

Dipl.-Ing. Ariane Middel

VISUALIZING URBAN FUTURES

A Framework for Visualizing Multidimensional Geospatial Data

DISSERTATION

Vom Fachbereich Informatik
der Technischen Universität Kaiserslautern
zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
genehmigte Dissertation
von

Dipl.-Ing. Ariane Middel

Dekan: Prof. Dr. Karsten Berns

Prüfungskommission

Vorsitz: Prof. Dr. Karsten Berns

Erster Berichterstatter: Prof. Dr. Hans Hagen

Zweiter Berichterstatter: Prof. Dr. Subhrajit Guhathakurta

Datum der wissenschaftlichen Aussprache: 11. April 2008

© 2008 by Ariane Middel. All rights reserved.

Abstract

In urban planning, sophisticated simulation models are key tools to estimate future population growth for measuring the impact of planning decisions on urban developments and the environment. Simulated population projections usually result in large, macro-scale, multivariate geospatial data sets. Millions of records have to be processed, stored, and visualized to help planners explore and analyze complex population patterns.

We introduce a database driven framework for visualizing geospatial multidimensional simulation data based on the output from *UrbanSim*, a software for the analysis and planning of urban developments. The designed framework is extendable and aims at integrating empirical-stochastic methods and urban simulation models with techniques developed for information visualization and cartography.

First, we develop an empirical model for the estimation of residential building types based on demographic household characteristics. The predicted dwelling type information is important for the analysis of future material use, carbon footprint calculations, and for visualizing simultaneously the results of land usage, density, and other significant parameters in 3D space. Our model uses multinomial logistic regression to derive building types at different scales. The estimated regression coefficients are applied to *UrbanSim* output in order to predict residential building types.

The simulation results and the estimated building types are managed in an object-relational geodatabase. From the database, density, building types, and significant demographic variables are visually encoded as scalable, georeferenced 3D geometries and displayed on top of aerial photographs in a *Google EarthTM* visual synthesis. The geodatabase can be accessed and the visualization parameters can be chosen through a web-based user interface. The geometries are encoded in *KML*, Google's markup language, as ready-to-visualize data sets. The goal is to enhance human cognition by displaying abstract representations of multidimensional data sets in a realistic context and thus to support decision making in planning processes.

Acknowledgements

Although this dissertation is a piece of individual work, I could never have finished it without the help and support of numerous people. I would like to take this opportunity and thank all those who have contributed in any way to the completion of this dissertation.

First and foremost, I would like to express my sincere thankfulness to my PhD advisors for their support and guidance over the last three years. I would like to thank Professor Dr. Hans Hagen of the University of Kaiserslautern for providing the vision, encouragement, and advice necessary for me to progress through the PhD program and complete my dissertation. I deeply appreciate his valuable suggestions in both scientific and non-scientific matters. Additionally, I am very grateful to him for giving me the opportunity to work in an interdisciplinary and international research group. I would also like to thank my co-advisor of the Arizona State University, Professor Dr. Subhrajit Guhathakurta. He made important contributions to the early stages of my dissertation and was always readily available for discussion. His comments were always extremely perceptive, appropriate, and helpful.

Furthermore, I would like to acknowledge the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG Bonn) for funding this research within the International Research Training Group “Visualization of Large and Unstructured Data Sets”, grant no. 1131.

Special thanks go to all my friends and colleagues at the University of Kaiserslautern and the Arizona State University. I am particularly thankful to my fellow doctoral candidate Tom for plenty of enlightening and productive discussions over coffee, tea, or via ICQ. Further thanks go to Peter and Floys who helped implementing the prototype when time was running short. I would also like to give my sincerest thanks to Inga for her encouragement, her receptiveness to all my concerns, and for easing my organizational workload. Above all, I wish to express my gratitude to Kerstin for proof-reading parts of my thesis. She provided many invaluable

suggestions and incisive comments to help me improve my presentation and clarify my arguments. I would like to express my special thanks to the Digital Phoenix research group in the Herberger Center for Design at the Arizona State University. It has been a pleasure working with my colleagues from Phoenix in a congenial and supportive work environment. I am deeply grateful to Dianne for investing time and energy discussing ideas with me. Furthermore, I would like to thank Kartik for sharing his GIS knowledge. I am especially thankful to Robert who provided useful input and comments for the visualization chapter. Many of the ideas in my dissertation originated from fruitful discussions with him.

I gladly express my gratitude to my invaluable network of supportive, generous and loving friends from across the Atlantic who shared the best and worst moments of my dissertation journey in spite of geographical separation. Most notably, I thank Elke for numerous calls at three in the morning and for keeping some pink ambiance around all the time. Over the last three years, Elke has witnessed many ups and downs of my dissertation and no one could ask for a better friend in life. Furthermore, I deeply appreciate Markus for holding the fort in Bonn while I lived overseas. I also wish to thank Luise and Lio for their encouraging words and thoughtful criticism as well as Saskia, Kai, and Ruth for the support they have lent me over all these years. Special thanks go out to Nicole for her sense of humor and to Julia for providing a pair of warming doctor fish socks. I would also like to take the opportunity to thank my friend Björn for enduring many moods and working nights in the weeks before my defense and for his invaluable support and encouragement.

Finally, and most importantly, I would like to thank my parents, Klaus-Peter and Monika, for their love, constant support, and patience during the long years of my education. Without their ongoing encouragement and confidence, this dissertation would not have been possible.

Contents

List of Figures	vii
List of Tables	ix
1. Introduction	1
1.1. Related Work	3
1.2. Research Objective	8
1.3. Selected Approach	9
1.4. Thesis Structure	11
2. Procedural City Modeling - A Survey	13
2.1. Grammar-based Models	13
2.2. Agent-based Models	16
2.3. Statistical Models	17
2.4. Real-time Procedural Modeling	18
2.5. Classification and Comparison of Procedural Modeling Techniques	19
2.6. Discussion of Application Areas	21
3. Key Framework Applications and Components	23
3.1. Data Modeling with <i>UrbanSim</i>	24
3.2. Data Management with <i>PostGreSQL</i> and <i>PostGIS</i>	25
3.3. Data Visualization with <i>Google EarthTM</i>	26
4. Estimation of Residential Building Types	29
4.1. Multinomial Logistic Regression (MNL)	30
4.2. Mapping the Real World	32
4.3. Block Group Based Estimation	36
4.3.1. Clustering	37
4.3.2. Empirical Analysis and Statistical Tests	38
4.3.3. Results	42
4.4. Grid Cell Based Estimation	44

4.4.1.	Estimation using Synthesized Data from <i>UrbanSim</i>	44
4.4.1.1.	Empirical Analysis and Statistical Tests .	45
4.4.1.2.	Results	47
4.4.2.	Estimation Using Census Data	48
4.4.2.1.	Empirical Analysis and Statistical Tests .	48
4.4.2.2.	Results	50
4.5.	Final Prediction with Regression Coefficients	52
4.6.	Discussion of Results	53
5.	Visualization Framework	55
5.1.	Geovisualization	56
5.2.	Prototypical Implementation	57
5.2.1.	Data Revisited	58
5.2.2.	GUI	59
5.2.3.	Data Scales of Measurement and Visual Variables .	61
5.2.4.	Scalable Geometries	63
5.2.5.	The Keyhole Markup Language <i>KML</i>	64
5.2.6.	Using PHP and MySQL to Generate KML	66
5.2.7.	Coordinate System Transformations	67
5.2.8.	Visualization Results in <i>Google EarthTM</i>	70
5.3.	Discussion of Results	74
6.	Conclusions	77
A.	Parameter Estimates	81
	Bibliography	85
	Curriculum Vitae	96

List of Figures

1.1.	Procedural Models of Pompeii and Beverly Hills [MWH ⁺ 06]	4
1.2.	Example of GIS Visualization: Population Density in Phoenix	5
1.3.	Examples for <i>InfoVIS</i> : Treemap on character data in a World of Warcraft realm, scatterplot comparing CO_2 levels against temperature [VWvH ⁺ 07]	5
1.4.	Visualization of monthly health data by means of 3D icons on a map [TSWS05]	7
1.5.	<i>GeoDOVE</i> [HJF06]	7
1.6.	Architecture of Geovisualization Framework	10
2.1.	Street networks generated with L-Systems	14
2.2.	Architectural models generated with shape grammars	15
2.3.	Office district and Pompeii, modeled with CGA SHAPES [MWH ⁺ 06]	16
2.4.	Agent-based modeling of virtual cities [LWR ⁺ 04]	17
2.5.	Statistical model of Manhattan [YBH ⁺ 02]	18
2.6.	Real-time generation of 'pseudo-infinite' city [GPSL03a]	19
2.7.	Predominant characteristics of procedural modeling techniques	20
3.1.	Data integration process (c.f. [Wad02])	24
3.2.	OGC Simple Feature Specification (cf. [Ope07])	26
3.3.	A Placemark in <i>Google Earth</i> TM	27
4.1.	Mapping the real world	32
4.2.	Maps of Arizona and Maricopa County	33
4.3.	Single family dwellings, very small lots (a) and small lots (b)	35
4.4.	Single family dwellings, medium lots (a) and large lots (b)	35
4.5.	Apartments, 5 units (a), 25 – 100 units (b), > 100 units (c)	35
4.6.	Percentages of different building types in each block group	36
4.7.	Predicted vs. observed neighborhood categories	43
4.8.	Difference picture of correct vs. incorrect predictions	44

4.9. Predicted building types	53
5.1. Access to <i>PostgreSQL</i> database via PgAdmin	59
5.2. Web-based data access and geometry choice	60
5.3. Web-based parameter assignment for scalable 3D box	60
5.4. Visual variables and their effectiveness (c.f. [Mac95])	62
5.5. Geometries	64
5.6. <i>KML</i> sample code	64
5.7. <i>KML</i> elements (c.f. [Goo07a])	65
5.8. Figures of the Earth	68
5.9. Geocentric and local <i>SRS</i>	69
5.10. Geographical Coordinate System	69
5.11. Scaled 3D boxes represent population density (2000) in Phoenix Downtown	70
5.12. Building type geometries: (a) apartments, (b) small single family, (c) medium single family, and (d) large single family	71
5.13. Visualization of (a) average income (color-coded grid cells), (b) building types (geometry), (c) population density (size of footprint), and (d) uncertainty of building type predic- tion (transparency) [MGH ⁺ 08]	72
5.14. Close-up of icons for different building types [MGH ⁺ 08]	73
5.15. Phoenix Downtown, color-coded grid cells display distance to nearest highway [MGH ⁺ 08]	74

List of Tables

4.1. Building types	34
4.2. Final cluster centers	37
4.3. Case processing summary	38
4.4. Pseudo R^2	39
4.5. Model fitting information	40
4.6. Likelihood ratio statistics, reduced model	41
4.7. Classification table	41
4.8. Parameter estimates for cluster 5 (single family, XS & XXS lots)	42
4.9. Case processing summary	46
4.10. Pseudo R^2	46
4.11. Model fitting information	46
4.12. Likelihood ratio statistics, reduced model	47
4.13. Classification table	48
4.14. Pseudo R^2	49
4.15. Model fitting information	49
4.16. Likelihood ratio statistics, reduced model	50
4.17. Classification table	50
A.1. Block group based estimation (see 4.3)	81
A.2. Grid cell based estimation (see 4.4.1)	82
A.3. Grid cell based estimation (see 4.4.2)	83
A.4. Prediction results for sample grid cells	84

1. Introduction

Technological innovation and economic development are crucial for an environmentally sound future and a sustainable society. Therefore, sophisticated simulation models are now extensively used in urban planning to measure the impact of planning decisions on future urban environments. These simulation models support decision-making in planning processes by providing assessment of alternative planning scenarios. For example, alternative planning scenarios could examine the impacts of future transportation networks on the type and location of new housing and on air quality. Other environmental aspects to assess are water supply and change of temperature.

Simulation models take into account past and present developments to draw conclusions on future trends. Advanced urban simulation models provide estimates on future population and job distributions, demographic household profiles, and predicted travel behavior at a fine grained spatial scale.

Yet, current planning tools lack empirical models for predicting building types. Knowledge of future building structures is crucial for various sustainability metrics associated with prospective urban developments. Future building type distributions are important parameters in material flow analyses. For example, we can calculate future construction materials from the number and types of residential dwellings that will be built in future years. Moreover, waste generation from the life cycle of building materials can be quantified when building types are known. Future homes also play an important role in calculating carbon footprints and are critical for the spatial analysis and simulation of 3-dimensional phenomena like air pollution, noise, and earthquake risks. Although there are urgent needs for the estimation of future building types, to our knowledge current urban planning tools do not provide this information.

Another notable aspect of urban simulation models is the large, macro-scale, and multidimensional geospatial data sets that they require and

generate as output. Millions of records have to be processed, stored, and visualized to help planners explore and analyze complex patterns. Decision-making in the urban context now requires developers, planners, stakeholders, and the public to deal with highly attributed, often time-varying data. Data sets from urban simulations are useful for decision making, but they are difficult to analyze. The interpretation of abstract multidimensional simulation data requires expertise that often cannot be expected from decision-makers. Therefore, a comprehensive and intuitive visualization tool is crucial for supporting non-experts in their decision-making process and for communicating results to the general public.

Unfortunately, some current urban simulation tools such as *UrbanSim* do not provide visual output for output data tables. Future land usage, density, or other significant environmental indicators and demographic characteristics are most commonly visualized as 2D thematic maps in a geographic information system. Realistic-looking virtual cityscapes on urban futures can be generated procedurally, but to date, procedural city models lack an underlying empirical framework for urban simulations.

Given the lack of adequate means for visualizing and communicating projected urban data sets, there is an immediate demand for intuitive, comprehensive visualization tools to support planning and informed decision making about the future. As already pointed out, simulation output often exhibits multidimensionality and spatial correlation. This raises the demand for new visual representations which are beyond classical 2D maps, particularly to visualize the interrelationships among multiple dimensions of the complex underlying data structure. Visual data mining, especially in large spatial databases, is a key technique for interactive analysis and explorative visualization of the parameter dependencies. Using an appropriate data representation and visualization framework, those engaged in visual analysis can rapidly perceive important patterns, identify previously unknown patterns, or find relationships between socioeconomic and demographic variables.

All things considered, we perceive considerable demand for an integrated planning tool to simulate and visualize future urban developments in the domains of urban and environmental planning. Simulation outputs of this planning tool have to include or allow for the estimation of future building types, since building structures are crucial for many sustainability measures. Furthermore, an integrated visualization framework has to display resulting multidimensional simulation data in an intuitive way to effec-

tively enhance visual thinking and knowledge discovery. Such a platform will make simulation outcomes accessible to planners, key decision-makers, and the general public and will aid participants in planning processes to reach decisions.

1.1. Related Work

Modeling urban futures has been of considerable interest of late to computer graphics and urban planning [BX94, Tor06, Wad02, KB01, Klo01]. The interest is driven by a number of developments, in particular by increased computing power and significant advances in the field of integrated urban and environmental modeling [Guh03]. Growing awareness of sustainability amplifies the demand for planning support systems to analyze and visualize future carbon footprints, air quality, and other indicators of urban quality of life and livability. Visualizing how the future of cities and urban developments may look communicates efficiently the impacts of today's planning decisions on future environments.

Recent efforts in computer graphics aim at automating the complex and expensive task of generating and visualizing realistic cityscapes. Various tools and techniques are constantly being developed to automatically detect and reconstruct buildings from remotely sensed imagery for generating 3D city models [För99, HYN03, Bre05] and to undertake building energy requirement calculations [NS04]. Yet, photogrammetric reconstruction approaches only allow for the generation of existing and past stages of urban developments. They lack the ability to provide information about future cityscapes. In this context, procedural modeling techniques have been subject to active research lately. Procedural methods algorithmically generate arbitrary geometries from a predefined rule-set. Key techniques for procedural visualization include agent-based modeling approaches [LWWF03] and grammar-based approaches. Procedural models based on grammars have been developed to generate architecture [WWSR03], building facades [MZWG07], ancient Roman sites [MWH⁺06], and large-scale 3D cityscapes [PM01, dSM06].

Procedural modeling of cities allows for the generation of realistic-looking urban environments, but to date it lacks underlying empirically-based models to generate meaningful projections of urban growth patterns that can be calibrated to simulate real conditions. However, population growth



Figure 1.1.: Procedural Models of Pompeii and Beverly Hills [MWH⁺06]

and patterns of distribution of urban activities in future cities are of great interest to urban planners. Planning agencies in most industrialized countries are now mandated to provide official projections of future urbanization patterns to monitor land development. Therefore, these planning agencies develop and maintain urban futures simulation models pursue research in modeling complex urban systems to analyze land-use decisions and evaluate alternate growth management strategies. Commercially available GIS-based planning support systems are *WhatIf* [Klo01] and *CommunityVIZ* [KB01]. *UrbanSim* [Wad02] is a sophisticated open-source planning tool for analyzing long-term effects of land use and transportation policies. It provides a platform for generating different urban scenarios based on current trends and specified policy choices, e.g., to model low-density urban sprawl or to evaluate the sustainability of transportation plans [JGK⁺06].

Typically, projections of urban simulation models are visualized with the help of Geographic Information Systems (GIS) as color-coded polygons or icons on 2D maps (see Figure 1.2). Since prediction models usually result in large-scale, highly-attributed spatial data sets, the challenge arises to find a multidimensional representation for the simulation results. The multi-attributed visualization technique could enable planners to easily compare different planning scenarios and to evaluate simulated impacts of different land use policies.

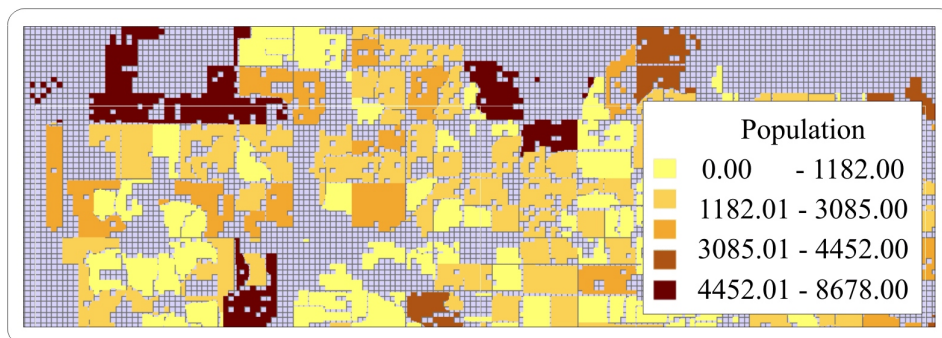


Figure 1.2.: Example of GIS Visualization: Population Density in Phoenix

Visualizing multivariate data sets has long been a key issue in Information Visualization (*InfoVIS*). In the early 70's, Chernoff presented a technique to visualize trends in highly dimensional data by relating data to facial features [CR75]. Gradually over the years, new information visualization techniques were introduced, ranging from 2D scatterplots to 3D treemaps (see Figure 1.3). For a comprehensive overview of developments in Information visualization we refer to [SCM99] and [Tuf90].

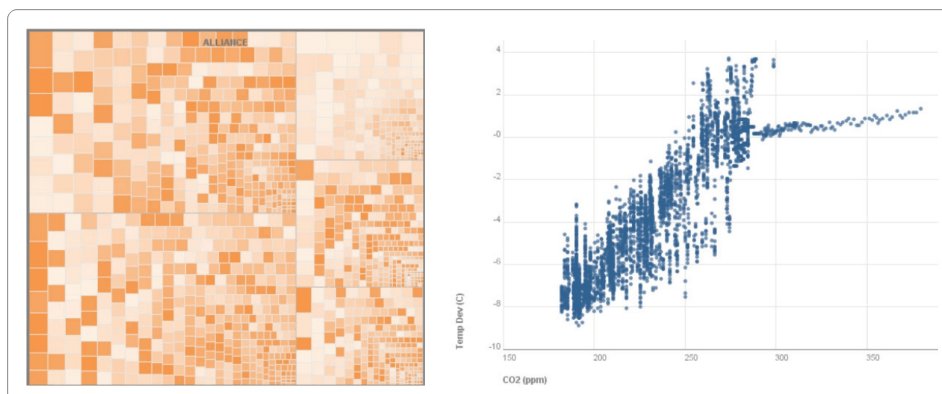


Figure 1.3.: Examples for *InfoVIS*: Treemap on character data in a World of Warcraft realm, scatterplot comparing CO_2 levels against temperature [VWvH⁺07]

Whereas *InfoVIS* primarily deals with the display of large multivariate data sets, cartography is mainly concerned with representations constrained to a spatial domain. The cartographer Bertin established a basis for designing maps in his classical work “Semiology of Graphics” [Ber67]

where he identified a set of fundamental visual variables and defined graphical rules for their appropriate use. Since then, Bertin's concepts have been constantly modified and extended. Modern cartography transfers design knowledge from 2D paper maps to new media [FAA⁺01]. On-screen interactive maps are designed to assist in visual data exploration and analyses [AA99]. Cartographic visualization is also extended to abstract and non-geographic data by spatialization [SF03].

Recently, efforts have emerged to combine techniques from both cartography and information visualization [Sku00, FS04]. Geographic visualization (*GeoVIS*) is a new, rapidly evolving domain, especially since the availability of geodata is increasing. In 1998, MacEachren compiled a research agenda entitled "Visualization - Cartography for the 21st century" [Mac98] and addressed *GeoVIS* research challenges. Since then, cartographic and *InfoVIS* techniques have frequently been applied to design integrated geovisualization tools. Latest advances include multivariate analyses with self-organizing maps [GGMZ05, SH03], studies on human activity patterns using 3D space-time paths [MPJ04], visual data mining in large-scale 3D city models [BD05], and bivariate maps for public health studies [MGP⁺04]. Most recent activities in geovisualization research are discussed in [Kra06].

Pinnel et al. [PDBB00] conducted a study on visualization designs for urban modeling. They found out that map-centered visualizations are the most useful portrayals for urban planning and analysis, since map layout encodes location information, which is crucial for decision-making. A map-based visualization approach, "The Indicator Browser", was designed by Schwartzman et al. [SB07] to display *UrbanSim* simulation results. The browser uses comparative visualizations of 2D maps to satisfy multivariability. This approach often impedes human vision to recognize complex patterns across many dimensions.

Instead of encoding n -dimensional data in n 2D maps, Tominski et al. [TSWS05] follow a different approach to display monthly health data (see Figure 1.4). They visualize time dependent multivariate disease information as 3D pencil and helix icons geocoded on a base map. Their research helps to analyze complex patterns across multivariate, spatial, and temporal dimensions, but lacks a powerful database and GIS functionality to manage, process, and distribute data. This problem is tackled by the web-based Geospatial Database Online Visualization Environment *GeoDOVE*. In [HJF06], Fairbairn et al. present a prototype visualization tool for integrating geodatabase servers with 3D geodata visualization methods using

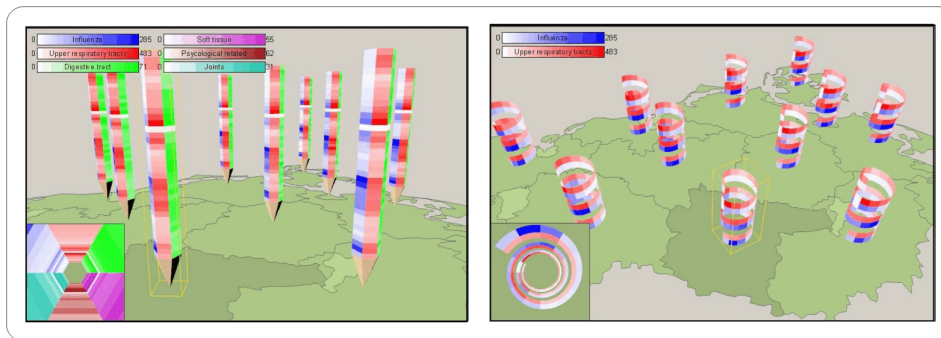


Figure 1.4.: Visualization of monthly health data by means of 3D icons on a map [TSWS05]

Java3D (see Figure 1.5). Similar to this approach, Steiner et al. [SMG01] developed a geovisualization tool for the internet where data is accessed from an Oracle commercial database through a map-based Flash interface. Also easily accessible to the public is *gCensus* [Imr07], a website that uses dynamic, high-resolution maps from *Google EarthTM* to visualize US Census 2000 data. Other 2D *Google EarthTM* visualizations are reported by Pezanowski et al. [PTM07] who created an application to support crisis management and by Wood et al. [WDSC07] in the context of interactive visual exploration of a large spatio-temporal data set.

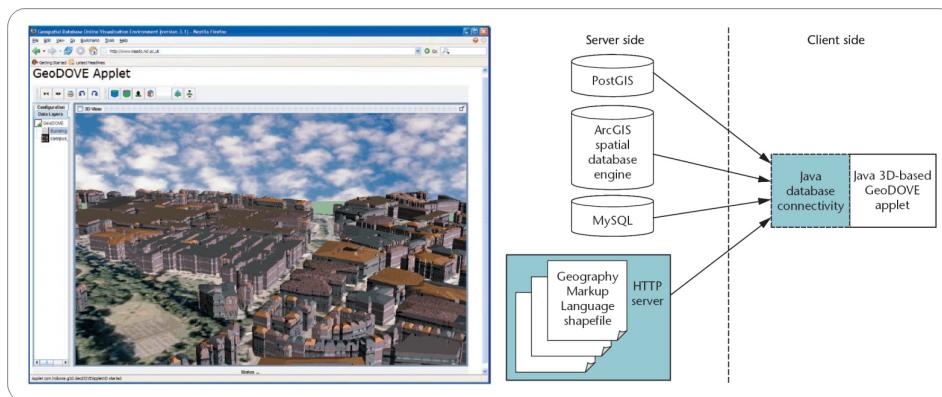


Figure 1.5.: *GeoDOVE* [HJF06]

From the review of literature we note that many different approaches for simulating and visualizing multidimensional geodata in the domains of urban planning, computer science, information visualization, cartography,

and geovisualization exist. However, to date no integrated framework has been established to combine the advantages of the reviewed approaches. Although urban simulation tools like *UrbanSim* provide plausible predictions and have been widely used to assist urban planners, they do not allow for a prediction of future residential building types, which is crucial for the analysis of some urban and environmental factors such as future carbon footprints or material use. Also, planning tools lack an integrated 3D map-based visualization framework as presented in [TSWS05] that can handle multidimensional geospatial data sets. Finally, a geodatabase as reported in [HJF06] facilitates storing, processing, and distributing data. *Google EarthTM* based visualization tools like *gCensus* [Imr07] or the *Mashup* introduced by [WDSC07] increase data accessibility to the general public, but both approaches do not make use of *Google EarthTM*'s ability to visualize data in 3D.

In this thesis, we develop a geovisualization framework to overcome above stated drawbacks. Our framework is capable of visualizing large-scale multidimensional geodata as 3D scalable geometries, superimposed on *Google EarthTM*. In the remainder of this chapter, we state our research objectives, provide an overview of the geovisualization framework, and discuss the outline of this thesis.

1.2. Research Objective

Urban simulation models are frequently used for decision making, since they communicate future impacts of planning decisions and allow envisioning alternative futures. Yet, current simulation data on future urban developments lack information on residential building types, which are crucial for many sustainability metrics. Moreover, simulation results cannot be visualized adequately with available planning systems. These systems encounter serious visualization limitations due to increasing data set size and multidimensionality. Consequently, participants in planning processes are often confronted with the problem of making decisions without sufficient knowledge about possible futures.

The underlying goal of this research is to improve decision-support in planning systems by developing a concept for an integrated simulation and visualization platform. The wider scope of our contribution impacts knowledge from many research domains including statistics, economics, ur-

ban and environmental planning, cartography, information visualization, and geovisualization.

First, we derive future residential building types from predicted demographic household profiles. Dwelling types are estimated using multinomial logistic regression analysis and can be used to advance analyses of future material use, carbon footprint calculations, and other indicators of urban quality of life and livability.

Then, we close the gap between the simulation and visualization of multidimensional geospatial data on future urban environments. We set up a powerful geodatabase to store, manage, and process simulation data. The geodatabase provides the technical basis for our integrated geovisualization framework. The focus of research here is on designing visualizations applied specifically to the multidimensional simulation data at hand in order to incorporate knowledge into planning processes. We aim at facilitating exploratory visual analyses of simulation data by creating a web-based interactive geovisualization tool. Our framework uses *Google EarthTM* as effective and easy-to-use visual interface to display abstracted 3D graphical representations of data attributes on top of aerial photographs. This contribution leads to a comprehensive, intuitive, and flexible environment for visual data mining. Our tool is of relevance to both experts and non-experts and will enhance understanding of the impacts of policy choices on resulting population growth patterns.

1.3. Selected Approach

The geovisualization framework architecture mainly consists of three architectural components (cp. Figure 1.6): a data processing layer, an object-relational data base management system (*ORDBMS*), and a geovisualization layer.

Within the data processing layer, demographic data is joined, intersected, and aggregated in *ArcGIS* from different sources. The Assessors' file makes detailed property information available at parcel level and the Census of Population and Housing provides aggregated statistical data on the number of persons as well as selected social, economic, and financial characteristics. The US census data is reported at different aggregation levels, census blocks being the smallest geographic subdivision for which the

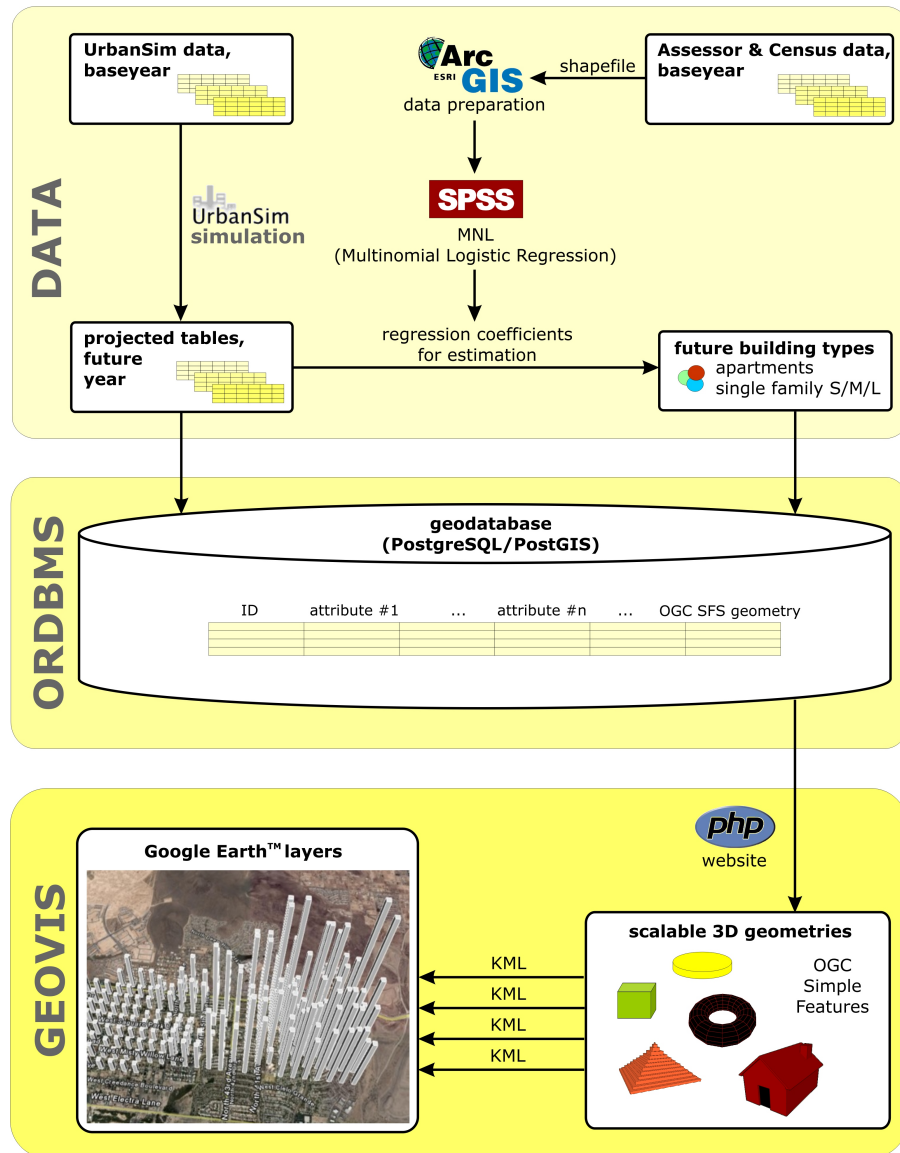


Figure 1.6.: Architecture of Geovisualization Framework

Census Bureau tabulates data. For a user defined base year, future residential building types are estimated based on from assessors' parcel and census data by multinomial logistic regression (*MNL*) in *SPSS* [Mid07b]. The regression analysis provides estimation coefficients, which are applied to simulated demographic data sets in order to predict future dwelling types. Data on household characteristics for future years is simulated in

UrbanSim, a sophisticated modeling tool for policy driven forecasts of regional employment, population, and land-use change in metropolitan regions. The demographic simulation results and the estimated building types are stored and managed in an object-relational *PostgreSQL* geodatabase. Within the database, demographic and geometric attributes can be retained in a geographic reference frame.

From the *PostgreSQL* database, density, building types, and significant demographic variables are visualized in 3D space on top of a map with scalable, georeferenced 3D geometries [Mid07a, MGH⁺08]. The variables are mapped to visual attributes of different geometric objects and visualized through a Google Earth client interface. The idea is to enhance human cognition by displaying abstract representations of multidimensional data sets in a realistic context. Access to the geodatabase and choice of visualization parameters is gained through a web-based *PHP* user interface. The scalable georeferenced 3D geometries are encoded in *KML*, Google's own markup language, as ready-to-visualize data sets.

The data pipeline preparation - simulation - estimation - visualization is separated in two high level conceptual components: data processing/storage and data visualization. This separation is also reflected in the organization of this thesis. Chapter 4 is dedicated to data-related framework components, chapter 5 focuses on the visualization module. In the next section, we provide an outline of the organization of this thesis.

1.4. Thesis Structure

The first chapter introduced the reader to the demands and challenges of visualizing urban futures. We explored the theoretical and methodological advances made in the domains of urban simulation and computer graphics. After reviewing central visualization ideas related to procedural modeling, urban simulation, cartography, information visualization, and geovisualization, we proposed a *GeoVIS* framework for processing, managing, and visualizing highly attributed geospatial data. The remainder of the thesis is organized as follows:

In Chapter 2, we provide an overview of existing procedural modeling techniques for generating virtual cities. We review and compare state-of-the-art research projects including grammar-based and agent-based mod-

els, statistical approaches, and real-time procedural modeling techniques. A subsequent discussion reveals potential applications for procedurally generated city models and explains the need for a different, more abstract visualization approach to analyze multidimensional urban simulation data.

Chapter 3 briefly sketches the main components of the visualization framework to provide a context of how each of these tools is employed to simulate, store, and visualize data on urban futures.

In Chapter 4, we develop an empirical model for the estimation of residential building types based on demographic characteristics. Our model uses multinomial logistic regression to derive dwelling types, first from *UrbanSim* data at a neighborhood scale, then from census data at a smaller grid cell scale. Finally, we apply the estimated regression coefficients to *UrbanSim* output in order to predict future residential building types.

Chapter 5 introduces the prototype of our interactive geovisualization tool. The prototype integrates *UrbanSim* output data and the empirical results on estimated residential building types from the previous chapter for superimposing on top of *Google EarthTM*. We describe how data is stored and accessed using a *PostgreSQL* geodatabase. For the visualization, multiple data attributes are represented by 3D scalable geometries which are generated from 2D geometric primitives. The geometries are output as *KML* file by means of *PHP* and *SQL*. Concluding, we show final visualization results and identify central advantages and disadvantages of our geovisualization approach.

Chapter 6 gives a synopsis of our new and comprehensive geovisualization framework. We highlight contributions of this thesis, draw conclusions and identify directions for further research.

2. Procedural City Modeling - A Survey

The problem of modeling large-scale virtual urban environments has remained a challenging task, especially for integrating the promising advances in computer graphics with the critical decision-making tasks of urban planners. Cities are difficult to model in detail, since buildings embody diverse and complex geometries. Modeling large-scale 3D city models by means of photogrammetric reconstruction is time and resource intensive, often semi-automatic process and does not provide data for visualizing future cityscapes.

Recently, significant research in the area of computer graphics has been dedicated to developing alternative visualizations in the field of procedural modeling. Procedural modeling uses algorithms to generate 3D geometries rather than storing an enormous amount of low-level primitives [Ebe96]. Creating formalized models is an efficient and flexible way to reduce the amount of stored data. Procedural modeling yields good results for repeating and random processes as well as for self-similar features like fractals [Bat05].

In the following sections, we address different techniques for procedural city modeling, compare the classified techniques, and discuss their usefulness for visualizing multidimensional projection data of urban simulation models.

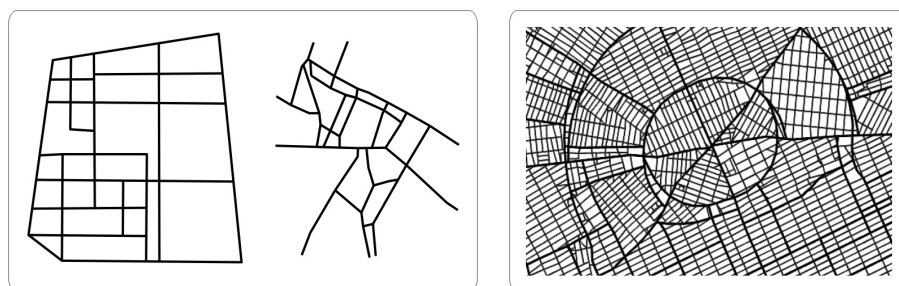
2.1. Grammar-based Models

In the context of procedural city modeling, formal grammars have proven to be a powerful modeling tool. Recent approaches use so-called L-Systems for generating a variety of geometric elements in 3D city models [CdSF05]. An L-System or Lindenmayer-System is a parallel string rewriting mecha-

nism that alters a string iteratively according to specified production rules. The resulting string can be interpreted geometrically to produce graphical output. L-Systems were conceived by Aristid Lindenmayer [Lin68] to describe the development of multicellular organisms. In the 90's, they became a sophisticated computer graphics tool for simulating and visualizing plant geometry [PL90].

Kato et al [KOO⁺98] are the first to reveal a substantial similarity between the growth of branching structures and the development of street networks. They introduce a virtual city modeling technique using stochastic parametric L-Systems to generate varying road networks. Their technique supports hierarchical street systems and can produce both linear flow systems and cellular networks (see Figure 2.1(a)).

The city modeling system CITYENGINE by Parish and Müller [PM01] incorporates an advanced street generation algorithm based on extended L-Systems. Unlike previous Lindenmayer-Systems, the enhanced grammar allows for the creation of closed loops and intersecting road branches. This is accomplished by adding self-sensitiveness to the nature of L-Systems. CITYENGINE employs a hierarchical set of production rules and enables the generation of streets that follow superimposed patterns. To derive a large-scale road map (compare Figure 2.1(b)), geographical image maps of elevation, vegetation, and land-water boundaries as well as geostatistical maps on population density, zones, and land-use serve as input data. Since the introduction of CITYENGINE, L-Systems have been widely used for both reproducing existing street networks [GMB06] and creating fictional road maps [HMFN04]. Yet for modeling geometrically detailed buildings, L-Systems are difficult to adapt since they emulate growth-like processes in open spaces, but realistically, building structures are bounded.



(a) Map and Tree L-Systems [KOO⁺98] (b) Self-sensitive L-Systems [PM01]

Figure 2.1.: Street networks generated with L-Systems

CITYENGINE implements an L-System to generate simple buildings consisting of translated and rotated boxes. In this way, large urban environments emerge, but with a low resulting Level of Detail (*LoD*).

INSTANT ARCHITECTURE is a procedural technique developed by Wonka et al. [WWSR03] for automatic modeling of geometrically detailed buildings. Their approach uses parametric split grammars, derived from the concept of shape grammars [Sti80] which have been successfully applied in architecture to construct and analyze architectural designs. Split grammars operate with production rules consisting of geometric split operations. The idea is to generate geometrically rich 3D building layouts by hierarchically subdividing building facades into simple attributed shapes (see Figure 2.2(a)). Wonka et al. set up a large grammar rule database to model various buildings in different architectural styles. INSTANT ARCHITECTURE yields to high LoD buildings, but is only applicable for small-scale urban areas.

Inspired by INSTANT ARCHITECTURE is the virtual 3D model of Roman housing architecture presented in [MVUG05]. Müller et al. reproduce ancient sites by extending the functional range of the CITYENGINE system to shape grammars similar to those introduced in [WWSR03]. The production rules for the shape grammars are deduced from archaeological and historical data to ensure a faithful reproduction of Roman architecture. Plausibility is further enhanced by importing real building footprints and streets as ground truth.

Also integrated in the CITYENGINE framework is a sophisticated technique for procedural modeling of computer graphics architecture evolved



(a) Instant Architecture [WWSR03]



(b) Roman Housing [MVUG05]

Figure 2.2.: Architectural models generated with shape grammars

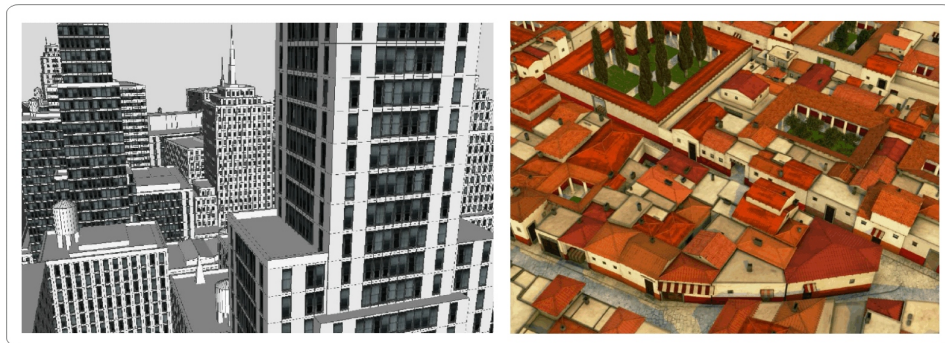


Figure 2.3.: Office district and Pompeii, modeled with CGA SHAPES [MWH⁺06]

by Müller et al. [MWH⁺06]. Their approach uses extended set grammars, so-called CGA SHAPES, combining the benefits of [WWSR03] and [PM01]. CGA SHAPES are suited for creating large-scale and at the same time geometrically detailed 3D cityscapes. Intrinsic context sensitive shape rules are applied sequentially to building footprints, the axioms of the productions, in order to generate mass models of the buildings. A mass model is a union of simple volumes and can consist of highly complex polygonal faces. In the next step, 2D building facades are extracted from the 3D shapes and structured into their elements. Here, CGA SHAPES re-use the volumetric information to solve intersection conflicts between adjacent facades. After adding details for ornaments, doors and windows to the facades, the buildings are finally roofed with different types of house tops. In this manner, CGA SHAPES are applicable to model diverse urban areas like office districts, suburban environments and ancient cities (see Figure 2.3).

2.2. Agent-based Models

Agent-based models are computational models for simulating real world phenomena and are closely related to cellular automata and multi agent systems. The main modules of agent-based models are rule based agents, situated in space and time. They reside in artificial environments, e.g., virtual cities, are free to explore their surroundings, and dynamically interact with their environment and other agents. Simple transition rules drive the agents' behavior and result in purposeful, intelligent, far more

complex reactions. For a detailed introduction to computational agent-based modeling, particularly regarding multi-agent systems, and a broad introduction to automata-based urban modeling the reader is referred to [BT04].

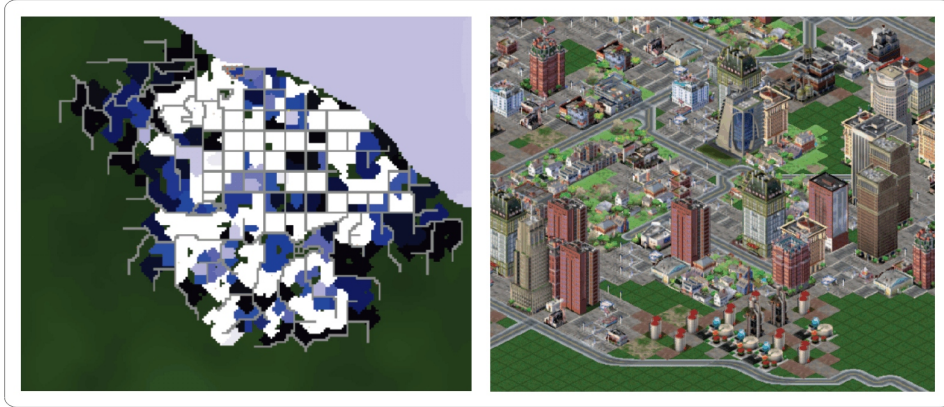


Figure 2.4.: Agent-based modeling of virtual cities [LWR⁺04]

Lechner et al. [LWR⁺04] present an approach to procedurally generate virtual cities using an agent-based simulation. Their system depends on a terrain description as low-level input and accepts optional parameters like water level and road density. Three types of agents are responsible for generating the road network by exploring the virtual space. Primary road agents connect highly populated regions, extenders expand the existing street network to urban areas not serviced by a road, and connectors interlink poorly accessible areas with streets. To output a land usage map (compare Figure 2.4) developer agents generate parcels and land use for residential, commercial, and industrial zones. Finally, the map is visualized using the SimCity 3000 graphics engine, as shown in Figure 2.4.

2.3. Statistical Models

As opposed to grammar-based and agent-based approaches, statistical models utilize statistical propagation techniques to procedurally generate urban environments. Statistical models are often employed to complement other procedural modeling techniques and are rarely used stand-alone. An example for the automatic generation of large geometric models based on

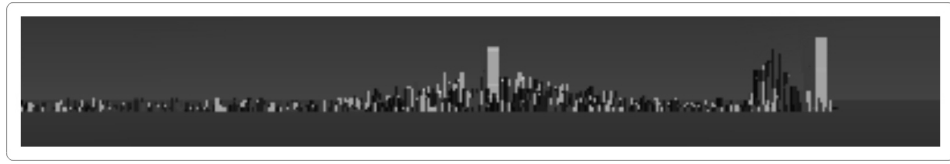


Figure 2.5.: Statistical model of Manhattan [YBH⁺02]

statistical parameters is “A Different Manhattan Project” [YBH⁺02]. Yap et al. reconstruct the city of Manhattan based on the TIGER data set and diverse physical parameters like average size and height of buildings, zoning classification of land use, and dominant architectural styles. The parameters are statistically propagated over the city, using different parameter scripts for varying districts in order to capture the uniqueness of each neighbourhood. Landmarks such as the Empire State Building are hand-coded into the geometric model to accomplish a realistic view of the city skyline (see Figure 2.5).

2.4. Real-time Procedural Modeling

Research goals of computer graphics with respect to procedural models are high realism and fast rendering of complex scenes. Other important characteristics for procedural models in computer graphics are real-time and interactivity. While the modeling approach by DiLorenzo et al. [DZT04] is focused on the interactive animation of evolving cities, the technique introduced in [GPSL03a, GSL04] concentrates on dynamic geometry rendering in real-time. Greuter et al. have developed a method for generating pseudo-infinite virtual cities on-the-fly. Randomly generated regular polygons are merged into floor plans, extruded to parameterized buildings and textured, resulting in high LoD office districts. The model’s building geometry is generated dynamically as needed inside the user’s cone of sight while the user is interactively exploring the city. This view frustum filling (see Figure 2.6) provides for the generation of pseudo-infinite cityscapes in real-time that would take a lifetime to explore.

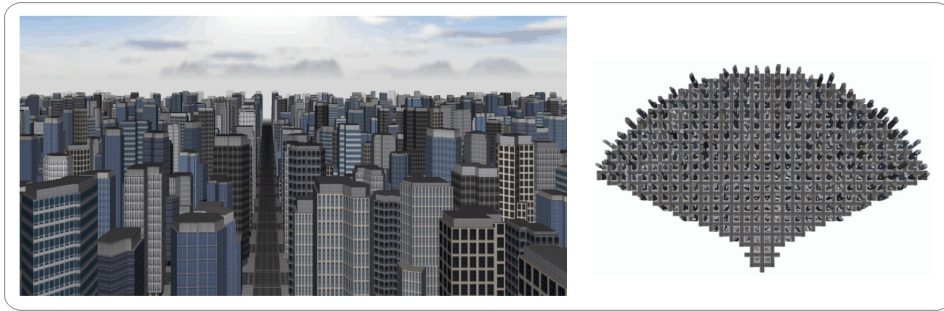


Figure 2.6.: Real-time generation of 'pseudo-infinite' city [GPSL03a]

2.5. Classification and Comparison of Procedural Modeling Techniques

Previous sections reviewed state-of-the-art techniques in procedural modeling, including grammar-based, agent-based, statistical, and dynamic models. The presented approaches mainly differ in the following characteristics (see Figure 2.7 for comparison):

- scalability/LoD (large-scale models vs. architectural models)
- realness (existent cities vs. fictitious cities)
- dynamics (static vs. dynamic cities)
- input data (extensive vs. little input data)

These characteristics affect the believability of the procedural model, although authenticity has a slightly different focus in different application areas. Besides visual fidelity, the realistic projection of economic input parameters is a decisive factor for applicability in planning situations. In contrast, gaming industry attaches importance to interactive, dynamic procedural modeling in real-time.

Greuter et al. [GPSL03a] mainly contribute to the applicability of procedural city modeling in computer graphics and visualization applications. They dynamically generate visually interesting and complex buildings in real-time with view frustum filling, still their interactive model is inapplicable for planning purposes. It only supports one building type, office skyscrapers, and the transportation network does not correspond to a realistic city, since streets are uniformly gridded. The procedural model

procedural modeling approaches	main characteristics	N. Kato, T. Okano, A. Okano, H. Kanoh, & S. Nishihara, 1998	Y. I. H. Parish & P. Müller, 2001	P. Wonka, M. Wimmer, F. Sillion & W. Ribarsky, 2003	P. Müller; T. Vereenooghe, A. Ulmer & L. Van Gool, 2005	P. Müller, P. Wonka, S. Haegler, A. Ulmer & L. Van Gool, 2006	T. Lechner, B. Watson, P. Ren, U. Wilensky, S. Tisue & M. Felsen, 2004	C. Yap, H. Biermann, A. Hertzman, C. Li, J. Meyer, H.-K. Pao & S. Faxia, 2002	S. Greuter, J. Parker, N. Stewart & G. Leach, 2003
large-scale urban model	architectural model	□	□	■	■	■	■	■	■
reconstruction of existing city	construction of fictional city	□	□	■	■	■	■	■	■
dynamic model	static model	□	□	■	■	■	■	■	■
extensive input data	little or no input data	□	□	■	■	■	■	■	■

Figure 2.7.: Predominant characteristics of procedural modeling techniques

of Yap et al. [YBH⁺02] overcomes this drawback by including transportation ground truth, the TIGER data set. On that account, the implemented method only works for existing cities, not for future scenarios or fictitious cities. Apart from this, the “Different Manhattan Project” does not take into account the economic conditions of the modeled cityscape as well as the dynamics of residents’ activities. In considering these aspects, the agent-based approach by Lechner et al. [LWR⁺04] simulates the development of different land use zones. Their model is suitable for planning applications where no socio-economic data is required. The believability of the agent-based simulation rests on more realistic urban layout, but its application is limited due to an authentic road network with street patterns and visual resemblance to real cities.

While Lechner et al. focus on land usage and building distribution, most grammar-based approaches aim at colonizing road networks. Kato et al. [KOO⁺98] generate road maps with two different street patterns, whereas

Parish et al. [PM01] enhance L-systems and implement production rules that are easily extendible to several different street patterns. Their method is applicable for both reproducing existing cities and creating fictitious or future cities. Nevertheless, the L-system production rules do not allow a proper representation of the buildings' functionality. This shortcoming is somewhat overcome by INSTANT ARCHITECTURE [WWSR03], offering a variety of different architectural styles and designs for individual buildings. The complex geometric representation is at the expense of scalability since the approach only focuses on architecture and disregards urban layout and streets. Procedural modeling techniques combining large-scale models with geometrically detailed architecture are introduced by Müller et al. In [MVUG05] grammar-based techniques are combined to recreate ancient Roman cities, and in [MWH⁺06] CGA SHAPES generate extensive urban models with up to a billion polygons. Both approaches offer high scalability as well as strong visual fidelity and are suited for the generation of ancient, existing, and future cityscapes. Yet, to visualize potential future impacts of planning decisions in urban planning applications, additional input data is needed. The models have to incorporate correctly projected demographic and economic data.

2.6. Discussion of Application Areas

As a result of ever-growing hardware performance and increasing user demands, automatic 3D city modeling has become more and more important for a number of application areas. Procedural content generation is a promising technique in domains where the visual believability of a model is more important than its socio-statistical validity.

Procedural models are useful for visualization and computer graphics applications, e.g., 3D computer games. Entertainment industry also benefits from 3D virtual cityscapes in animated movie productions. In general, procedural models are suitable for applications where the visualization of look-alike cities is adequate and the empirical validity of the environmental model is less relevant. To date, procedural approaches are not suitable for the visualization of urban simulation data since they lack the capability to model the impacts of human behavior on urban developments.

To overcome issues of empirical validity, we suggest combining agent-based simulation tools like *UrbanSim* with grammar-based approaches like CGA

SHAPES. Linking the parameters of a procedural model with data from urban simulations will result in an integrated tool for the visualization of existing and developing cities.

However, the usability of those models for the analyses of future urban developments remains questionable. Procedural models digitally create a highly complex visual reality that might impede visual data mining. A realistic representation helps analysts to perceive urban environments in a more tangible way, but at the same time it distracts from the actual task of analyzing and comprehending the underlying complex data structures. The real world is too complex to assimilate at once, we need abstraction to help us interpret it [SBJ⁺01]. Realism can be distracting depending on the specific nature of visualization and cognition is often enhanced through the use of abstract symbols. As MacEachren points out in [Mac01], we have to understand the relative advantages of realism and abstraction for different visualization tasks, users, and kinds of information representation.

For the visualization of multidimensional urban simulation data, we propose an integration of abstract and realistic 3D visualization. To effectively enhance visual thinking, multidimensional data has to be visualized in a way that the analyst can easily detect hidden patterns and relationships. In Chapter 5, we will present an approach that makes uses of abstract visualizations of geospatial data in a realistic 3D context. We will encode multidimensional urban simulation data as specified graphic variables of different geometries and map those geometries to a *Google EarthTM* environment. In the following chapter, we introduce key applications and components for the proposed geovisualization framework.

3. Key Framework Applications and Components

Basically, our framework works with any grid cell based urban simulation model. We decided to integrate *UrbanSim* as simulation tool into the framework, since it implements a very disaggregated microsimulation approach at a fine spatial scale. *UrbanSim* takes individual households and jobs into account at a fine spatial resolution (usually $150m \times 150m$). Furthermore, the microsimulation runs on an annual basis. To date, no other simulation model performs at this level of detail in time, space, and in the range of agents whose behaviors are modeled [WBN⁺03]. Section 3.1 will provide a brief introduction to the *UrbanSim* model.

The inherent geospatial nature of *UrbanSim* grid cell based data requires the implementation of a spatial database. So-called geodatabases extend the database concept to storage, query, and editing of georeferenced objects. Accordingly, the projected demographic data from *UrbanSim* and the estimated building type data is stored in a geodatabase implemented in *PostgreSQL* (see Section 3.2).

Finally, we chose the geobrowser *Google EarthTM* (see Section 3.3) to visualize our *UrbanSim* simulation output. *Google EarthTM* offers an exploratory interface to visually synthesize and display information from multiple data sources such as high resolution aerial photos, road networks, and place names. Since it is possible to superimpose georeferenced 3D buildings and customized cartographic symbols on the aerial photos, *Google EarthTM* is an attractive environment for geovisualization.

3.1. Data Modeling with *UrbanSim*

UrbanSim [Wad02, BW04] is a spatially disaggregated land use and transportation simulation software package for modeling the possible long-term effects of different policies on urban developments. More precisely, it simulates the interactions among the different decision-making actors such as households, employers, developers, and policy-makers to determine their collective impact on future transportation and land use.

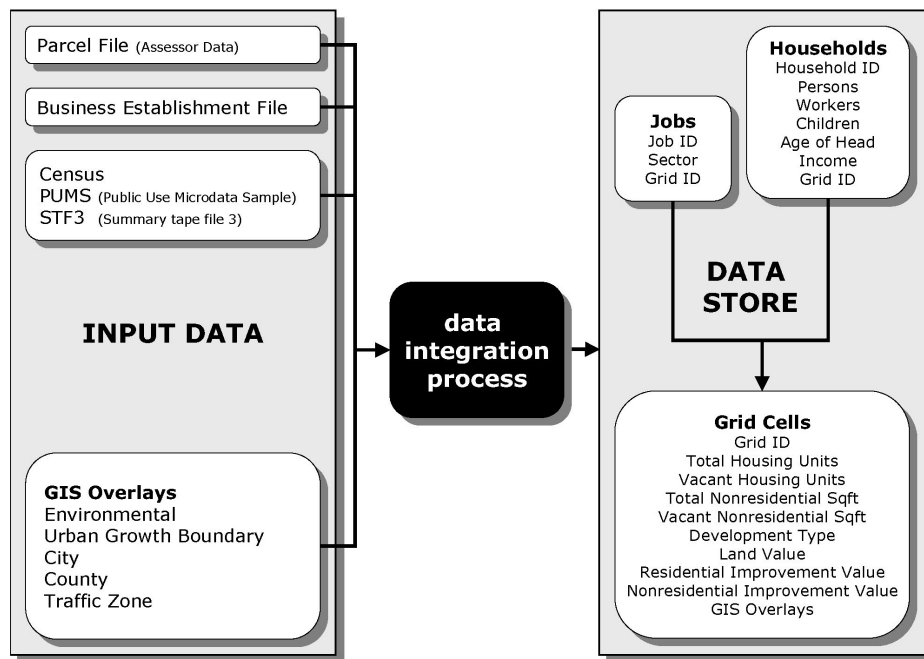


Figure 3.1.: Data integration process (c.f. [Wad02])

UrbanSim consists of several model components that simulate different actors in the urban development process, including discrete choice models for household moving and residential location, for business choice of employment location, and for developer choices of locations and types of real estate development. An overview of the *UrbanSim* evolution as well as a detailed description of implemented model components is given in [BWF06] and [WBN⁺03] respectively. The open source simulation model was developed by a research group in Washington and is currently integrated in the Digital Phoenix project [GHK⁺07] at ASU, Arizona State University

to predict population and job distributions in the Phoenix Metropolitan area.

UrbanSim input data is aggregated from various sources (see Figure 3.1) and spatially mapped to a georeferenced grid cells file. Base year data includes information on parcels from the Assessor’s office, employment data, and Census data. Additional input data on city, county, and urban growth boundaries as well as environmental and traffic information is overlaid in ArcGIS. The *UrbanSim* data store contains a grid cells table, a jobs table holding information on each job and its employment sector in the grid cells, and a household table. The latter is synthesized probabilistically and compiles demographic characteristics for each household in the metropolitan area.

Subsequent to the data integration process, the *UrbanSim* simulation is run for a predefined number of years. The output projection tables include data on future households with grid cell location and demographic characteristics, future jobs, and exogenous input data. The projection results can be integrated into a variety of analyses, e.g., the analysis of future population density, material use, or carbon footprints.

3.2. Data Management with *PostgreSQL* and *PostGIS*

PostgreSQL [Pos07] is an open source object-relational database management system (*ORDBMS*) that allows for managing data. In contrast to relational database management systems, an *ORDBMS* is not limited to a pre-defined set of data types, which raises the level of abstraction. *PostgreSQL* natively supports i.a. arbitrary precision numerics, unlimited length text, arrays, and geometric primitives (points, lines, and polygons). In addition, it provides support for the integration of custom data types and methods in the database.

As geospatial extension to the *PostgreSQL* backend server, we use the *PostGIS* module [Ref05]. *PostGIS* is an open source add-on developed by Refrations Research under the GNU General Public License and enables *PostgreSQL* to integrate spatial data structures into the database. *PostGIS* follows the Simple Features for SQL specification (*SFS*) from the Open Geospatial Consortium (*OGC*). The *OGC* is an international

non-profit organization that is leading the development of standards for geospatial and location based services [Ope07]. As a Simple Features for SQL compliant spatial database, *PostGIS* includes the geometry types diagrammed in Figure 3.2 and provides functionality for spatially enabled SQL queries like distance between geographic objects, unions, calculation of perimeter, and buffering.

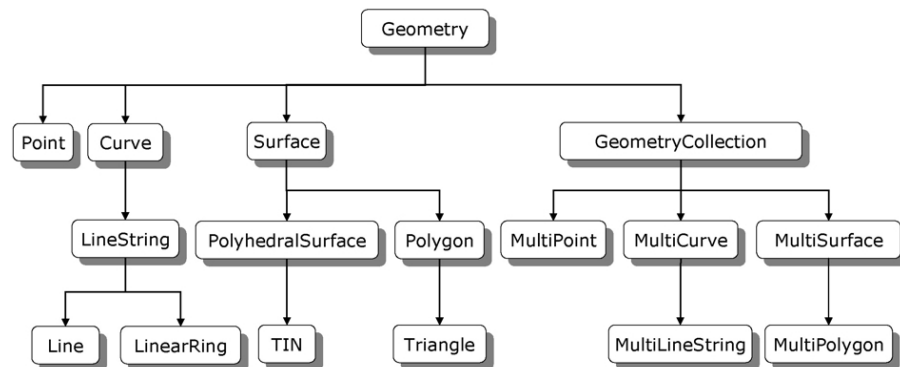


Figure 3.2.: OGC Simple Feature Specification (cf. [Ope07])

3.3. Data Visualization with *Google EarthTM*

Created by Keyhole Inc. and originally named Earth Viewer, the geobrowser *Google EarthTM* [Goo07b] maps the earth by superimposing satellite images and aerial photos on a virtual globe. *Google EarthTM* allows the user to interactively browse the globe in a 3D view and to zoom from space into street level views. Data is streamed from Google's Server upon request to the client computer. To date, most parts of the earth's surface are covered with images having a minimum resolution of 15 meters and a maximum age of 3 years.

The aerial photographs are mapped onto a digital elevation model (*DEM*) provided by *NASA*. *Google EarthTM* uses a map projection called General Perspective to show the earth as it appears from space. This cartographic projection resembles an orthographic projection, but instead of an infinite point of perspective it has a finite point of perspective near the globe.

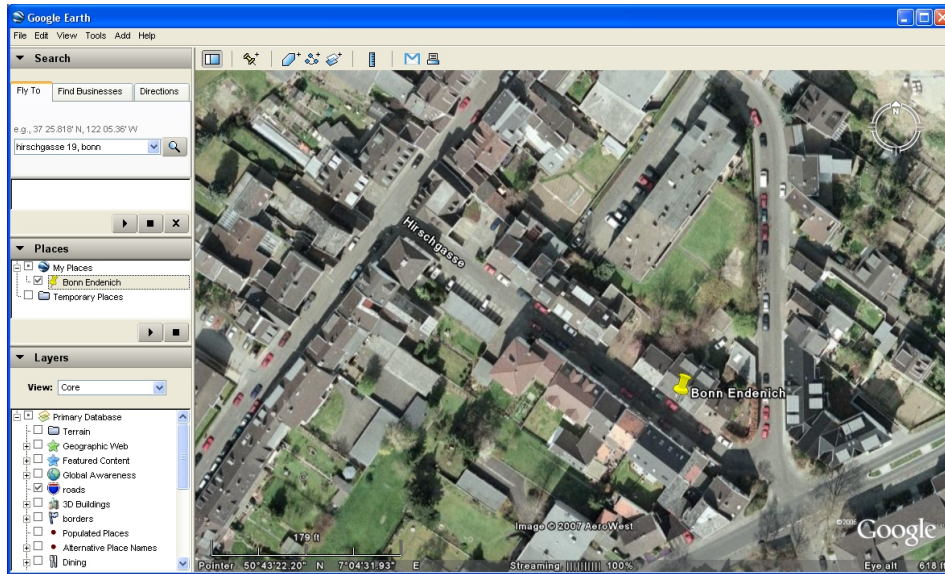


Figure 3.3.: A Placemark in *Google Earth*TM

*Google Earth*TM has the capability to combine the multi-resolution raster image representation of the planet with any kind of georeferenced data. The geobrowser allows overlaying a wide set of geographic features like streets, freeways, and county borders. The user can filter those data sets by space, time, and layer. Furthermore, *Google Earth*TM supports an overlay of points of interests (Placemarks, see Figure 3.3) as well as 3D buildings and structures onto the surface. Layers containing geospatial data are managed through the so-called Keyhole Markup Language *KML* (see Section 5.2.5).

All in all, *Google Earth*TM is a free, easy-to-use, interactive application to visually explore high-resolution aerial photographs, terrain, and superimposed geographic data in an integrated manner. This visualization approach highlights spatial relationships between real-world locations and overlaid geodata.

In Chapter 5, we will present how more abstract data and demographic variables can be visualized on top of *Google Earth*TM to support the analysis of urban simulation data. The following chapter introduces a statistical method to derive building types from demographic data, which then provides the three-dimensional attributes for visualization.

4. Estimation of Residential Building Types

As stated in Section 1.3, future building types and their spatial distribution is derived from *UrbanSim* output data. The predicted building type information is important for the analysis of future material use, carbon footprint calculations, and for visualizing simultaneously the results of land usage, density, and other significant parameters in 3D space.

The mapping between dwelling types and *UrbanSim* household characteristics is realized with regression analysis. Regression maps the values from a predictor variable in such a way that the prediction error is minimized. The analysis involves more than one predictor, since our input data (see Section 4.2) is given as multi-dimensional vectors with one dimension for every demographic variable. In order to investigate the effect of more than one independent variable on the discrete outcome, we use multinomial logistic regression (*MNL*). This kind of regression can handle several predictors as input variables as well as polytomous response variables with more than two output categories.

First, we estimate building type distributions based on demographic household information at a census block group scale (Section 4.3). For this purpose, we form clusters with typical building type distributions by *k*-means to establish nominal categories for the regression model (Section 4.3.1). Then, the log odds of the clustered neighborhood category predictors are modeled as a linear function of the categories' covariates in the estimation process (Section 4.3.2). This approach gives good results, but implies a possible correlation of the clustered building type categories. Therefore, we refine our model in Section 4.4 and estimate building types at spatial scale of grid cells used in *UrbanSim* projections. Estimating coefficients for a set of $150m \times 150m$ grid cells with demographic data synthesized from Census (Section 4.4.2) gives slightly better results than using synthesized *UrbanSim* household data obtained from the households table (Section

4.4.1). Finally, the estimated coefficients are applied to *UrbanSim* simulation output (Section 4.5). Prediction results are discussed in the last section of this chapter.

4.1. Multinomial Logistic Regression (MNL)

Multinomial logit analysis has a wide range of applications in behavioral assessment and categorical data analysis. *MNL* is widely-used in social sciences and has a long tradition in the economics of consumer choice. Multinomial logit refers to the conditional discrete choice model, first introduced and most notably influenced by McFadden [McF73, McF76, McF97]. Since then, a lot of research has been conducted in econometrics to develop multinomial logit models of residential location choice. John Quigley investigated consumers' qualitative choice behavior of residential location and building type in Pittsburgh, using a nested *MNL* [Qui76], a generalization of McFadden's conditional logit model. Weisbrod assessed household location choice based on tradeoffs between accessibility and other housing and location characteristics [WBAL80]. Multinomial logit models were more recently applied to estimate the relationship between the overall level of housing prices and the mix of building types [Ska99]. Furthermore, discrete choice analysis was successfully adopted in transportation planning to examine travel demand [BAL85].

The process of predicting values of multilevel responses with unordered qualitative categorical outcomes by means of *MNL* is described below (compare [PX99]).

Consider the polytomous outcomes y_i for a dependent variable with J categories where i denotes the i th respondent. Let the response probability P_{ij} represent the probability of a particular outcome, thus the chances that the i th respondent falls into category j . Consider that \mathbf{x}_i is the vector of predictors storing values for the independent variables and $\boldsymbol{\beta}$ the regression parameter vector. Then, the linear relationship

$$z_{ij} = \mathbf{x}_i' \boldsymbol{\beta}_j = \sum_{k=0}^K \beta_{jk} x_{ik} = \alpha_j + \sum_{k=1}^K \beta_{jk} x_{ik} \quad (4.1)$$

renders the parameters of the model estimable, and we can calculate the set of coefficients β_{jk} which correspond to the covariates x_{ik} of the response

probabilities. For the outcome y_i , the probability of selecting j is:

$$P(y_i = j | \mathbf{x}_i) = P_{ij} = \frac{e^{z_{ij}}}{\sum_{j=1}^J e^{z_{ij}}} = \frac{e^{\mathbf{x}'_i \boldsymbol{\beta}_j}}{\sum_{j=1}^J e^{\mathbf{x}'_i \boldsymbol{\beta}_j}} = \frac{e^{\mathbf{x}'_i \boldsymbol{\beta}_j}}{1 + \sum_{j=2}^J e^{\mathbf{x}'_i \boldsymbol{\beta}_j}} \quad (4.2)$$

This probability statement has the constraint that all probabilities must sum up to 1:

$$\sum_{j=1}^J P_{ij} = 1 \quad (4.3)$$

The multinomial logistic regression is a generalized discrete choice model for nominal response variables [Agr02]. Discrete choice multinomial logit models are widely used in economics and differ from the standard model in a way that the explanatory variables vary not only by outcome but also by individual. It is assumed that the individual has preferences over a set of alternatives, e.g., travel modes, and chooses the alternative which maximizes utility:

$$y_i = j \text{ if } u_{ij} \geq \max(u_{i1}, \dots, u_{iJ}) \quad \mathbf{u}_{ij} = \mathbf{r}'_{ij} \boldsymbol{\beta} + \epsilon_{ij} \quad (4.4)$$

$\mathbf{r}'_{ij} \boldsymbol{\beta}$ is the systematic component of alternative j that considers the characteristics of the choice as well as the preferences of the individual and ϵ is a stochastic term.

The multinomial logistic model simultaneously describes log odds for all $\binom{J}{2}$ category pairs. The log-odds of membership in one category of the dependents versus an arbitrary baseline category, normally the first category, are fitted as a linear function of covariates \mathbf{x}_i :

$$\log \left(\frac{P_{ij}}{P_{i1}} \right) = \mathbf{x}'_i \boldsymbol{\beta}_j \quad (4.5)$$

The odds between two arbitrary categories j and j' are calculated as follows:

$$\frac{P_{ij}}{P_{ij'}} = e^{\mathbf{x}'_i (\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})} \quad (4.6)$$

Regression parameters are estimated using maximum likelihood. This is a stable calculation, since the log-likelihood of the probabilities is convex.

4.2. Mapping the Real World

In this section, we present how a mapping between building types and demographics is accomplished and introduce the input data for our multinomial logistic regression model.

Let D denote a convex set of demographic household characteristics \mathbf{d} , B denote a convex set of building types \mathbf{b} and C be a set of contexts \mathbf{c} . We assume that demographic household characteristics within a predefined context have an impact on the residents' choice of building types. Then, the mathematical mapping function Φ from the input sets D and C to the output set B can be derived by discrete choice modeling from the relationship $(D, C) \xrightarrow{\Phi} B$ with given input and output sets (see Figure 4.1).

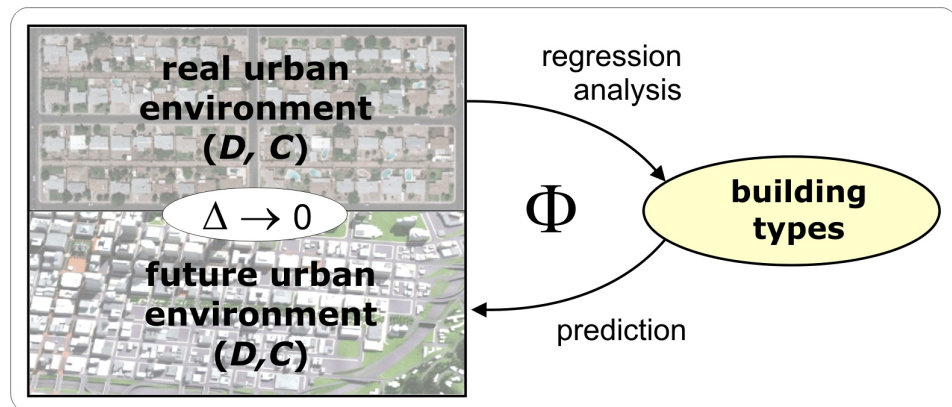


Figure 4.1.: Mapping the real world

In the remainder of this work, we will focus on data from Maricopa County to illustrate our approach. Maricopa is located in the center of the U.S. state of Arizona (see Figure 4.2) with county seat in Arizona's largest city and capital Phoenix. As of 2000, Maricopa County had 3,072,149 residents in 1,132,886 households [US 07].

For the simulation in *UrbanSim* and the final estimation of building types, we superimpose a regular, georeferenced grid on Maricopa County with a resolution of $150m \times 150m$. These grid cells serve as reference units for the prediction of future demographics and for the multinomial logistic regression.

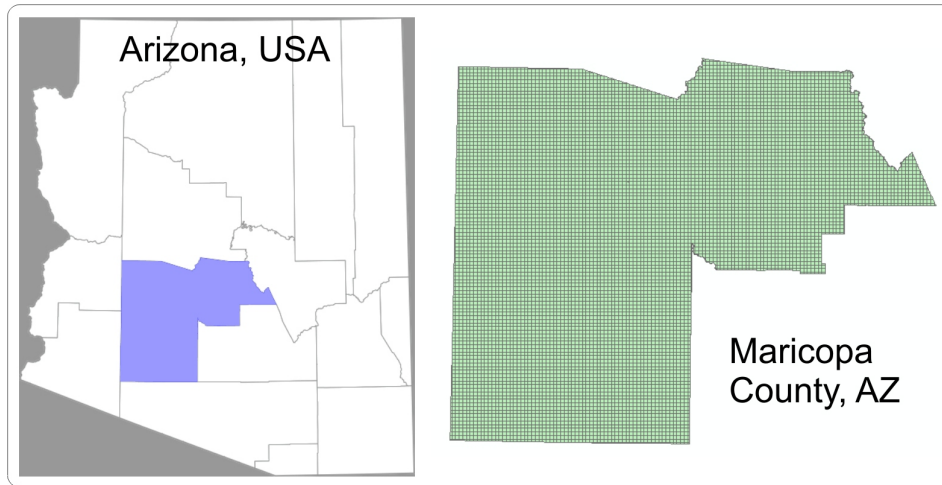


Figure 4.2.: Maps of Arizona and Maricopa County

Demographics

In our model, the feature vectors \mathbf{d} store aggregated household information on demographic backgrounds. The following measures of household characteristics are included into \mathbf{d} for the subsequent regression analysis:

- population density
- median age
- average household size
- median household income
- average number of children per household
- average number of cars per household
- percentage of Hispanics and minorities
- distance to nearest highway
- average age of structures built

Context

The context vectors \mathbf{c} are discretely time dependent and contain the social, legal, and political framework associated with the spatial location of the modeled environment. The context can be used to calibrate the model and to generate different scenarios. In the remainder of this analysis, C will be treated as exogenous and not part of the model. For the regression analysis, C is predetermined to match the current context of Maricopa County to simplify the model.

Building Types

The building type vector \mathbf{b} is a container for information on the physical characteristics of dwellings. In the current model, \mathbf{b} is restricted to residential buildings. We derive the building type vectors from the Maricopa County Assessors data for 2000, which is an extensive database providing detailed property information at parcel level.

We use *ArcGIS* to intersect the Assessors file with the Census data, to prepare and store the output table as a shape file. According to the Primary Use Code, we define single family dwellings and apartments as the two main residential building type categories. The single family category is subclassified into lots with different sizes, the apartments are subdivided according to the number of housing units. Altogether, we obtain nine different building type categories:

Apartments	Single family dwellings
2 – 24 units	lot size < 6,742 sqrft (XXS)
25 – 99 units	6,742 < lot size < 7,986 sqrft (X)
> 100 units	7,986 < lot size < 9,801 sqrft (S)
	9,801 < lot size < 12,705 sqrft (M)
	12,705 < lot size < 18,150 sqrft (L)
	lot size > 18,150 sqrft (XL)

Table 4.1.: Building types

After establishing the building type categories, we obtain a frequency count of each building type for each reference unit (block groups in Section 4.3 and grid cells in Section 4.4) and transform the count data into percentages. The result is a building type composition where all building

types add up to 100 percent in each reference unit. Figures 4.3 to 4.5 show *Google EarthTM* aerial photographs of representative building types in the Phoenix metropolitan area.



Figure 4.3.: Single family dwellings, very small lots (a) and small lots (b)



Figure 4.4.: Single family dwellings, medium lots (a) and large lots (b)



Figure 4.5.: Apartments, 5 units (a), 25 – 100 units (b), > 100 units (c)

4.3. Block Group Based Estimation

Before disaggregating the demographic information to the spatial scale of grid cells in Section 4.4.2, we will run the regression analysis with demographic data at census block group level. A block group is an aggregation of census blocks and generally contains between 600 and 3,000 people. Due to the relatively coarse spatial resolution, most block group assembles a mix of different building types rather than being a homogeneous composition of dwellings (see Figure 4.6). Consequently, we have to find similar building type distributions and classify them into different typical neighborhood categories. To build these categories, the percentages of building types in each block group are calculated from frequency count data and then grouped together into classes of similar neighborhoods by means of clustering. Clustering is an unsupervised learning technique used in statistical data analysis to determine the inherent grouping in a collection of unclassified data. For an overview of classical data mining techniques and a detailed description of different clustering methods see [TSK05].

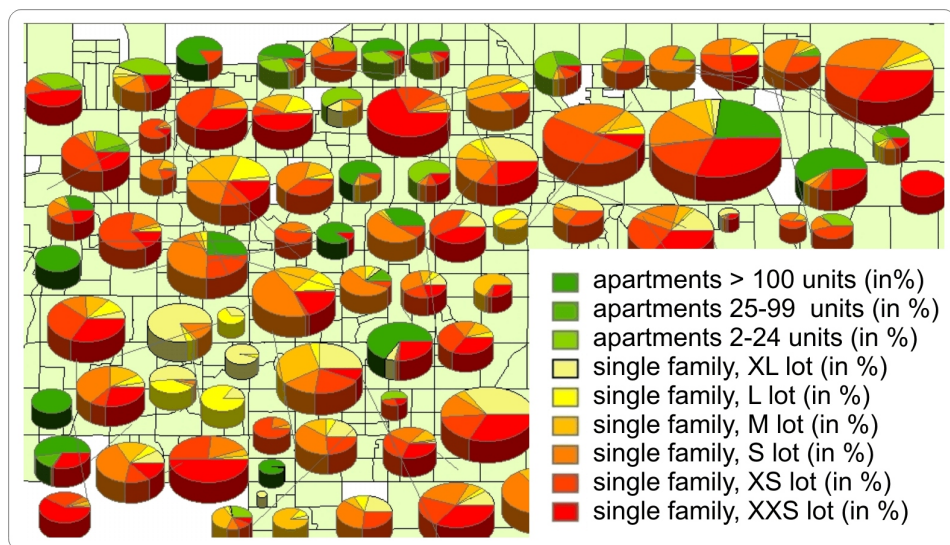


Figure 4.6.: Percentages of different building types in each block group

4.3.1. Clustering

To find significant neighborhood patterns in the building type data, we use a standard k -means algorithm. K -means iteratively partitions m data points x_1, x_2, \dots, x_m into a user-specified number of k clusters c_1, c_2, \dots, c_k , assigning all data points to their closest cluster centroid. In our case, the data to be clustered is a set of real-valued vectors in n -dimensional feature space where each building type adds a new dimension. Distances between those vectors are defined by the Euclidean metric.

The a priori choice of the optimal number of clusters k is essential for reasonable clustering results. Forming homogeneous clusters is especially difficult for high-cardinality data. Choosing a large k will reduce the impact of noise on the categorization, but will also result in fuzzier cluster boundaries. The challenge is to create homogeneous clusters and at the same time minimize k . Table 4.2 shows the clustering result of the

	Cluster					
	1	2	3	4	5	6
apartments 2-24 units (in %)	3.6	39.2	2.0	2.2	3.4	2.8
apartments 25-99 units (in %)	5.2	9.5	0.1	0.9	1.0	1.2
apartments >100 units (in %)	69.9	4.4	0.0	3.4	4.0	2.7
single family, XXS lots (in %)	5.8	11.5	2.8	8.4	52.2	5.5
single family, XS lots (in %)	5.9	15.7	2.1	30.6	23.3	5.3
single family, S lots (in %)	4.1	8.6	2.1	38.6	10.0	16.1
single family, M lots (in %)	1.8	4.2	2.8	10.8	4.3	36.4
single family, L lots (in %)	1.0	2.5	6.4	4.3	1.8	19.3
single family, XL lots (in %)	1.1	2.6	83.0	1.9	1.6	12.1

Table 4.2.: Final cluster centers

k -means algorithm after 14 iterations, presenting the centroids for every cluster. The optimal number of clusters $k = 6$ was determined empirically. From the most representative cluster prototypes we see that apartments with over 100 units and single family dwellings with very large lots clearly emerge as the prominent building types in clusters number 1 and 3. The typical building type distributions of the other clusters are more heterogeneous with variable dwelling types.

After finding the cluster centroids, each block group in the data set is assigned to the nearest cluster centroid based upon the dwelling type

distribution in the block group. The assigned building type cluster is then estimated from demographic variables using *MNL* in the following section.

4.3.2. Empirical Analysis and Statistical Tests

The building type distributions derived in the previous section are used as dependent variables x_i in our multinomial regression model. The six neighborhood category outcomes provided below are interpreted according to the cluster centers (see table 4.2):

- 1 = mainly apartments
- 2 = balance between apartments and single family, majority of apartments < 25 units
- 3 = mainly single family, XL lots
- 4 = mainly single family, S and XS lots
- 5 = mainly single family, XS and XXS lots
- 6 = mainly single family, majority of M lots

		N	Marginal Percentage
Cluster Number of Case	1	283	13.9%
	2	327	16.0%
	3	117	5.7%
	4	539	26.4%
	5	555	27.2%
	6	219	10.7%
Total		2040	100.0%

Table 4.3.: Case processing summary

Table 4.3 shows the number and percentage of block groups in every cluster that are included in the analysis. The demographic vectors \mathbf{d} of aggregated household information at the block group level, introduced in Section 4.2, are the explanatory variables considered in the estimation process. To predict the outcome categories in the multinomial logit model, we incorporate 2040 block groups of Maricopa County with a total number of 1,115,570 households. The rounded average household size per block group is 3 persons, and the average median income per household is \$48,745. About 17% of the block groups have a majority of Hispanics, whereas 75% of the block groups have predominantly White population. The number of

children per household in every block group averages 0.7 with a rounded number of one car per household. After building the *MNL* model with the described input data, we run the regression and assess the fit of our model to the data with different statistical tests as described below.

Goodness of Fit Test

Pearson and Deviance are the most prevalent goodness-of-fit statistics used to validate the model in multinomial logistic regression. They test the null hypothesis that the model adequately fits the data. In our logit model, we have several predictors with continuous values as covariates causing many subpopulations with zero frequencies. Because of the many cells with expected zero values, the test statistics lack large sample properties and a dependable goodness-of-fit test is not provided.

Pseudo R^2

In multinomial logistic regression, a direct analog to the R-Squared statistic as used in ordinary least-squared regression does not exist. R^2 cannot be computed, since R^2 measures the variability in the dependents, but the variance of a categorical predictor is a function of the variable's frequency distribution. To summarize the strength of the association between the dependent and independent variables, *MNL* uses pseudo R^2 statistics which are designed to have similar characteristics to the R^2 statistic. Regression output table 4.4 shows three pseudo R^2 estimates: Cox and Snell R^2 , Nagelkerke's R^2 , and McFadden's R^2 . Larger values between 0 and 1 indicate a better explanation of the variation by the model, which means that our model performs reasonably well.

Cox and Snell	0.736
Nagelkerke	0.763
McFadden	0.398

Table 4.4.: Pseudo R^2

Model	-2 Log Likelihood	χ^2	DoF	Significance
Intercept Only	6841.390			
Final	4121.008	2720.382	55	0.000

Table 4.5.: Model fitting information

Likelihood-Ratio Tests

Before examining individual coefficients, we will assess the significance of the *MNL* by showing that the model fits the data better than a null model. The overall test of the null hypotheses that the regression coefficients β for all of the variables \mathbf{x} in the model are 0 is called likelihood ratio test. Likelihood denotes the probability that the estimated values of the dependent may be predicted from the independents. The likelihood ratio is a function of log likelihood and makes a statement about the significance of the unexplained variance in the outcome. As shown in Table 4.5, the chi-square distributed difference between the -2 log likelihood (-2LL) values for the null hypothesis and the final model has an observed significance level of 0.000 (rounded). Therefore, we can reject the null hypothesis that the *MNL* model without predictors performs as well as the model with predictors. The results provide strong support for the overall model significance, but the previous test does not assure that every predictor variable is significant for the prediction. In order to test individual model parameters, we have to check the contribution of each predictor to the model separately. Therefore, we create a reduced model by omitting one independent variable at a time and use a likelihood ratio test to analyze the differences in -2LL between the overall model and the nested model. This process tests whether the coefficient for the omitted effect can be treated as zero if the effect does not have an influence on the regression. Table 4.6 shows that the test clearly rejects the null hypothesis for almost all independents included in the analysis. Yet, the average number of cars and children per household do not contribute to the model at a very high level of significance (level of significance > 0.05). For this reason, the variables will be excluded from the *MNL* model.

Classification

An indicator for how well the multinomial logistic regression model predicts categories of the polytomous dependent variable is the so-called clas-

Effects	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood Reduced Model	χ^2	DoF	Sig.
Intercept	4464.185	343.178	5	0.000
Population Density	4368.175	247.168	5	0.000
Median Age	4216.168	95.160	5	0.000
Median Household Income	4257.356	136.348	5	0.000
Av. #Children p. Household	4128.715	7.707	5	0.173
Av. #Cars p. Household	4128.971	7.963	5	0.158
Minorities in %	4150.439	29.432	5	0.000
Hispanics in %	4139.113	18.106	5	0.003
City	4132.266	11.259	5	0.046
Av. Household Size	4242.518	121.510	5	0.000
Distance to Highway	4156.048	35.041	5	0.000
Av. Age of Structure Built	4432.104	311.096	5	0.000

Table 4.6.: Likelihood ratio statistics, reduced model

sification table. Table 4.7 shows the classification table for our *MNL* model. This $J \times J$ table crosstabulates observed categories with predicted categories and helps measuring correct and incorrect estimates. The diagonals contain correct predictions, whereas cells off the diagonal are falsely predicted. For example, 198 of the 283 block groups observed to be in neighborhood category 1 (mainly apartments) were classified correctly. The

Observed	Predicted						Percent Correct
	1	2	3	4	5	6	
1	198	43	1	12	27	2	70.0%
2	39	216	4	23	43	2	66.0%
3	1	3	86	6	8	13	73.5%
4	15	32	7	331	129	25	61.4%
5	27	58	8	151	301	10	54.2%
6	2	7	16	84	17	93	42.5%
Overall Percentage	13.8%	17.6%	6.0%	29.8%	25.7%	7.1%	60.0%

Table 4.7.: Classification table

table shows good results in terms of correct predictions. In most of the cases, the percentage of correctly predicted categories exceeds 60 percent.

The null model which classifies all cases according to the modal category classifies correctly for only 13.9% of the cases (compare 4.7). Furthermore, the classification table proves that our *MNL* has homoscedasticity, since the percentage of correct predictions is approximately the same for every row.

4.3.3. Results

The model parameters of our multinomial logistic regression are compiled in the Appendix (see A.1). Exemplarily, the parameter estimates for neighborhood category 5 are summarized in table 4.8.

For each cluster except the reference category, we get estimated logit coefficients β associated with each predictor as well as a value for the intercept. The $(J - 1)$ logits β can be used in prediction equations to generate logistic scores and thus are the key to predicting future building types. The

	β	Std. Error	DoF	Sig.	Exp(β)
Intercept	-16.6966	1.3311	1	0.0000	
Population	-0.3195	0.0725	1	0.0000	0.7265
Median Age	0.1195	0.0150	1	0.0000	1.1269
Median Income	0.0001	0.0000	1	0.0000	1.0001
Minorities in %	-0.0387	0.0157	1	0.0137	0.9620
Hispanics in %	0.0004	0.0179	1	0.9809	1.0004
Av. Household Size	4.8590	0.3622	1	0.0000	128.8943
Dist. to Highway	0.0001	0.0001	1	0.3375	1.0001
Age of Head	0.0251	0.0087	1	0.0037	1.0254

Table 4.8.: Parameter estimates for cluster 5 (single family, XS & XXS lots)

algebraic sign of the coefficients determines the effect of each predictor on the model. Positive parameter estimates like the average household size increase the likelihood of the response category with respect to reference category 1. Responses with significant negative coefficients as the percentage of minorities reduce the likelihood of that category. In general, the effect of the predictors is strongest for category 3 versus 1 and weakest for category 2 versus the reference category. The table also shows the standard error of the coefficients and the odds ratio, labeled as $\text{Exp}(\beta)$. Looking at the significance of the explanatory variables, average household size, median age, average age of structure built, and population density

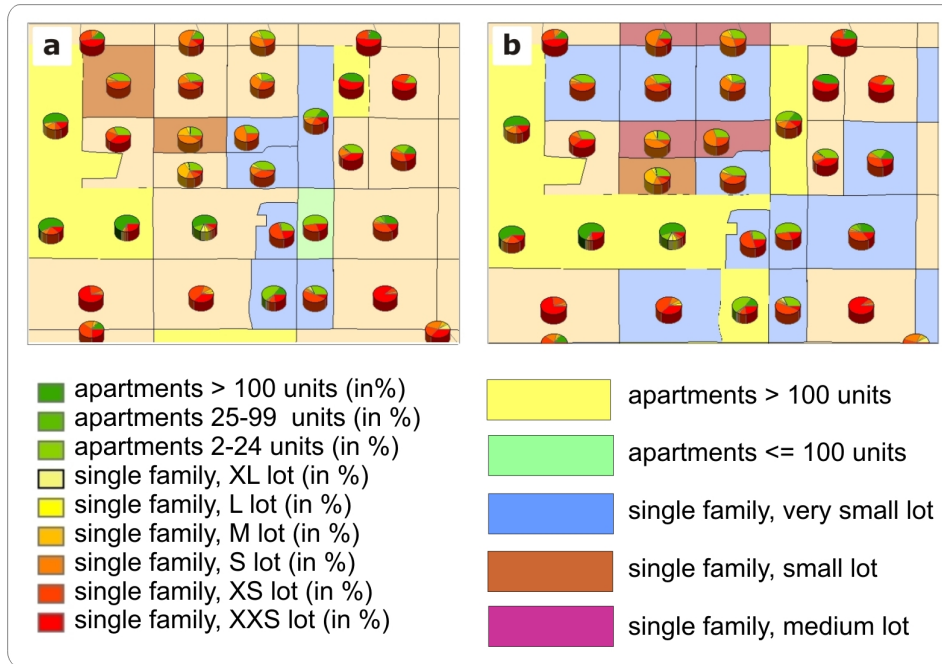


Figure 4.7.: Predicted vs. observed neighborhood categories

are highly significant for predicting building types. The significance level of the variable depicting the percentage of minorities varies but shows an overall significance whereas the percentage of Hispanics per block group does not seem to be important for this prediction. Finally, the distance to the nearest Highway is not significant for three of the categories, but is very important for predicting categories 3 and 6. Figure 4.7 shows a comparative visualization of the predicted and observed categories for sample block groups in Maricopa County. The prediction is based on the logit coefficients calculated earlier. Additionally, a difference picture of correctly and incorrectly predicted categories can be seen in Figure 4.8. Here, block groups with matching response and observed category are colored blue, whereas neighborhoods with mismatching categories appear in red. Both visualizations affirm the prediction rate of about 60 percent as stated in the classification table (see Section 4.7).

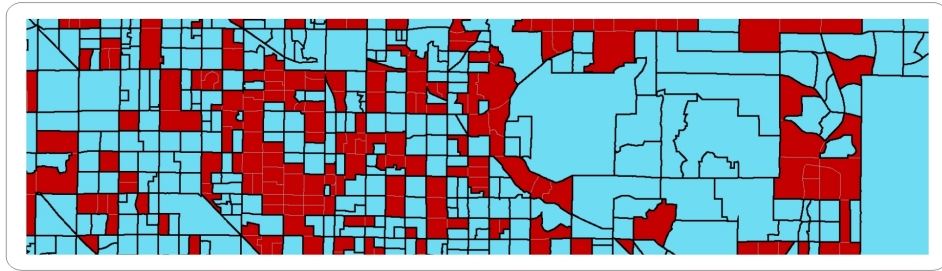


Figure 4.8.: Difference picture of correct vs. incorrect predictions

4.4. Grid Cell Based Estimation

All tests in the previous section provided strong support that our model fits the data reasonably well and that logit regression is a coherent framework for assessing the relationship between demographic characteristics and the type of residential neighborhoods people live in. So far, we predicted different neighborhood categories rather than distinct dwelling types due to the coarse block group resolution available from Census. A major drawback of this approach is that a correlation of the response variables cannot be excluded. A correlation violates the Independence of Irrelevant Alternatives (*IIA*) assumption within the *MNL* framework. Multinomial logistic regression assumes that the odds of predicting one outcome over any other outcome are not dependent on the number or characteristics of the other responses. If the *IIA* assumption does not hold, the logistic regression will yield biased estimates because of correlated error terms.

Our clustering produces typical building type distribution categories that are as distinct as possible. Nevertheless, to fully satisfy independence across the model's response categories, we will reduce the reference unit size from block groups to grid cells. This will allow us to estimate independent homogeneous building types on a grid cells level instead of potentially correlated building type distributions on a block group level.

4.4.1. Estimation using Synthesized Data from *UrbanSim*

For the second *MNL* model we use synthesized grid cell based household data from *UrbanSim* to reduce the heterogeneity of building types within

our defined spatial units. Since demographic data at the household level are generally not available, *UrbanSim* uses a Household Synthesis Utility to synthesize demographic data available in the U.S. Census tables and the Public Use Microdata Samples (*PUMS*) from Census. The utility generates a synthetic profile of each individual household at the census block level. The households are allocated to census blocks so that marginal totals are preserved and the demographic profile of each census block is reflected by the household characteristics reported for the census block. These households are then intersected with the grid cell boundary file to determine the location of each household by a grid cell identifier. For a detailed description of methods to create baseline synthetic populations of households using Census data we refer to [BBM96].

4.4.1.1. Empirical Analysis and Statistical Tests

In our grid cell based multinomial regression model for Maricopa County, we assign a dominant building type to each grid cell by identifying the dwelling type with most frequency counts per grid cell weighted by residential units. The nine building type categories established in Section 4.3.2 are aggregated to four major building type categories in order to avoid small case numbers (see case processing summary, Figure 4.9). The smaller set of building type categories avoids underrepresented categories with relatively low marginal percentages that are difficult to estimate.

The resulting building types for the grid cell based *MNL* analysis are categorized as follows:

- 1 = single family, small lots (formerly XS and XXS lots)
- 2 = single family, medium lots (formerly S and M lots)
- 3 = single family, large lots (formerly L and XL lots)
- 4 = apartments

Table 4.9 shows the number of grid cells for each building type, summing up to a total of 34,013 grid cells. Again, the explanatory variables for our regression model are the feature vectors $\mathbf{d} \in D$. This time, they contain synthesized demographic data from *UrbanSim*, i.e., population density, income, percentage of minorities and Hispanics, household size, distance to nearest highway as well as number of cars and children. In the following paragraphs, we review the estimation results of the refined regression analysis using statistical tests to evaluate the model.

		N	Marginal Percentage
Cluster Number of Case	1	6,339	18.6%
	2	3,872	11.4%
	3	22,704	66.8%
	4	1,098	3.2%
Total		34,013	100.0%

Table 4.9.: Case processing summary

Pseudo R^2

The pseudo R^2 measures below are lower compared to the measures obtained in the block group based model. This indicates a weaker association between the variables in the grid cell based unit model when *UrbanSim* synthesized data is used. However, it must be noted that pseudo R^2 values tend to be smaller than R^2 (from ordinary least-squared regressions) and values of 0.2 to 0.4 for McFadden's R^2 are considered satisfactory.

Cox and Snell	0.318
Nagelkerke	0.375
McFadden	0.203

Table 4.10.: Pseudo R^2

Likelihood-Ratio Tests

Table 4.11 displays the results of a likelihood ratio test comparing our model to a reduced model with the constant only. The likelihood ratio test yields a test statistic of 51012.743. This statistic is distributed χ^2 with 27 degrees of freedom and allows a rejection of the null hypothesis at a significance of much smaller than $1 \cdot 10^{-4}$. The rejection of this null

Model	-2 Log Likelihood	χ^2	DoF	Significance
Intercept Only	64020.221			
Final	51012.743	13007.478	27	0.000

Table 4.11.: Model fitting information

hypothesis implies that our model fits significantly better than the baseline model. Again, to test individual model parameters on significance, we use

the likelihood ratio test to create a nested reduced model and drop one independent variable at a time. Test results are compiled in table 4.12. In contrast to the block group based estimation, all predictors in the grid cell based estimation are highly significant for the multinomial logistic regression analysis.

Effects	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood Reduced Model	χ^2	DoF	Sig.
Intercept	51389.51	367.77	3	0.00
Population Density	51049.55	36.80	3	0.00
Median Age	51285.49	272.75	3	0.00
Median Household Income	51735.22	722.48	3	0.00
Av. #Children p. Household	51074.41	61.67	3	0.00
Av. #Cars p. Household	51375.07	362.32	3	0.00
Minorities in %	52510.23	1497.48	3	0.00
Hispanics in %	52032.74	1020.00	3	0.00
Av. Household Size	51392.30	379.62	3	0.00
Distance to Highway	53459.21	2446.47	3	0.00

Table 4.12.: Likelihood ratio statistics, reduced model

Classification

The classification table (see table 4.13) shows the correspondence between observed and predicted building types by grid cells. The model shows an overall estimation improvement with 70.7% compared to the overall percentage of 60.0% from the block group based regression analysis in Section 4.3.2. Regardless, building type categories 2 (single family, medium lot) and 4 (apartment) were poorly predicted and the overall percentage of 70.0% is achieved solely by an outstanding 95.1% prediction of dwelling type category 3 (single family, large lot).

4.4.1.2. Results

The unbalanced classification table and the low pseudo R^2 measures suggest that a different approach to aggregating the feature vectors $\mathbf{d} \in D$ is desirable for predicting building types. It seems reasonable to assume

Observed	Predicted				Percent Correct
	1	2	3	4	
1	2253	58	3957	71	35.5%
2	518	99	3229	26	2.6%
3	1034	15	21601	54	95.1%
4	781	3	284	93	8.5%
Overall Percentage	13.3%	0.5%	85.5%	0.7%	70.7%

Table 4.13.: Classification table

that the process of household synthesis does not accurately represent the underlying data at the micro-spatial level (although its estimates may be adequate for the purpose of urban simulation). In the next section, we adopt a different approach that uses the grid cell based geography together with actual census information rather than the synthesized demographic data.

4.4.2. Estimation Using Census Data

In order to assign census data to grid cells, we intersect the Maricopa County block group shape file with the *UrbanSim* grid cell base file and assign census demographics to each household in the corresponding grid cell. Other model input parameters remain unmodified.

4.4.2.1. Empirical Analysis and Statistical Tests

Since building type vectors and the reference unit persist from the previous analysis, the case processing summary with marginal percentages for building type categories equals table 4.9. After re-running the regression, we once again test the results of the multinomial logistic regression model for data fit and significance of coefficients.

Pseudo R^2

In contrast to the pseudo R^2 measures from the previous regression analysis, pseudo R^2 values improve significantly when census demographic data

is used. The pseudo R^2 values in Table 4.14 compare well with the satisfactory measures obtained in the block group based regression model (see Section 4.3.2).

Cox and Snell	0.523
Nagelkerke	0.617
McFadden	0.393

Table 4.14.: Pseudo R^2

Likelihood-Ratio Tests

The test statistic -2LL confirms the significance of the overall logistic regression model at a rounded 0.000 level (see table 4.15). This suggests that our model is well-fitting. The likelihood ratio has decreased by more

Model	-2 Log Likelihood	χ^2	DoF	Significance
Intercept Only	64020.221			
Final	38864.010	25156.213	24	0.000

Table 4.15.: Model fitting information

than 20% which means that the model has a better fit than the previous model using synthesized data. A Likelihood ratio test of individual model parameters reveals that all independent variables are linearly related to the log odds of the dependent, also at a high significance level (see table 4.16).

Classification

A comparison of observed and predicted frequency values as displayed in table 4.17 strongly supports the predictive efficiency of model. The percentage of correctly calculated building types exceeds the result from the previous regression analysis. Even categories weakly predicted beforehand (single family medium lots and apartments) now reach double-digit percentages. In general, single family dwellings (1, 2, 3) are rarely estimated as apartments (4), whereas apartments are mixed up with single family buildings on a small lot in about 50% of the cases. A possible explanation

Effects	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood Reduced Model	χ^2	DoF	Sig.
Intercept	39071.00	207.00	3	0.00
Population Density	50861.50	11997.50	3	0.00
Median Household Income	39506.34	642.33	3	0.00
Av. #Children p. Household	38967.63	103.62	3	0.00
Av. #Cars p. Household	38987.17	123.16	3	0.00
Minorities in %	39534.05	670.04	3	0.00
Hispanics in %	39093.03	229.01	3	0.00
Av. Household Size	39062.42	198.41	3	0.00
Distance to Highway	39391.50	527.50	3	0.00

Table 4.16.: Likelihood ratio statistics, reduced model

for this confusion might be a high similarity between incorporated demographic characteristics of households choosing to live in apartments and households living in small single family dwellings.

Observed	Predicted				Percent Correct
	1	2	3	4	
1	3942	240	2100	75	61.9%
2	1451	457	1960	4	11.8%
3	700	43	21961	0	96.7%
4	548	7	37	506	46.1%
Overall Percentage	19.5%	2.2%	76.6%	1.7%	78.9%

Table 4.17.: Classification table

4.4.2.2. Results

A complete listing of model parameters for the multinomial logistic regression is tabulated in Appendix A.3. The estimated coefficients β reflect the effect of our demographic variables on the likelihood of living in an apartment (reference category 4) relative to living in a single family dwelling (categories 1-3). In general, a positive sign of β means an increase

of likelihood and a negative sign will reduce the choice probability of the corresponding building type.

The explanatory variable 'population density' has an overall negative effect. As population density decreases the likelihood of predicting single family dwellings when compared to apartments becomes higher. The fact that an increase in population density lowers the likelihood of predicting single family buildings is obvious, since building types are inherently related to population and housing density. The impact of the discussed explanatory variable is confirmed by the high significance of the corresponding coefficient β . The variables income and distance to highway have a positive effect $\beta > 0$ which is smaller than $1 \cdot 10^{-4}$. Nevertheless, while income is highly significant for predicting building types distance to highway is not significant for most of the outcome categories. The percentage of minorities has a negative effect on the model, which means that odds of predicting single family buildings over apartments decrease with increasing proportion of minorities. The results also indicate that the demographic variable 'Hispanics' has a slightly positive effect on all associated alternatives. The sign of the coefficient for the average number of children varies within single family building type categories. Furthermore, the explanatory variable children does not appear to be significant at all for choosing building type category 2 over apartments. A relatively large corresponding standard error indicates a lower precision with which the parameters are estimated. Finally, we find the average number of cars and the average household size highly important for the prediction of building types. As expected, negative parameter estimates are associated with the variable cars in categories 1-3 whereas the household size exerts a significant positive effect on the model.

In summary, the high significance of demographic explanatory variables in our model indicates they are important as determinants of building type choice. Especially population density is a prime parameter for distinguishing different dwelling types. In the following section, we apply the estimated logit coefficients to an *UrbanSim* data set to predict building types in Maricopa County.

4.5. Final Prediction with Regression Coefficients

As explained in Chapter 4.1, the probability of choosing building type category j for the polytomous outcome y_i is given by:

$$P(y_i = j | \mathbf{x}_i) = P_{ij} = \frac{e^{\mathbf{x}'_i \boldsymbol{\beta}_j}}{1 + \sum_{j=2}^J e^{\mathbf{x}'_i \boldsymbol{\beta}_j}} \quad (4.7)$$

We substitute the previously calculated parameter estimates $\boldsymbol{\beta}$ into (4.7) and apply the equation to demographic vectors composed of *UrbanSim* household characteristics.

Table A.4 in the Appendix summarizes prediction results for 51 sample grid cells located in Maricopy County. Each grid cell constitutes a table row and is associated with corresponding values for the following explanatory variables which are incorporated in the multinomial logistic regression analysis:

- average number of cars (Av. # Cars)
- average number of children (Av. # Chl.)
- average income (Av. Inc.)
- average household size (Av. Hh size)
- average number of households (Av. # Hh)
- percentage of Hispanics (His. %.)
- percentage of minorities (Min. %.)
- distance to the nearest highway (Dist. to Hw)

Moreover, the table lists calculated probabilities P_{ij} for predicting categories 1, 2 or 3 (single family dwellings) with respect to reference category 4 (apartments). The probabilities for outcome category 4 are not listed, they can be obtained by subtraction since the sum of all P_{ij} yields 1 (compare equation 4.3). The category with the highest outcome probability is stored as predicted building type category in table column **C** for related sample grid cells.



Figure 4.9.: Predicted building types

The prediction results for parts of Maricopa County are visualized in Figure 4.9 as color-coded 2D map. The map illustrates the distribution of estimated building types, but it does not provide information on the geographic context and the values of our explanatory demographic variables. We will present a comprehensive visualization framework in the next chapter.

4.6. Discussion of Results

In this chapter we developed a model to predict building types from demographic characteristics with the help of multinomial logistic regression. By examining different model implementations, we determined that the grid cell based regression approach using census data delivers best results. The final model's predictive accuracy is reasonably good for all outcome categories, although the model has a weakness in distinguishing apartments from single family dwellings with small lots. Future research is needed to find variables that help differentiate between these categories.

In addition, further research is necessary to refine dwelling type categories. Additional building attributes should be taken into account, such as the

number of stories. Up to the time of our analysis, the number of floors was not available for parcel level data.

Our model provides a basis for estimating future commercial and industrial buildings as well using a similar *MNL* technique. At the moment, our model lacks the ability to predict building types that are different from residential units or are for mixed use. Furthermore, our approach is unsuitable for the prediction of building types other than the types defined a priori.

Finally, more advanced research is necessary to assess model precision. The overall uncertainty of the estimation given the uncertain demographic input data remains to be examined in detail and should be a focus of future research.

5. Visualization Framework

After illustrating the data processing layer of our framework in the previous chapter, we now focus on the object-relational data base management system and the geovisualization layer (see Figure 1.6, Section 1.3). In the following, we integrate results from Chapter 4 into our geovisualization framework: estimated residential building types and *UrbanSim* predicted household characteristics for future years are visualized as 3D geometries on top of a map. The geometries are scaled according to multiple attributes from the geodatabase. The goal is to enhance human cognition by visualizing abstract representations of multidimensional data sets in a realistic context, namely *Google EarthTM* aerial photographs. Our framework will allow planners and key decision makers to interactively explore the data in order to find previously invisible patterns of economic development.

First, we define geovisualization and explain the concept of *Mashups*. Section 5.2 introduces our prototypical implementation of the proposed geovisualization framework in context of the Digital Phoenix Project at Arizona State University. First, we build a geodatabase of predicted demographic household data from *UrbanSim* and estimated future residential building types on a dedicated *PostgreSQL* database server (Section 5.2.1). Storage and access of geospatial data is realized using *Open GIS* standards. Our geovisualization framework is distributed over the internet and accessible through a web-based *GUI* (Section 5.2.2) where users can interactively customize visualization parameters. To visualize multiple data attributes on top of *Google EarthTM*, 3D georeferenced geometries are generated from 2D geometric primitives and scaled by attribute values. Thereby, we take into account different data scales of measurement and visual variables (Section 5.2.3). In Section 5.2.4, we give examples of suitable geometric shapes for representing multidimensional data. Afterwards, the basic concepts of Keyhole Markup Language *KML* are elaborated in further detail (Section 5.2.5). We explain how geometries can be encoded dynamically in *KML* using *PHP* scripts and *SQL* commands in Section 5.2.6. In order to convert georeferenced attributes from our *PostgreSQL* database to the

Google EarthTM spatial reference frame, we have to transform coordinate systems. Therefore, Section 5.2.7 gives a thorough introduction to geodetic datum transformations. Finally, we present our visualization results (Section 5.2.8), implementing iconized building types as scalable geometries that are superimposed on top of *Google EarthTM* aerial photographs. We conclude by highlighting central characteristics of our integrated geovisualization framework in Section 5.3.

5.1. Geovisualization

Geovisualization or *GeoVIS* (short for geographic visualization) is a multidisciplinary domain gaining from research in a variety of disciplines. Most notably, *GeoVIS* was influenced by cartography. Cartography has a long and successful tradition using abstraction and generalization to visualize data on maps. The International Cartographic Association (ICA) Commission on Visualization and Virtual Environments was the first to provide a comprehensive definition for Geovisualization:

“Geovisualization integrates approaches from visualization in scientific computing (ViSC), cartography, image analysis, information visualization, exploratory data analysis (EDA), and geographic information systems (GISystems) to provide theory, methods, and tools for visual exploration, analysis, synthesis, and presentation of geospatial data (any data having geospatial referencing).” [MK01]

This definition is widely accepted today. Other definitions advance a more human-centered view and see Geovisualization as a tool for knowledge construction [MGP⁺04, Mac95, LGMR05]. In [Mac01], MacEachren states:

“Geovisualization, from my perspective, is about the use of visual geospatial displays to explore data and through that exploration to generate hypotheses, develop problem solutions, and construct knowledge.” [Mac01]

This definition points out that geovisualization goes beyond static map-centered information communication. In fact, geovisualization creates visual representations of spatially referenced data to facilitate thinking and understanding about human and physical environments.

Effective geovisualization techniques can help explore, understand, and communicate spatial patterns. A popular and approved *GeoVIS* technique is the *Mashup* concept. A *Mashup* is a web application which integrates data sets from multiple sources into a single tool. The result is a new application tailored to a specific task. *Mashups* often use *XML*-based standards (see Section 5.2.5) to 'mark up' data so that the data can be used again in a different context. Usually, *Mashup* allow for the definition of visualization styles and for specifying semantic information. Moreover, they incorporate server-side technologies, e.g., servlets and *PHP* (see Section 5.2.6) for dynamic content generation.

In the next section, we will present a *Mashup* built by including data from simulations and empirical analyses into *Google EarthTM* to create an integrated geovisualization application for visual data mining. We will superimpose simulated demographic data and estimated dwelling type data on top of aerial photographs from *Google EarthTM*. For that purpose, data will be mapped to graphics variables of georeferenced scalable geometries.

5.2. Prototypical Implementation

The developed geovisualization framework is implemented as prototype within the Digital Phoenix Project [GHK⁺07] at Arizona State University. The Digital Phoenix Project is developing a multidimensional digital representation of the Phoenix metropolitan area in time and space. The goal of the project is to create a dynamic planning tool with an integrated visualization platform. Such a tool will help planners and policy-makers to assess the impacts of relevant policy decisions on urban growth and on environmental factors like air quality and urban heat island effect.

The Digital Phoenix Project uses *UrbanSim* to create complex scenarios of future urban developments. We will integrate those *UrbanSim* results and the building type data estimated through *MNL* (see Chapter 4) into a geodatabase. From there, data can be selected to generate multidimensional geovisualizations.

5.2.1. Data Revisited

As explained in Chapter 3.2, we store and manage our data in a *PostgreSQL* database with *PostGIS* extension. A geodatabase facilitates to associate data with temporal, spatial, and geometric information for visualization and analysis. The open source object-relational database management system *PostgreSQL* has high potential for fast aggregation of heterogeneous data sets and is considered the glue between our simulated *UrbanSim* output and the multidimensional data visualization. Contemplating multidimensional data sets, we refer to a set of objects where each object is associated with a feature vector storing discrete, continuous or nominal values (see Section 5.2.3).

In general, each spatial table in the geodatabase represents a separate PostGIS layer. An ancillary table contains meta-data on the associated geodetic datum (see Section 5.2.7). Each distinct geographic object constitutes a record in a spatial table and associated attribute information is stored in data columns. PostGIS provides a dedicated geometry column which contains geometric information for each feature in the form of point, line, or polygon data types. We will use this geometry column to store the scalable geometries generated from attribute data for the geovisualization.

First, we set up a database for *Digital Phoenix* on a dedicated *PostgreSQL* web server. Aggregated *UrbanSim* household data at grid cell resolution is imported from an *ESRI* shape file and is passed on to the backend database for processing. Furthermore, the geodatabase is populated with empirical results from our building type estimation (see Section 4.5).

In the next step, we use *PostGIS* functions to calculate the centroids for each grid cell in the geodatabase. The centroids are stored as 2D points and provide a basis for anchoring the scalable geometries. Calculations are accomplished in the built-in SQL Console of PgAdmin III [pgA07] (see Figure 5.1), a free administration and development platform for the *PostgreSQL* database.

Once the estimated parameters and the calculated centroids of the grid-cells have been coded in, *UrbanSim* projected demographic variables and the associated estimated building type categories are ready for geometry generation. Geospatial features defined by the *OGC SFS* (see Section 3.2) and supported by *PostGIS* are points, lines, and polygons. To date, *PostGIS* lacks support for 3D primitives, i.e., all basic vector data types

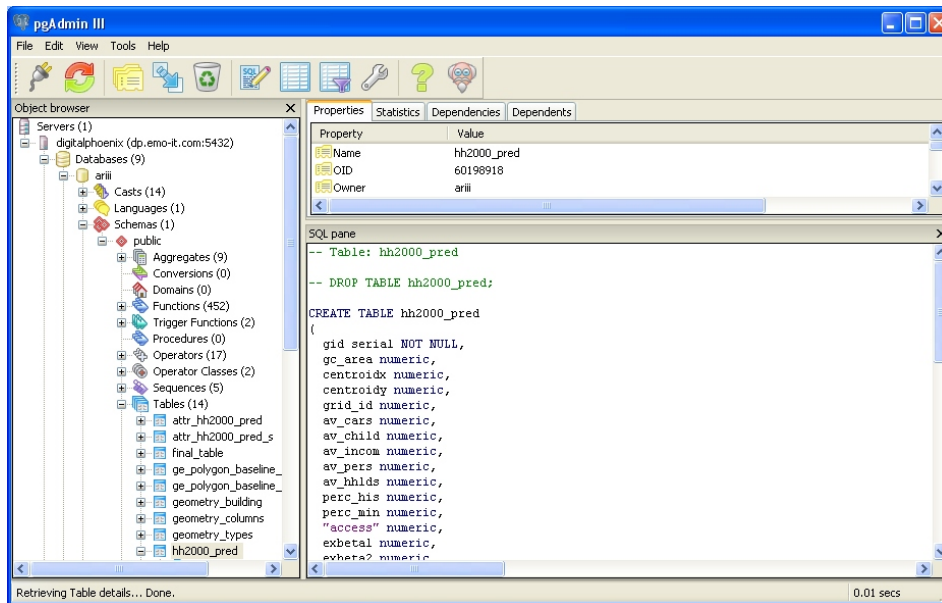


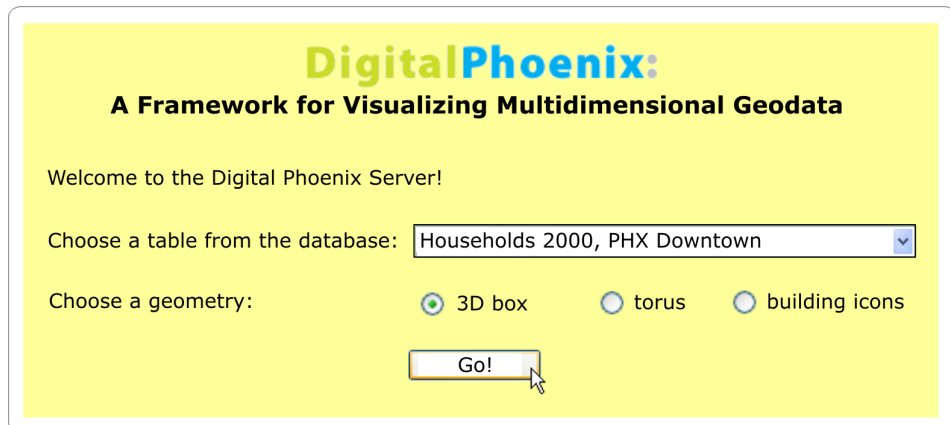
Figure 5.1.: Access to *PostgreSQL* database via PgAdmin

are defined in the plane. However, 3D coordinates can be specified to represent 2D entities. Consequently, we can integrate 3D objects into the geodatabase by defining 2D primitives and assembling them to represent 3D geometries. For example, a 3D box can be represented by 6 attached polygons, one polygon for each face of the box. Geometry selection, scaling, and visualization is managed by the user through a web-interface, described in the following section.

5.2.2. GUI

Access to the geodatabase is granted online through a website in order to ensure broad immediate accessibility and to reach the largest possible audience. The graphical user interface (*GUI*) provides a login screen to the data server and several options for customizing the multidimensional data visualization.

After logging in, the user is asked to choose a table from the geodatabase and to select a scalable geometry (see Figure 5.2) for superimposing in *Google EarthTM*. At the moment, the user has three options, namely boxes, tori, and building icons. More options will be implemented soon.



DigitalPhoenix:
A Framework for Visualizing Multidimensional Geodata

Welcome to the Digital Phoenix Server!

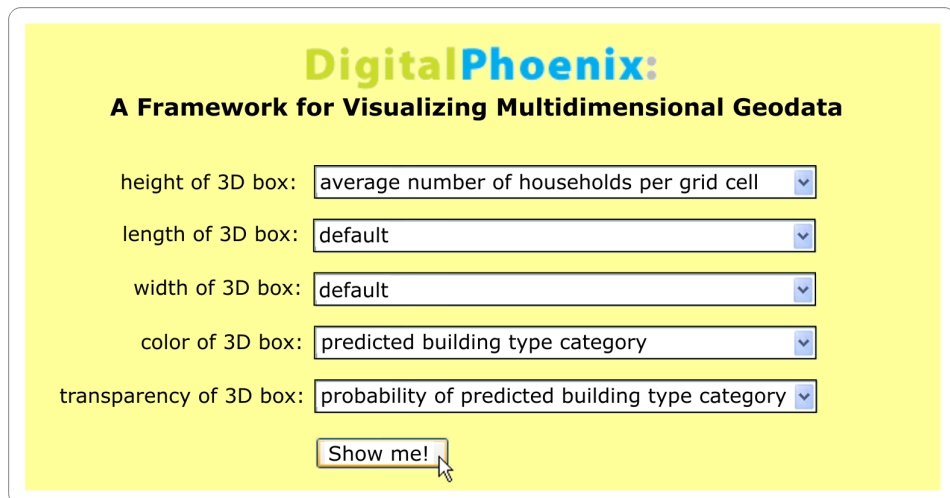
Choose a table from the database: Households 2000, PHX Downtown

Choose a geometry: 3D box torus building icons

Go!

Figure 5.2.: Web-based data access and geometry choice

In the next step, the user can interactively assign attributes from the database table to parameters of the selected geometry (compare Figure 5.3). Thus, the user is capable of choosing a meaningful visual representation for the multidimensional dataset. We will discuss visual variables in detail in Section 5.2.3.



DigitalPhoenix:
A Framework for Visualizing Multidimensional Geodata

height of 3D box: average number of households per grid cell

length of 3D box: default

width of 3D box: default

color of 3D box: predicted building type category

transparency of 3D box: probability of predicted building type category

Show me!

Figure 5.3.: Web-based parameter assignment for scalable 3D box

The *GUI* lists available columns from the selected table and lets the user attribute those to the scalable geometry's degrees of freedom. In case of the 3D box, available attributes might allow the user to choose scale

height, size of the footprint, color, and transparency of the body. For each grid cell, a scaled geometry is generated by server-side *PHP* scripts and encoded in *KML* for visualization.

5.2.3. Data Scales of Measurement and Visual Variables

Before describing the implementation of the geovisualization framework, we have to theoretically contemplate different data types and possible visual variables to convey information to the user. Our goal is not to provide the optimal cognition enhancing visualization for multidimensional geodata, since there is not one best visualization for all kinds of data. We aim at establishing a generic and extendable visualization framework that is appropriate for a variety of different visualizations.

Stevens [Ste46] defines four categories for measuring data scales: nominal, ordinal, interval, and ratio. This data type taxonomy is broadly accepted and in practice reduced to three scales: nominal, ordinal, and continuous data.

Nominal and ordinal scales only delineate different categories of information and are therefore used for categorical data. A nominal scale defines a set of identifiers with no intrinsic order, e.g., color, race or sex. The relationship between those variables is transformation invariant. Binary nominal variables have only two categories (e.g., male/female, yes/no) and often appear in surveys. In contrast, non-binary variables have more than two categories. Even though category labels can be numerical, qualitative data does not allow for arithmetic operations. Still, categorical data can be used for statistical analyses and is generally summarized in the form of frequencies or percentages. On the other hand, ordinal scale involves data with an inherent order, e.g., sizes categorized into small, medium, and large. Nevertheless, ordinal categories lack numerical properties and prohibit arithmetic operations.

Unlike qualitative scales, quantitative data generally results from measurements with specified ratios between similar increments in measures. Descriptive statistics for this kind of data can be summarized by mean, median, and variability statistics. Quantitative data can either be discrete (e.g., population) or continuous (e.g., distance). Continuous and discrete data can be classified by discrete categories and treated as categorical

data. Similarly, categorical variables can be handled as continuous when sample sizes are large.

Output of *UrbanSim* simulations is mainly quantitative data. We obtain data on numbers of children, cars, and workers per household as well as income. For the visualization, we will categorize income into different classes and establish categories for various population densities within grid cells (see Section 5.2.8). Estimated building types resulting from our empirical analysis are inherently categorical with four distinct classes. Next, we examine which visual variables are available for representing quantitative and qualitative data.

Jacques Bertin’s “Semiology of Graphics” [Ber67] systematically classifies the use of visual elements for data visualization. Bertin differentiates between seven visual variables: position, form, orientation, color, texture, value, and size. According to Bertin, the eye is sensitive to all these ‘retinal properties’ of graphics, independent of the position of the object. Consequently, manipulation of these visual variables can enhance the understanding of visualization.

Bertin’s classification was originally developed for paper maps. In the age of digital cartography, new visual variables were added such as transparency, resolution, and crispness. MacEachren extended and adapted Bertin’s original framework to twelve visual variables [Mac95] and matched these to the data types mentioned above using three degrees of effectiveness (see Figure 5.4).

visual variable	nominal	ordinal	quantitative
location	good	good	good
size	good	good	good
texture	good	good	good
color hue	good	good	good
orientation	good	good	good
shape	good	good	good
color value	poor	good	good
color saturation	poor	good	good
resolution	poor	good	poor
crispness	poor	good	poor
transparency	poor	good	poor
arrangement	poor	poor	poor

effectiveness	good	marginal	poor
---------------	------	----------	------

Figure 5.4.: Visual variables and their effectiveness (c.f. [Mac95])

MacEachren argues that only location and size are suited for visualizing quantitative data. Shape is exclusively appropriate for depicting nominal attributes. The more a shape resembles the represented data attribute, the easier it is for users to understand the visualization. With decreasing size and increasing number of geometric shapes, symbol recognition is reduced. Depicting the same data attribute as a combination of visual variables, e.g., size and color, increases the representational power of the visualization for that particular attribute.

In a manner of speaking, geovisualization is mapping spatial data values and attributes to visual variables or a combination of graphical entities. Figure 5.4 provides a heuristic for linking data attributes to visual elements that we will adapt in our visualization framework. Our geovisualization process is user controlled, meaning the user can interactively define the mapping between attributes and visual variables in the web-based interface presented in Section 5.2.2. Degrees of freedom for choosing appropriate visual variables include size, color, geometry, and transparency of graphical entities. We will discuss possible symbologies for *UrbanSim* output data and estimated building type data in the next section.

5.2.4. Scalable Geometries

The main information carrier for data attributes in our geovisualization framework is the visual variable shape. The scalable geometries generated by our *PHP* scripts can assume any kind of discrete or discretized shape. To visualize future urban structures and household demographics, an implementation of iconized building types seems obvious. Building type symbols comprehensively convey a visual sense of density. Visual encodings and shape parameters for dwelling type geometries include footprint, building height, roof type, roof height, ridge height, the number of chimneys, color, and transparency for instance (see Figure 5.5). Other possible scalable geometries for representing multidimensional data are pyramids, discretized cones, and tori. The geometric shape is chosen according to the user's need, generated for each database entry and geovisualized in a map context using Keyhole Markup Language.

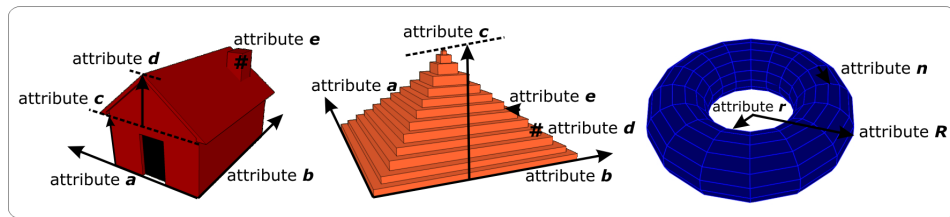


Figure 5.5.: Geometries

5.2.5. The Keyhole Markup Language *KML*

Keyhole Markup Language (*KML*) [Goo07a] is an eXtensible Mark-up Language (*XML*) grammar and file format. Like all *XML*-based language schemas, *KML* uses text to define elements that represent entities and to specify their hierarchical relationships. In particular, *KML* models geographic features like points, lines, polygons or images for visualization in web-based online 2D maps and 3D geobrowsers such as *Google Earth*TM. The Keyhole Markup Language has a tag-based structure, i.e., elements are encoded by the convention:

```
<elementName attributeName = "value"> element </elementName>
```

Elements are enclosed by tags defining name and optional associated attributes. They can be nested, that is, child elements inherit the characteristics of their parents.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.1">
  <Placemark>
    <description> Herberger Lab, Arizona State University </description>
    <name> Lab </name>
    <Point>
      <coordinates> -111.937182, 33.421084 </coordinates>
    </Point>
  </Placemark>
</kml>
```

Figure 5.6.: *KML* sample code

Placemark elements provide the basis for visualizing spatial entities; they contain a geometrical description and coordinates for the represented entity.

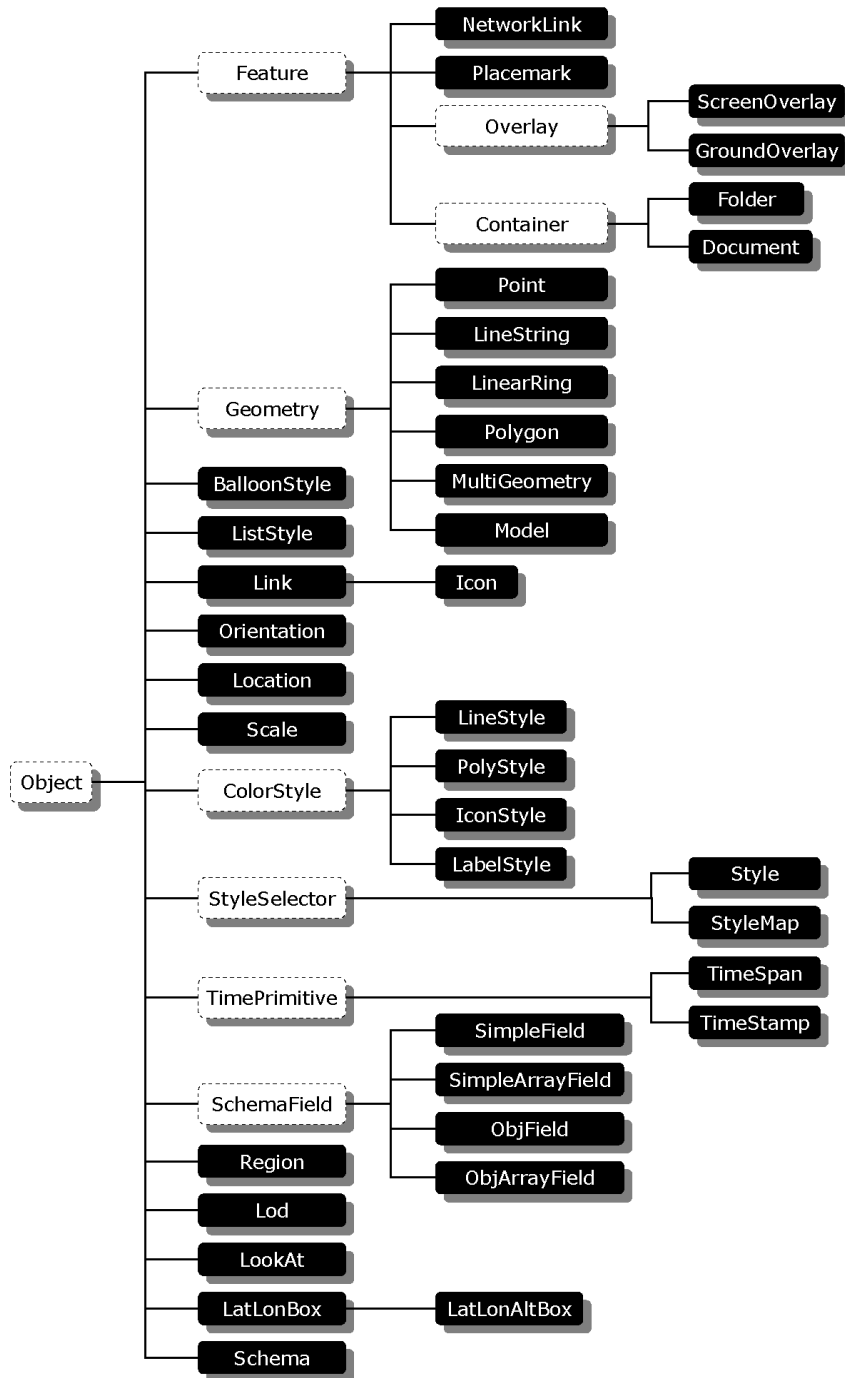


Figure 5.7.: *KML* elements (c.f. [Goo07a])

Figure 5.6 shows sample code of a KML file with a single *Placemark* element. The first two code lines specify that the file is *XML*-based and state where the language schema is defined. As root element, *Placemark* encloses three children: the *Placemark* name, a descriptive text, and a *Point* element with geographic coordinates. *KML* uses latitude and longitude in the World Geodetic System of 1984 (*WGS84*), but it lacks support for other geodetic reference systems. Therefore, it may be required to convert the observed geodetic datum of spatial entities to *WGS84* before implementing the *KML* encoding to avoid discrepancies. We will discuss geographic coordinate system transformations in Section 5.2.7.

Placemarks can be associated with style elements to define the symbolism of corresponding geometries. *Style* elements allow varying visual variables like hue, lightness, saturation, and size. They are either associated with individual *Placemarks* as child elements or defined globally and then linked by a number of *Placemark* elements via *URL*.

Other useful *KML* elements are *Overlay* (for superimposing images on the ground or the screen), *NetworkLink* (for streaming content over the internet), and *TimeSpan* (for specifying a visualization time period). An overview of *KML* elements is given in Figure 5.7. Note that abstract elements which cannot form entities are marked as white boxes.

Altogether, *KML* provides high flexibility to model and display customized discrete geometries on specified locations of the Earth's surface. In the following section, we will elaborate on how to dynamically generate *KML* files. Using *PHP* and *SQL*, we will visually encode demographic data from a database in *KML* for display in *Google Earth*TM.

5.2.6. Using PHP and MySQL to Generate KML

As explained in the last section, we use *KML* to encode the position, shape, and visual characteristics of geometries. The geometries, representing multidimensional geodata on demographics and residential building types, are scaled with attribute values from the geodatabase. *KML* files are dynamically generated from a *PostgreSQL* database by *PHP* scripts. *PHP* (version 5 used here) [PHP07] is a widely-used general-purpose scripting language. It is highly flexible and especially suited for web development since it can be embedded into *HTML* code. *PHP* runs server-side and can be used to connect with and query databases from *HTTP* requests.

The object relational database management system *PostGIS* allows access through standard *SQL* commands. In our visualization framework, we incorporate these commands into *PHP* scripts. First, user data is submitted via *PHP* forms to the geodata server for access. Then, the user can query the database for tables, choose a preferred geometry to represent the data, and select attributes for scaling.

Finally, a *PHP* script creates children of *KML Placemark* elements for each grid cell in the geodatabase. Every grid cell is associated with an instance of the previously chosen geometry class. Geometry is scaled and calculated on the fly in a dedicated *PHP* script based on the extracted attribute information from the corresponding database row. For scaling, the attributes have to be normalized with respect to minimum and maximum values. Alternatively, the attribute values can be classified into different categories.

The geometries suggested for geovisualization in Section 5.2.4 consist of a set of multiple 2D polygons (*MultiPolygons*) forming a 3D body. These *MultiPolygons* are attributed with a *Style* element to define the visual variables color and transparency of the geometry. The final *KML* file is generated by creating an array of strings holding basic *KML* elements as well as a *Placemark* for each grid cell (see Section 5.2.5). Thereby, the coordinates of the *MultiPolygons* within the *Placemarks* have to be transformed to a geodetic reference frame understood by *Google EarthTM*.

5.2.7. Coordinate System Transformations

Each object in a geospatial database has a predefined spatial reference system (*SRS*). The geometries in our *PostgreSQL* database are georeferenced in a coordinate system that differs from the geocentric reference frame *Google EarthTM* uses. In this section, we briefly explain why different coordinate systems exist in parallel and how to transform them from one system to another.

The figure of the Earth is difficult to model, since the Earth's surface is highly complex. Mountains, valleys, and oceans render the terrestrial topography unsuitable for an exact mathematical computation. The science of geodesy is the core discipline for modeling the Earth's surface as precisely as possible. An accurate representation is crucial for other geosciences like cartography, geography, and surveying.

To approximate the shape of the Earth mathematically, its figure is simplified assuming that the surface is identical with the mean sea level of the ocean. This approximation is called geoid. A geoid constitutes an equipotential gravity surface and serves as reference surface for leveling. Due to gravitational anomalies in the Earth's interior, the hypothetically defined equipotential surface undulates. Therefore, the geoid has irregular shape and the Earth is in fact potatoe-shaped (see Figure 5.8).

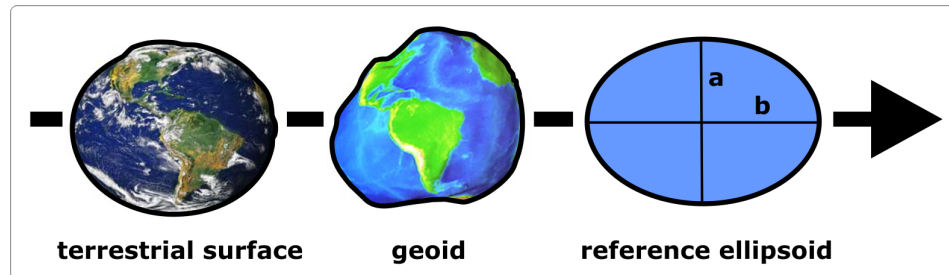
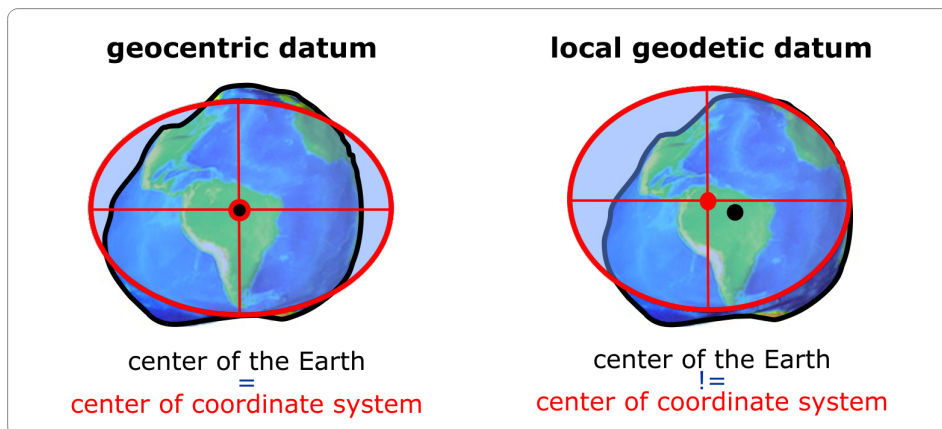


Figure 5.8.: Figures of the Earth

Since the geoid cannot be measured directly, the figure of the Earth is further simplified and mathematically approximated by an oblate ellipsoid of revolution. Depending on the application area, the reference ellipsoid has to be globally or locally best-fitting with minimal deviation from the geoid (see Figure 5.9). The ellipsoid is used as basis for defining a geodetic datum. Geodetic datums determine the figure of the Earth as well as the orientation and origin of the *SRS* which is used to map the Earth.

All over the world, nations use different datums with different locally best-fitting reference ellipsoids as basis for coordinate systems. Consequently, a point on Earth has differing coordinates in different geodetic datums. Since assigning coordinates to the wrong reference datum can result in location errors of hundreds of meters, a transformation between coordinate systems is required. This datum conversion rests upon seven parameter transformations: three translations along the x-, y-, z-axis, three rotations, and scaling.

Common reference frames in North America are *NAD27*, *NAD83*, and *WGS84*. The North American Datum of 1927 has a locally best-fitting reference ellipsoid, whereas the North American Datum of 1983 and the World Geodetic System 1984 have earth-centered reference systems with best fit for the entire Earth. The geocentric ellipsoidal models were derived from satellite measurements and are based on the Geodetic Reference

Figure 5.9.: Geocentric and local *SRS*

System 1980 (*GRS80*) with minor differences in the reference ellipsoid parameters. *WGS84* is the standard geodetic datum for *GPS* satellite navigation and provides a worldwide geodata basis. The *WGS84* coordinates are given in latitude and longitude (compare Figure 5.10). The geodetic

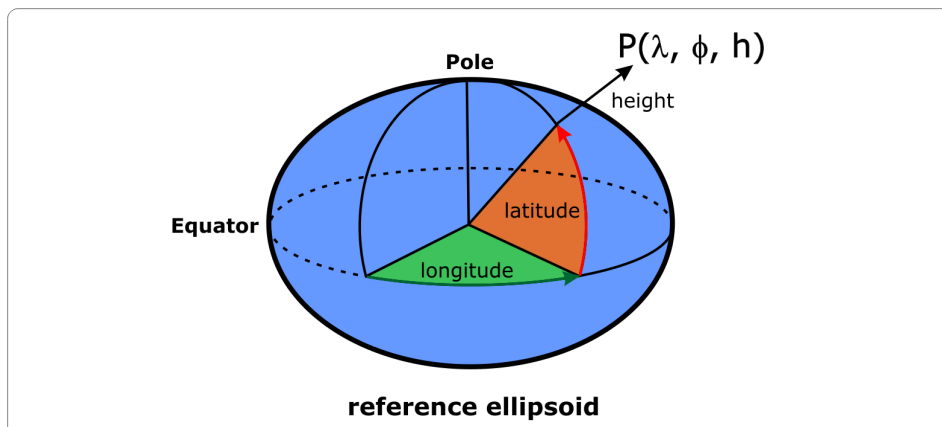


Figure 5.10.: Geographical Coordinate System

latitude of a point is defined as the angle between the equatorial plane and the direction vector to the point from the geocenter. The angle between a reference plane (the Prime Meridian) and a plane perpendicular to the Equator passing through the point specifies the geodetic longitude. The scalable geometries calculated by our *PHP* scripts live in a coordinate system that is predefined by the geodetic reference frame of *UrbanSim*

grid cells, the *NAD83*. Thus, the inherent Cartesian coordinates of the geometries have to be transformed into latitude and longitude *WGS84* coordinates for visualization in *Google Earth™*. This transformation is implemented in a dedicated *PHP* class. Detailed information about the defining parameters of the *WGS84* reference frame and its relationship with *NAD83* can be obtained from the National Geospatial Intelligence Agency [Nat97]. The transformed coordinates of our scaled geometries are directly encoded in *KML* and made available to the user for download and visualization.

5.2.8. Visualization Results in *Google Earth™*

The *Google Earth™ Mashup* integrates the empirical results from multinomial regression results for building types and model inputs as discussed in Chapter 4. Two illustrative visual encodings have been developed to show the potential and flexibility of the integrated visualization framework. The aim was not to develop the optimal information visualization technique, but to provide a generic and extendable framework for visualizing multidimensional geodata.

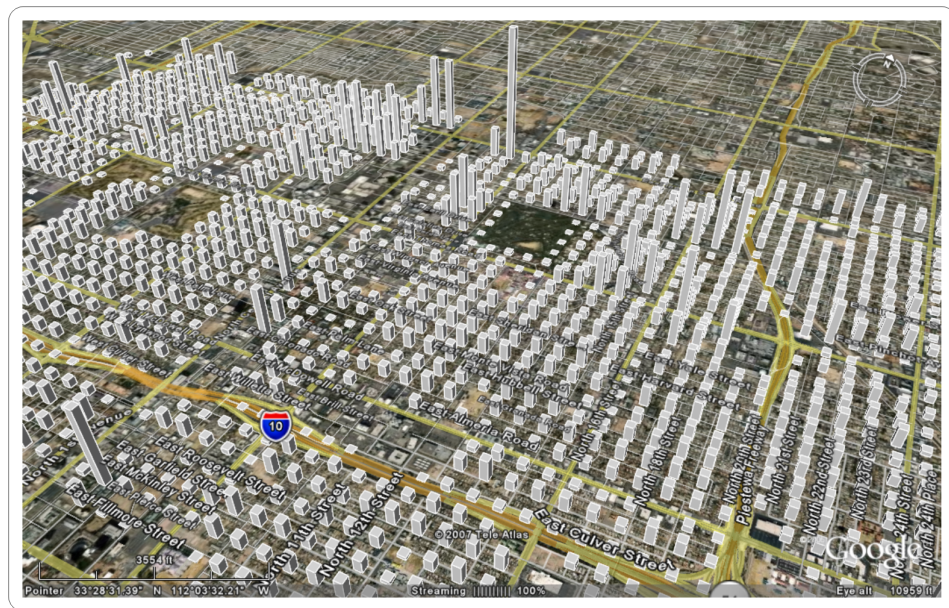


Figure 5.11.: Scaled 3D boxes represent population density (2000) in Phoenix Downtown

The first approach displays a single attribute from the database by scaling the height of simple 3D boxes. The boxes are linked to the grid cell centroids from the *UrbanSim* 150m grid and represent population density in 2000. The encoded attribute is normalized according to minimum and maximum population values. Figure 5.11 shows that displaying population by scaled 3D boxes comprehensively conveys a sense of density.

In the following, we will focus on 3D building icons as main components through which demographic information is visualized. We choose building type geometries since they most intuitively represent the data at hand. Figure 5.12 illustrates the applied symbology for residential buildings. Apartments are depicted as cubes with a shed roof whereas single family dwellings are equipped with saddle roofs consisting of one, two, or three ridges. The geometry for each building is defined by a single set of polygonal faces (*MultiPolygon*). More specifically, dwelling cubatures are composed as follows:

- apartment - 8 points, 6 polygons
- single family (small) - 18 points, 11 polygons
- single family (medium) - 14 points, 9 polygons
- single family (large) - 9 points, 7 polygons

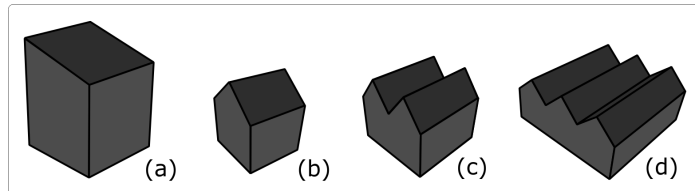


Figure 5.12.: Building type geometries: (a) apartments, (b) small single family, (c) medium single family, and (d) large single family

For the visualization of each grid cell, an icon is chosen according to the predicted building type category, which was calculated by means of *MNL* and stored in the geodatabase earlier. Then, associated demographic data is mapped to specified visual variables of the model's geometry and appearance. The visual variables considered for our building icons include building height, footprint size, color, and transparency.

Population density is encoded in different building footprint sizes. Five different density classes were established ($900m^2$, $16.00m^2$, $3.600m^2$, $6.400m^2$,

and $14.4400m^2$) which could then be visualized by scaling the building footprint in relation to expected density. To amplify the visual density effect, building height is increased proportionally. The probability of predicting the correct building type based on our empirical estimates is visualized in three different transparency levels:

- 0% - 33% uncertainty = 100% opacity
- 34% - 66% uncertainty = 75% opacity
- 67% - 100% uncertainty = 50% opacity

We decided not to color-code building cubatures based on demographic variables in order to avoid visual overload. Instead, each grid cell is associated with a colored polygon which is mapped to the ground. In Figure 5.13, the variable average income is color-coded via grid cells. The legend, superimposed as *Overlay* element in *Google EarthTM*, is the key to classification thresholds.



Figure 5.13.: Visualization of (a) average income (color-coded grid cells), (b) building types (geometry), (c) population density (size of footprint), and (d) uncertainty of building type prediction (transparency) [MGH⁺08]

Figure 5.14 is a close-up of the situation displayed in Figure 5.13. Each grid cell contains a scaled and stylized *MultiPolygon* building geometry

and a color-coded polygon. Combined, all grid cell related geometries form a *Placemark* element in the *KML* file. The name and description of each *Placemark* is listed under the *Temporary Places* folder in *Google Earth™*.

The grid cells classification system established for the average income attribute in Figures 5.13 and 5.14 can be applied to any other demographic variable in our geodatabase, such as access, population share, average number of cars, and average number of children. Screenshot 5.15 displays color-coded grid cells in birds eye view representing distances to nearest highway. Observers familiar with Phoenix Downtown can clearly identify the courses of *I-10* and *I-17*.



Figure 5.14.: Close-up of icons for different building types [MGH⁺08]

The four provided screenshots serve as an example of flexible and interactive geometry layers designed to work in conjunction with *Google Earth™*. Each *Google Earth™ Mashup* facilitates visualizing multidimensional aspects of the data and offers a means of recognizing relative patterns and relationships between different characteristics embedded in the information. Given that our visualization framework uses open-source software and open geospatial data standards, it also offers an inexpensive, yet powerful tool for spatial data visualization in three or four dimensions.

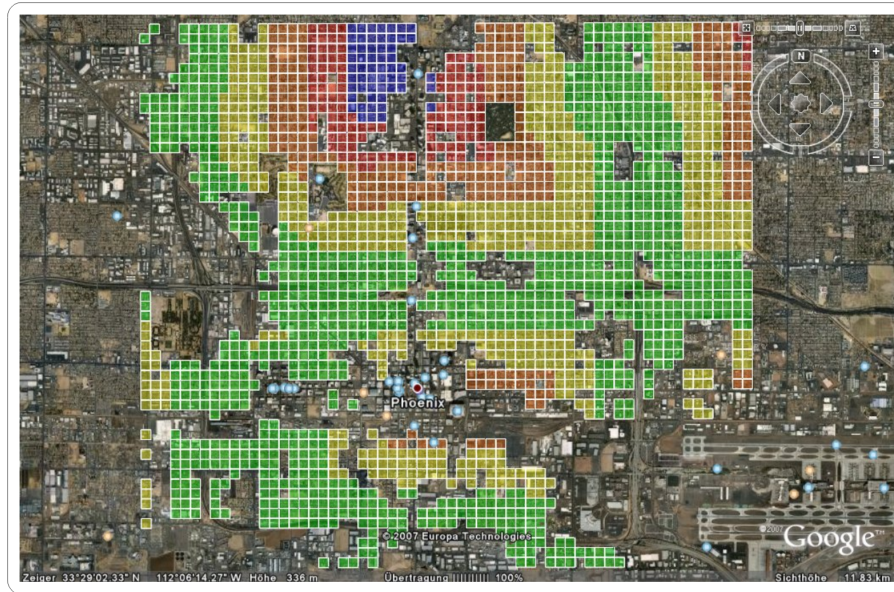


Figure 5.15.: Phoenix Downtown, color-coded grid cells display distance to nearest highway [MGH⁺08]

5.3. Discussion of Results

This chapter introduced a 3D geovisualization framework for displaying multidimensional geodata. The framework implements a *PostgreSQL* object relational geodatabase to store and maintain building type data and demographic household characteristics generated with *UrbanSim*. Data are accessed through an easy-to-use PHP driven web-interface. The visualization environment is based on the geobrowser *Google EarthTM*. Multidimensional database attributes are visualized on top of *Google EarthTM* maps as georeferenced geometries that are scaled and colored according to attribute values. In the symbolization process, we concentrate on building icons as main geometries through which demographic information is displayed. We designed and implemented *PHP* scripts to encode different geometries in *KML* files for visualization in *Google EarthTM*. The *KML* files are generated on-the-fly based on visualization parameters specified by user input.

Our approach has a number of advantages over classical 2D thematic maps and hitherto existing geovisualization frameworks. *Google EarthTM* is free for personal use software and the incorporated database *PostgreSQL* as

well as the spatial *PostGIS* extension are open-source software packages that use open geospatial data standards, which are *OGC* compliant. This makes the presented system architecture widely available to both laymen and expert users. Web-based access of geospatial information further increases availability due to growing ubiquity of internet use.

The *Google EarthTM Mashup* has an intuitive and easy-to-use interface to dynamically and interactively explore data. It allows for browsing data layers with a variety of interactive pan and zoom operations. The user is empowered to retrieve spatial detail by zooming in and to filter geodata by space, time, and attribute details. In this manner, the geobrowser fulfills the visual information-seeking paradigm “overview first, zoom and filter, then give details on demand” [SCM99].

Our geovisualization framework supports information overlay to enhance visualization by additional informative data layers from the geodatabase or from the *Google EarthTM* server. A visual synthesis of heterogeneous data sets helps the user to spot hidden correlations within the data. That way, our *Mashup* serves as geospatial data mining tool for recognizing relative patterns and relationships between different characteristics embedded in the information.

Our visualization framework is not based on realism because realism is unnecessary and inherently inaccurate for simulated future neighborhoods even when the underlying data has a reasonable degree of certainty. Also, realism can be distracting, since too much visualization details might hide insight and cause cognitive overload. The presented framework incorporates abstraction methods from cartography and Information Visualization to amplify cognition. Demographic and building type data is communicated to users as abstract geometries, scaled by attribute values, in 3D real context on *Google EarthTM*. Our approach enhances spatial cognition, facilitates thinking and supports decision making in integrated urban and environmental analyses.

These tasks are further supported by incorporating the 3rd dimension into our visualization. A 3D model enhances the spatial characteristics of geodata and provides an additional dimension for encoding multidimensional data. On the other hand, the 3rd dimension causes problems not existent in 2D map-based visualizations. To date, little research has been dedicated to analyzing the effect of perspective on the perception of 3D visualizations.

The presented geovisualization framework is highly generic, that is we can display any kind of thematic data from various application domains and choose any kind of discrete geometry as symbolization. Our generic approach uses KML to specify scalable 3D geometries and to map data attributes to color, shape, and transparency of the specified geometries.

Regardless, encoding geometries in *KML* via *PHP* scripts poses a bottleneck, which may be overcome eventually with increasing computational power. At the moment, on-the-fly generation of *KML* files turns out to be difficult with large data sets, since a high number of polygon calculations causes the entire process to slow down. Calculating geometries with 6 polygons each for 65,000 grid cells in Maricopa County takes up to 30 seconds depending on server capabilities. This conceived limitation can be bypassed through aggregated grid cells. Aggregation is equivalent to the concept of generalization in cartography and facilitates consideration of spatial data at various scales for different viewpoints. A so-called level of detail implementation decreases the complexity of the 3D scene in the geobrowser as the scene moves away from the viewer and thus decreases the number of polygons to be rendered.

Another drawback of using *Google EarthTM* is the lack of built-in geo-analysis capabilities. Users who need to apply sophisticated GIS functionalities to their data have to revert to GIS systems like Esri's *ArcGIS* [Esr07] or the open-source software *Quantum GIS* [Qua07].

Google EarthTM's time series function has not yet been implemented in our framework but offers an enormous scope for further research. Extending the visualization to 4 dimensions by mapping data evolution over time empowers the user to visualize and analyze trends.

6. Conclusions

The motivation for this thesis was to develop an inexpensive, accessible, and comprehensive planning tool for the simulation and visualization of future urban developments. Our research aims at supporting participants in planning processes to better understand the impacts of decisions made today on the development of future urban environments.

We presented a framework that integrates the *UrbanSim* simulation package for modeling the possible long-term effects of different policies on urban developments. On top of that, we developed an empirical method for estimating future residential building types at different scales and a geovisualization tool based on *Google EarthTM* to communicate multidimensional attributes of a complex data set. Thus, our framework allows for the visualization of future built environments and the characteristics of their inhabitants based on *UrbanSim* output.

UrbanSim, like most other urban simulation tools, offers very limited ability to visualize the output. Moreover, almost no land use change model offers comprehensive visualizations on built forms in 3D. Our research bridges this gap in the current visualization literature dealing with urban forms. It also provides a path beyond the rule-based methods of procedural modeling towards a more empirically-based framework for developing future environments. Our system architecture consists of three main components, all combined in an integrated framework:

- data processing (simulation of future urban environments with *UrbanSim* and estimation of building types)
- geodatabase for storage (*PostgreSQL* with *PostGIS* extension)
- geovisualization (scalable 3D geometries, superimposed on *Google EarthTM*)

In the context of this framework, we presented a statistical method for estimating residential building types on different scales from demographic

data. The mapping between dwelling types and household characteristics was realized with multinomial logistic regression. In our first modeling approach, clusters with typical building type distributions were formed by k -means to establish nominal categories for the regression model at a neighborhood scale. Subsequently, the log odds of the clustered neighborhood category predictors were modeled as a linear function of the categories' covariates in the estimation process. Afterwards, the model was refined to a spatial scale of $150m \times 150m$ grid cells. Thereby, estimating distinct residential building types based on synthesized census data gave best results. Finally, the regression results were tested for data fit significance. The high significance of demographic explanatory variables as well as the variable 'population density' indicated they are important as determinants of building type choice.

All tests provided strong support that the model fits the data reasonably well and that logit regression is a coherent framework for assessing the relationship between demographic characteristics and the building types people live in. From a theoretical point of view, our research results are novel and can be used to solve crucial problems in urban planning that require information on future residential building types. Our established *MNL* model is also extendable to commercial and industrial dwelling types and provides an important basis for carbon footprint calculations and material flow analyses of future urban developments. Thus, we make a relevant contribution to the reliability of sustainability metrics.

The tie connecting simulation and visualization in our framework is a *PostgreSQL* geodatabase with *PostGIS* extension in the backend. *PostgreSQL* is based on *OGC* data standards and is open source software, i.e., it constitutes a widely available platform for storing and retrieving large spatial data sets. Our implemented prototype stores *UrbanSim* projections and estimated building types for Maricopa County in the geodatabase. However, any geospatial data can be added to the *PostgreSQL* database, which provides the potential to record heterogeneous data rapidly from various sources. In addition, *PostgreSQL* can be easily and tightly integrated with web services. We developed a front-end web-interface in *PHP* to access and process data for visualization. Using *PHP* scripts offers high flexibility in terms of user-friendliness, interactivity, and accessibility. Furthermore, server-side scripting has huge potential for rapid prototyping and takes away workload from the client. Data query and processing tasks are executed by the web server and the results are served to the client as

ready-to-visualize files. This approach offers a clear advantage over client-side applications and is therefore also suitable for mobile clients.

Driven by the demand for an intuitive and comprehensive integrated visualization framework for urban simulation data, we proposed a geovisualization *Mashup*. Our *Mashup* re-uses existing functionality and data of *Google EarthTM* to create a tool for the integrated visualization of multi-dimensional data in urban planning. To visualize density, estimated residential building types, and significant demographic attributes, data set columns from the geodatabase were mapped to visual variables of scalable 3D geometries. The abstract data representations were encoded in *KML* and superimposed on top of *Google EarthTM* aerial photographs. Useful scalable 3D geometries range from simple 3D boxes to more complex discrete icons. For the implementation of our prototype decision support tool, we concentrated on building type symbols as main geometries through which demographic data should be displayed. Dwelling type icons are meaningful representations for the simulation data set in our geodatabase and convincingly demonstrate the visual expressiveness of our geovisualization *Mashup*.

Google EarthTM is a data rich application that offers access to a wide range of ancillary data sets like road networks, places of interest, and georeferenced panoramic images. This allows the user to visually synthesize the abstract representations of attributes from our database with other geospatial data in real world context. That way, our *Google EarthTM Mashup* combines human cognitive skills and technology to visualize hidden information patterns and trends.

In this thesis, we tackled the need for an integrated simulation and visualization framework for multidimensional geospatial data. We conclude that *Google EarthTM* is an inexpensive but powerful tool to visualize geodata in three or four dimensions. Our *Mashup* overcomes the shortcomings of classic static mapping concepts which are inadequate for analyzing and visualizing urban growth dynamics. *Google EarthTM* is user-friendly, has an intuitive interface, is interactive, browsable and offers easy access to geospatial information. Integrating an empirical framework for simulating urban futures results in a geo-analytically powerful environment that offers keen insights into urban dynamics. Our method provides realistic information without invoking high degree of photorealism which can distract attention from the complexity of information that needs to be communicated. It provides a 3-dimensional representation of abstract data in real

context to avoid cognitive overload and to amplify cognition, and facilitate thinking, problem solving, and decision making. Thus, our thesis provides a relevant contribution towards developing an integrated framework for a comprehensive urban planning tool that can support decision-making of planners, politicians, and the general public.

A. Parameter Estimates

# of Case		β	Std. Error	DoF	Significance	Exp(β)
2	Intercept	-8.9665	1.3334	1	0.0000	
	Population	-0.2266	0.0646	1	0.0005	0.7973
	Median Age	0.0504	0.0153	1	0.0010	1.0517
	Median Income	0.0000	0.0000	1	0.0015	1.0000
	Minorities in %	-0.0639	0.0164	1	0.0001	0.9381
	Hispanics in %	0.0270	0.0177	1	0.1281	1.0274
	Average Household Size	3.2540	0.3504	1	0.0000	25.8931
	Distance to Highway	0.0001	0.0001	1	0.2865	1.0001
Age of Head	0.0971	0.0092	1	0.0000	1.1020	
3	Intercept	-32.8111	2.6217	1	0.0000	
	Population	-4.0936	0.3705	1	0.0000	0.0167
	Median Age	0.2142	0.0271	1	0.0000	1.2389
	Median Income	0.0001	0.0000	1	0.0000	1.0001
	Minorities in %	-0.2323	0.0532	1	0.0000	0.7927
	Hispanics in %	0.1137	0.0539	1	0.0348	1.1204
	Average Household Size	8.5353	0.6094	1	0.0000	5091.1159
	Distance to Highway	0.0002	0.0001	1	0.0004	1.0002
Age of Head	0.1713	0.0167	1	0.0000	1.1868	
4	Intercept	-21.1080	1.4310	1	0.0000	
	Population	-0.3909	0.0864	1	0.0000	0.6764
	Median Age	0.1576	0.0151	1	0.0000	1.1706
	Median Income	0.0001	0.0000	1	0.0000	1.0001
	Minorities in %	-0.0564	0.0179	1	0.0016	0.9451
	Hispanics in %	-0.0305	0.0205	1	0.1357	0.9699
	Average Household Size	5.9174	0.3884	1	0.0000	371.4462
	Distance to Highway	0.0000	0.0001	1	0.9594	1.0000
Age of Head	0.0743	0.0092	1	0.0000	1.0771	
5	Intercept	-16.6966	1.3311	1	0.0000	
	Population	-0.3195	0.0725	1	0.0000	0.7265
	Median Age	0.1195	0.0150	1	0.0000	1.1269
	Median Income	0.0001	0.0000	1	0.0000	1.0001
	Minorities in %	-0.0387	0.0157	1	0.0137	0.9620
	Hispanics in %	0.0004	0.0179	1	0.9809	1.0004
	Average Household Size	4.8590	0.3622	1	0.0000	128.8943
	Dist. to Highway	0.0001	0.0001	1	0.3375	1.0001
Age of Head	0.0251	0.0087	1	0.0037	1.0254	
6	Intercept	-25.7086	1.8542	1	0.0000	
	Population	-1.1585	0.1451	1	0.0000	0.3139
	Median Age	0.1579	0.0190	1	0.0000	1.1710
	Median Income	0.0001	0.0000	1	0.0000	1.0001
	Minorities in %	-0.0750	0.0301	1	0.0127	0.9277
	Hispanics in %	-0.0373	0.0349	1	0.2852	0.9634
	Average Household Size	6.2972	0.4645	1	0.0000	543.0533
	Distance to Highway	0.0002	0.0001	1	0.0082	1.0002
Age of Head	0.1334	0.0118	1	0.0000	1.1427	

Table A.1.: Block group based estimation (see 4.3)

A. Parameter Estimates

# of Case		β	Std. Error	DoF	Significance	Exp(β)
1	Intercept	-4.250	0.295	1	0.000	
	Population	0.000	0.000	1	0.068	1.000
	Median Income	0.000	0.000	1	0.000	1.000
	Minorities in %	-0.041	0.004	1	0.000	0.960
	Hispanics in %	0.036	0.005	1	0.000	1.037
	Average Household Size	1.020	0.107	1	0.000	2.773
	Distance to Highway	0.000	0.000	1	0.000	1.000
	Children per Household	-0.239	0.096	1	0.013	0.788
Cars per Household	1.099	0.122	1	0.000	3.001	
2	Intercept	-5.082	0.323	1	0.000	
	Population	0.000	0.000	1	0.180	1.000
	Median Income	0.000	0.000	1	0.000	1.000
	Minorities in %	-0.065	0.006	1	0.000	0.937
	Hispanics in %	0.033	0.007	1	0.256	1.033
	Average Household Size	1.271	0.113	1	0.000	3.563
	Distance to Highway	0.000	0.000	1	0.000	1.000
	Children per Household	-0.336	0.103	1	0.001	0.715
Cars per Household	0.788	0.129	1	0.000	2.200	
3	Intercept	-5.704	0.304	1	0.000	
	Population	0.000	0.000	1	0.920	1.000
	Median Income	0.000	0.000	1	0.000	1.000
	Minorities in %	-0.188	0.006	1	0.000	0.828
	Hispanics in %	0.167	0.007	1	0.000	1.181
	Average Household Size	1.617	0.108	1	0.000	5.038
	Distance to Highway	0.000	0.000	1	0.000	1.000
	Children per Household	-0.500	0.097	1	0.006	0.607
Cars per Household	1.668	0.124	1	0.000	5.301	

Table A.2.: Grid cell based estimation (see 4.4.1)

# of Case		β	Std. Error	DoF	Significance	Exp(β)
1	Intercept	1.196	0.177	1	0.000	
	Population	-0.076	0.177	1	0.000	0.927
	Median Income	0.000	0.000	1	0.000	1.000
	Minorities in %	-0.057	0.005	1	0.000	0.945
	Hispanics in %	0.030	0.007	1	0.000	1.031
	Average Household Size	1.098	0.161	1	0.000	2.999
	Distance to Highway	0.000	0.000	1	0.889	1.000
	Children per Household	0.737	0.284	1	0.010	2.089
	Cars per Household	-0.851	0.174	1	0.000	0.427
2	Intercept	0.696	0.205	1	0.001	
	Population	-0.114	0.003	1	0.000	0.893
	Median Income	0.000	0.000	1	0.000	1.000
	Minorities in %	-0.079	0.007	1	0.000	0.924
	Hispanics in %	0.010	0.008	1	0.256	1.010
	Average Household Size	2.165	0.176	1	0.000	8.716
	Distance to Highway	0.000	0.000	1	0.929	1.000
	Children per Household	-0.038	0.311	1	0.904	0.963
	Cars per Household	-1.694	0.187	1	0.000	0.184
3	Intercept	2.354	0.187	1	0.000	
	Population	-0.283	0.004	1	0.000	0.754
	Median Income	0.000	0.000	1	0.000	1.000
	Minorities in %	-0.158	0.007	1	0.000	0.854
	Hispanics in %	0.095	0.008	1	0.000	1.100
	Average Household Size	1.650	0.171	1	0.000	5.205
	Distance to Highway	0.000	0.000	1	0.000	1.000
	Children per Household	1.567	0.303	1	0.006	4.792
	Cars per Household	-1.272	0.188	1	0.000	0.280

Table A.3.: Grid cell based estimation (see 4.4.2)

A. Parameter Estimates

Grid ID	Av. # Cars	Av. # Chl.	Av. Inc.	Av. Hh Size	Av. # Hh	His. [%]	Min. [%]	Dist. to Hw	P_{41}	P_{42}	P_{43}	C
472100	1.6	0.4	77129	2.2	14.0	7.1	0.0	2815	0.15	0.11	0.74	3
472101	1.8	0.4	74806	2.4	13.0	7.7	0.0	2906	0.13	0.10	0.76	3
472102	1.8	0.2	70916	2.2	10.0	0.0	0.0	3001	0.15	0.15	0.70	3
472106	1.3	0.0	38533	2.3	3.0	0.0	0.0	3424	0.07	0.16	0.77	3
472107	1.5	0.6	64150	2.3	14.0	0.0	7.1	3537	0.31	0.23	0.45	3
472108	1.5	0.3	104792	2.1	14.0	0.0	0.0	3653	0.13	0.14	0.73	3
472111	1.4	0.3	36766	1.9	19.0	5.3	15.8	4015	0.59	0.23	0.07	1
472112	1.4	0.1	45903	1.7	14.0	14.3	7.1	4140	0.34	0.16	0.49	3
472113	1.4	0.7	36380	2.1	69.0	30.4	5.8	4266	0.42	0.02	0.00	4
472119	1.6	0.6	42934	2.2	62.0	19.4	6.5	4643	0.47	0.03	0.00	4
472120	1.4	0.6	35519	2.2	50.0	18.0	4.0	4700	0.65	0.09	0.00	1
472122	2.0	0.0	60000	1.0	1.0	0.0	0.0	4825	0.08	0.03	0.89	3
472129	1.5	0.4	51161	1.9	54.0	9.3	5.6	4554	0.53	0.06	0.00	1
472130	1.4	0.3	45565	1.7	65.0	4.6	16.9	4478	0.14	0.01	0.00	4
472262	1.3	0.0	54933	2.0	3.0	0.0	0.0	5161	0.05	0.09	0.85	3
472263	1.8	0.8	98599	2.8	17.0	0.0	0.0	5262	0.13	0.13	0.74	3
472264	2.3	0.7	63467	2.7	3.0	0.0	0.0	5366	0.03	0.03	0.94	3
472269	1.4	0.8	79259	2.7	16.0	6.3	0.0	5896	0.10	0.11	0.79	3
472270	1.8	1.0	74700	3.0	5.0	0.0	0.0	6006	0.02	0.03	0.96	3
472271	2.5	0.5	36950	3.5	2.0	0.0	0.0	6117	0.03	0.06	0.91	3
472272	2.0	0.0	90300	2.0	1.0	0.0	0.0	6227	0.02	0.03	0.95	3
472276	2.0	0.7	62500	2.7	3.0	0.0	0.0	6669	0.02	0.03	0.95	3
472284	2.5	0.5	112360	2.5	2.0	0.0	0.0	7554	0.01	0.01	0.98	3
472285	1.9	1.3	90766	3.3	7.0	14.3	14.3	7667	0.02	0.01	0.97	3
473196	2.2	0.8	82924	2.6	5.0	20.0	0.0	1906	0.01	0.01	0.98	3
473198	2.0	0.0	54900	2.0	1.0	0.0	0.0	1910	0.06	0.07	0.87	3
473199	3.0	2.0	50100	4.0	1.0	0.0	0.0	1913	0.01	0.01	0.99	3
473200	2.0	0.0	63000	3.0	1.0	0.0	0.0	1925	0.03	0.10	0.87	3
473241	1.5	0.2	55333	2.0	6.0	0.0	16.7	1739	0.32	0.26	0.40	3
473242	2.1	0.0	110440	2.0	10.0	0.0	30.0	1887	0.56	0.24	0.19	1
473244	2.2	0.0	57053	1.8	6.0	0.0	0.0	2120	0.17	0.11	0.72	3
473245	2.1	0.3	82871	2.3	7.0	0.0	0.0	2123	0.08	0.07	0.84	3
473246	2.5	0.0	74700	2.0	2.0	0.0	0.0	2124	0.06	0.05	0.89	3
473247	1.7	0.0	71889	2.0	7.0	0.0	0.0	2125	0.12	0.13	0.75	3
473248	2.0	0.0	23500	2.0	1.0	0.0	0.0	2123	0.11	0.11	0.77	3
473249	1.8	0.0	51808	2.0	5.0	0.0	0.0	2126	0.12	0.13	0.74	3
473250	1.7	0.2	72047	2.1	11.0	0.0	18.2	2127	0.45	0.28	0.25	1
473251	1.4	0.3	64232	2.2	15.0	0.0	0.0	2128	0.26	0.27	0.46	3
473252	2.0	0.3	78918	2.3	16.0	6.3	6.3	2133	0.36	0.18	0.45	3
473253	1.6	0.0	45176	1.8	5.0	0.0	0.0	2137	0.14	0.14	0.72	3
473254	2.1	0.3	66171	2.3	7.0	0.0	0.0	2142	0.11	0.09	0.79	3
473255	1.7	0.0	64977	1.8	6.0	0.0	0.0	2146	0.12	0.12	0.76	3
473256	1.5	0.0	53405	2.0	10.0	0.0	10.0	2150	0.34	0.32	0.32	1
473257	1.8	0.1	47980	1.9	18.0	0.0	0.0	2155	0.48	0.29	0.19	1
473258	1.8	0.2	67959	2.1	21.0	4.8	0.0	2160	0.46	0.26	0.27	1
473259	1.8	1.0	77715	3.2	13.0	23.1	15.4	2168	0.10	0.06	0.84	3
473260	1.9	1.2	79560	3.2	20.0	15.0	15.0	2184	0.34	0.16	0.49	3
473261	2.1	0.9	101395	3.2	14.0	0.0	21.4	2203	0.33	0.28	0.39	3
473262	2.1	1.1	98853	2.8	13.0	23.1	0.0	2231	0.02	0.01	0.97	3
473280	1.5	0.3	48148	2.8	4.0	0.0	25.0	2671	0.26	0.40	0.33	2
473281	2.2	1.6	79157	3.4	18.0	22.2	16.7	2675	0.17	0.05	0.78	3

Table A.4.: Prediction results for sample grid cells

Bibliography

- [AA99] Gennady L. Andrienko and Natalia V. Andrienko. Interactive Maps for Visual Data Exploration. *International Journal Geographic Information Science*, 13(4):355–374, June 1999.
- [Agr02] A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc., New York, 2002.
- [BAL85] M. E. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, Ma, 1985.
- [Bat05] Michael Batty. *Cities and Complexity*. Blackwell Publishing, September 2005. ISBN 0262025833.
- [BBM96] R. J. Beckman, K. A. Baggerly, and M. D. McKay. Creating Synthetic Baseline Populations. *Transportation Research. Part A, Policy and Practice*, 30(6):415–429, 1996.
- [BD05] Henrik Buchholz and Jürgen Döllner. Visual Data Mining in Large-Scale 3D City Models. In *Proceedings of the II International Conference and Exhibition on Geographic Information*, Estoril Congress Center, 2005.
- [Ber67] Jacques Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1967. ISBN 0299090604.
- [Bre05] Claus Brenner. Building Reconstruction from Images and Laser Scanning. *International Journal of Applied Earth Observation and Geoinformation*, 6(3-4):187–198, March 2005.
- [BT04] Itzhak Benenson and Paul M. Torrens. *Geosimulation: Automata-Based Modeling of Urban Phenomena*. John Wiley & Sons, July 2004. ISBN 0-470-84349-7.
- [BW04] Alan Borning and Paul Waddell. Integrated Land Use, Transportation, and Environmental Simulation: UrbanSim Project

- Highlights. In *dg.o '04: Proceedings of the 2004 Annual National Conference on Digital Government Research*, pages 1–2. Digital Government Research Center, 2004.
- [BWF06] Alan Borning, Paul Waddell, and Ruth Fórster. Urban-Sim: Using Simulation to Inform Public Deliberation and Decision-Making. In Hsinchun Chen, editor, *Digital Government: Advanced Research and Case Studies*. Springer-Verlag, 2006. in press.
- [BX94] M. Batty and Y. Xie. From Cells To Cities. *Environment And Planning B-Planning & Design*, 21:31–38, 1994.
- [CdSF05] António Fernando Coelho, António Augusto de Sousa, and Fernando Nunes Ferreira. Modelling Urban Scenes for LBMS. In *Proceedings of the tenth international conference on 3D Web technology*, pages 37 – 46, 2005. ISBN 1-59593-012-4.
- [CR75] H. Chernoff and M. H. Rizvi. Effect on Classification Error or Random Permutations of Features in Representing Multivariate Data by Faces. *Journal of American Statistical Association*, 70:548–554, 1975.
- [dSM06] Luiz Gonzaga da Silveira and Soraia Raupp Musse. Real-time Generation of Populated Virtual Cities. In *VRST '06: Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, pages 155–164, New York, NY, USA, 2006. ACM Press. ISBN 1-59593-321-2.
- [DZT04] Paul DiLorenzo, Victor B. Zordan, and Duong Tran. Interactive Animation of Cities Over Time. In *17th International Conference on Computer Animation and Social Agents*, Geneva, Switzerland, 2004.
- [Ebe96] David S. Ebert. Advanced Modeling Techniques for Computer Graphics. *ACM Computing Surveys*, 28(1):153–156, March 1996.
- [Esr07] Esri. ArcGIS: The Complete Enterprise GIS. Available online at: <<http://www.esri.com/software/arcgis/>>, 2007. (Last Accessed: 12-03-2007).
- [FAA⁺01] David Fairbairn, Gennady Andrienko, Natalia Andrienko, Gerd Buziek, and Jason Dykes. Representation and its

- Relationship with Cartographic Visualization: A Research Agenda. *Cartography and Geographic Information Science*, 28(1):1–29, 2001.
- [För99] Wolfgang Förstner. 3D City Models: Automatic and Semi-automatic Acquisition Models. In D. Fritsch and R. Spiller, editors, *Photogrammetric Week '99*, pages 291–303, Heidelberg, 1999. Wichmann Verlag.
- [FS04] Georg Fuchs and Heidrun Schumann. Visualizing Abstract Data on Maps. *IV '00*, 00:139–144, 2004.
- [GGMZ05] Diansheng Guo, Mark Gahegan, Alan M. MacEachren, and Biliang Zhou. Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach. *Cartography and Geographic Information Science*, 32(2):113–132, April 2005.
- [GHK⁺07] Subhrajit Guhathakurta, Janet Holston, Yoshi Kobayashi, Tim Lant, and Mookesh Patel. Digital Phoenix Project. Available online at: <<http://www.digitalphoenix-asu.net>>, 2007. (Last Accessed: 10-21-2007).
- [GMB06] Kevin R. Glass, Chantelle Morkel, and Chaun D. Bangay. Duplicating Road Patterns in South African Informal Settlements Using Procedural Techniques. In Stephen N. Spencer, editor, *Proceedings of the 4th International Conference on Virtual Reality, Computer Graphics, Visualisation and Interaction in Africa, Afrigraph 2006*, Cape Town, South Africa, January 2006. ACM. ISBN 1-59593-288-7.
- [Goo07a] Google Inc. Google Earth KML 2.1 Reference. Available online at: <<http://code.google.com/apis/kml/documentation>>, 2007. (Last Accessed: 12-08-2007).
- [Goo07b] Google Inc. Google Earth website. Available online at: <<http://earth.google.com/>>, 2007. (Last Accessed: 10-23-2007).
- [GPSL03a] Stefan Greuter, Jeremy Parker, Nigel Stewart, and Geoff Leach. Real-time Procedural Generation of ‘Pseudo Infinite’ Cities. In *Proceedings of the 1st international conference on Computer graphics and interactive techniques in Australasia*

- and South East Asia*, Melbourne, Australia, February 2003. ACM Press. ISBN 1-58113-578-5.
- [GPSL03b] Stefan Greuter, Jeremy Parker, Nigel Stewart, and Geoff Leach. Undiscovered Worlds - Towards a Real Time Procedural World Generation Framework. In *5th International Digital Arts and Culture Conference in Melbourne, Australia*. RMIT, May 2003.
- [GSL04] Stefan Greuter, Nigel Stewart, and Geoff Leach. Beyond the Horizon: Computer-generated, Three-dimensional, Infinite Virtual Worlds without Repetition. In *Image Text and Sound Conference 2004*. RMIT University, 2004.
- [Guh03] Subhrajit Guhathakurta. *Integrated Urban and Environmental Models: A Survey of Current Research and Applications*. Springer-Verlag, New York, NY, 2003.
- [HJF06] Gobe Hobona, Philip James, and David Fairbairn. Web-based Visualization of 3D Geospatial Data Using Java3D. *IEEE Computer Graphics and Applications*, 26(4):28–32, Jul/Aug 2006.
- [HMFN04] Masanobu Honda, Kaznori Mizuno, Yukio Fukui, and Seiichi Nishihara. Generating Autonomous Time-Varying Virtual Cities. In *Third International Conference on Cyberworlds (CW'04)*, pages 45–52, 2004.
- [HYN03] Jinhui Hu, Suya You, and Ulrich Neumann. Approaches to Large-Scale Urban Modeling. *Computer Graphics and Applications, IEEE*, 23(6):62–69, November 2003.
- [Imr07] Imran Haque. gCensus. Available online at: <http://gencensus.stanford.edu/gcensus/index.html>, 2007. (Last Accessed: 12-28-2007).
- [JGK⁺06] Himanshu Joshi, Subhrajit Guhathakurta, Goran Konjevod, John Crittenden, and Ke Li. Simulating Impact of Light Rail on Urban Growth in Phoenix: an Application of Urbansim Modeling Environment. In *Proceedings of the 7th Annual International Conference on Digital Government Research, DG.O 2006*, pages 135–141, 2006.
- [KB01] M. Kwartler and R.N. Bernard. CommunityViz: an Inte-

- grated Planning Support System. In Richard K. Brail and Richard E. Klosterman, editors, *Planning Support Systems: Integrating Geographic information Systems, Models, and Visualization Tools*, pages 285–308, Redlands, CA: Environmental Systems Research Institute and New Brunswick, NJ: Rutgers Center for Urban Policy Research, 2001. ESRI Press. ISBN 978-1589480117.
- [Klo01] Richard E. Klosterman. The What if? In Richard K. Brail and Richard E. Klosterman, editors, *Planning Support Systems: Integrating Geographic information Systems, Models, and Visualization Tools*, pages 263–284, Redlands, CA: Environmental Systems Research Institute and New Brunswick, NJ: Rutgers Center for Urban Policy Research, 2001. ESRI Press. ISBN 978-1589480117.
- [KOO⁺98] Nobuko Kato, Tomoe Okuno, Aya Okano, Hitoshi Kanoh, and Seiichi Nishihara. An Alife Approach to Modeling Virtual Cities. *IEEE International Conference on Systems, Man, and Cybernetics*, 2:1168–1173, October 1998.
- [Kra06] Menno-Jan Kraak. Visualization Viewpoints: Beyond Geovisualization. *IEEE Computer Graphics and Applications*, 26(4):6–9, Jul/Aug 2006.
- [LGMR05] Paul A. Longley, Michael F. Goodchild, David J. Maguire, and David W. Rhind. *Geographic Information Systems and Science*. John Wiley and Sons, 2005. ISBN 0470870001.
- [Lin68] Aristid Lindenmayer. Mathematical Models for Cellular Interaction in Development, parts i and ii. *Journal of Theoretical Biology*, 18:280–315, March 1968.
- [LWR⁺04] Thomas Lechner, Ben Watson, Pin Ren, Uri Wilensky, Seth Tisue, and Martin Felsen. Procedural Modeling of Land Use in Cities. Technical Report NWU-CS-04-38, Northwestern University, August 2004.
- [LWWF03] Tom Lechner, Ben A. Watson, Uri Wilensky, and Martin Felsen. Procedural City Modeling. In *1st Midwestern Graphics Conference*, St. Louis, MO, 2003.
- [Mac95] A. M. MacEachren. *How Maps Work: Representation, Visu-*

- alization, and Design*. Guilford Press, 1995.
- [Mac98] Alan M. MacEachren. Visualization - Cartography for the 21st century. In *Proceedings of the 7th Annual Conference of Polish Spatial Information Association*, Warsaw, Poland, May 1998.
- [Mac01] Alan M. MacEachren. An Evolving Cognitive-Semiotic Approach to Geographic Visualization and Knowledge Construction. *Information Design Journal*, 10(1):26–36, May 2001.
- [McF73] Daniel McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1973.
- [McF76] Daniel McFadden. Properties of the Multinomial Logit (MNL) Model. Urban Travel Demand Forecasting Project Institute of Transportation Studies, Working Paper No. 7617, 1976.
- [McF97] Daniel McFadden. Modelling the Choice of Residential Location. *The Economics of Housing*, 1:531–552, 1997.
- [MGH⁺08] Ariane Middel, Subhrajit Guhathakurta, Hans Hagen, Peter-Scott Olech, and Florian Hoepel. Visualizing Future 3-Dimensional Neighbourhoods in Phoenix: An Application Incorporating Empirical Methods with Computational Graphics. In *Virtual Geographic Environments*, Hong Kong, 2008. accepted.
- [MGP⁺04] Alan M. MacEachren, Mark Gahegan, William Pike, Isaac Brewer, Guoray Cai, Eugene Lengerich, and Frank Hardisty. Geovisualization for Knowledge Construction and Decision Support. *IEEE Computer Graphics and Applications*, 24(1): 13–17, Jan/Feb 2004.
- [Mid06] Ariane Middel. Procedural 3D Modeling of Cityscapes. In Hans Hagen, Andreas Kerren, and Peter Dannenmann, editors, *Visualization of Large and Unstructured Data Sets*, pages 133–142, Kaiserslautern, 2006. GI-Edition, Lecture Notes in Informatics (LNI), Seminars Vol. S-4. ISBN 978-3-88579-438-7.

-
- [Mid07a] Ariane Middel. A Framework for Visualizing Multivariate Geodata. In *Visualization of Large and Unstructured Data Sets*, Kaiserslautern, 2007. GI-Edition, Lecture Notes in Informatics (LNI). in press.
- [Mid07b] Ariane Middel. Estimating Residential Building Types from Demographic Information at a Neighborhood Scale. In *Sustainable Planning*. Springer, 2007.
- [MK01] Alan M. MacEachren and M.-J. Kraak. Research Challenges in Geovisualization. *Cartography and Geoinformation Science*, 28(1):3–12, 2001.
- [MPJ04] Kwan Mei-Po and Lee Jiyeong. Geovisualization of Human Activity Patterns Using 3D GIS: A Time-Geographic Approach. In Michael F. Goodchild M and Donald G. Janelle, editors, *Spatially Integrated Social Science: Examples in Best Practice*, pages 48–66, Oxford, 2004. Oxford University Press.
- [MVUG05] Pascal Müller, Tijn Vereenoghe, Andreas Ulmer, and Luc Van Gool. Automatic Reconstruction of Roman Housing Architecture. In Baltsavias et al., editor, *Recording, Modeling and Visualization of Cultural Heritage*, pages 287–297. Balkema Publishers (Taylor & Francis group), 2005.
- [MWH+06] Pascal Müller, Peter Wonka, Simon Haegler, Andreas Ulmer, and Luc Van Gool. Procedural Modeling of Buildings. *Proceedings of ACM SIGGRAPH 2006 / ACM Transactions on Graphics*, 25(3):614–623, 2006.
- [MZWG07] Pascal Müller, Gang Zeng, Peter Wonka, and Luc Van Gool. Image-based Procedural Modeling of Facades. *Proceedings of ACM SIGGRAPH 2007 / ACM Transactions on Graphics*, 26(3), 2007.
- [Nat97] National Imagery & Mapping Agency. Department of Defense World Geodetic System 1984, Its Definition and Relationships With Local Geodetic Systems. Technical Report TR8350.2, July 1997.
- [Nat03] National Research Council. *IT Roadmap to a Geospatial Future*. National Academies Press, 2003. ISBN 0-309-08738-4.
- [NS04] Hauke Neidhart and Monika Sester. Creating a Digital Ther-

- mal Map Using Laser Scanning and GIS. In *Proceedings of the District Heat and Cooling Symposium*, Hannover, Germany, 2004.
- [Ope07] Open Geospatial Consortium Inc. OGC Simple Feature Specification. Available online at: <http://www.opengeospatial.org/standards>, 2007. (Last Accessed: 10-30-2007).
- [PDBB00] L. Denise Pinnel, Matthew Dockrey, A. J. Bernheim Brush, and Alan Borning. Design of Visualizations for Urban Modeling. In *VisSym '00: Joint Eurographics – IEEE TCVC Symposium on Visualization*, Amsterdam, The Netherlands, May 2000.
- [pgA07] pgAdmin Development Team. pgAdmin PostgreSQL Tools. Available online at: <http://www.pgadmin.org>, 2007. (Last Accessed: 12-18-2007).
- [PHP07] PHP Group. PHP website. Available online at: <http://www.php.net>, 2007. (Last Accessed: 12-12-2007).
- [PL90] Przemyslaw Prusinkiewicz and Aristid Lindenmayer. *Algorithmic Beauty of Plants*. Springer-Verlag New York, Inc., 1990. ISBN 0-387-97297-8.
- [PM01] Yoav I. H. Parish and Pascal Müller. Procedural Modeling of Cities. In Eugene Fiume, editor, *Proceedings of ACM SIGGRAPH 2001*, pages 301–308, New York, NY, USA, 2001. ACM Press. ISBN 1-58113-374-X.
- [Pos07] PostgreSQL Global Development Group. PostgreSQL website. Available online at: <http://www.postgresql.org>, 2007. (Last Accessed: 10-23-2007).
- [PTM07] Scott Pezanowski, Brian Tomaszewski, and Alan M. MacEachren. An Open GeoSpatial Standards-Enabled Google Earth Application to Support Crisis Management. In *Lecture Notes in Geoinformation and Cartography, Geomatics Solutions for Disaster Management*, pages 225–238, Sacramento, CA, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-72106-2.
- [PX99] D. A. Powers and Y. Xie. *Statistical Methods for Categorical*

-
- Data Analysis*. Academic Press, 1999.
- [Qua07] Quantum GIS. Quantum GIS Documentation. Available online at: <<http://qgis.org/>>, 2007. (Last Accessed: 12-03-2007).
- [Qui76] J. M. Quigley. Housing Demand in the Short Run: An Analysis of Polytomous Choice. *Explorations in Economic Research*, 3(1):76–102, 1976.
- [Ref05] Refrations Research. PostGIS website. Available online at: <<http://postgis.refrations.net>>, 2005. (Last Accessed: 10-23-2007).
- [SB07] Yael Schwartzman and Alan Borning. The Indicator Browser: A Web-Based Interface for Visualizing UrbanSim Simulation Results. In *HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, page 92, Washington, DC, USA, January 2007. IEEE Computer Society. ISBN 0-7695-2755-8.
- [SBJ+01] T. Slocum, C. Blok, B. Jiang, A. Koussoulakou, D. Montello, S. Fuhrmann, and N. Hedley. Cognitive and Usability Issues in Geovisualization. *Cartography and Geographic Information Science*, 28:61–75, 2001.
- [SCM99] Ben Shneiderman, Stuart K. Card, and Jock D. Mackinlay. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, CA, 1999. ISBN 1558605339.
- [SF03] André Skupin and Sara I. Fabrikant. Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization. *Cartography and Geographic Information Science*, 30(2):95–115, 2003.
- [SH03] André Skupin and Ron Hagelman. Attribute Space Visualization of Demographic Change. In *11th ACM GIS Symposium (GIS'03)*, New Orleans, Louisiana, USA, November 7-8 2003. ACM Press.
- [Ska99] A. Skaburskis. Modelling the Choice of Tenure and Building Type. *Urban Studies*, 36(13):2199–2215, 1999.

- [Sku00] André Skupin. From Metaphor to Method: Cartographic Perspectives on Information Visualization. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVIS 2000)*, pages 91–97, Salt Lake City, UT, USA, October 9-10 2000. Los Alamitos: IEEE Computer Society. ISBN 0-7695-0804-9.
- [SMG01] Erik B. Steiner, Alan M. MacEachren, and Diansheng Guo. Developing and Assessing Light-Weight Data-Driven Exploratory Geovisualization Tools for the Web. In Richard K. Brail and Richard E. Klosterman, editors, *Workshop on Geovisualization for the Web*, Taupo, New Zealand, 2001. ICA Commission on Visualization & Virtual Environments.
- [Ste46] S. Stevens. On the Theory of Scales of Measurement. *Science*, 103:677–680, 1946.
- [Sti80] George Stiny. Introduction to Shape and Shape Grammars. *Environment and Planning B: Planning and Design*, 7:343–351, 1980.
- [Tor06] Paul M. Torrens. Geosimulation and its Application to Urban Growth Modeling. *Complex Artificial Environments*, pages 119–134, 2006.
- [TSK05] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson Addison Wesley, Boston, 2005.
- [TSWS05] Christian Tominski, Petra Schulze-Wollgast, and Heidrun Schumann. 3D Information Visualization for Time Dependent Data on Maps. In *IV '05: Proceedings of the Ninth International Conference on Information Visualisation (IV'05)*, pages 175–181, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2397-8.
- [Tuf90] Edward Tufte. *Envisioning Information*. Graphics Press, 1990. ISBN 0961392118.
- [US 07] US Census Bureau. Census 2000. Available online at: <<http://www.census.gov>>, 2007. (Last Accessed: 11-06-2007).
- [VWvH⁺07] Fernanda B. Viégas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. Many Eyes: A Site for Visu-

- alization at Internet Scale. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVIS 2007)*, Sacramento, CA, 2007.
- [Wad02] Paul Waddell. UrbanSim: Modeling Urban Development for Land Use, Transportation and Environmental Planning. *Journal of the American Planning Association*, 68(3):297–314, 2002.
- [WBAL80] Glen Weisbrod, Moshe Ben-Akiva, and Steven Lerman. Tradeoffs in Residential Location Decisions: Transportation Versus Other Factors. *Transportation Policy and Decision-Making*, 1(1):13–16, 1980.
- [WBN⁺03] Paul Waddell, Alan Borning, Michael Noth, Nathan Freier, Michael Becke, and Gudmundur Ulfarsson. Microsimulation of Urban Development and Location Choices: Design and Implementation of UrbanSim. *Networks and Spatial Economics*, 3(1):43–67, 2003.
- [WDSC07] Jo Wood, Jason Dykes, Aidan Slingsby, and Keith Clarke. Interactive Visual Exploration of a Large Spatio-Temporal Dataset: Reflections on Geovisualization Mashup. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVIS 2007)*, Sacramento, CA, 2007.
- [WWSR03] Peter Wonka, Michael Wimmer, Francois Sillion, and William Ribarsky. Instant Architecture. *Transactions on Graphics, SIGGRAPH 2003*, 22(3):669 – 677, July 2003.
- [YBH⁺02] Chee Yap, Henning Biermann, Aaron Hertzman, Chen Li, Jon Meyer, Hsing-Kuo Pao, and Salvatore Paxia. A Different Manhattan Project: Automatic Statistical Model Generation. In *IS&T SPIE Symposium on Electronic Imaging*, San Jose, California, January 2002.

Curriculum Vitae



Name: Ariane Middel
Date of Birth: October 28, 1977
Place of Birth: Düsseldorf, Germany
Nationality: German

Work Experience and Technical Expertise

- 09/2004–10/2004 Media Tenor, institute for media analysis (Bonn, Germany)
Optimizing the institute's internet presentation, programming dynamic web sites
- 10/2003–09/2004 Freelance IT consultant and digital media designer (Bonn, Germany)
Computer training, IT services, design and implementation of websites
- 10/1999–06/2003 Student research assistant, Institute of Cartography and Geoinformation (University of Bonn, Germany)
Development of multimedia CD-ROMs, digital video editing, supervision of JAVA tutorials, pedestrian navigation and routing in cities by means of videos and SMIL, georeferencing of maps via ArcInfo
- 10/1997–12/1998 Student research assistant, Institute of Theoretical Geodesy (University of Bonn, Germany)
EXCEL training for first-year students
- 07/1997–10/1997 Surveying office "ÖbVI Dipl.-Ing. Eicker" (Haan, Germany)
Cadastral surveying, levelling

Education

- 01/2005–12/2007 Graduate Student of the International Research Training Group “Visualization of Large and Unstructured Data Sets”, funded by the German Science Foundation (TU Kaiserslautern in cooperation with ASU)
PhD in Computer Science, Thesis: Visualization of Urban Futures - A Framework for Visualizing Multidimensional Geospatial Data
- 10/2004–01/2005 Studies in Business Information Systems (University of Applied Sciences Bonn-Rhein-Sieg, Germany)
- 10/1997–10/2003 Dipl.-Ing. in Geodetic Engineering (University of Bonn, Germany)
Majors: Geographic Information Systems, Cartography, and Photogrammetry
Minors: Statistics, Physical and Mathematical Geodesy, Surveying Engineering
Thesis: Virtual Signposts in Panoramic Images
- 08/1988–06/1997 Grammar School “Städt. Gymnasium Haan” (Haan, Germany)