simulation domain is split into cells with an edge length that can be defined in the configuration. Each cells is loaded individually for processing to avoid memory leaks. For the visibility analysis the cell, that is currently being evaluated and its eight neighbor cells are included.

9.4.2 Results

The output estimates of the application are the direct, diffuse and global irradiation energy values and the SVF. All of them are presented in different spatial and temporal aggregation levels. First, all results are computed for each building point by summing up the values for direct and diffuse irradiation and the fraction of the area of the hemisphere respectively for each non intersected ray. The results per point are aggregated per month (kWh/m^{2}month) and are written to a new database table in the 3DCityDB. They can now be evaluated based on spatial and attributive criteria using Structured Query Language (SQL). The global irradiation is calculated by summing up direct and diffuse irradiation. An example of the yearly sum of the global irradiation per point is shown in Fig. 9.4.

Besides the point results, aggregates for each building surface and building are computed in monthly (kWh/month) and yearly (kWh/year) time resolution using the Java 8 Streams API for parallelization (Urma et al. 2015). Therefore, the results per point are summed up for each surface or building respectively. The aggregated parameters are stored in the city model with the features they belong to using the Generic Attributes extension mechanism of the CityGML standard. Thus, they are available for visualization and analysis tools of the CityGML framework. The thematic surface and building instances of the city model are *persistently semantically enriched* with the results of the solar potential analysis allowing for data fusion with other information like the energy demand estimation describes in Sect. 9.6 increasing the utility value of the city model.

9.4.3 Evaluation and Discussion

For the accuracy evaluation of the method two solar potential analysis have been conducted based on the CityGML city models for Weihenstephan near Munich and Potsdam. As part of a Bachelor's Thesis by Benjamin Eberle (2015) their results have been compared to ground truth data series from pyranometers by Deutscher Wetter Dienst (DWD) ranging from 1983 to 2005 having a maximum measurement inaccuracy of ≤±5% for an hourly and monthly temporal resolution.

The monthly resolution shows relatively low deviations between measured and estimated solar irradiation. The direct irradiation is overestimated during winter and underestimated during summer, resulting in an absolute underestimation of the global irradiation sum per year of ~25 kWh/m^2/year for both test cases which corresponds to a relative deviation of less than ~3%. Comparing the NASA data to the DWD data has shown, that these deviations correlate with the deviations of the transition model compared to the DWD data. Hence, they are likely caused by the calibration of the transition model with the NASA data. A calibration of the transition model with high quality data from ground measurements could further improve the accuracy of the model.

The comparison of the hourly resolution between DWD and transition model shows significant deviations. Generally, the solar irradiation is underestimated in the morning and evening and overestimated at noon. Those inaccuracies are caused by an incorrect calculation of the relative optical path length ($m\ (\theta)$), especially for low sun positions with azenit angle (θ) of more than 80°, as Eberle found in his study (Eberle 2015). Hence, the transition model does currently not deliver accurate estimates for a temporal resolution of 1 h, in contrast to daily, monthly or yearly aggregated values. A model including correction factors for the refraction of the sunlight for low sun positions could help to increase the quality of the results.

Another factor influencing the quality of the solar irradiation estimation is related to the current practice of data acquisition for 3D city models. Today, most models are derived in automated processes from Light Detection and Ranging (LiDAR) point clouds and official geographic base data using predefined roof shapes. The application of these roof shapes may cause a slight change of roof inclinations, which can strongly influence the amount of radiation power a surface receives. Additionally, installations and building infrastructure on roofs and facades are not included in the model, which decrease the usable area for solar energy generation significantly.

The approach for the estimation of solar irradiation based on CityGML city models described in this section delivers reliable results for a wide range of applications. It can be applied to models of arbitrary size. In a test scenario, the roofs of the CityGML model of the London Borough of Barking and Dagenham containing 89,000 buildings have been calculated successfully. The work of Kausika et al. (2016) presents a case study for the city of Utrecht, Netherlands where the tool has been used to support decision making in the planning of the cities new railway station. The simulation can be controlled with a Graphical User Interface (GUI) and the results can be visualized and analyzed using a cloud-based 3D web client (Yao et al. 2014, 2016).

This enables non expert users to perform, visualize and evaluate the sun potential analysis.

9.5 Simulation of Detonations in Urban Space Based on Semantic 3D City Models

The second detailed usage example introduced in this section is about the simulation of explosions in urban space using a semantic 3D city model as data exchange platform. The following summary is based on the results published in Willenborg et al. (2016).

9.5.1 Introduction

Urban regions are characterized by dense population and a high concentration of infrastructure and businesses. Thus, they are highly vulnerable to destructive events cause by humans or nature. One of the most threatening scenarios endangering those regions are explosions caused by catastrophic events, accidents, or terrorism. Computational Fluid Dynamics (CFD) simulation tools support planning and decision making in the field of explosive safety and building construction and allow strategic and conceptual preparation for individual blast scenarios (Trometer and Mensinger 2014). These applications are tailored to efficient simulation of explosions and blast waves, but do not provide interactive access to simulation results. 3D city models and their frameworks offer comprehensive tools for visualization and result analysis even for non-expert users. They represent a reliable and growingly available data source for both geometries and semantics in an urban context.

When trying to perform CFD simulations based on semantic 3D city models we encounter two substantially different modeling paradigms: 3D city models, on the one hand, are modeled *object based*. According to the ISO 19109 definition of geographic objects (ISO 19109:2005(E) 2005), the city model objects are decomposed following logical criteria which can be observed in the real world. Their shape, orientation and location in the model is derived from their real world counterpart. CFD simulation tools on the other hand, operate on *field-based* models. The simulation domain is subdivided into e.g. a regular grid of finite volume elements or volume pixels (voxels). Real world objects are approximated by an accumulation of these cells.

The central challenge is to allow information exchange between both models and to develop an automated workflow that allows non-expert users to configure, perform, visualize and analyze blast simulations based on semantic 3D city models according to the international standard City Geography Markup Language (CityGML). An example of the desired information flow is given in Fig. 9.5. A field based representation of the city objects needs to be derived that allows the usage of the semantic information of the city model for the simulation and the back referencing of the simulation results to their corresponding city model entities.

9.5.2 Derivation of a Voxel Model from CityGML Geometries

The derivation of the voxel representation from CityGML city models is performed using the Open Source 3D geodatabase 3DCityDB running on a Post greSQL/PostGIS installation. The process is implemented as Procedural Language-/PostgreSQL (PL/pgSQL) database functions, hence only lightweight function calls have to be transferred between the main application and the database. User interaction and workflow control is implemented in Java as a plugin for the 3DCityDB Importer/Exporter. The application comes with an easy-to-use Graphical User

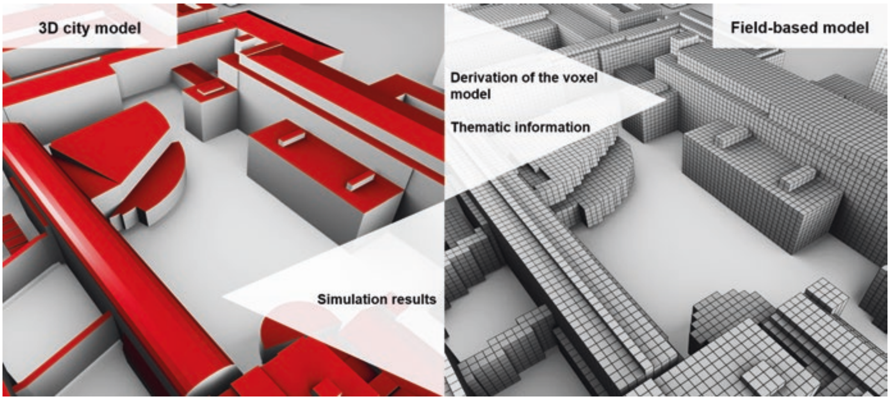


Fig. 9.5 Information exchange between the semantic 3D city model and the voxel representation of the simulation tool

Interface (GUI), where all required configuration can be handled interactively. For the derivation of the voxel model the desired CityGML layers and the simulation domain as a Bounding Box (BBox) need to be selected. Furthermore, the edge length of the voxels needs to be specified. This parameter is crucial for the performance of the application and should be adjusted depending on the quality requirements of the scenario.

The voxel model is computed in two steps. First, the voxels located in the simulation domain need to be created. Therefore, the observed area is divided into a regular grid according to the voxel edge length, starting from its lower left bottom. The resulting integer IJK coordinate system for the grid cells can now be used for the voxel creation. Based on the origin of the grid, the voxel edge length and the grid coordinates a PL/pgSQL functions creates all voxels in the domain as Post Geographical Information System (PostGIS) spatial objects of type PolyhedalSurfaceZ in the Coordinate Reference System (CRS) of the city model using PostGIS spatial operations. Second, each voxel is queried for spatial relation with the city model objects using the PostGIS 3D intersection test procedure. Thereby, the GiST index structures provided by the database system are used (Hellerstein et al. 1995). They implement an R-Tree spatial index which increases query performance using a tree data structure for bounding box comparisons (Guttman and Stonebraker 1983). All voxels having a spatial relation to a city object are added to the field-based representation. An example of a voxel model derived from a CityGML model of the campus of the Technical University of Munich is illustrated in Fig. 9.5.

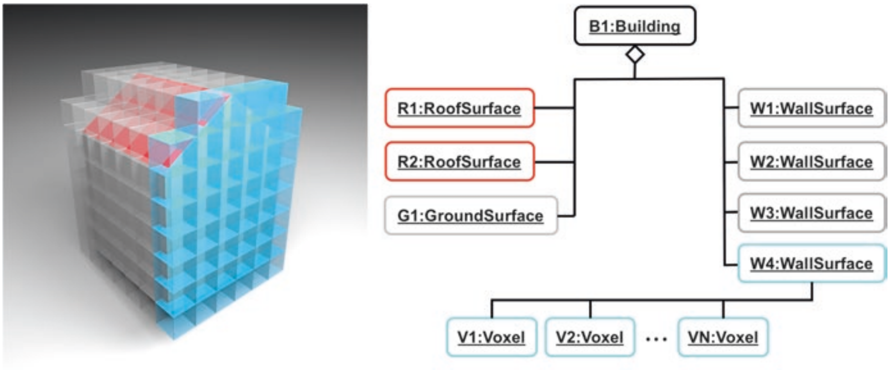


Fig. 9.6 Field based voxel approximation of an object based CityGML building. The logical relations between WallSurface W4 and the voxel model are highlighted in *blue* color

9.5.3 Information Exchange Between City and Voxel Model

Besides its geometric representation, for each voxel overlapping a CityGML geometry its logical reference to the intersecting city model object is stored in the database, where a voxel is uniquely identified by its grid coordinates and a CityGML object by its GMLID. The resulting n:m relationship between voxels and city model objects can be used to exchange information between both systems using standard database join operations.

Figure 9.6 shows an illustration of a CityGML building consisting of four WallSurface, two RoofSurface and one GroundSurface objects, its overlain voxel representation and the logical links between both of them. The semantic information (e.g. material, color, area) attached to the highlighted WallSurface W4 is linked to the highlighted voxels and can be utilized by the simulation tool. The other way round, results delivered by the simulation software in the field based model (e.g. highlighted voxels) can be referenced to their corresponding city objects (e.g. WallSurface W4).

Consequently, simulation results can be aggregated per city model object. Using the Generic Attributes extension mechanism of the CityGML standard the simulation results can be stored with their corresponding objects in the city model. The *persistent semantical enrichment* of the city model objects makes the simulation results available for visualization and analysis tools of the CityGML framework. Moreover, the information generated by the simulation can be combined with other data enhancing the analytic capabilities of the model and therefore increases its value.

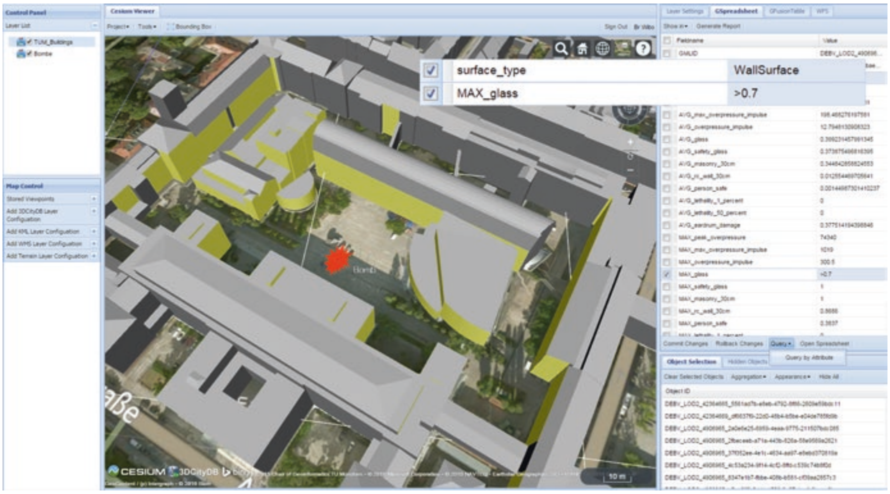
9.5.4 Example Usage Scenario: Blast Simulation with the Apollo Blastsimulator

In the following section the proposed approach for the integration of CFD simulation tools and 3D city models is evaluated for the example of a blast simulation with the *APOLLO Blastsimulator*. The work described in this section is based on the Master's Thesis of Willenborg (2015). The Apollo Blastsimulator is a CFD simulation tool developed at Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institute (EMI) in Freiburg, Germany. It is mainly used for risk analysis and combines good usability, versatility and computational efficiency by tailoring the methodological concepts to the application of explosions and blast waves (Klomfass 2016).

For evaluation purpose a fictive test scenario has been created. We assume, that an unexploded bomb from World War II has been uncovered during ground working in the inner court of the Technical University of Munich. Only the CityGML building layer has been used, all buildings have been translated to a plane.

First, the computational mesh the Apollo Blastsimulator operates on needs to be generated with the method described above. It is passed to the application in the form of a text file. After a simulation run, the Apollo Blastsimulator returns two types of results. Besides physical quantities (e.g. overpressure, overpressure impulse) a set of probability values for various damage categories (e.g. glass, masonry or concrete wall, eardrum damage, lethality) is provided. Using the logical link between the voxel and the city model the simulation results are aggregated and stored as Generic Attributes with the wall and roof surface objects in the city model.

Visualization and analysis tasks can now be performed with the cloud-based 3D web client developed at the Chair of Geoinformatics of the Technical University of Munich. The browser based application uses the Cesium Virtual Globe Viewer to visualize the 3D city model using state of the art WebGL technology for rendering and the glTF format for exchanging 3D visualization files (Yao et al. 2016). The well known interface of the 3D globe allows intuitive navigation and exploration of city models. Thematic information is distributed via cloud services like Google Spreadsheet TM or Google Fusion Tables TM, that allow analytic tasks with spreadsheet calculations. To demonstrate the analytic capabilities of the 3D web client we will identify all walls, where windows are likely to break if the bomb cannot be defused and needs to be detonated on site. First, we need to setup the query in the attribute panel of the web client. As shown in Fig. 9.7, we enter the required filter criteria to query only wall surfaces with a maximum glass breakage probability of >70 % (see enlarged entries). After issuing the query, all matching surfaces are highlighted (yellow) in the 3D view. Further analytic tasks on the selected objects can be performed directly in the client with its aggregation operations. For example, by summing up the area of all currently selected surfaces we are able to determine the total affected wall surface area. By multiplication with factors for window area per wall and window price we can perform a rough cost estimation for broken windows.



☑	surface_type	WallSurface
☑	MAX_glass	>0.7

Fig. 9.7 Evaluation of a fictive blast scenario on the campus of the Technical University of Munich with the cloud-based 3D web client: wall surfaces with a glass breakage probability >70 % are highlighted in *yellow color* in the 3D view

9.6 Estimation of Building Heat Energy Demands

In Sect. 9.4 the importance of energy policies in the context of CO2 saving potential was discussed. An essential component of the global energy demand, which is responsible for a huge amount of emitted CO_2, is required for the heating of living spaces. The goal is to reduce the energy demand by appropriate planning actions and, thus, to reduce the emission of greenhouse gases. To initiate these actions in terms of optimization and refurbishment of residential buildings and to frame political funding instruments it is essential to simulate the current and future energy demand, so it is possible to virtually play through different scenarios and compare their impacts on the build environment. In Kaden et al. (Kaden and Kolbe 2013; Kaden et al. 2013) the authors have shown that virtual semantic 3D city models combined with other data from official statistics serve as an ideal information base to support the calculation of heating, electricity, and hot water energy demands. The following summary is based on these publications. The calculation of heating energy demand of residential buildings is presented and the added value of semantic 3D city models is shown. The current German Energy Saving Regulation envisages the building simulation methods according to DIN V 18599 (2010) for calculating the heat energy demand of buildings. This standard specifies the method for calculating the monthly net, final, and primary energy demand for heating, cooling, ventilation, domestic hot water, and lighting. Besides information about the type of use of the buildings and their refurbishment state, especially information on the building geometry and building construction are crucial for the calculation of heat energy

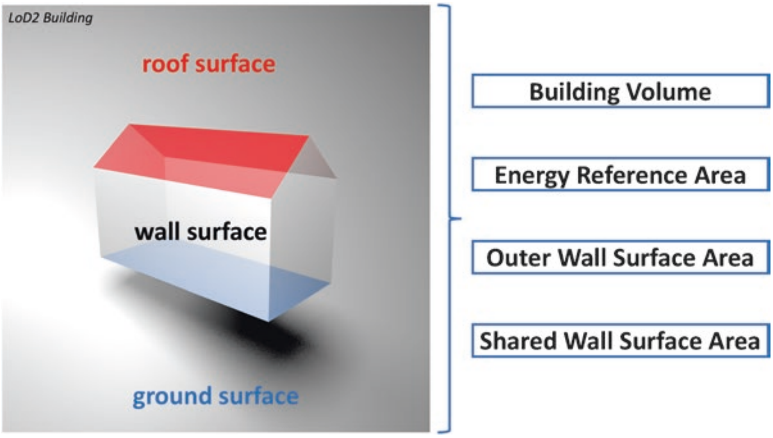
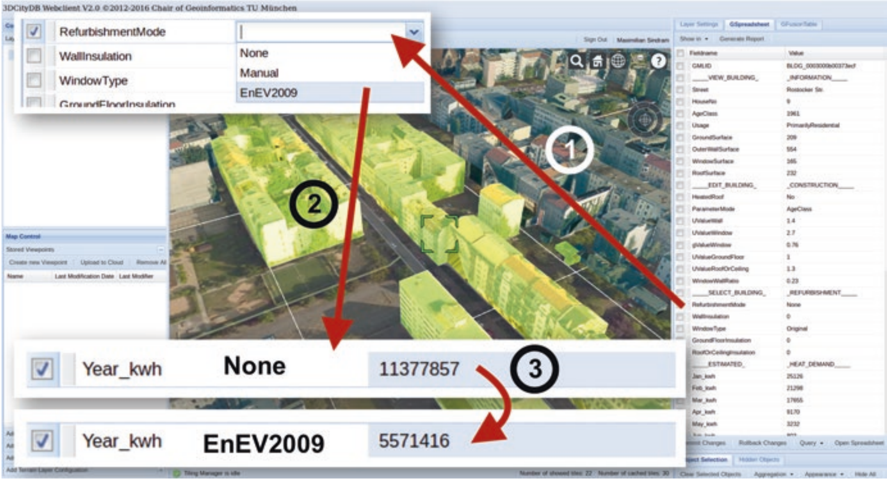


Fig. 9.8 Relation between building characteristics and required parameter for heat energy demand estimation

demands. By component-related calculations, it is possible to identify saving potentials of refurbishment measures.

Figure 9.8 illustrates the coherence between the geometric properties of the virtual building models and the input values needed for the energy demand estimation. On the left of the figure a building is shown in City Geography Markup Language (CityGML) Level of Detail 2 (LoD2). The building is subdivided into its parts (roof surface, wall surface and ground surface). Thus, it is possible to meet the requirements of DIN V 18599 (2010) to use it for a component-related calculation. On the right side of the figure the main parameters that can be calculated based on the geometric and semantic representation of the building are listed. Besides the building volume the energy reference area can be calculated using the ground surface area and the number of full stories. The shared wall surface area is a significant parameter for calculating the heat energy demands of buildings. Heat losses through walls require a drop of temperature from the inside to the outside of the building. The shared wall surface area is the portion of the total wall surface area that is adjacent to another surface that belongs to a heated building or building part and, thus, is not affected by heat losses. This ratio can be calculated using the topology of the virtual 3D city model.

Information on the building construction and the renovation state of a building are not officially provided by administration departments in Germany, thus this information has to be linked to the buildings by the integration of statistical information from statistics agencies. The heat transfer coefficients are determined based on the age class of a building. These coefficients can be adopted from the values of the predominant building type in each age class. It is however possible, if accurate values for the building parts are available, to use the precise values instead of the



☑ RefurbishmentMode | None / Manual / EnEV2009
□ WallInsulation
□ WindowType
□ GroundFloorInsulation

Fieldname	Value
GMLID	BLDG_0003000b00373ecf
____VIEW_BUILDING_	_INFORMATION____
Street	Rostocker Str.
HouseNo	9
AgeClass	1961
Usage	PrimarilyResidential
GroundSurface	209
OuterWallSurface	554
WindowSurface	165
RoofSurface	232
____EDIT_BUILDING_	_CONSTRUCTION____
HeatedRoof	No
ParameterMode	AgeClass
UValueWall	1.4
UValueWindow	2.7
gValueWindow	0.76
UValueGroundFloor	1
UValueRoofOrCeiling	1.3
WindowWallRatio	0.23
____SELECT_BUILDING_	_REFURBISHMENT____
RefurbishmentMode	None
WallInsulation	0
WindowType	Original
GroundFloorInsulation	0
RoofOrCeilingInsulation	0
____ESTIMATED_	_HEAT_DEMAND____
Jan_kwh	25126
Feb_kwh	21298
Mar_kwh	17655
Apr_kwh	9170
May_kwh	3232

☑ Year_kwh None 11377857

☑ Year_kwh EnEV2009 5571416

Fig. 9.9 Virtually improving the energy efficiency of all buildings in a road according to the German Energy Saving Regulation 2009 using a cloud-based 3D web client

estimated values. Another important step towards more agile urban planning is the ability to simulate planning scenarios. Figure 9.9 depicts how the 3D city model can be used in combination with an interactive cloud-based 3D web client (Yao et al. 2014, 2016) for the evaluation of refurbishment measures. The example shows all buildings in a street in Berlin, which have been previously selected by the attribute *street name*. To estimate the impact of an energetic refurbishment measure according to the Energy Saving Regulation 2009 we summed up the heating energy demand in kWh per year for all buildings. For the example in Fig. 9.9 the estimated total heat energy demand for all buildings is 11.38 GWh per year prior to the refurbishment measure. Adjusting the values of the heat transmission coefficients according to the requirements of the German Energy Saving Regulation 2009 (see step 2 in Fig. 9.9) triggers an immediate recalculation of the heat energy demand for each individual building. Summing up the heat energy demand values of all buildings in the street leads to a total estimated heat energy demand of 5.57 GWh per year. This corresponds to a reduction by half.

9.7 Discussion and Outlook

This article provides a review of what 3D city models are and how especially semantic 3D city models contribute to a better understanding of the city as a complex system. After giving an overview on the current state of 3D city model applications and how their use cases can be categorized, three practical examples were described in detail, covering the estimation of solar irradiation, the simulation of detonations, and the estimation of building heating energy demand.

As demonstrated with these use cases, semantic 3D city model are an ideal integration platform for many kinds of applications around the city system supporting decision making and planning. They combine a detailed geometric representation of the real world with rich thematic information for the most common features of cities and rural areas. This facilitates the virtual mapping of complex processes of the city system that need to be comprehensible and predictable to maintain good living conditions in the urban area in the future.

As, delineated in Sect. 9.3, 3D city models are widely used today and there are many future applications to come. Recent advances in augmented and virtual reality, the integration of Geographical Information System (GIS) and Building Information Modeling (BIM) and advances in procedural modeling appear as a promising sources for future use cases and applications (Biljecki et al. 2015).

However, semantic 3D city models need further developments for future challenges. Regarding the examples discussed in Sects. 9.4 and 9.5 a significant improvement would be the inclusion of dynamic attributes to enable the storage of the time dependent result data of the simulations directly within the city model. This issue is currently researched by Chaturvedi et al. (Chaturvedi and Kolbe 2015).

The quality of 3D city model data available today is still an issue. A frequent problem for instance, is that the outer shell of volumetric city model geometries is not closed, avoiding sound volume computation, which is a relevant spatial operation for many applications (see Sect. 9.6). Research is done on that topic by e.g. Sindram et al. (2016) and Steuer et al. (2015).

Another important challenge in the context of complex city systems is the coupling of planning actions and the analysis of their effects. While Sindram and Kolbe (2014) are developing a model for describing planning actions, Elfouly et al. (2015) are working on a framework for evaluating their effects on Key Performance Indicators (KPIs). To compare different scenarios Chaturvedi et al. (2015) are currently working on a concept for the versioning of entire 3D city models that will, for instance, allow the comparison of different planning stages.

References

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV): Produktblatt – 3D-Gebäudemodelle LoD2 (2016) http://www.adv-online.de/AdV-Produkte/ Standards-und-Produktblaetter/binarywriterservlet? imgUid=e9e60187-4fe3-2b41-6ad4-1fd3072e13d6& uBasVariant=11111111-1111-1111-1111-111111111111

Becker T, Nagel C, Kolbe TH (2011) Integrated 3D modeling of multi-utility networks and their interdependencies for critical infrastructure analysis. In: Advances in 3D geo-information sciences. Springer, pp 1–20. 10.1007/978-3-642-12670-3_1

Becker T, Nagel C, Kolbe TH (2013) Semantic 3D modeling of multi-utility networks in cities for analysis and 3D visualization. In: Pouliot J, Daniel S, Hubert F, Zamyadi A (eds) Progress and new trends in 3D geoinformation sciences, lecture notes in geoinformation and cartography. Springer, Berlin, pp 41–62. doi:10.1007/978-3-642-29793-9_3

Biljecki F, Stoter J, Ledoux H, Zlatanova S, Çöltekin A (2015) Applications of 3D city models: state of the art review. ISPRS Int J Geo-Inform 4(4):2842–2889. doi:10.3390/ijgi4042842

Brito MC, Gomes N, Santos T, Tenedório JA (2012) Photovoltaic potential in a Lisbon suburb using LiDAR data. Sol Energy 86(1):283–288. doi:10.1016/j.solener.2011.09.031

Catita C, Redweik P, Pereira J, Brito MC (2014) Extending solar potential analysis in buildings to vertical facades. Comput Geosci 66:1–12. doi:10.1016/j.cageo.2014.01.002

Chaturvedi K, Kolbe TH (2015) Dynamizers – modeling and implementing dynamic properties for semantic 3D city models. In: 3rd Eurographics workshop on urban data modelling and visualisation. TU Delft; Eurographics, Delft. 10.2312/udmv.20151348. URL https://diglib.eg.org/handle/10.2312/udmv20151348

Chaturvedi K, Smyth CS, Gesquière G, Kutzner T, Kolbe TH (2015) Managing versions and history within semantic 3D city models for the next generation of CityGML. In: Rahman AA (ed) Selected papers from the 3D GeoInfo 2015 conference, lecture notes in geoinformation and cartography. Springer, Kuala Lumpur

Chen LC, Wu CH, Shen TS, Chou CC (2014) The application of geometric network models and building information models in geospatial environments for fire-fighting simulations. Comput Environ Urban Syst 45:1–12. doi:10.1016/j.compenvurbsys.2014.01.003

Czerwinski A, Kolbe TH, Plümer L, Stöcker-Meier E (2006) Spatial data infrastructure techniques for flexible noise mapping strategies. In: Proceedings of the 20th international conference on environmental informatics-managing environmental knowledge. Graz

Czerwinski A, Sandmann S, Stöcker-Meier E, Plümer L (2007) Sustainable SDI for EU noise mapping in NRW–best practice for INSPIRE. Int J Spat Data Infrastruct Res 2(1):90–111

DIN V 18599 (2010) Energetische Bewertung von Gebäuden. Standard, Deutsches Institut für Normung, Berlin

Eberle BT (2015) Validierung von geschätzten Sonneneinstrahlungswerten anhand von Messdaten des Deutschen Wetterdienstes für die Standorte Potsdam und Weihenstephan. Bachelor's thesis, Technical University of Munich (TUM). URL https://mediatum.ub.tum.de/download/1296103/1296103.pdf

Elfouly M, Kutzner T, Kolbe TH (2015) General indicator modeling for decision support based on 3D city and landscape models using model driven engineering. In: Buhmann E, Ervin SM, Pietsch M (eds) Peer reviewed proceedings of digital landscape architecture 2015 at Anhalt University of Applied Sciences. Wichmann, Dessau, pp 256–267

Fu P, Rich PM (1999) Design and implementation of the solar analyst: an ArcView extension for modeling solar radiation at landscape scales. In: Proceedings of the nineteenth annual ESRI user conference, vol 1, pp 1–31

Grena R (2012) Five new algorithms for the computation of sun position from 2010 to 2110. Sol Energy 86(5):1323–1337

Gröger G, Kolbe TH, Nagel C, Haefele KH (2012) OGC City Geography Markup Language (CityGML) encoding standard. Open Geospatial Consortium. URL http://www.opengis.net/spec/citygml/2.0

Gröger G, Kutzner T, Kolbe TH (2013) A CityGML-based encoding for the INSPIRE data specification on buildings. In: INSPIRE conference 2013

Guttman A, Stonebraker M (1983) R-trees: a dynamic index structure for spatial searching, Memorandum, vol no. UCB/ERL M83/64. Electronics Research Laboratory, College of Engineering, University of California, Berkeley

Hellerstein JM, Naughton JF, Pfeffer A (1995) Generalized search trees for database systems. In: Proceedings of the 21. international conference on very large data bases

Herring J (2001) OGC abstract specifications topic 1 – feature geometry (ISO 19107 Spatial Schema)

Hildebrandt D, Timm R (2014) An assisting, constrained 3D navigation technique for multiscale virtual 3D city models. GeoInformatica 18(3):537–567. doi:10.1007/s10707-013-0189-8

INSPIRE: EU Directive (2007) Tech. rep., Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). Official Journal of the European Union, L 108/1 (2007)

ISO 19109:2005(E) (2005) Geographic information—rules for application schema. Standard, international organization for standardization

Janssen W, Blocken B, van Hooff T (2013) Pedestrian wind comfort around buildings: comparison of wind comfort criteria based on whole-flow field data for a complex case study. Build Environ 59:547–562

Kaden R, Kolbe TH (2013) City-wide total energy demand estimation of buildings using semantic 3D city models and statistical data. In: Proceedings of the 8th international 3D GeoInfo conference, vol II-2/W1

Kaden R, Prytula M, Krüger A, Kolbe TH (2013) Energieatlas Berlin: Vom Gebäude zur Stadt – Am Beispiel zur Abschätzung der Wärmeenergiebedarfe von Gebäuden. In: Geoinformationssysteme 2013 – Beiträge zum 18. Münchner Fortbildungsseminar Geoinformationssysteme, pp 17–32

Kausika B, Moshrefzadeh M, Kolbe TH, van Sark W (2016) 3D solar potential modelling and analysis: a case study for the city of Utrecht. In: 32nd European photovoltaic solar energy conference and exhibition, EUPVSEC 2016 (accepted)

Khan AA, Kolbe TH (2012) Constraints and their role in subspacing for the locomotion types in indoor navigation. In: Indoor Positioning and Indoor Navigation (IPIN). 10.1109/IPIN.2012.6418872. http://ieeexplore.ieee.org/xpls/abs_all. jsp?arnumber=6418872&tag=1

Khan AA, Yao Z, Kolbe TH (2015) Context aware indoor route planning using semantic 3D building models with cloud computing. In: Breunig M, Al-Doori M, Butwilowski E, Kuper PV, Benner J, Häfele KH (eds) 3D geoinformation science, lecture notes in geoinformation and cartography. Springer, Cham. doi:10.1007/978-3-319-12181-9_11

Khan Aftab A, Kolbe Thomas HA (2013) Subspacing based on connected opening spaces and for different locomotion types using geometric and graph based representation in multilayered space-event model (MLSEM). In: Proceedings of the 8th 3D GeoInfo conference. Istanbul, Turkey. http://www.isprs-ann-photogramm-remote-sens-spatial-inf-sci.net/II-2-W1/173/2013/

Klomfass A (2016) Improved explosion consequence analysis with combined CFD and damage models. Chem Eng Trans 48:109–114. doi:10.3303/CET1648019

Kolbe TH (2009) Representing and exchanging 3D city models with CityGML. In: Lee J, Zlatanova S (eds) 3D geo-information sciences, lecture notes in geoinformation and cartography. Springer, Berlin, pp 15–31. doi:10.1007/978-3-540-87395-2_2

Kolbe TH, Burger B, Cantzler B (2015) CityGML goes to broadway. In: Fritsch D (ed) Photogrammetric week ’15. Institute for Photogrammetry, University of Stuttgart, Wichmann, Stuttgart, pp 343–356

Kubiak J, Ławniczak R (2015) The propagation of noise in a built-up area (on the example of a housing estate in Poznań). J Maps 12(2):231–236. doi:10.1080/17445647.2014.1001801

Kutzner T, Kolbe TH (2016) Extending semantic 3D city models by supply and disposal networks for analysing the urban supply situation. In: Kersten TP (ed) Lösungen für eine Welt im Wandel, Dreiländertagung der SGPF, DGPF und OVG, 36. Wissenschaftlich-Technische Jahrestagung der DGPF, Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation (DGPF) e.V, vol 25. Deutsche Gesellschaft für Photogrammetrie, Fernerkundung und Geoinformation e.V, Bern, pp 382–394. http://www.dgpf.de/src/tagung/jt2016/ proceedings/papers/36_DLT2016_Kutzner_Kolbe.pdf

Kwan MP, Lee J (2005) Emergency response after 9/11: the potential of real-time 3D GIS for quick emergency response in micro-spatial environments. Comput Environ Urban Syst 29(2):93–113. doi:10.1016/j.compenvurbsys.2003.08.002

Liu L, Zlatanova S (2011) A ‘door-to-door’ path-finding approach for indoor navigation. In: Proceedings Gi4DM 2011: GeoInformation for Disaster Management, Antalya, Turkey, 3–8 May 2011

Mao B, Ban Y, Harrie L (2015) Real-time visualization of 3D city models at street-level based on visual saliency. Sci China Earth Sci 58(3):448–461. doi:10.1007/s11430-014-4955-8

Maragkogiannis K, Kolokotsa D, Maravelakis E, Konstantaras A (2014) Combining terrestrial laser scanning and computational fluid dynamics for the study of the urban thermal environment. Sustain Cities Soc 13:207–216

Möller T, Trumbore B (2005) Fast, minimum storage ray/triangle intersection. In: ACM SIGGRAPH 2005 Courses. ACM, pp 1–7. 10.1145/1198555.1198746

NASA Langley Research Center (2016) Surface meteorology and Solar Energy (SSE) data and information website. https://eosweb.larc.nasa.gov/project/sse/sse_table. Accessed 26 July 2016

Nedkov S (2012) Knowledge-based optimization of 3D city models for car navigation devices. Master's thesis, Delft University of Technology. http://repository.tudelft.nl/islandora/object/uuid:b429e899–9955-4a23-9ceb-66ffb6210b30?collection=education

Nouvel R, Kaden R, Bahu JM, Kaempf J, Cipriano P, Lauster M, Benner J, Munoz E, Tournaire O, Casper E (2015) Genesis of the CityGML energy ADE. In: Proceedings of international conference CISBAT 2015 future buildings and districts sustainability from nano to urban scale, EPFL-CONF-213436. LESO-PB, EPFL, pp 931–936

Oracle: Java 3D Project Website (2016) https://java3d.java.net/. Accessed 26 July 2016

Oulasvirta A, Estlander S, Nurminen A (2009) Embodied interaction with a 3D versus 2D mobile map. Pers Ubiquit Comput 13(4):303–320. doi:10.1007/s00779-008-0209-0

Redweik P, Catita C, Brito MC (2013) Solar energy potential on roofs and facades in an urban landscape. Sol Energy 97:332–341. doi:10.1016/j.solener.2013.08.036

Schilling A, Coors V, Laakso K (2005) Dynamic 3D maps for mobile tourism applications. In: Meng L, Reichenbacher T, Zipf A (eds) Map-based mobile services. Springer, Berlin, pp 227–239. doi:10.1007/3-540-26982-7_15

Schulte C, Coors V (2008) Development of a CityGML ADE for dynamic 3D flood information. In: Joint ISCRAM-CHINA and GI4DM conference on information systems for crisis management

Sindram M, Kolbe TH (2014) Modeling of urban planning actions by complex transactions on semantic 3D city models. In: Ames D, Quinn N, Rizzoli A (eds) Proceedings of the International Environmental Modelling and Software Society (iEMSs). International Environmental Modelling and Software Society (iEMSs), San Diego

Sindram M, Machl T, Steuer H, Pültz M, Kolbe TH (2016) Voluminator 2.0 – speeding up the approximation of the volume of defective 3D building models. In: ISPRS annals of photogrammetry, remote sensing and spatial information sciences, vol III-2

Slingsby A, Raper J (2008) Navigable space in 3D city models for pedestrians. In: Cartwright W, Fendel EM, Gartner G, Meng L, van Oosterom P, Penninga F, Peterson MP, Zlatanova S (eds) Advances in 3D geoinformation systems, lecture notes in geoinformation and cartography. Springer, Berlin, pp 49–64. doi:10.1007/978-3-540-72135-2_3

Steuer H, Machl T, Sindram M, Liebel L, Kolbe TH (2015) Voluminator – approximating the volume of 3D buildings to overcome topological errors. In: Bação F, Santos MY, Painho M (eds) AGILE 2015 – geographic information science as an enabler of smarter cities and communities, lecture notes in geoinformation and cartography. Springer, Lisbon, pp 343–362

Tashakkori H, Rajabifard A, Kalantari M (2015) A new 3D indoor/outdoor spatial model for indoor emergency response facilitation. Build Environ 89:170–182. doi:10.1016/j.buildenv.2015.02.036

Thill JC, Dao THD, Zhou Y (2011) Traveling in the three-dimensional city: applications in route planning, accessibility assessment, location analysis and beyond. J Transp Geogr 19(3):405–421. doi:10.1016/j.jtrangeo.2010.11.007

Trometer S, Mensinger M (2014) Simulation von Detonationsszenarien im urbanen Umfeld. In: Kolbe TH, Bill R, Donaubauer A (eds) Geoinformationssysteme 2014: Beiträge zur 1. Münchner GI-Runde. Wichmann, Berlin, pp 150–164

Ujang U, Anton F, Rahman AA (2013) Unified data model of urban air pollution dispersion and 3D spatial city model: groundwork assessment towards sustainable urban development for Malaysia. J Environ Prot 4(7):701–712

Urma RG, Fusco M, Mycroft A (2015) Java 8 in action: lambdas, streams, and functional-style programming. Manning, Shelter Island

Vanhorn JE, Mosurinjohn NA (2010) Urban 3D GIS modeling of terrorism sniper hazards. Soc Sci Comput Rev 28(4):482–496. doi:10.1177/0894439309360836

Varduhn V, Mundani RP, Rank E (2015) Multi-resolution models: recent progress in coupling 3D geometry to environmental numerical simulation. In: 3D Geoinformation science. Springer, pp 55–69

Willenborg B (2015) Simulation of explosions in urban space and result analysis based on CityGML City models and a cloud based 3D Web client. Master's thesis, Technical University of Munich (TUM). https://mediatum.ub.tum.de/download/1250726/1250726.pdf

Willenborg B, Sindram M, Kolbe TH (2016) Semantic 3D city models serving as information hub for 3D field based simulations. In: Kersten T (ed) Lösungen für eine Welt im Wandel, 3-Ländertagung der SGPF, DGPF und OVG in Bern, vol 25. Deutsche Gesellschaft für Photogrammmetrie, Fernerkundung und Geoinformation e.V, Bern, pp 54–65. http://www.dgpf.de/src/tagung/jt2016/ proceedings/papers/06_DLT2016_Willenborg_et_al.pdf

Yaagoubi R, Yarmani M, Kamel A, Khemiri W (2015) HybVOR: a Voronoi-based 3D GIS approach for camera surveillance network placement. ISPRS Int J Geo-Inf 4(2):754–782. doi:10.3390/ijgi4020754

Yao Z, Sindram M, Kaden R, Kolbe TH (2014) Cloud-basierter 3D-Webclient zur kollaborativen Planung energetischer Maßnahmen am Beispiel von Berlin und London. In: Kolbe TH, Bill R, Donaubauer A (eds) Geoinformationssysteme 2014. Wichmann, Heidelberg

Yao Z, Chaturvedi K, Kolbe TH (2016) Browser-basierte Visualisierung großer 3D-Stadtmodelle durch Erweiterung des Cesium Web Globe. Geoinformationssysteme 2016:77–90

Zahn W (2015) Sonneneinstrahlungsanalyse auf und Informationsanreicherung von großen 3D–Stadtmodellen im CityGML-Schema. Master's thesis, Technical University of Munich (TUM). URL https://mediatum.ub.tum.de/download/1276236/1276236.pdf

Chapter 10
An Automatic Approach for Generalization of Land-Cover Data from Topographic Data

Frank Thiemann and Monika Sester

Abstract The paper presents an approach for the automatic generalization of large land-cover datasets from topographic data using fast generalization algorithms. The generalization approach is composed of several steps consisting of topologic cleaning, aggregation, feature partitioning, identification of mixed feature classes to form heterogeneous classes and simplification of feature outlines. The workflow will be presented with examples for generating CORINE Land Cover (CLC) features from the high resolution German authoritative land-cover dataset of the whole area of Germany (DLM-DE). The results will be discussed in detail.

10.1 Introduction

10.1.1 Project Background

The European Environment Agency (EEA) collects the Coordinated Information on the European Environment (CORINE) Land Cover (CLC) dataset to monitor the land-cover changes in the European Union. The member nations have to deliver this data every few years. Traditionally this dataset was derived from remote sensing data. However, the classification of land-cover from satellite images in shorter time intervals becomes more cost intensive.

Therefore, together with the federal mapping agency (BKG) an approach of deriving the land cover data from topographic information was investigated. The BKG collects the digital topographic landscape models (ATKIS Base DLM) from all federal states. The topographic base data contains up-to-date land-use information; the update rate being one year. This data is transformed to a high resolution land-cover dataset called DLM-DE. After this transformation there are still some

F. Thiemann (✉) • M. Sester
Institute of Cartography and Geoinformatics, Leibniz Universität Hannover,
Appelstraße 9a, 30167 Hannover, Germany
e-mail: frank.thiemann@ikg.uni-hannover.de; monika.sester@ikg.uni-hannover.de

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_10

Table 10.1 Comparison of ATKIS DLM and CORINE Land Cover

Dataset	CORINE LC	ATKIS basis DLM
Scale	1:100 000	>1:10 000
Source	Satellite images	Aerial images, cadastre
Min. area size	25 ha	<1 ha
Min. width	100 m	<10 m
Classes of heterogeneous agricultural cover	4/2 relevant	Marginal, mostly separated in its homogeneous components
# of classes	46/37 relevant	155 relevant

differences between DLM-DE and CLC, mainly concerning the resolution in geometry and semantics. Table 10.1 summarizes the main characteristics of the two datasets.

10.1.2 CORINE Land Cover (CLC)

CORINE Land Cover is a polygon dataset in the form of a planar partitioning (or tessellation): polygons do not overlap and cover the whole area without gaps. The scale is 1:100 000. Each polygon has a minimum area of 25 ha and a minimum width of 100 m. There are no adjacent polygons with the same land-cover class as these have to be merged.

Land cover is classified hierarchically into 46 classes in three levels, for which a three digit numerical code is used. The first and second level groups are:

1xx	artificial (urban, industrial, mine)
2xx	agricultural (arable, permanent, pasture, heterogeneous)
3xx	forest and semi-natural (forest, shrub, open)
4xx	wetland (inland, coastal)
5xx	water (inland, marine)

In CLC there are four aggregated classes for heterogeneous agricultural land-cover. Such areas are composed of small areas of different agricultural land-cover. In Germany only two of these four classes occur. Class 242 is composed of alternating agricultural covers (classes 2xx). Class 243 is a mixture of agricultural and (semi-) natural areas.

10.1.3 From Basis DLM to DLM DE-LC

The land cover (LC) layer of the Digital Landscape Model (DLM) of Germany (DE) is a product of Germany’s national mapping agency (BKG). DLM-DE LC is derived by a semantic (model) generalization of the Authoritative Topographic Cartographic

Information System (ATKIS) which is Germany's large scale topographic landscape model (Arnold 2009). After selecting all relevant features from ATKIS (e.g. water-bodies, vegetation, settlement areas but not administrative bodies or areas on bridges or in tunnels) the topological problems like overlaps and gaps are solved automatically using appropriate algorithms. The reclassification to the CLC nomenclature is done using a translation table which takes the ATKIS classes and their attributes into account. In the cases where a unique translation is not possible, a semi-automatic classification from remote sensing data is used. The scale of the DLM-DE is approx. 1:10 000. The minimum area for polygons is less than one hectare.

10.1.4 Automatic Derivation of CLC from DLM-DE LC

The aim of the project is the automated derivation of CLC data from ATKIS. This derivation can be considered as a generalization process, as it requires both thematic selection and reclassification, and geometric operations due to the reduction in scale. Therefore, the whole workflow consists of two main parts. The first part is a model transformation and consists of the extraction, reclassification and topological correction of the data. The derived model is called DLM-DE LC. The second part, the generalization part, which will be described in more detail in this paper, is the aggregation, classification and simplification for the smaller scale. For that purpose a sequence of generalization operations is used, which will be executed in a fully automatic way. The operators are dissolve, aggregate, split, simplify and a heterogeneous class filter. The program computing the generalization is called CLC-generator.

The classification of agricultural heterogeneous areas to 24x-classes in the case that a special mixture of land-covers occurs is one of the main challenges. The difficulty is to adequately model these classes and separate these areas from homogeneous as well as from other heterogeneous classes.

10.1.5 Scalability

Another challenge of the project is the huge amount of data. The DLM-DE LC contains ten million polygons. Each polygon consists in average of thirty points, so one has to deal with 300 million points, which is more than a standard PC can store in main memory. While fast algorithms and efficient data structures reduce the required time for the generalization, we have developed a partitioning and composition strategy in order to overcome problems due to memory limitations when processing large data-sets (Thiemann et al. 2011). We store the source data for the generalization process in a spatial database system and divide it into smaller partitions, which can efficiently be handled by the CLC-generator on standard computers. The resulting CLC-datasets for the individual tiles are then composed into one dataset within the database.

The generalization operations typically have local or regional effects, which lead to different results at the boundaries of the tiles. To ensure consistency, i.e. to get identical results from partitioned and un-partitioned execution, some redundancy is added to the partitions in the form of overlapping border regions. This redundancy is removed in the composition phase and geographic objects residing at the border of different partitions are reconciled.

The amount of redundancy added can be controlled by the width of the border regions. As bigger regions cause longer running times of the generalization, we are interested in using values as small as possible while still ensuring consistency. Another parameter influencing performance is the number of partitions. The tiles have to be small enough to avoid memory limitations but a fine-granular partitioning leads to more composition overhead.

10.2 Related Work

CORINE Land Cover (Büttner et al. 2006) is being derived by the European States (Geoff et al. 2007). In order to link the topographic database with the land-use data the Federal Agency of Cartography and Geodesy has developed a mapping table, including transformation rules between CLC and ATKIS objects (Arnold 2009). In this way, the semantic mapping has been established by hand, introducing expert knowledge. There are approaches to automate this process, e.g. Kuhn (2003) or Kavouras and Kokla (2008). Jansen et al. (2008) propose a methodology to integrate land-use data.

As described above, the approach uses different generalization and interpretation steps. The current state of the art in generalization is described in Mackaness et al. (2007). The major generalization step needed for the generalization of land-cover classes is aggregation. The classical approach for area aggregation was given by van Oosterom (1995), the so-called GAP-tree (Generalized Area Partitioning). In a region-growing fashion areas that are too small are merged with neighboring areas until they satisfy the size constraint. The selection of the neighbor to merge with depends on different criteria, mainly geometric and semantic constraints, e.g. similarity of object classes or length of common boundary. This approach is implemented in different software solutions (e.g. Podrenek 2002). Although the method yields areas of required minimum size, there are some drawbacks: a local determination of the most compatible object class can lead to a high amount of class changes in the whole dataset. Also, objects can only survive the generalization process, if they have compatible neighbors. The method by Haunert (2008) is able to overcome these drawbacks. He is also able to introduce additional constraints e.g. that the form of the resulting objects should be compact. The solution of the problem has been achieved using an exact approach based on mixed-integer programming (Gomory 1958), as well as a heuristic approach using simulated annealing (Kirkpatrick 1983). However, the computational effort for this global optimization approach is very high.

Collapse of polygon features corresponds to the skeleton operation, which can be realized using different ways. A simple method is based on triangulation; another is medial axis or straight skeleton (Haunert and Sester 2008). Displacement is needed to allow all object to be perceived as separate. The problem has been solved using optimization methods (Sester 2005).

The identification of mixed classes is an interpretation problem. Whereas interpretation is predominant in image understanding where the task is to extract meaningful objects from a collection of pixels (Lillesand and Kiefer 1999), also in GIS-data interpretation is needed, even when the geo-data are already interpreted. E.g. in our case although the polygons are semantically annotated with land-cover classes, however, we are looking for a higher level structure in the data which evolves from a spatial arrangement of polygons. Interpretation can be achieved using pattern recognition and model based approaches (Heinzle and Anders 2007).

Partitioning of spatial data has extensively been investigated in the area of parallel spatial join processing. In Zhou et al. (1998) a framework for partitioning spatial join operations in a parallel computer environment is introduced and the impact of redundancy on performance is studied. Other work (Meng et al. 2007) presents an improved join method for decomposing spatial datasets in a parallel database system. Spatial joins only need to collect partition-wise results, maybe including elimination of duplicates. An approach for partitioning in a distributed processing framework using Hadoop has been presented (Thiemann et al. 2013), which takes adequate context dependence into account.

10.3 Generalization Approach

10.3.1 Data and Index Structures

An acceptable run time for the generalization of ten million polygons can only be reached with efficient algorithms and data structures. For topology depending operations a topologic data structure is essential. For spatial searching a spatial index structure is needed; furthermore, also structures for one-dimensional indexing are used.

In the project we use an extended Doubly Connected Edge List (DCEL) as topologic structure. A simple regular grid (two-dimensional hashing) is used as spatial index for nodes, edges and faces. For the DLM-DE a grid width of 100 m for points and edges (<10 features per cell) and 1000 m for faces (40 faces per cell) leads to nearly optimal speed.

10.3.2 Topological Cleaning

Before starting the generalization process, the data has to be imported into the topological structure. In this step we also look for topological or semantic errors. Each polygon is checked for a valid CLC class. Small sliver polygons with a size under a threshold of e.g. 1 m^2 will be rejected. A snapping with a distance of 1 cm is done for each inserted point. With a point in polygon test and a test for segment intersection overlapping polygons are detected and also rejected. Holes in the tessellation can be easily found by building loops of the half-edges which not belong to any face. Loops with a positive orientation are holes in the dataset.

10.3.3 Generalization Operators

Dissolve
The dissolve operator merges adjacent faces of the same class. For this purpose the edges which separate such faces will be removed and new loops are built.

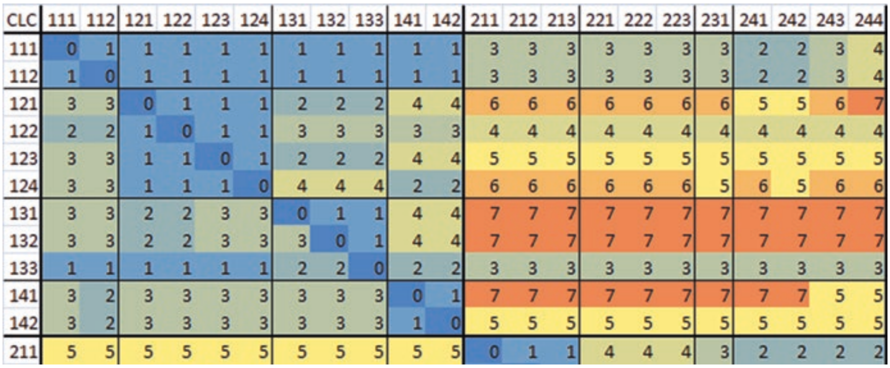
Aggregate
The aggregation step aims at guarantying the minimum size of all faces. The aggregation operator in our case uses the simple greedy algorithm by van Oosterom (1995). It starts with the smallest face and merges it to a compatible neighbor. This fast algorithm is able to process the dataset sequentially. There are different options to determine compatible neighbors. The criterion can be:

- the semantic compatibility (semantic distance),
- the geometric compactness
- or a combination of both.

The semantically nearest partner can be found using a priority matrix. We use the matrix from the CLC technical guide (Bossard et al. 2000) (Fig. 10.1). The priority values are from an ordinal scale, so their differences and their values in different lines should not be compared. The matrix is not symmetric, as there may be different ranks when going from one object to another than vice versa (e.g. settlement → vegetation). Priority value zero is used if both faces have the same class. The higher the priority value, the higher is the semantic distance. Therefore the neighbor with the lowest priority value is chosen.

As geometric criterion the length of the common edge is used. A shorter perimeter leads to better compactness. So the maximum edge length has to be reduced to achieve a better compactness.

The effects of using the criteria separate are shown in a real example in Fig. 10.2. The semantic criterion leads to non-compact forms, whereas the geometric criterion is more compact but leads to a large amount of class change. The combination of both criteria allows merging of semantically more distant objects, if the resulting form is more compact. This leads to Eq. 10.1.



CLC	111	112	121	122	123	124	131	132	133	141	142	211	212	213	221	222	223	231	241	242	243	244
111	0	1	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2	3	4
112	1	0	1	1	1	1	1	1	1	1	1	3	3	3	3	3	3	3	2	2	3	4
121	3	3	0	1	1	1	2	2	2	4	4	6	6	6	6	6	6	6	5	5	6	7
122	2	2	1	0	1	1	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	4
123	3	3	1	1	0	1	2	2	2	4	4	5	5	5	5	5	5	5	5	5	5	5
124	3	3	1	1	1	0	4	4	4	2	2	6	6	6	6	6	6	5	6	5	6	6
131	3	3	2	2	3	3	0	1	1	4	4	7	7	7	7	7	7	7	7	7	7	7
132	3	3	2	2	3	3	3	0	1	4	4	7	7	7	7	7	7	7	7	7	7	7
133	1	1	1	1	1	1	2	2	0	2	2	3	3	3	3	3	3	3	3	3	3	3
141	3	2	3	3	3	3	3	3	3	0	1	7	7	7	7	7	7	7	7	7	5	5
142	3	2	3	3	3	3	3	3	3	1	0	5	5	5	5	5	5	5	5	5	5	5
211	5	5	5	5	5	5	5	5	5	5	5	0	1	1	4	4	4	3	2	2	2	2

Fig. 10.1 Small extract of the CLC priority matrix

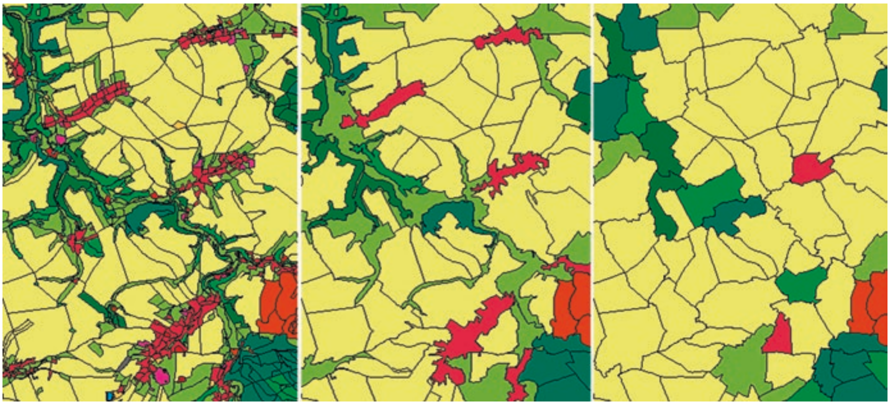


Fig. 10.2 (*left* to *right*) Original situation, the result of the semantic, and geometric aggregation

$$distance\left(A,B\right)=\frac{b^{priority}}{length} \tag{10.1}$$

The equation means that a b-times longer shared edge allows a neighbor with the next worse priority. The base b allows to weight between compactness and semantic proximity. A value of b = 1 leads to only compact results, a high value of b leads to semantically optimal results. Using the priority values is not quite correct; it is only a simple approximation for the semantic distance.

Another application of the aggregation operation is a special kind of dissolve that stops at a defined area size. It merges small faces of the same class to bigger compact faces using the geometric aggregation with the condition that only adjacent faces of the same class are considered.

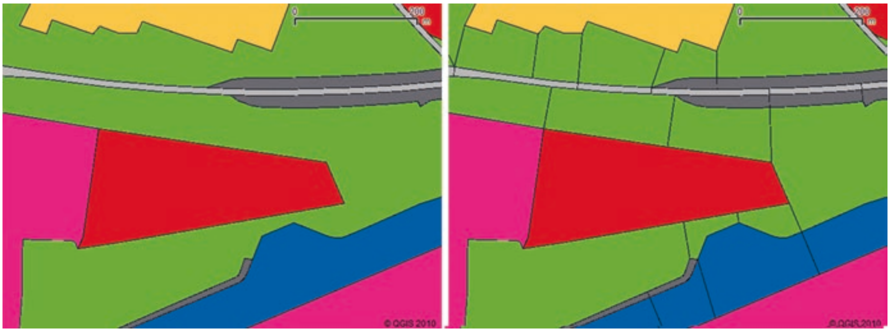


Fig. 10.3 Data before and after a 100 m split operation

Split

In addition to the criterion of minimal area size also the extent of the polygon is limited to a minimum distance. That demands for a collapse operator to remove slim, elongated polygons and narrow parts. The collapse algorithm by Haunert and Sester (2008) requires buffer and skeleton operations that are time consuming. Therefore – as faster alternative – a combination of splitting such polygons and merging the resulting parts with a geometric aggregation to other neighbors is used. Instead of shrinking the slim parts to their medial axes we split it at suited points and use the aggregation step to merge the slim polygons with another neighbor.

To find the narrows we use a constrained Delaunay triangulation of the polygon. Each triangle is checked for edges and heights smaller than a threshold. These edges or heights will be used for splitting (see Fig. 10.3).

"24x-Filter" for Identification of Heterogeneous Classes

In CORINE land-cover there is a group of classes which stands for heterogeneous land-covers. The classes 242 and 243 are relevant for Germany. Class 242 (complex cultivation pattern) is used for a mixture of small parcels with different cultures. Class 243 is used for land that is principally occupied by agriculture, with significant areas of natural vegetation.

Heterogeneous classes are not included in the DLM-DE. To form these 24x-classes an operator for detecting heterogeneous land-cover is needed. The properties of these classes are that smaller areas with different, mostly agricultural land-cover alternate within the minimum area size (actually 25 ha in CLC). For the recognition of class 242 only the agriculture areas (2xx) are relevant. For 243 also forest, semi- and natural areas (3xx, 4xx) and lakes (512) have to be taken into account.

The algorithm calculates some neighborhood statistics for each face. All adjacent faces within a distance of the centroid smaller than a given radius and with an area size smaller than the target size are collected by a deep search in the topological structure. The fraction of the area of the majority class and the summarized fractions of agricultural areas (2xx) and (semi-) natural areas (3xx, 4xx, 512) are calculated.

In the case the majority class dominates (>75%) then the majority class becomes the new class of the polygon. Otherwise there is a check, if it is a heterogeneous area or only a border region of larger homogeneous areas.

For that purpose the length of the borders between the relevant classes is summarized and weighted with the considered area. A heterogeneous area is characterized by a high border length, as there is a high number of alternating areas. To distinguish between 242 and 243 the percentage of (semi-natural) areas has to be significant (>25%).

Simplify
The simplify-operator removes redundant points from the loops. A point is redundant, if the geometric error without using this point is lower than an epsilon and if the topology does not change. Therefore we implemented the algorithm of Douglas and Peucker (1973) with an extension for closed loops and a topology check.

Enlargement and Displacement
Displacement is an operation aiming at generating graphical clarity. Objects are shifted apart and/or deformed in order to guarantee that the can be perceived as separate objects. This operation mostly is needed, when small objects had to be enhanced to allow their visibility, such as narrow objects that have to be widened (roads, rivers), after which they may overlay neighboring objects. In the case of land-use/land-cover classes, this can occur in the presence of small, elongated objects.

10.3.4 Process Chain

In this section the use of the introduced operators and their orchestration in the process chain is shown. The workflow for a target size of 25 ha is as follows:

1. import and clean data and fill holes
2. dissolve faces <25 ha
3. split faces <100 m
4. aggregate faces <1 ha geometrically (base 1.2)
5. reclassify faces with 24x-filter (r = 282 m)
6. aggregate faces <5 ha weighted (base 2)
7. aggregate faces <25 ha semantically
8. simplify polygons (tolerance 20 m)
9. dissolve all

During the import step (1) semantic and topology is checked. Small topologic errors are resolved by a snapping. Gaps are filled with dummy objects. These objects will be merged to other objects in the later steps.

A first dissolve step (2) merges all faces with an adjacent face of the same CLC class which are smaller than the target size (25 ha). The dissolve is limited to 25 ha to prevent polygons from being too large (e.g. rivers that may extend over the whole partition). This step leads to many very non-compact polygons. To be able to remove them later, the following split-step (3) cuts them at narrow internal parts (smaller than 100 m). Afterwards an aggregation (4) merges all faces smaller than 1 ha (100 × 100 m) to geometrically fitting neighbors. As small isthmuses were eliminated in the split-step, all objects were wide enough to be visualized appropriately, thus, no displacement was applied in this process.

The proximity analysis of the 24x-filter step (5) re-classifies agricultural or natural polygons smaller than 25 ha in the 25 ha (corresponding to a radius of 282 m) surrounding as heterogeneous (24x class).

The next step aggregates all polygons to the target size of 25 ha. First we start with a geometric/semantically weighted aggregation (6) to get more compact forms, second only the semantic criterion is used (7) to prevent large semantic changes of large areas.

The simplify step (8) smoothes the polygon outlines by reducing the number of nodes. As geometric error tolerance 20 m (0.2 mm in the map) is used. The finishing dissolve step (9) removes all remaining edges between faces of same class.

10.4 Experiments and Results

10.4.1 Runtime and Memory Use of the Generalization Step

The implemented algorithms are very fast but require a lot of memory. Data and index structures need up to 160 Bytes per point on a 32 bit machine.

The run-time of the generalization routines was tested with a 32 bit 2.66 GHz Intel Core 2 processor with a balanced system of RAM, hard disk and processor (windows performance index 5.5). The whole generalization sequence for a 45 × 45 km dataset takes less than 2 min. The most time consuming parts of the process are the I/O-operations which take more than 75% of the computing time. We are able to read 100 000 points per second from shape files while building the topology. The time of the writing process depends on the disk cache. In the worst case it is the same as for reading.

10.4.2 Semantic and Geometric Correctness

To evaluate the semantic and geometric correctness we did some statistics comparing input, result and a CLC 2006 reference dataset, which was derived from remote sensing data.

Figure 10.4 shows the input data (DLM-DE), our result and the CLC 2006 of the test area Dresden. The statistics in Fig. 10.5 verifies that our result matches with DLM-DE (75% of area) better than the reference dataset (60%). This is not surprising as for CLC 2006 different data sources were used. Because of the removing of

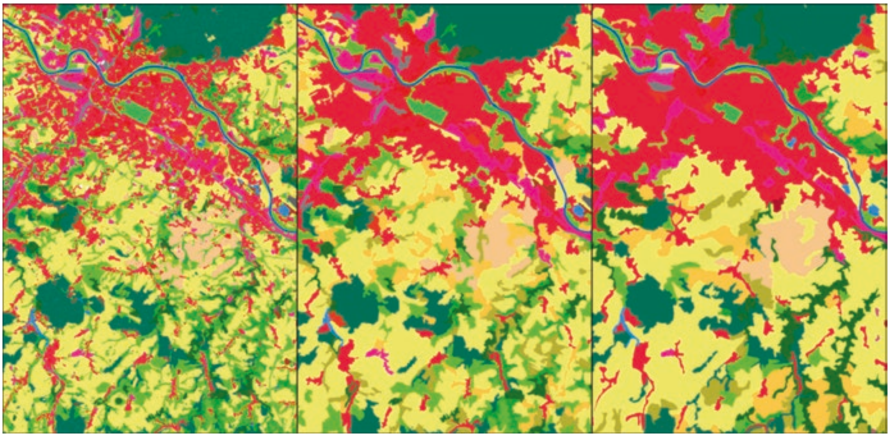


Fig. 10.4 Extract (20 × 25 km) of test dataset Dresden from left to right: input DLM-DE, our result and CLC 2006 as reference

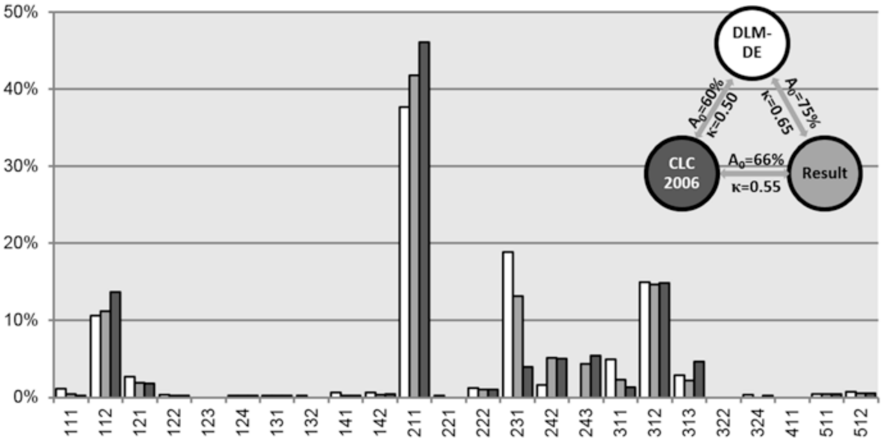


Fig. 10.5 Percentage of area for each CLC class (bars) and percentage of matching area ($A_{0,}$ area with the same class) and κ-values for the Dresden dataset

the small faces our generalization result is a bit more similar to CLC 2006 (66%) than CLC 2006 to the input dataset.

The aim of the geometric generalization is to reduce the number of vertices and polygons while preserving the structure. Table 10.2 shows some metrics on the datasets. The number of polygons was reduced during generalization, but is 50% higher than the number of polygons in the reference dataset. The complexity of the polygons (number of points per polygon) is a bit smaller (62 vs. 75); also the compactness of the polygons is smaller (29% vs. 33%), which means that they are a bit more elongated.

Table 10.2 Statistics of the test dataset Dresden (45 × 45 km)

Dataset	DLM-DE	Result	CLC 2006
# Polygons	91,717	1244	876
# Points per polygon	23	62	75
Avg. compactness (C)	50%	29%	33%
Diversity (H)	2.8	2.7	2.6
Homogeneity (E)	60%	61%	57%

The percentage of the CLC classes is similar in all datasets (Fig. 10.5). This is also indicated by the structure indices diversity and homogeneity, which means that the structure was well preserved. Diversity (*H*) is calculated by Shannon's index and also known as entropy. Its smallest value is zero for only one land use class ($k = 1$). Its maximum H_{max} for k classes would be reached when all classes have the same probability p_i. While the diversity is decreasing with the number of classes, homogeneity or equitability *E* is the normalized diversity to values between zero and one.

$$\text{Compactness} \quad C = 4\pi A / P^2 \tag{10.2}$$

$$\text{Diversity} \quad H = -\sum_{k}^{i=1} p_i \log_2\left(p_i\right) \tag{10.3}$$

$$\text{Diversity} \quad H_{\max} = \log_2\left(k\right) \tag{10.4}$$

$$\text{Homogeneity} \quad E = H / H_{\max} \tag{10.5}$$

There are some significant differences between the DLM-DE and CLC 2006 within the classes 211/234 (arable/grass land) and also between 311/313 (broad-leaved/mixed forest) and 111/112 (continuous/discontinuous urban fabric). We assume that this comes from different interpretations and different underlying data sources. The percentages in our generated dataset are mostly in the middle. The heterogeneous classes 242 and 243 are only marginally included in the input data. Our generalization generates a similar fraction of these classes. However, the automatically generated areas are often not at the same location as in the manually generated reference dataset. We argue though that this is the result of an interpretation process, where different human interpreters would also yield slightly different results.

Input (DLM-DE) and the result match with 75%. This means that 25% of the area changes its class during generalization process. This is not an error; it is an unavoidable effect of the generalization. The κ-values 0.5–0.65 which stand for a moderate up to substantial agreement should also not be interpreted as bad results, because it is not a comparison with the real truth, or with a defined valid generalization, respectively.

10.4.3 Stability of Generalization Results

To test the influence of the generalization parameters to the result we made some experiments with our test datasets. To get an impression of its influence and to optimize the generalization, we changed each parameter separately in small steps. The result of the changed generalization was then compared with the input data and the CLC reference dataset. Also the statics (Table 10.2) was taken into account.

To simulate an update process and its effects on the generalized data, we used two different versions of the DLM-DE (a test version with data from 2006 and a refined version with data from 2009) (see Fig. 10.6). The land-cover of these two datasets differs in nine percent of the area (ground truth). Both datasets were generalized with the same parameters; the land-cover of the generalization results differs in 13% of the area. Twenty percent of these differences in generalized data are

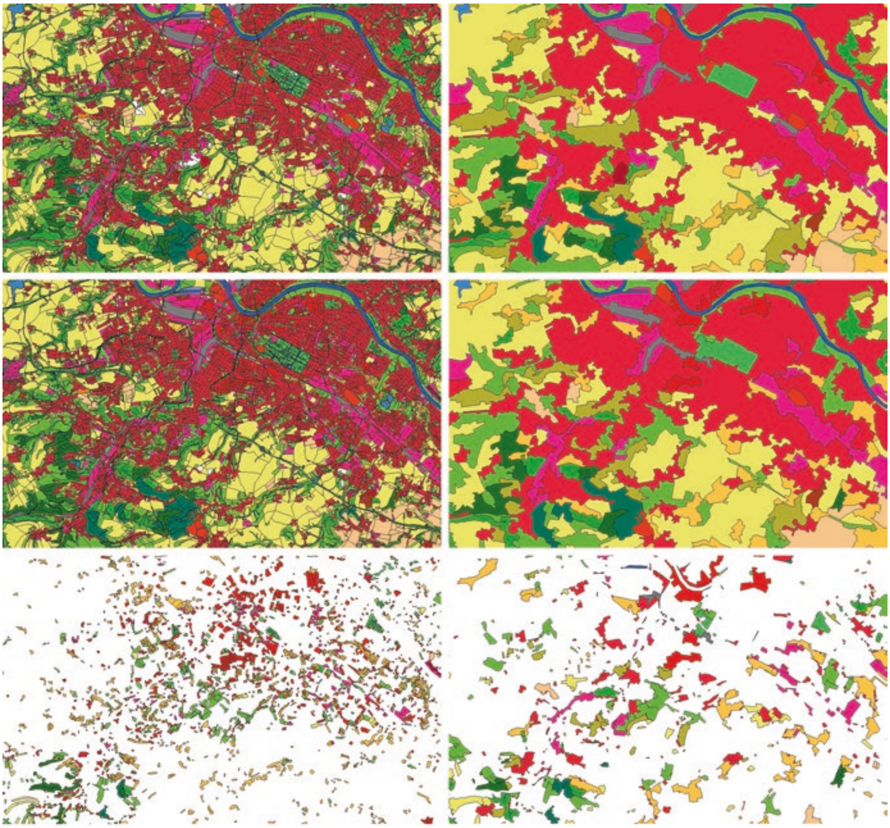


Fig. 10.6 Two versions of input DLM-DE (*left*), their generalization results (*right*) and the differences between the versions (*below*). Nine percent changes of the input data produce 13% differences in the generalized data

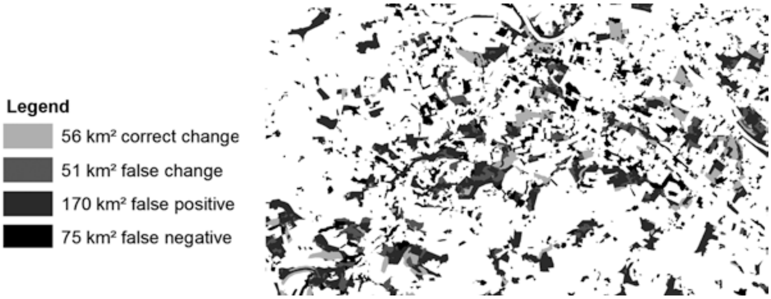


Fig. 10.7 Overlay of the differences between input and output data

correct and 20% are false (different classes). The other 60% are false positive – they occur at areas where no differences are in ground truth. Thirty percent of the real changes are missing (false negative) (see Fig. 10.7).

This example shows that changes in the input data produce more and also different changes in the generalized data. The causes of these changes are the classification and the aggregation step. In these generalization operations decisions are made based on thresholds. A small change can switch between the states under or over the threshold and produce a very different result. Because of the local decisions of the generalization algorithms this often leads also to changes in the local environment. Changes of the input data have only an influence in a limited environment.

10.5 Conclusions and Outlook

In this article, the so-called CLC-generator was described, which allows for a completely automatic production of CORINE LC data from a topographic data set, DLM-DE. The necessary operations and their suitable sequencing were described.

The classification of heterogeneous areas, in particular their demarcation from homogeneous areas, proved to be difficult with our rule-based approach. Better results may be achieved through supervised machine leaning.

Another challenge is to derive the CORINE land cover *change layer* from different versions of DLM-DE. The change layer cannot be generated by intersecting CORINE land cover datasets, due to the minimum mapping unit of the change layer, which is only five hectare in contrast to 25 ha for the land cover dataset. The EEA is only interested in real changes and not in so called technical changes (changes that are produced by the generalization). Resulting from our experiments in Sect. 4.3, we plan to intersect versions of the high resolution data DLM-DE and then to filter and aggregate the detected changes.

References

Arnold S (2009) Digital landscape model DLM-DE – deriving land cover information by integration of topographic reference data with remote sensing data. In: Proceedings of the ISPRS workshop on high-resolution Earth imaging for geospatial information, Hannover

Bossard M, Feranec J, Otahel J (2000) EEA CORINE land cover technical guide – addendum 2000. – Technical Report No. 40, Kopenhagen

Büttner G, Feranec G, Jaffrain G (2006) EEA CORINE land cover nomenclature illustrated guide – addendum 2006. – European Environment Agency

Douglas D, Peucker T (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Can Cartographer 10(1973):112–122

Geoff B et al (2007) UK land cover map production through the generalisation of OS MasterMap®. Cartogr J 44(3):276–283

Gomory R (1958) Outline of an algorithm for integer solutions to linear programs. Bull Am Math Soc 64(5):274–278

Haunert J-H (2008) Aggregation in map generalization by combinatorial optimization, vol. Heft 626 of Reihe C. Deutsche Geodätische Kommission, München

Haunert J-H, Sester M (2008) Area collapse and road centerlines based on straight skeletons. GeoInformatica 12(2):169–191

Heinzle F, Anders K-H (2007) Characterising space via pattern recognition techniques: identifying patterns in road networks. In: Mackaness W, Ruas A, Sarjakoski LT (eds) Generalization of geographic information: cartographic modelling and applications. Elsevier, Oxford, pp 233–253

Jansen LJM, Groom G, Carrai G (2008) Land-cover harmonisation and semantic similarity: some methodological issues. J Land Use Sci 3(2–3):131–160

Kavouras M, Kokla M (2008) Semantic integration of heterogeneous geospatial information. In: Li Z, Chen J, Baltsavias E (eds) Advances in photogrammetry, remote sensing and spatial information sciences 2008 ISPRS congress book. CRC Press/Balkema, London. 2008

Kirkpatrick S, Gelatt CD Jr, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671

Kuhn W (2003) Semantic reference systems. Int J Geogr Inf Sci Guest Editorial 17(5):405–409

Lillesand TM, Kiefer RW (1999) Remote sensing and image interpretation, 4th edn. Wiley, New York

Mackaness WA, Ruas A, Sarjakoski LT (2007) Generalisation of geographic information – cartographic modelling and applications. Elsevier Applied Science, Oxford

Meng L, Huang C, Zhao C, Lin Z (2007) An improved Hilbert curve for parallel spatial data partitioning. Geo-spatial Inf Sci 10(4.), 2007):282–286

Pondrenk M (2002) Aufbau des DLM50 aus dem Basis-DLM und Ableitung der DTK50 – Lösungsansatz in Niedersachsen. in: Kartographische Schriften, Band 6, Kartographie als Baustein moderner Kommunikation. Bonn, pp 126–130

Sester M (2005) Optimizing approaches for generalization and data abstraction. Int J Geogr Inf Sci 19(8–9):871–897

Thiemann F, Warneke H, Sester M, Lipeck U (2011) A scaleable approach for generalization of land cover data. In: Advancing geoinformation science for a changing world. Springer, Berlin, pp 399–420

Thiemann F, Werder S, Globig T, Sester M (2013) Investigations into partitioning of generalization processes in a distributed processing framework. In: Proceedings of the 26th international cartographic conference, Dresden Germany, 25–30 August 2013

van Oosterom P (1995) The GA ppP-tree, an approach to 'on-the-fly' map generalization of an area partitioning. In: Müller J-C, Lagrange J-P, Weibel R (eds) GIS and generalization – methodology and practice. Taylor & Francis, London, pp 120–132

Zhou X, Abel DJ, Truffet D (1998) Data partitioning for parallel spatial join processing. GeoInformatica 2(2):175–204

Chapter 11
Epilogue

Denise Pumain

This book marks a major milestone on the way towards a clever use of geographical data for solving various urban and regional problems. This is important because in many circumstances scientists do have an immense responsibility when facing choices for improving the human way of inhabiting the planet and especially when advising planners and stakeholders.

The authors in this book share strong intellectual and ethical requirements for a sound scientific approach in proposing new analytic methods and modeling approaches. Martin Behnisch and Gotthard Meinel in chapter 1 give a broad overview of the challenges and opportunities that are brought by a new era of immense data availability and computing power. They first underline the importance of starting from relevant theoretical perspectives for extracting information from data as well as in model building. The recent surge and explosion of geo-referenced data is both a great boon for spatial analysts but also a real challenge to transform them in meaningful knowledge and inject them in useful tools. On this topic, the book introduces new methods of data mining and exploration of new sources of data. Methods for playing with available data are becoming more effective, such as presented in Chap. 2 by Galen Maclaurin and Stefan Leyk who plea for developing processes aiming at improving the extraction of the information contained in remote sensing images. They suggest using machine learning tools for spatial extrapolation or temporal extension of land use data. In Chap. 3, Bin Jiang reminds us about the fundamental heterogeneity of geospatial data. Power laws and lognormal distributions are the rule because they are generated by the dynamics of non linear processes in complex systems. Since long geographers have admitted in an implicit or explicit way the spatial dependence in socio-spatial interactions that sustains the emergence of accumulations of very unequal sizes. The chapter is an invitation to consider

D. Pumain (✉)
Université Paris 1 – Panthéon-Sorbonne, Paris, France
e-mail: pumain@parisgeo.cnrs.fr

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8_11

scaling effects and fractal organizations as the ordinary references for any spatial distributions and to use related methods which were elaborated during the last decades for processing these data.

The question of relying on a safe theoretical background for meeting the new challenges about what to do with big data has become an even more acute issue when not only material traces but social characteristics and practices as well are involved. Facing the proliferation of geo-tagged communication traces on a variety of media, mixing quantitative and qualitative approaches is as usual very promising, including combinations of spatial and temporal with a semantic analysis for data mining. That is why a specific attention can be dedicated to Chap. 4 where Quan Yuan et al. propose a rich state of the art demonstrating that much progress is currently made in this research area. They enumerate a large number of different situations where knowledge about various uses of these new functionalities can be extracted.

It is however interesting to remark that traditional data sets and old ways of collecting information such as censuses have not gone completely out of style. In Chap. 5 Fernando Bação et al. illustrate the usefulness of census data at fine geographical resolution for enriching the concept of quality of life in a "smart city". They use algorithms of self-organizing networks for analysing the recent urban changes at block level in Lisboa, observing some hundred variables and thousands of local units at two dates and classifying them in reduced meaningful categories. This methodology in which both statistical similarities and spatial proximities are combined for a significant grouping of neibourhoods is also used by Julian Hagenauer and Marco Helbich in Chap. 6 for a socio-economic analysis of the city of Chicago. Indeed, the urban concepts that are behind the clustering do not differ much from those that were in use during the 1960s and 1970s when Brian Berry and John Kasarda tried to formalize the spatial distribution of the major socio-economic and demographic or ethnic processes identified by the patient field work of the previous "Chicago school" of sociologists deciphering the "urban ecology". But the methodological improvement brought about by powerful computing and visualization tools is now of considerable help for exploring spatial data, especially when different algorithms can be compared.

The next decades certainly will celebrate the arrival of interactive simulation models on the desk of urban planners and stakeholders. Solutions are actively investigated to face problems of reproducibility of simulation experiments and reliability of model predictions. Jonatan Almagor, Itzhak Benenson and Daniel Czamanski imagine in Chap. 7 a competitive process between urban developers whose ability to concentrate urban rent may be affected by the way planners make their decisions, whereas in Chap. 8 Andreas Koch tests on the city of Salzburg an agent based model on the location decision of residents leading to more or less segregated patterns. The resulting pattern at macro-scale is analyzed according to a variety of realistic behavioral rules (including interdependences between residents and with urban institutions) that have been added to the classical Schelling model, which makes the results more detailed and meaningful.

No less than 14 authors have collaborated in Chap. 9 to summarize the lessons from comparative experiments of land use change analysis. They certainly will create unanimity while reviewing the main challenges in mapping as being data selection, choice of resolution and categories identification. However, their recommendation to separate calibration and validation exercises in modeling could be discussed in the light of new computing developments enabling to couple both, such as made available on the OpenMOLE simulation platform for instance. However, the authors propose a very useful review of many models that are in use for that field of spatial analysis and we applaud their strong recommendation to provide precise measurements of their degree of uncertainty with any result given to the final users of the models. They wisely insist on the ability to learn with models which is an essential dimension of their usefulness.

Great advances in modeling also emerge from the improvement of visualization tools, such as the applications presented in Chap. 10 by Bruno Willenborg, Maximilian Sindram and Thomas H Kolbe for the estimation of solar irradiation, the simulation of detonations and the estimation of building heating energy demand. The 3D mapping techniques together refine the computation of indicators related to the physical dimensions of buildings and enrich the models with a powerful sense of realism that is more liable to ensure firmness of conviction. Another question that frequently arises is about the generalization procedure enabling to simplify the details of mapping according to the scale of documents. Frank Thiemann and Monika Sester propose in Chap. 11 a method enabling to automatize this delicate exercise in the case of producing land cover maps from a topographic database in the area of Dresden and explain all steps that have to be recognized in the process.

Making more explicit the very sophisticated methods that are now in use for collecting, processing, modeling and visualizing the old and new geo-tagged data is an urgent need for all planners and territorial stakeholders. This book really paves the way for improving our confidence in the present and future ability of spatial scientists to meet the challenge.

Index

M. Behnisch, G. Meinel (eds.), *Trends in Spatial Analysis and Modelling*,
Geotechnologies and the Environment 19, DOI 10.1007/978-3-319-52522-8