Simulating Residential Heat Consumption with Spatially Referenced Synthetic Microdata

A dissertation
submitted to HafenCity University Hamburg
in fulfillment of the requirements for the degree of
Doctor-Engineer

Dissertation
vorgelegt im Promotionsausschuss der
HafenCity Universität Hamburg
zur Erlangung des akademischen Grades
Doktor-Ingenieur

by/von

Marcelo Esteban Muñoz Hidalgo

of/aus

Quito, Ecuador

First advisor / Erstgutachter: Prof. Irene Peters, Ph.D.

Second advisor / Zweitgutachter: Prof. Robert Tanton, Ph.D.

18. Mai 2016

This manuscript was generated with help of KOMA-Script and LaTeX

To my grandmother **Waltraud** for drawing a fantastic path to follow. To my wife **Denise** for being always there to listen and comment about my latest algorithm. To my daughter **Lea** for forcing me to always take a play-break. To my daughter **Amelie** for always having a smile for me. To my parents **Ariane** and **Fernando** for supporting me all these years, to my father for pushing me to submit my dissertation, to my mother for not pushing me. To my sister **Alexandra** for marching with me against fossil fuels. To my sister **Claudia** for being a fantastic artist and letting me be part of her art. To my "Doktormutter" **Irene** for guiding me throughout this adventure, for all the interesting discussions, for letting me pursue a wide research spectrum and for introducing me to the modeling world, especially to the modeling of complex systems. To my second advisor **Robert** for hosting me at NATSEM, for all the interesting Fridays discussions and for guiding me through the wonderful world of Spatial Microsimulation. To **Yogi** for all the interesting discussions "in between". To the other two musqueteers **Fabio** and **Lukas** for all the fantastic discussions. To **Juan Pablo** for all the support from the other side of the Atlantic. To my fellow PhD student **Christoph** for all his advice on fishing and other matters. To the research assistants: Daniel, Mostafa, Aparajita, Hsiao-Hui, Fabian and Katharina for helping me with my research. To all the people I fail to mention on this short note that have helped me with my work.

An important part of this research was made possible thanks to the support of the *Federal Ministry of Economic Affairs and Energy (BMWi)* in the framework of the research project: *EnEff:IBA - 03ET1081A*. And by the German Academic Exchange Service (DAAD) in the framework of the program: *Year Stipend for Doctoral Students 2014/15 (57044990).* I would like to extend my thanks to both institutions for their economic support.

Where we are coming from:

> "There is an inherent difficulty, if not practical impossibility, in aggregating anything but absurdly simple relationships about elemental decision-making units into comprehensible relationships between large aggregative units such as industries, the household sector, and the government sector."
>
> (Orcutt, 1957)

Where we are heading to:

> "A simple behavioral model, with agents gradually switching to better performing heuristics, explains individual, as well as emergent, macro behavior in these laboratory economies."
>
> (Battiston et al., 2016)

Abstract:

Motivation for the thesis is the need for coordinating urban planning and energy planning. The emissions of the building sector can be reduced by the spatial coordination of building energy demand and building energy supply, e.g. via the spatial prioritization of building energy retrofits, the creation of heating grids and the integration of waste heat sources into the supply. As heat transport is associated with losses, the spatial constellation of heat sinks and sources are highly relevant for supply concepts. That is why a better understanding of the spatial distribution of heat consumption is needed – that is, real heat consumption, not computed heat demand based on physical building characteristics. Heat consumption can differ as much as 50% or more from pre-calculated heat demands. It is the building users and their behavior and presence times that account for this spread. As heat supply and electric power supply are increasingly integrated, and as there is a trend towards smaller heating grids, the temporal distribution of urban heat demand, too, becomes more interesting. There is little systematic analysis of this as of yet. This is where the present dissertation is making a contribution?

Aim of the thesis is to develop a model capable of simulating heat demand at the micro scale (i.e. for individual buildings) in its spatial-temporal spread for large urban agglomerations. This requires: (1) individual building models, (2) the explicit consideration of human behavior in estimating heat consumption and (3) the simulation of a large number of different buildings defining an urban area.

Towards this objective, the thesis uses **methods from different disciplines :** (a) **spatial microsimulation** (microsimulation originated in economics and has been taken up in geography and the social sciences which enriched it with the spatial dimension), (b) **building simulation** which traces energy flows in buildings (traditionally anchored in building physics and technical thermodynamics) and (c) the **classification of the building stock, and allocation of types onto buildings** in various data sets (practiced by many authors, e. g. in engineering and economics, for evaluation of policy measures addressing the building stock).

The main result of this thesis is a model which can (1) generate a synthetic building stock with households residing in buildings (whereby "synthetic" does not mean "invented", but rather, satisfying data on buildings and socio-demographics available at a certain (low) level of aggregation, but not at the micro-level) (2) simulate heat demand of individual buildings at high temporal resolution, taking explicit consideration of socio-demographic characteristics of building users. The model developed in the thesis represents a simple urban environment containing geo-referenced buildings enriched with energy relevant parameters. The synthetic households which are allocated to the buildings do match, in sum and distribution of socio-demographic parameters, household data available at a certain level of aggregation. These household data are enriched with national German time use survey data in order to specify activities and occupancy times of household members.

The model is validated internally, meaning that the results of the mechanism which generates the synthetic stock and the synthetic population is checked against the aggregated data on which the building and population allocation process started. The primary metric used for this validation is the *Total Absolute Error* (TAE). It was not possible, at this point, to validate the mechanism externally, as there were too few heat consumption data for this purpose. Consumption data, however, are urgently needed in this field, and there is work in the scientific community to produce this type of data and make it accessible to the research community.

Using the spatial microsimulation approach made it possible, in this thesis, to project the building stock into the year 2030, without major effort on data collection or model structure. The same applies for scaling the model to the national level, which was also done in this thesis. The applicability of the model to other geographical areas and research questions (e. g., concerning water consumption or building sanitation technology) seems promising.

Keywords: Spatial Microsimulation, Digital Cadastre, Building Simulation, Building Stock Models, Urban Heat Demand, User Behavior, Activity Based Models

Zusammenfassung:

Anlass der Arbeit ist die Notwendigkeit der Koordination von städtebaulicher und Energieplanung. Die Emissionen des Gebäudesektors können durch räumliche Koordination von Energienachfrage und –angebot gesenkt werden, z.B. durch die räumliche Priorisierung energetischer Sanierung, die Schaffung von Wärmeverbünden und die Einbindung von Abwärmequellen. Wegen der mit Wärmetransport verbundenen Verluste ist die räumliche Konstellation von Quellen und Senken für die Wärmeversorgung hoch relevant. Deshalb braucht es ein besseres Verständnis der räumlichen Verteilung der Wärmenachfrage, d.h. nicht der **Bedarfe** (ein terminus technicus für die mit Hilfe bautechnischer Information vorausberechnete Nachfrage), sondern der tatsächlichen **Verbräuche**, die von Bedarfen um oft mehr als 50% differieren und im Wesentlichen von Nutzern bestimmt sind. Mit der zunehmenden Verzahnung von Strom- und Wärmesektor und der Schaffung kleinerer Wärmeverbünde wird nicht nur das räumliche, sondern auch das zeitliche Muster der Wärmverbräuche zunehmend interessant. Hierzu gibt es noch kaum systematische Analysen. In diese Lücke stößt die Arbeit vor.

Ziel der Arbeit ist die Entwicklung eines Modells, das den Wärmeverbrauch auf der Mikro-ebene (d.h. für einzelne Gebäude) in seiner räumlich-zeitlichen Verteilung für große städtische Agglomerationen simulieren kann. Dies erfordert: (1) Modelle einzelner Gebäude, (2) die explizite Berücksichtigung des menschlichen Verhaltens in der Bestimmung der Wärmenachfrage und (3) die Simulation einer großen Anzahl unterschiedlicher Gebäude, die ein Stadtgebiet definieren.

Dafür kombiniert die Arbeit **Methoden aus mehreren Disziplinen:** (a) die **räumliche Mikrosimulation** (in den Wirtschaftswissenschaften entstanden, heute auch von Geografie und Sozialwissenschaften vorangetrieben), (b) die **Gebäudesimulation**, die Energieflüsse im einzelnen Gebäude nachzeichnet (traditionell in Bauphysik und technischer Thermodynamik verankert) und (c) die **Typisierung des Gebäudebestands und die Allokation der Typen auf Einzelgebäude**, die von vielen Autoren (meist aus den Ingenieur- und Wirtschaftswissenschaften) benutzt wird, um Politikmaßnahmen für den Gebäudesektor zu evaluieren.

Das wichtigste Ergebnis dieser Arbeit ist ein Modell, das (1) einen synthetischen geo-refenzierten Gebäudebestand mit darin wohnenden Haushalten erzeugen kann („synthetisch“ heißt nicht etwa „frei erfunden“, sondern Daten über Gebäude und Soziodemografie genügend, die auf relativ niedrigem räumlichen Niveau, aber nicht immer auf Mikro-ebene vorliegen) und (2) die Wärmenachfrage der einzelnen Gebäude mit einer hohen zeitlichen Auflösung unter Berücksichtigung von Haushaltscharakteristika simulieren kann. Das in der Arbeit entwickelte Modell stellt eine einfache städtische Umgebung mit Daten zu einzelnen Gebäuden dar, angereichert mit energetisch relevanten Parametern. Die synthetischen Haushalte, die auf die Gebäude alloziiert werden, entsprechen in Summe und Verteilung vorhandenen soziodemografischen Daten auf höherer Aggregationsebene. Diese Daten werden mit der bundesweiten deutschen Zeitbudgeterhebung angereichert, um Aktivitäten und vor allem Aufenthaltszeiten der Haushaltsmitglieder im Gebäude zu beschreiben.

Das Modell ist intern validiert, d.h. der Mechanismus, der den synthetischen Gebäudebestand und die synthetische Population erzeugt, wird in seiner Wirkung mit den aggregierten Ausgangsdaten verglichen. Das primäre Maß für diese Validierung ist der *Total Absolute Error* (TAE). – Eine externe Validierung des Modells war aufgrund fehlender Verbrauchsdaten in der Breite noch nicht möglich. Die Forschergemeinde arbeitet jedoch daran, solche Datenbanken zu erzeugen und zur Verfügung zu stellen.

Die Nutzung der Methode der *Spatial Microsimulation* ermöglichte die Projektion des Gebäudebestands ins Jahr 2030 ohne großen modell- und datentechnischen Aufwand. Dasselbe gilt für die in dieser Arbeit vorgenommene Skalierung des Modells auf die nationale Ebene. Die Übertragbarkeit des Modells auf andere geografische Gebiete sowie seine Anwendung auf andere Forschungsfragen (z.B. Wasserverbräuche oder Sanitärtechnik) scheint vielversprechend.

Stichworte: Räumliche Mikrosimulation, Digitale Kataster, Gebäudesimulation, Modellierung des Gebäudebestands, Städtischer Wärmeverbrauch, Gebäudenutzerverhalten, activity-based Modell

Contents

List of Figures

List of Tables

HCU | HafenCity Universität Hamburg

Die Vorsitzende des Promotionsausschusses
Überseeallee 16
D-20457 Hamburg
Telefon:040 42827-4585

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbständig sowie ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die von der Deutschen Forschungsgemeinschaft (DFG) festgelegten Standards guter wissenschaftlicher Praxis wurden eingehalten. Bei der Auswahl und Auswertung folgenden Materials haben mir die nachstehend aufgeführten Personen in der jeweils beschriebenen Weise entgeltlich/unentgeltlich geholfen:

1. ...
2. ...

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich hierfür nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder anderer Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.
Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.
Vor Aufnahme der obigen Versicherung an Eides statt wurde ich über die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung belehrt.

....
(Ort, Datum)   (Unterschrift)

Unterschrift des die Versicherung an Eides statt aufnehmenden Beamten:

....
(Ort, Datum)   (Unterschrift)

BELEHRUNG

§156 StGB: Falsche Versicherung an Eides Statt

Wer vor einer zur Abnahme einer Versicherung an Eides Statt zuständigen Behörde eine solche Versicherung falsch abgibt oder unter Berufung auf eine solche Versicherung falsch aussagt, wird mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft.

§161 StGB: Fahrlässiger Falscheid; fahrlässige falsche Versicherung an Eides Statt

(1) Wenn eine der in den §§ 154 bis 156 bezeichneten Handlungen aus Fahrlässigkeit begangen worden ist, so tritt Freiheitsstrafe bis zu einem Jahr oder Geldstrafe ein.

(2) Straflosigkeit tritt ein, wenn der Täter die falsche Angabe rechtzeitig berichtigt. Die Vorschriften des § 158 Abs. 2 und 3 gelten entsprechend.

1 The Need for a Multidisciplinary Approach to Simulate Urban Heat Consumption

The development of integrated, activity based urban models able to simulate complex systems through emergent processes is the umbrella under which this thesis has been developed. The presented method aims to expand some of the methods developed by Peters, Brassel und Spörri (2002).

The focus of this thesis is on the simulation of heat demand, the topic of energy in Germany is high positioned on its national research agenda. The scope of this thesis is the interaction between the development of urban agglomerations and the needed energy supply infrastructure supporting this development. Germany has a long tradition of using district heating networks for the provision of heat, providing a clear link between urban development and the planning of heat supply infrastructures.

During the performed research we developed many simulation libraries, two of them build the core of this thesis. The first library is an implementation of the German DIN norm regulating the computation of heat demand (Muñoz H., 2015; Muñoz H., Dochev, Seller & Peters, 2016), the second library is an implementation of the GREGWT algorithm used for the generation of synthetic populations (Muñoz H., Vidyattama & Tanton, 2015a; Muñoz H., Tanton & Vidattama, 2015), commonly used for the development of spatial microsimulation models.

On this thesis we show how we combined these two methods in order to generate a geo-referenced synthetic building stock. This synthetic building stock describes the physical properties of the buildings and the residents living on them. One of the innovations of the model is its ability to benchmark the synthetic building stock to multiple aggregation levels.

With this model we have been able to compute the heat demand at a micro level for the entire country. This method has proved to be robust and fast. We have also experimented with thermal dynamic models for the estimation of heat demand with a high temporal resolution. The integration of user behavior into heat demand models required a more elaborated heat demand model, for this we made use of the well established EnergyPlus model, feeding it with occupational patterns of households. This integration to a thermal-dynamic model allow us to simulate heat demand of Hamburg at a 15 minute resolution. With help of the GREGWT library we also projected the population (aging) and the building stock (retrofits) into the future for the analysis of urban heat densities in the city of Hamburg.

For the future development of models we want to integrate external data sources for the simulation of agents interaction. In order to create an activity based model we have used the national time-use survey, we believe that by using other type of big-data-sets like cellphone usage or tweeter data we can develop more realistic activity based urban models for the simulation of many urban phenomena.

The biggest challenge of this thesis has been the incorporation of different methods coming from different disciplines. This incorporation of different method was essential for the development of the model presented in this thesis. The main aim of this thesis is to develop a *heat consumption model for large urban agglomerations at a low aggregation level.* In the following section we will explain in detail what this means and what kind of methods where used to achieve this goal. After this explanation we will describe the main structure of the thesis and how this structure fits with the model architecture and its underlying methods.

Occupant behavior: The small distinction between **consumption** and **demand** is normally used to distinguish between estimated demand and monitored consumption (see Sub-Section 2.1.2 for a more detailed differentiation between these two definitions). This distinction is very important when talking about the estimation of heat demand because current building simulation models can't simulate the interaction of human behavior (see Sub-section2.1.3 for an overview of heat demand models integrating human behavior in their simulation routines). In the last decade the simulation of human interaction with the building has gained attention in the building simulation community. This is because new building regulations around the globe are making new constructed buildings more energy efficient. The more efficient a building becomes the higher its sensitivity towards occupant influence becomes, making it essential to include user behavior into energy simulation models.

The integration of user influence within the developed model emerge during the initial development of the model. With a detail description of the synthetic population and the link to the building stock the ability to simulate the influence of users on heat demand at a low aggregation level was evident. In order to capture this influence we enriched the used survey for the creation of synthetic families with a time-used survey. By enriching the German micro census with a time use survey we were able to describe a detail schedule of every individual in my model. This information is used to generate building specific schedules which are used as input in a thermal simulation model.

Urban model: The developed model in the framework of this thesis aims to simulate heat consumption of **urban areas**. This means that the model has to be able to simulate many buildings with: (a) as little input as possible, because data regarding individual buildings of urban areas is hard to retrieve; (b) integrate user interaction with the buildings of the building stock; and (c) maintain the spatial reference throughout the computation process of the model.

A major problem with this data intensive model structure was the proper data abstraction. The model had to deal not only with large amounts of data but with complex data structures. The interpretation of building geometries, needed for the computation of heat demand within a thermal simulation model, needed to be simplified in order to achieve; (a) manageable computational times; and (b) transferability within urban models and within urban agglomerations.

Used methods: A list of the used methods to construct the developed model is:

1. Classification of the building stock through the use of **building typologies** (see Section 2.3 for an overview of the different typologies and Section 4.2 for a detail description of the automatic classification of the building stock);

2. Generate individual occupation schedules for each family using **time use data** (see Section 5.4 for a description of this process);

3. Generation of a **synthetic population** for each area to be allocated into the building stock (see Section 2.2 for an overview of methods regarding the reweighting of surveys and the generation of synthetic populations and Section 6.11 for the implemented method into de model); and

4. Simulation of heat consumption through the use of a **heat demand model** (see Section 7.1 for the implementation of a heat balance model and Section 5.6 for the implementation of a thermal simulation model).

Used and generated data: Below is a synthesis of all the data used as input to the developed model and the output data generated by the model.

The following data is used as input to the model:

1. Small area **census** for the city of Hamburg for the year 2010 (see Section 3 for a description of this dataset and Section 3.3 for the description of its integration into the model);

2. The German **micro-census** for the year 2010 (see Section 3 for a description of the micro census and Section 7.4 for the description of the method used to reweight and synthesize this survey);

3. The 2011 German census is used as benchmarks at a national level. We use available statistics at a NUTS-3 level "Kreise/Kreisfreie Städte". We use this data in order to present the transferability of the developed method. For the reweighting of the micro census we use the same method as with the small area statistics for Hamburg (see Section 3). The results from the reweighting of the micro census to the German NUTS-3 areas is presented and discussed in Section 6.

4. The German **time use survey** for year 2002 (see Section 5.4 for the use of this dataset for the generation of family specific schedules);

5. The **digital cadastre** of the city of Hamburg for the year 2010, see Section 3 for a description of this dataset; and

6. The IWU **building typology** tables (see Section 2.3 for a description of this and other typologies and Section 4.1 for a description of the classification of the building stock)

The model output delivers following data:

1. A Simplified **digital cadastre** optimized for urban simulations. This digital cadastre represents the building stock with a simplified geometry (LOD -1).

2. A **synthetic population** allocated to the individual buildings on the digital cadastre.

3. A classification of the building stock on a user defined **building typology**.
4. An enriched building stock with detail information on the **materials** used for all construction elements.
5. A high temporal resolution of simulated **heat consumption** at a building level.

Structure of this thesis:

Chapter 2 Presents a **literature review** of:

1. Energy demand models (see Section 2.1)
2. Occupant behavior on these models (see Section 2.1.2)
3. Construction of synthetic population (see Section 2.2)
4. Classification of the building stock (see Section 2.3)

Chapter 3 Describes of the **model structure**, its components and the data schema underneath the model.

Chapter 4 Describes the classification of the building stock through the use of **building typologies**.

Chapter 5 Describes the use of **time use data** to generate occupant schedules of each individual of the micro census.

Chapter 6 Presents the method used to generate a **synthetic population** for each small area and the allocation of the individual families into the building stock.

Chapter 7 Describes two **heat demand models** and their application on the simulation model. The heat demand models are:

1. A simple heat balance method; and
2. The use of a thermal simulation model.

Chapter 8 Present the **Transferability** of the model to national level, the consequences and challenges of this transfer, Description of the main **simulation results**, **Conclusions** and **policy implications** and **Further development** strategies and application possibilities.

The use of the author's we: Throughout the thesis I make an extensive use of the author's we out of respect towards my collogues that have contributed to this PhD (see footnotes on the different chapters).

2 People, Space and Energy: A Literature Review

2.1 Energy Demand Models

In the following section we present a literature review of: (a) energy models (see Sub-Section 2.1.1); (b) Socio-demographic parameters influencing heat demand (see Sub-Section 2.1.2); and (c) Integration of user behavior into building simulation models (see Sub-Section 2.1.3). The literature review concludes with a selection of the parameters influencing heat demand in the residential sector (see Sub-Section 2.1.4).

The aim of this section is to present different approaches used to estimate urban heat demand. Most of the approaches used for this estimation do not have any spatial reference and often simulate heat demand at an aggregated level. The biggest problem of energy simulation model at a low level of aggregation is the lack of data to: (a) develop the models; and (b) validate these models. Required data for a validation of these models is very hard to get. This is because the necessary resolution and amount of data required for a validation is vast. In order to identify behavioral patterns of energy consumption on datasets, these datasets would have to contain a high temporal resolution. Another challenge of energy estimation models is to control for all other variables, requiring such a large data set of both records and variables. The number of variables affecting the consumption of energy demand are vast. This literature review aims to get the reader an insight about the type of models developed to estimate energy demand and the type of variables used in the different type of models.

This literature review is by no means exhaustive as the available methods for the estimation of energy demand covers a vast set of disciplines and countries. The aim of this literature review is to get a sense of the available methods, different approaches and findings regarding the development of energy demand models around the world.

2.1.1 Overview of Recent and Current Energy Demand Modeling

Section 2.1.1 makes a review of current energy demand models specifically used for the estimation of energy demand in urban environments. For a description of "energy demand models" used for the estimation of individual buildings see Section 2.1.3, which explores the integration of user behavior into building simulation models. This review focuses on models working at a disaggregated level, but also considers its counterpart: model developed at a high aggregation level. The aim of this review is to gain an insight into current methods and practices for the development of energy models among the scientific community. The review also makes a small retrospective to models developed two decades ago, in order to acquire an inside on the evolution of this type of models.

Many of the papers found on through the literature use available data at a household level (Micklewright, 1989) for the analysis of energy demand patterns.

The development of energy demand models has a long tradition in economic studies, many of the papers produced from this community aim to estimate energy expenditure rather that energy consumption of energy demand (Micklewright, 1989). This fact emerges, not only because of the research scope of the community but because of the available data. The collection of expenditure is easier than the monitoring of energy consumption, because of the needed technical infrastructure for the monitoring of energy flows. In recent years we have seen an increase of monitoring efforts of energy demand, we hope to see more energy consumption data in the future.

Schenk, Moll und Schoot Uiterkamp (2007) discuss the need for a meso-level approach for a systematic policy analysis of the energy sector. Swan und Ugursal (2009) distinguish between statistical and engineering models, we believe that a trend towards a combination of both methods is emerging. This thesis uses many statistical methods, mainly for the generation of a synthetic population and building stock and engineering models for the estimation of heat demand. The scientific community is aware of the need to develop integrated urban models (Kavgic et al., 2010). Keirstead, Jennings und Sivakumar (2012) propose the development of "activity based models" for the integration of different modeling approaches.

Most recent approaches introduce a spatial constraint to the models (Mavrogianni, Davies, Kolokotroni & Hamilton, 2009; Caputo, Costa & Ferrari, 2013). This added dimension to the models allows, not only a result allocated in space, but allow the authors to recover data through the use of (1) automatic recovery methods (Mavrogianni et al., 2009); (2) available geographical referenced databases (this thesis); or (3) on-site collected data (Caputo et al., 2013).

The topic of models estimating heat demand with a spatial reference is covered on Section 2.3 of this chapter as well as the use of archetypes for the classification of the building stock. Caputo et al. (2013) presents a modeling approach using building typologies for the classification of the building stock.

From the literature review we see a clear trend of energy demand models towards: (a) models working at a low aggregation level; (b) a consideration of socio-demographic data and (c) the utilization of GIS systems or geo-referenced data for the construction of urban models. All these trends are needed for the planning of distributed energy systems (Manfren, Caputo & Costa, 2011).

2.1.2 Socio-demographic Parameters Influencing Heat Demand

Section 2.1.2 reviews the influential literature about behavioral and socio-demographic parameters that may change the energy demand of households. This section concludes with a table listing the main parameters found in the literature. The aims of this section are: (1) Identify methods used in this kind of analysis; (2) List the type and source of data used for the analysis; (3) List parameters identified in the literature as "relevant"; and In his work Scott (1980) identifies the problem of combining all these parameters in one model. The amount of data needed and the quality of the data required, makes such an approach difficult.

Van Raaij und Verhallen (1983a) intend in their work to define influencing parameters in the household energy demand and to establish relations between them. The focus of the work relays on behavioral parameters. Another important factor that the authors describe is the constant or even rising energy demand caused by a retrofit due to climate control mechanisms; this effect is further analyzed as the "Rebound-Effect" The authors also describe other parameters besides behavior. In their work, building specific parameters considered are: (a) building type; and (b) location (neighbors).

As part of the socio-demographic parameters, the authors try to incorporate the influence of *income* correlating it to *household size* and *floor space*. The influence of *energy price* is considered by Van Raaij und Verhallen (1983a) to play an important role. A correlation between the energy consumption and the energy price cannot be demonstrated, one of the problems in correlating these parameters is the time differences between them, as the payments are done on a yearly basis.

Douthitt (1986) based their work on a sample of 174 Canadian households, this dataset can be classified by previous retrofits and energy conservation activities to save energy, Douthitt (1986) intends to quantify the influence of parameters on the energy consumption of these households. In order to show this influence a model for the consumption of natural gas for space heating and water heating is created. In order to estimate the influence of *energy price* measured consumption data of two years are analyzed.

The main findings of the work are: (a) households, which were retrofitted react with high price elasticity to price variation; and (b) influencing factors, such as thermal efficiencies, floor space and building shape are essential variables for an energy consumption model.

J. G. Anderson und Kushman (1987) provide a model approach for modeling households heating energy demand. Their goal is to create an instrument to analyze the responses to energy-saving incentives. The starting point of the model is the *room temperature*, which is seen as a direct relation to the *user behavior*. The data sets used in this work are: phone interviews and data from previous studies. For the evaluation of the model a sample of 629 households is used. This sample has technical as well as socio-demographic parameters.

Wirl (1987) developed his work in a mixed economic-technical model for the heating demand of households. His model is based on the assumption that an individual consumer claims a service (thermal comfort). In order to satisfy this claim the individual chooses from durable (buildings) and non-durable (fuels) goods, which are entirely substitutable. The model calculates an optimal solution for that consumer. Wirl (1987) notice in his work the problem of separating durable goods from a financial point of view.

Douthitt (1989) analyzed the energy consumption of 370 Canadian households and developed an econometric model in which economic, sociological and structural parameters were taken into account. Douthitt (1989) does not find significant dependencies between personal variables and internal temperatures. The author underlines the importance of the found correlation between energy consumption and energy prices.

Van Raaij und Verhallen (1983b) examined the behavioral influences on energy consumption based on 145 Dutch households. The households are cluster on five households' types based on following characteristic: "cold", "warm", "wasteful", "savers" and "average". The "hot" and "wasteful" households type have high room temperature, the "cold" and "wasteful" are characterized by a high level of interior ventilation.

Van Raaij und Verhallen (1983b) prove that the "savers" group use significantly less energy, and the group of "wasteful" significantly more energy than the "average" household consumes. As a result, the authors attempt to combine socio-demographic characteristics and show the differences in the energy-related household personal attitude of the inhabitants.

Cuijpers (1995) uses micro census data from about 2000 Belgian households as primary data set. Cuijpers (1995) develops a model that considers households, both, as energy-consumer and supplier. Cuijpers (1995) defines the parameter *"Room-Temperature-Hour"* in order to measure the thermal comfort in the household. This parameter presents a clear advantage as this parameter includes not only internal temperature but volume of the room and time.

Baker, Blundell und Micklewright (1989) discuss how individual households have adjusted their consumption patterns by implementing a model that joins determination of appliance demand and use with micro data on households. The authors take into consideration the difficulties arising from the often severe data requirements and the assumptions about price expectations and the housing market. They also contemplate the possibility of households whose durable choice is constrained by their tenure type, instead of the aforementioned which are restricted to owner occupiers. Given the lack of micro level research into energy demand in the United Kingdom, they concentrate on modeling expenditure on different fuels conditional on durable ownership. This allows them to consider the conditional demands in more detail than typically occurs in the joint models of appliance choice and use. They particularly allow for the marginal rate of substitution across disaggregated energy demands to differ across households with different durable stocks. As a result, although energy demands are modeled conditional on durable stocks, they are not assumed to be separable from these stocks. Data is drawn from the annual Family Expenditure Survey (FES) of Great Britain.

Yamasaki und Tominaga (1997) argue in their work that the development of an aged society is expected to increase residential energy demand in Japan. This study is designed to analyze various factors which determine the energy demand of elderly households and aims to predict future trends in residential energy demand.

Boonekamp (2007) propose the use of a bottom-up simulation model for the analysis of user response to changing energy prices. In order to investigate this effect the authors, analyze trends in the period between 1990 and 2000. The authors define a set of "energy functions" for the analysis of energy demand. Each of this "energy functions" has sets of driving factors that control the function and so the energy consumption. In a following step the consumption of each system or appliance is defined by three factors: (1) item the ownership rate; (2) item the intensity of use; and (3) item the energy efficiency of the system or appliance From these three factors ownership rate and intensity of use (number 1 and number 2), are driven by Socio-economic-demographic parameters, see (Boonekamp, 2007, p. 135; Fig. 1). The author doesn't find any relationship between energy demand and income.

Rehdanz (2007) makes use of the German socio-economic panel for the identification of determinants influencing the heating expenditures in Germany. Rehdanz (2007) concludes with the confirmation of the postulated hypothesis, showing a significant difference in expenditures for heating and hot water supply between rented and owner occupied dwellings. The author points out that the results suggest that this difference is likely to become smaller over time.

Guerra Santin, Itard und Visscher (2009) start arguing the importance of the building occupant while analyzing the energy demand of building. The authors argue that through tighter building regulation the quality of thermal properties of buildings is improving and therefore overall energy use associated with building characteristics is decreasing, making the role of the occupant more important. The data used in this study comes from the Kwalitatieve Woning Registratie (KWR) of the Ministry of Housing of the Netherlands (VROM). The dataset used includes 15,000 houses across the Netherlands. This dataset includes housing quality. Guerra Santin et al. (2009) conclude in their paper that a temperature setting in dwelling units is an important predictor of energy use. The authors also present a small correlation between temperature setting and occupant characteristics. The authors also conclude that the continuous presence of people at home increases energy use in comparison to cases where the users are almost never at home or their presence is very variable. The study showed that occupant characteristics and behavior significantly affect energy use (4.2%), but building characteristics still determine a large part of the energy use in a dwelling (42%).

Zhun Yu, Benjamin C.M. Fung, Fariborz Haghighat, Hiroshi Yoshino und Edward Morofsky (2011) present in this paper a new method for examining the influence of occupant behavior on building energy consumption. The analysis method used in this method is a cluster analysis of consumption data from a set of residential buildings. The data used in this analysis comes from a survey and monitoring program from the architecture institute of Japan. The survey was carried on between 2002 and 2004. This data set consist of monitors values from 80 residential buildings located in six different districts in Japan. The monitored data consist of indoor temperature and consumption of electricity, gas, and/or kerosene. Zhun Yu et al. (2011) identify indoor temperature as one of the most important factors influencing energy use in the buildings. A significant difference between clusters can be seen, leading to the conclusion that the impact of the user can be significant.

2.1.3 Integration of Occupant Behavior Into Building Simulation Models

Section 2.1.3 makes a brief overview of current methods used for the integration of socio-demographic characteristics (and therefor the resulting behavior) into heat demand simulation models. The focus in this section is on current methods rather than on the development of such methods through history. It is therefore that the literature review makes an exclusive selection of papers published in the last decade.

Statistical analysis This type of analysis have been performed since the eighties, see (Scott, 1980; Bohi, op. 1981; Van Raaij & Verhallen, 1983b).

> "...it is difficult to completely identify the influences of occupant behavior and activities through simulation due to users' behavior diversity and complexity..."
> (Zhun Yu et al., 2011, S. 1409).

The trend on this type of analysis is that it is too complicated to find concrete relationships between all the relevant parameters. The big challenge of empirical studies is to control and monitor for all parameters influencing heat demand. Most of the performed empirical analyses on this issue lack the precision to pinpoint specific effects of individual behaviors'.

Special attention has been given in the literature to the occupation patterns of users as a proxy to estimate the influence of user behavior, Guerra Santin et al. (2009) observed a significant influence of this parameter in heat consumption. Page, Robinson, Morel und Scartezzini (2008) deliver interesting methods for the integration of this parameter into simulation models.

The monitoring of specific user-driven actions in the residential sector (e.g.: opening of windows) and the link to (1) demographic data and (2) the integration of these actions into simulation models is extremely complicated because of (a) lack of empirical data in order to establish a robust relationship between demographic characteristics and specific user behaviors'; and (b) the complexity of the required simulation architecture for the computation of these effects. The measurement of these actions have been reported in the scientific literature (Kah et al., 2010; Nouidui, Wetter & Zuo, 2012). Attempts to integrate the user behavior, exist (Hensen & Lamberts, 2011; Nouidui et al., 2012).

2.1.4 Distilling Parameters Influencing Heat Demand From the Literature

In Tables 2.1 through 2.5 an overview of the main parameters encountered in the literature is presented. The tables are separated based on five main categories:

1. Economic parameters;
2. User behavior characteristics;
3. Socio-demographic characteristics;
4. Building characteristics; and
5. Exogenous parameters.

This table can be seen as the synthesis from the performed literature review. The aim of this literature review is to use the identified parameters for further analysis. A clear signal from the literature review is that the hours spend at home may by a key parameter influencing energy demand. We consider this parameter for the further creation of synthetic data and of the creation of families. In Germany there is many information about the time budget of families and individuals, we want to generate the parameter "hours spend at home" based on available statistics.

Table 2.1: Parameters used in different studies, arranged by type of parameter: Economic parameters

Economic parameters:	
Price of energy	Scott (1980); Douthitt (1986)
Heating included in rent	Guerra Santin et al. (2009)
Price of services	Scott (1980)
Household income	Scott (1980); Van Raaij und Verhallen (1983b); Douthitt (1986); Guerra Santin et al. (2009); Lutzenheiser (2002)
Number of household members being officially registered as unemployed.	Rehdanz (2007)
Capital	Scott (1980)
Purchase behavior	Van Raaij und Verhallen (1983b)

Table 2.2: Parameters used in different studies, arranged by type of parameter: User behavior

User behavior:	
	J. G. Anderson und Kushman (1987)
Maintenance behavior	Van Raaij und Verhallen (1983b)
Bedroom temperature	Van Raaij und Verhallen (1983b)
Room temperature	Douthitt (1986); J. G. Anderson und Kushman (1987)
Temperature during the night, evening and day	Guerra Santin et al. (2009)
Thermostat adjustment during absence and presence	Van Raaij und Verhallen (1983b)
Adjustment and use of window shutter	Van Raaij und Verhallen (1983b)
Ventilation behavior	Van Raaij und Verhallen (1983b)
Use of specific rooms	Van Raaij und Verhallen (1983b)
Use of entrance door	Van Raaij und Verhallen (1983b)
Energy conservation attitude	J. G. Anderson und Kushman (1987)
Room-Temperature-Hour	Cuijpers (1995)
Ownership and usage of electric appliances	Sanquist, Orr, Shui und Bittner (2011)

Table 2.3: Parameters used in different studies, arranged by type of parameter: Socio-demographic characteristics

Socio-demographic characteristics:	
	Scott (1980)
Race / Ethnicity.	Lutzenheiser (2002)
Level of education.	Van Raaij und Verhallen (1983b); J. G. Anderson und Kushman (1987); Lutzenheiser (2002)
Number of persons per dwelling unit.	Van Raaij und Verhallen (1983b); Douthitt (1986); Lutzenheiser (2002); Rehdanz (2007); Boonekamp (2007); Guerra Santin et al. (2009); Sanquist et al. (2011)
Age structure in the household.	Van Raaij und Verhallen (1983b); J. G. Anderson und Kushman (1987); Lutzenheiser (2002); Boonekamp (2007)
Number of adults.	(Douthitt, 1986),
Number of persons under 18 years.	Douthitt (1986),
Average age of the adult household members.	Rehdanz (2007),
Age of respondent.	Guerra Santin et al. (2009); Lutzenheiser (2002)
Internal gains.	Douthitt (1986)
Always people during weekends.	Guerra Santin et al. (2009)
Always people during day.	Guerra Santin et al. (2009)
Occupation rate	Boonekamp (2007)

Table 2.4: Parameters used in different studies, arranged by type of parameter: Building characteristics

Building characteristics:	
	Scott (1980)
Presence of a pilot light.	Van Raaij und Verhallen (1983b)
Building type.	Van Raaij und Verhallen (1983b); Lutzenheiser (2002); Rehdanz (2007); Boonekamp (2007); Guerra Santin et al. (2009)
Location.	Van Raaij und Verhallen (1983b)
Floor space.	Van Raaij und Verhallen (1983b); Lutzenheiser (2002); Rehdanz (2007); Guerra Santin et al. (2009); Sanquist et al. (2011)
Number of rooms & Number of heated bedrooms.	Guerra Santin et al. (2009)
Thermal quality of buildings.	Douthitt (1986); Cuijpers (1995); Guerra Santin et al. (2009)
Building envelope surfaces.	Douthitt (1986)
Efficiency of the heating system.	Douthitt (1986),
Type of heating device.	Cuijpers (1995); Rehdanz (2007); Boonekamp (2007)
If a new heating system.	Rehdanz (2007)
Building age.	Cuijpers (1995); Rehdanz (2007); Guerra Santin et al. (2009); Sanquist et al. (2011)
State of renovation.	Rehdanz (2007)
New windows were installed.	Rehdanz (2007)
If a modernization took place in the previous year.	Rehdanz (2007)
A bath or shower.	Rehdanz (2007)
It is owner-occupied or a subsidized apartment.	Lutzenheiser (2002); Rehdanz (2007)

Table 2.5: Parameters used in different studies, arranged by type of parameter: Exogenous parameters

Exogenous Parameters:	
Extreme temperatures.	Scott (1980)
Cooling degree days.	Sanquist et al. (2011)
Heating degree days.	Sanquist et al. (2011)

2.2 Generation of Spatially Allocated Synthetic Population

In the following section of the thesis we present a short overview of common methods to generate a spatially allocated synthetic population and the mathematical expression of the models. This procedure is normally the first step in a microsimulation model as well as some agent based models. This section is structured into four sub-sections:

1. a discussion about some methods used to reweight a survey sample to match aggregated values of statistics with a geographic reference (see Section 2.2.1);

2. discussing the explicit generation of a synthetic population allocated in space (see Section 2.2.2);

3. The mathematical expression of some spatial microsimulation models (see Section 2.2.4 onwards); and

4. A list of available measures for the internal validation of spatial microsimulation models (see Section 2.2.5).

We have divided the discussion into these two categories as the method used to either achieves a reweighting of a survey or the generation of a synthetic population differs on its core. In Section 2.2.2 we also briefly introduce the sequential process of reweighting a sample to further create a synthetic population.

A classification of spatial microsimulation model can be found in Tanton (2014). In his paper Tanton classifies the different methods using the following model characteristics:

1. Static or Dynamic; and

2. Reweighting or Synthetic reconstruction

The method highlighted in red in Figure 2.1 (GREGWT) is the method implemented in this thesis for the generation of a synthetic population. As described in Section 2.2.2 the generation of a synthetic population for the sequential allocation of individuals to the building stock is a two-step process: (1) the initial population survey is reweighted— using the GREGWT algorithm— and a synthetic population is created using the new weights as selection probabilities for the new synthetic population.

This section of the thesis aims to briefly present the most common methods used by the microsimulation community to either create a synthetic population or reweight an initial survey to small area benchmarks. The most used method by community is the IPF algorithm or the Simulating Annealing algorithm. The first algorithm is a **Deterministic** algorithm while the former is a **Probabilistic** one. In this thesis we will describe three algorithms used by the community. In Section 2.2.1 we briefly describe the use of two deterministic algorithms: the IPF and the GREGWT algorithms, in Section 2.2.4 we describe the use of the combinatorial optimization method: simulating annealing. This brief description of the algorithms aims to make a small literature review of methods and applications of the algorithms. A detail description of the used algorithms is presented though the next Sections 2.2.4.

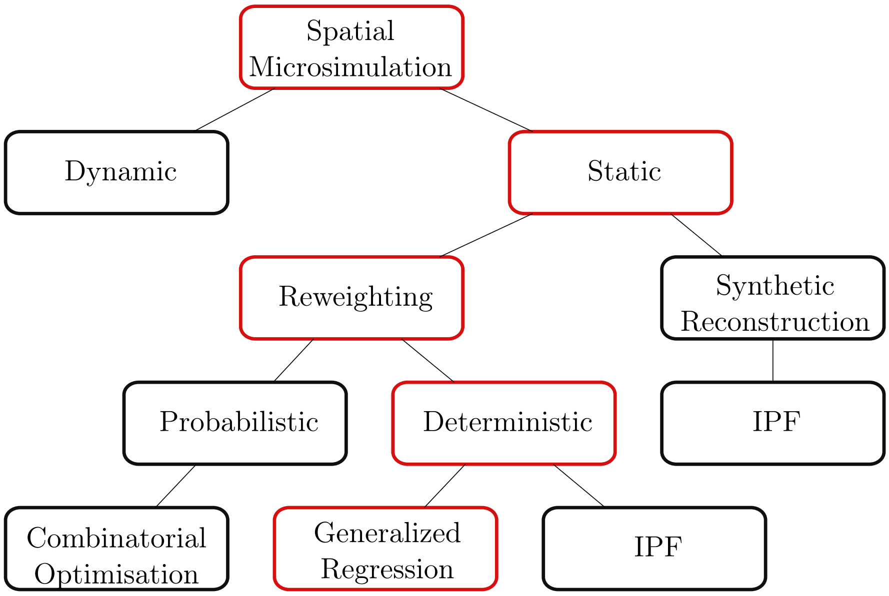


Figure 2.1: Different methods for small area estimation. After (Tanton, 2014)

The deterministic reweighting methods do not create a synthetic population by default. The result from these methods is a vector of new weights corresponding to each individual on the initial survey. These weights are not integer numbers. In order to generate a truly synthetic population an extra step is needed. This is not the case of the combinatorial optimization methods. In these algorithms individual records from the initial survey are taken in order to generate a new survey. In this case the resulting weights are all integer numbers, making it possible to interpret these file as a synthetic population.

The process and importance of a synthetic population in the scientific literature is highlighted on Section 2.2.2. A detail description of the algorithms used for the generation of synthetic population as well as the developed algorithm for generation of a synthetic population in combination with the GREGWT algorithm are described on Section 7.4.

An overview of methods used for the generation of synthetic populations and the underlying algorithms are presented by Harland, Heppenstall, Smith und Birkin (2012). The authors evaluate the different method by comparing the Total Absolute Error (TAE) as an internal validation of the algorithms. The authors compare a deterministic reweighting algorithm and two stochastic algorithms for the generation of synthetic populations. The results from their analysis shows that one of the stochastic algorithm (simulating annealing) outperforms the other two methods.

2.2.1 Deterministic Reweighting Methods

Deterministic reweighting methods aim to reweight a survey to match known aggregate values of small geographical areas. The size and available data of these geographical areas vary between countries.

For the European Union a standard incorporating the different national definitions exist. This is the Nomenclature of Territorial Units for Statistics (NUTS[1]) standard, This nomenclature describes four hierarchies: (0) national territories; (1) NUTS–1; (2) NUTS–2; and NUTS–3. A reweighting of a national survey could be implemented at any of the NUTS levels. Depending on the research question a suitable geographical area should be selected. In the case of Germany there are some exceptions to this hierarchy. The city of Hamburg is one of these examples. These geographical areas have different names on each country and the authors referee then according to the used case location. These areas are known as: (a) Summary Files in the U.S.; (b) Profile Tables or Basic Summary Tabulations (BSTs) in Canada; (c) and Small Area Statistics in the U.K. (Pritchard & Miller, 2012).

The city of Hamburg is both a city and a federal state. The German statistical office "Statistische Ämter des Bundes und der Länder" classifies Hamburg as one "Gemeinde", which corresponds to the European NUTS–3 aggregation level. Clearly we can use these geographical areas for the simulation of urban processes. The regional statistical office "Statistik-Nord" provides aggregated statistics for the city at different levels of aggregation (not regulated by the European Union NUTS nomenclature), for a detail description of data structure and geographical levels available for the city of Hamburg see Chapter 3.

In this thesis we will discuss two important deterministic methods used by the microsimulation community for the reweighing of survey to small geographical areas: (1) the Iterative Proportional Fitting (IPF) approach, and (2) the Generalized Regression (GREGWT) approach.

2.2.2 Generating Synthetic Populations

Some of the most common methods to reweight a population survey to geographical areas deliver non-integer weights. Depending on the aim of the simulation this can be a problem, this is the case for most agent based simulation, these simulations need to describe every single micro unit within the simulation. There are many methods developed for the integerization of reweighting results. Most of the developed methods to generate a synthetic population are a form on integerization of the estimated new weights (Beckman, Baggerly & McKay, 1996; Lovelace & Ballas, 2013). Pritchard und Miller (2012) propose a "Conditional Monte Carlo Synthesis Procedure" to create a synthetic population benchmarked to different aggregation units. Ma und Srinivasan (2015) developed another method to create a synthetic population: "fitness-based synthesis" (FBS). The method presented by Ma und Srinivasan proposed the computation of two fitness values expressing the adding and subtracting probability of individuals from a random selected population from the initial sample survey, this method is implemented in this thesis for the generation of synthetic populations (see Section 2.2.4).

2.2.3 Population Projections

An important use of microsimulation models is the construction and simulation of scenarios for the analysis of specific events. For the particular case of estimated heat demand of urban areas the projec-

[1] `http://ec.europa.eu/eurostat/web/nuts/overview`

tion of: (a) demographics and (b) the building stock, plays an important role. Because we integrate the influence of occupant behavior into our simulation model the change in demographic characteristics would have a direct effect on the resulting heat demand. This section will describe current methods used for the projection of demographic characteristics used in spatial microsimulation models. Section 6.7 describes the projection of aggregated statistics for the reweighting of the sample survey simulating an aging population and the retrofit of the building stock.

2.2.4 Algorithms of Spatial Microsimulation Models

In the literature there is some work describing and comparing these methods: Ballas, Clarke und Turton (1999) present an overview of methods and discuss in detail the use of the IPF method. Ballas, Rossiter, Thomas, Clarke und Dorling (2005) present a detailed description of available methods of static and dynamic spatial microsimulation models. Rahman, Harding, Tanton und Shuangzhe (2010) discuss different methods used in the microsimulation community, provides a clear example for the implementation of the methods and discusses the implementation of new methods for the community. Tanton und Edwards (2013) provide an historical overview of the methods used in the community. Williamson (2013) compares two methods on his contribution: (a) a synthetic reconstruction and (b) a reweighting methodology. O'Donoghue, Morrissey und Lennon (2014) offers an overview of methods and common applications of spatial microsimualation in the scientific community. Schmid und Münnich (2014) provide an overview of small area estimation methods not so common in the spatial microsimulation community. Tanton (2014) provides an up to date overview of methods used in the spatial microsimulation community and Tanton, Williamson und Harding (2014) provide a comparison of a GREGWT method and a combinatorial optimization method.

In this chapter we present: (1) the mathematical expression of the different methods available for the generation of synthetic populations, (2) a simple example for the comparison of some of the presented methods; and (3) a collection of measures for the performance of the internal validation of microsimulation models.

Simulated Annealing Section 2.2.4 describes the simulated Annealing method, this is the most common type of combinatorial optimization method in the spatial microsimulation community.

Iterative Proportional Fitting Section 2.2.4 describes the IPF method, commonly used my many spatial microsimulation models and other urban models for the reweighting of surveys.

Generalized Regression and Weighting of Sample Survey Results Section 7.4 gives an overview of the GREGWT method, Section 2.2.4 describes the GREGWT implementation of the Australian Bureau of Statistics (ABS).

Synthetic population presents the developed method for the creation of synthetic populations, because the IPF and GREGWT method compute non integer weights we need to post process this data for the representation of individual persons of families on the small areas. The used algorithm to achieve these are discussed on Section 2.2.2 and the implementation of the "Fitness based Model" is discussed on Section 2.2.4.

Internal validation of the model is presented on Section 2.2.5. On this section we describe and discuss some of the implemented measures for the internal validation of spatial microsimulation models available of the scientific literature.

Sample Population

The first step of most spatial microsimulation methods is the definition of a sample population. This sample population is normally a survey containing information about the individuals. These records can be categorical or numerical. Most of the microsimulation methods will assume parameters of categorical data.

The initial population sample is described by the number of records m. One important attribute of these records is the attached weight to every record. These weights describe how representative this record is to the population. The presented spatial microsimulation methods reweight these samples (modify the initial weight) to match some aggregated statistics available at a small area level.

After the reweighting of the population sample we get a new set of weights w for each area. The sum of the weights is equal to the population size of the small area. Each record on the sample will have different weights for each area, the weight describes how representative an individual is for a specific small area. In the case of remote areas the reweighting algorithm may fail to achieve convergence, these is because none of the records (or any combination of them) available in the original population sample is representative to these small area.

For the reweighting of the sample we define constraints. The constraints used in the simulation are given by T_i. The individual modeled elements are estimated as:

$$T_i = \sum_i w_{i,j} x_j \tag{2.1}$$

Where: T_i is the observed individual characteristics for small area i.

The aim of the model is to minimize the distance between observed marginal totals and estimated marginal totals. We minimize these by manipulating the initial weight of the sample survey. Depending on the implemented method the initial sample weights (also described as design weights) have a lower impact on the simulated output weights. The IPF method has a lower sensitivity to initial weights, in this case we can simple set all initial weights to 1. This is not the case for the GREGWT method. The GREGWT method aims to minimize the distance between initial and final weights (see description below). The used input weights, more specifically the weight distribution of the input survey sample has a strong influence on the simulation results.

A common measure of performance is the Total Absolute Error (TAE). This is not the only measure of model performance, see Section 2.2.5 for an overview of performance measures of spatial microsimulation models. The TAE is measures the total absolute difference between the observed individual characteristics T_i and the model based estimated individual characteristics $\hat{T}_i$. The TAE is expressed as:

$$TAE = \sum_i \|\hat{T}_i - T_i\| \tag{2.2}$$

Where: $\hat{T}_i$ is the model based estimate individual characteristics for small area i.

Simulated Annealing

In this section we describe the algorithm used for a combinatorial optimization model, more specifically we describe a simulating annealing algorithm. The idea behind this method is to start with a random population and change individuals at random, if the change reduces the distance between observed and estimated marginal totals TAE, this change is accepted, otherwise rejected. A simpler version of the simulated annealing is the hill climbing algorithm. The advantage of the simulating annealing algorithm is that this algorithm will take steps back in order to avoid suboptimal results, in order to achieve this, the algorithm defines "temperature steps".

Temperature steps:

- stating temperature t_0
- temperature steps of $t (t_0 >> t)$

Initial Random State: All counters for the initial state are set to zero. A zero value of the survey sample weights w means that the corresponding individual is not selected as part of the synthetic population.

$$w_{i,j} = d_j \tag{2.3}$$

$$w_{i,j} = 0 \quad \forall j = 1 \dots m \tag{2.4}$$

We select a random sample of M_i individuals. Set the weights of the selected individuals to 1. The value of M corresponds to the number of units (e.g.: persons) on a particular geographical areas i. Normally M is much smaller that the survey sample size. Nonetheless, each individual on the sample can be represented more than one time. Each time an individual is selected from the sample its weight is increased by 1. Notice that in this procedure the weights will always be integers.

$$\begin{aligned} q &= rnd(1:m) \\ w_{i,q} &= 1 \end{aligned} \tag{2.5}$$

Repeat equation 2.5 until the required number of individuals has been selected. The sum of the sample weights has to be equal to the small area units total M.

$$\sum_{j=1}^{m} w_{i,j} = M_i \tag{2.6}$$

Where:

m	Number of individuals in the sample
$w_{i,j}$	A weight for small area i in the j^{th} member of the population sample
M_i	Number of individuals in small area i

Compute the total absolute error. See Section 2.2.4 "Sample population":

$$\begin{aligned} \hat{T}_i &= \sum_i w_{i,j} x_j \\ TAE &= \sum_i \left| T_i - \hat{T}_i \right| \end{aligned} \tag{2.7}$$

Replace Individual at Random: Pick a random individual and subtract 1 from its weight.

$$\begin{aligned} k &= rnd(1:M_i) \\ w'_{i,k} &= w_{i,k} - 1 \end{aligned} \tag{2.8}$$

Pick a new random individual from the sample:

$$
\begin{aligned}
q &= rnd(1:m) \\
w^{'}_{i,q} &= w_{i,q} + 1
\end{aligned} \tag{2.9}
$$

Calculate the goodness of fit for the new solution:

$$
\begin{aligned}
\hat{T}^{'}_{i} &= \sum_{i} w^{'}_{i,j} x_j \\
TAE^{'} &= \sum_{i} \left| T_i - \hat{T}^{'}_{i} \right|
\end{aligned} \tag{2.10}
$$

Update the weights if the annealing threshold is exceeded:

$$
\text{if } (TAE^{'} - TAE) > e^{-kt_o} \text{ then } w_{i,j} = w^{'}_{i,j} \forall j \tag{2.11}
$$

Repeat Equations 6.7 to 2.11 until annealing threshold is zero:

$$
\begin{aligned}
t_o &= t_o - t \\
\text{if } (t_0 \leq 0) &\text{ then } Stop
\end{aligned} \tag{2.12}
$$

The simulating annealing is computational expensive by design. This algorithm needs to perform many iterations in order to find a good solution. Because the algorithm starts at a random state the time needed to find a representative population for the corresponding area can take a lot of time, specially is the difference between small area population size M and sample size m is large. The advantages of this method are: (1) accuracy, there are many papers reporting that the performance of this algorithm harvest better results that the IPF or GREGWT method; and (2) the representation of weights as integers. Similar to the application of the fitness based method in this thesis for the generation of integer weights (see Section 2.2.4) in combination with the GREGWT algorithm we could implement a combination of method in order to take advantage of the speed of GREGWT and the accuracy of simulating annealing methods. Instead of starting with a random population we could select the observations based on the pre-computed weight by the GREGWT algorithm.

Iterative Proportional Fitting

The IPF method is a commonly used method in the spatial microsimulation community and other urban simulation communities. This algorithm iterates through the individual parameters of the initial sample and fits them to the know marginal totals of the small areas. Similar to the GREGWT method the IPF method delivers non integer weights.

Initialize the Weights: Many of the IPF implementations initialize the weights before the iterative process. This step is not always necessary and in some cases can be counterproductive, Section 6.6 discuss the role of initial weights on spatial microsimulation models. The IPF method has a low sensitivity to initial weights and therefor the modification of the initial weights has a minimal impact to final results. The GREGWT algorithm has a higher sensitivity to the modification of initial weights, a modification of the initial weights would have a large impact on final results. The IPF method has a low sensitivity to initial weights and therefor the modification of the initial weights has a minimal impact to final results. The GREGWT algorithm has a higher sensitivity to the modification of initial weights, a modification of the initial weights would have a large impact on final results.

$$w_{i,j} = 1 \quad \forall j = 1 \dots m \tag{2.13}$$

Where: $w_{i,j}$ are the weights for small area i in the j^{th} member of the population sample and m is the number of individuals in the sample survey.

Implementing the Algorithm for Each of the Attributes in Turn: For each attribute (or benchmark) the algorithm will compute a new vector of weights based on the share of estimated cell sums of the attribute and the corresponding marginal sum of the small geographical areas.

Generate cell counts from the model using the current weights.

$$\hat{T}_{i,k} = \sum_j w_{i,j} x_{i,k} \tag{2.14}$$

Where: $T_{i,k}$ is the model based estimate individual for attribute k for small area i

Update the weights based on the comparison between the model cell counts $\hat{T}$ and the constraining tables T.

$$w_{i,j} = w_{i,j} \frac{T_{i,k}}{\hat{T}_{i,k}} \tag{2.15}$$

Repeat steps 2.14 to 2.15 until there is no further reduction in the total absolute error. For the computation of the total absolute error see section 2.2.4 “Sample population”.

$$TAE = \sum_{j,k} |T_{i,k} - \hat{T}_{i,k}| \tag{2.16}$$

Simple IPF example: Listing 2.1 shows a simplified version of the IPF procedure implemented in the synthpop python library[2] (Ye, Konduri, Pendyala, Sana & Waddell, 2009).

Listing 2.1: Simplified ipf procedure written in python as part of the synthpop library

```
1 while calc_diff(constraints, prev_constraints) > tolerance:
2     prev_constraints[:] = constraints
3     for loc, target in list_of_loc:
4         constraints[loc] *= target / constraints[loc].sum()
5     iterations += 1
6     if iterations > max_iterations: beak
7 return pd.Series(constraints, index=joint_dist.index), iterations
```

This implementation of the IPF algorithm iterated until the difference between observed and estimated constrains are below a predefine `tolerance` level. The function `calc_diff` is actually an implementation of the TAE. The algorithm iterated through each constrain (or benchmark) and computes the new weights as in equation 2.15. Code line 4 is a direct implementation of equation 2.15.

Generalized Regression and Weighting of Sample Survey Results (GREGWT)

In this section we describe the GREGWT algorithm as well as the implementation of the algorithm in the R language. The implementation of the algorithm has been put into an R package. Through this thesis we implement this algorithm for the creation of synthetic populations, see `https://github.com/emunozh/GREGWT`.

Constraints function: The model constrain we are trying to satisfy is define by Equation 2.17. We try to find a set of weights that multiply by the sample survey characteristics x is equal to the know small area marginal sums T.

$$T_i = \sum_j w_j x_j \tag{2.17}$$

Where: T_i are the true population total and w_j the estimated new weights

[2]`https://github.com/UDST/synthpop/blob/master/synthpop/ipf/ipf.py`

Distance Measure and Minimization: GREGWT ties find a set of new weights satisfying Equation 2.17 by minimizing the distance between initial/design weights and final/estimated new weights. This fact is important to keep in mind, GREGWT tries to keep the estimated weights very close to the original population weights. Contrary to the IPF we do not modify the design weights of the survey sample.

$$D = \sum_j G_j \left(w_j, d_j\right) \tag{2.18}$$

Where: D is the total distance over the sample and $G(w_j, d_j)$ is the distance between new weights w_j and sampling design weights d_j

Generation of New Weights: Now the problem consists of minimizing Equation 2.18 constrain to Equation 2.17. We want to minimize equation 2.18, subject to constraint equation 2.17. There are many ways to minimize this distance. The distance measure used in the GREGWT algorithm is known as truncated Chi-squared distance function, see Rahman et al. (2010).

$$L = D + \sum_{k=1}^{p} \lambda_k \left(T_i - \sum_j w_j x_j \right) \tag{2.19}$$

By implementing the Chi-squared distance into Equation 7.4:

$$L = \frac{1}{2} \sum_j \frac{\left(w_j - d_j\right)^2}{d_j} + \sum_{k=1}^{p} \lambda_k \left(T_i - \sum_j w_j x_j \right) \tag{2.20}$$

By differentiating 7.4 with respect to w_k and then applying the first order condition, we have:

$$\frac{\delta L}{\delta w_j} = \left(\frac{w_j - d_j}{d_j} \right) - \sum_{k=1}^{p} \lambda_k x_j = 0 \tag{2.21}$$

Where: λ_k are the Lagrange multipliers and p represents the p^{th} constraint condition

The new weights can be computed as:

$$w_j = d_j + d_j c_j^{'} \lambda \tag{2.22}$$

Where: $c_j^{'}\lambda = \sum \lambda_k x_j$

The estimated new weights are the design weights multiplied by the sum of the Lagrange multipliers times the sample survey values and summed to the original design weights. In the iterative section of the GREGWT algorithm for the truncation of negative values (see below) the design weight of observations causing a negative weight are set to zero. See code line 13 and 14 of Listing 2.4.

GRWGWT — The ABS Macro

The following section briefly describes the implementation of the GREGWT algorithm in the ABS SAS macros as described in Bell (2000) and Bell (2000).

$$T = \sum_j d_j x_j \tag{2.23}$$

Where:

T	Row vector of benchmark totals
d_j	Initial weight for unit j
x_j	Row vector of characteristic for unit j

$$w_j = d_j \left(1 + \left(\hat{T} - T \right) \left(\frac{\sum_i d_j x_i x_j^{'}}{c_j} \right)^{-1} \frac{x_j^{'}}{c_j} \right) \tag{2.24}$$

Where:

Tx	Row vector of benchmark totals
$\hat{t}x$	New resulting marginal totals
w_j	Estimates new weights
d_j	Initial weight for unit j
x_j	Row vector of characteristic for unit j
$x_j^{'}$	Inverse vector of x_j
c_j	Typically $c_j = 1$, giving an un-weighted least squares distance

$$
\begin{aligned}
Y^{''} &= \sum_j w_j y_j && \text{(2.25a)}\\
&= \sum_j d_j y_j + \left(T-\hat{T}\right)\left(\frac{\sum_j d_j x_j x_j^{'}}{c_j}\right)^{-1}\frac{x_j}{c_j} && \text{(2.25b)}\\
&= Y_a + \left(T-\hat{T}\right)\beta_a && \text{(2.25c)}
\end{aligned}
$$

(1) Initial state

$$
\begin{aligned}
m &= 0 && \text{(2.26a)}\\
d_j^{(0)} &= d_j \text{ for all unitis } j && \text{(2.26b)}\\
\hat{T}^{(0)} &= \sum_j d_j^{(0)} x_j && \text{(2.26c)}\\
A^{(0)} &= \sum_j d_j^{(0)} x_j x_j^{'} && \text{(2.26d)}
\end{aligned}
$$

$$
\lambda^{(0)} \text{ a solution of } (T-\hat{T}^{(0)}) = \lambda^{(0)} A^{(0)} \qquad \text{(2.27)}
$$

(2) Start iterations: Truncate to upper and lower bounds

$$
\begin{aligned}
m &= m+1 && \text{(2.28a)}\\
\text{For each unit of } j&: \\
w_j^{(*)} &= d_j\left(1+\lambda^{(m-1)}x_j^{'}\right) && \text{(2.28b)}\\
\text{if } w_j^{(*)} < L \text{ then } & \quad w_j^{(m)} = L;\ d_j^{(m)} = 0 && \text{(2.28c)}\\
\text{else if } w_j^{(*)} > U \text{ then } & \quad w_j^{(m)} = U;\ d_j^{(m)} = 0 && \text{(2.28d)}\\
\text{else } & \quad w_j^{(m)} = w_j^{(*)};\ d_j^{(m)} = d_j && \text{(2.28e)}
\end{aligned}
$$

(3) Find a solution for λ^*

$$T^{(m)} = \sum_j w_j^{(m)} x_j \tag{2.29a}$$

$$A^{(m)} = \sum_j d_j^{(m)} x_j x_j^{'} \tag{2.29b}$$

$$\lambda^{*}\text{a solution of} \quad \left(T - \hat{T}^{(m)}\right) = \lambda^{*} A^{(m)} \tag{2.29c}$$

$$\lambda^{(m)} = \lambda^{(m-1)} + \lambda^{*} \tag{2.29d}$$

(4) Convergence

$$|T - \hat{T}^{(m)}| < \epsilon_x \tag{2.30a}$$

$$|\lambda^{(m)} - \lambda^{(m-1)}| < \epsilon_A \tag{2.30b}$$

Where:

m	Iteration counter
a_j	Initial weight for unit j
w_j	Estimates new weights for unit j
x_j	Row vector of characteristic for unit j
T	Row vector of benchmark totals
L	Lower bound
U	Upper bound
λ	The Lagrange multipliers
$\hat{T}$	New resulting marginal totals

If the algorithm does not converge after step (2) and (3), the algorithm repeats the process from step (2) until either: (a) convergence is achieved; or (b) the predefined maximum number of iterations is achieved. Convergence is achieved **if either** equation 2.30a (boundary conditions are met) **or** equation 2.30b (no improvement) are true.

Equation 2.30a, representing the boundary conditions to define convergence of the algorithm is in fact the TAE used for the internal validation of the algorithm. In is important to mention that the algorithm itself does not use the TAE metric for the estimation of the new weights but only as a convergence measure. In many cases the computed TAE by the GREGWT algorithm will differ from the “real” TAE. Because the GREGWT algorithm uses a general regression for the estimation of the Lagrange multipliers (Equation 2.27) the input matrix can not be singular. We need to define one of the input matrix attributes as a reference category, the distance between estimated and observed values for this category is not computed for the convergence of the GREGWT algorithm, i.e. the internal computed TAE does not take reference categories into account.

The code: In this subsection we describe the GREGWT implementation on the R language. Listing 2.2 and Listing 2.3 present the simplest implementation of the GREGWT method. This computation can result in negative weights, for a spatial microsimulation this is not useful and therefor the GRWGWT method integrates an iterative process in order to truncate the computed weights (Equations 2.28), Listing 2.4 implements the truncation of the computed weights.

Listing 2.2: Pseudo-code GREGWT simple

```
hT <- colSums(X * dx)           # Sample totals
A <- crossprod(dx * X, X)       # Crossproduct matrix
A <- ginv(A)                    # Inverse (A) matrix (MASS library)
lambda = A %*% (T - hT)         # Lagrange multipliers
wx = dx * (1 + X %*% lambda)    # New weights
```

Listing 2.3: Pseudo-code GREGWT using the R equation solver

```
hT <- colSums(X * dx)           # Sample totals
A <- crossprod(dx * X, X)       # Crossproduct matrix
lambda <- solve(A, (Tx - hTx))  # Lagrange multipliers
wx = dx * (1 + X %*% lambda)    # New weights
```

Listing 2.4: Pseudo-code GREGWT using bound contains

```
hT <- colSums(X * dx)           # Sample totals
A <- crossprod(dx * X, X)       # Crossproduct matrix
lambda <- solve(A, (T - hT))    # Lagrange multipliers
convergence = F
number.iter = 0
while(!convergence){
    number.iter = number.iter + 1 # Iteration
    wx = dx * (1 + X %*% lambda)  # New weights
    # Truncate weights
    wx[wx<bounds[1]] <- bounds[1]
    wx[wx>bounds[2]] <- bounds[2]
    # Truncate initial wights
    dx[wx<bounds[1]] <- 0
    dx[wx>bounds[2]] <- 0
    # Recompute Tx , A and lambda
    hT <- colSums(X * wx)           # Sample totals
    A <- crossprod(dx * X, X)       # Crossproduct matrix
    lambdaS <- solve(A, (T - hT))   # Lagrange multipliers
    # Save lambda m-1
    lambda0 <- lambda
    # Compute new lambda
    lambda = lambda + lambdaS
    # Compute values for convergence
    delta.tx <- abs(T - hT)
    delta.a  <- abs(lambdaS-lambda0)
    convergence <- (
        all(delta.tx < epsilon) |
        all(delta.a < epsilon) |
        number.iter >= max.iter)}
```

Fitness-Based Synthesis (FBS)

In the following section we describe the implemented method for the construction of synthetic populations. The construction of synthetic populations consists of using the reweighted survey and using the estimated new weights for the construction of a synthetic population. In the framework of this thesis we refer to a synthetic population to a description of individuals within a geographical area, this type of description is necessary for any type of agent based simulation. The described method to generate this type of data implements the method developed by Ma und Srinivasan (2015). Ma und Srinivasan (2015) proposed the computation of two fitness values expressing the adding and subtracting probability of individuals from the random selected population from the reweighted population survey.

$$F_{I}^{i,n} = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(R_{j,k}^{n-1}\right)^2 - \left(R_{j,k}^{n-1} - HT_{j,k}^{i}\right)^2\right] \tag{2.31}$$

$$F_{II}^{i,n} = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(R_{j,k}^{n-1}\right)^2 - \left(R_{j,k}^{n-1} + HT_{j,k}^{i}\right)^2\right] \tag{2.32}$$

Where:

$R_{j,k}^{n-1}$	$= \mathrm{T}_{j,k} - CT_{j,k}^{n-1}$, represents the number of households/persons requires to satisfy the target for cell k in control table j after iteration $n-1$
J	is the total number of control tables
K	is the total number of cells in the corresponding j table
$T_{j,k}$	represents the k cell value in control table j
$CT_{j,k}^{n-1}$	represents the estimated k cell value in control table j after iteration $n-1$
$HT_{j,k}^{i}$	is the contribution of record i in the seed data to cell k in control table j

Equations 2.31 and 2.32 represent the "fitness" values to decide if a specific record of the input data is added or removed from the new sample:

1. $R_{j,k}^{n-1} - HT_{j,k}^{i}$ is the number of households required to achieve the target in cell k of control table j if household i is **added**; and

2. $R_{j,k}^{n-1} + HT_{j,k}^{i}$ is the number of households required to achieve the target in cell k of control table j if household i is **removed**.

The algorithm iterates until no record in the input data has positive values for either type I or type II fitness measure.

The above describe equations follow the notation of Ma und Srinivasan (2015), The equations are design to be used with cross tabulation data. This is not the case for the analysis presented in this thesis. Below we change Equations 2.31 and 2.32 to meet the thesis notation and data structure used in the analysis. Equations 2.33 and 2.34 describe the same process with a different notation.

$$FI_j = \sum_k R_k{}^2 - (R_k - x_{j,k})^2 \tag{2.33}$$

$$FII_j = \sum_k R_k{}^2 - (R_k + x_{j,k})^2 \tag{2.34}$$

Where:

$$\begin{aligned} R_k &= T_k - \textstyle\sum_j x_{j,k} \times \lfloor w_j \rceil \\ &= T_k - \hat{T}_k \end{aligned}$$

R is the difference between the small area totals T and estimated totals $\hat{T}$ (with an integer weight w) for benchmark category k. Both fitness measures are computed for each individual j of the sample.

On the algorithm implementation we introduced an extra constrain, the total absolute error TAE. Performing a spatial microsimulation at a low level of aggregation with just a few people on each area (10 individuals) is difficult. The FBS algorithm is able to reduce the TAE achieved by GREGWT. The extra constrain introduced makes sure that changing individuals results in a reduction of TAE (i.e. the change is only accepted if the achieved TAE is lower than the previous TAE value).

We implement this algorithm as an addition to the GREGWT method. This combination allows us to speed up the creation of synthetic families. The big disadvantage of the FBS method is on speed. Normally with the FBS method we start with a random sample of records, instead of starting with a random sample of records, we start with a sample selected with the selection probability vector defined by the GREGWT computed weights. With this implementation the FBS method just needs a couple of iteration to find the best population instead of performing twice the number of the input sample records iterations, as reported by Ma und Srinivasan (2015).

2.2.5 Performance Metrics

Weights Measures

(1) Total weight distance (D)

Measures the absolute distance between the initial design weights d and the estimated new weights w.

$$D_i = \sum_j^m |w_j - d_j| \tag{2.35}$$

(2) Mean weight distance (θD)

The mean weight distance measures the distance between the initial design weights d and the estimated new weights w. In this case the mean value is asses, this may be useful by comparing the performance of the algorithm while using samples with different size.

$$\theta D_i = \frac{\sum_j^m |w_j - d_j|}{m} \tag{2.36}$$

(3) Chi-squared distance (Chi)

With help of the chi-squared distance we have a more precise description of the individual distance measure between weights. This measure avoids for negative distances compensating positive ones.

$$Chi_i = \sum_j^m \frac{(w_j \times d_j)^2}{2d_j} \tag{2.37}$$

(4) Mean chi-squared distance (θChi)

Similar to the chi-squared distance the mean chi-squared distance aims to measure the distance between weights at an individual level. The difference with the previous measure is that this measure takes the size of the sample into account.

$$\theta Chi_i = \sum_j^m \frac{(w_j \times d_j)^2}{2d_j} \div m \tag{2.38}$$

(5) Total absolute distance (TAD)

The TAD measured the absolute distance between the actual population count and the estimated population count.

$$TAD = \sum_i^n \left| \sum_j^m w_{i,j} - pop_i \right| \tag{2.39}$$

(6) Error in margins (EM)

This measures the ratio between estimated and known number of individuals of households in the simulated small areas.

$$EM_i = \frac{\sum w_j - pop_i}{pop_i} \tag{2.40}$$

(7) Error in Distribution (ED)

This measures the ration between the absolute sum of residuals ($estimated - known$) and the actual count of individuals of households in the simulated area or areas.

$$ED_i = \frac{|\sum w_j - pop_i|}{pop_i} \tag{2.41}$$

Where:

w	new weights
d	sampling design weights
n	small area index
m	survey sample size
pop_i	actual population for simulation area i

Marginal Totals Measures

(8) Total absolute error (TAE)

Applied on: (Burden & Steel, 2015; B. Anderson, 2013; Edwards & Tanton, 2013; Harland et al., 2012; Huang & Williamson, 2001; Tanton & Vidyattama, 2010).

This measure is commonly used for the internal validation of spatial microsimulation models. The total absolute error measures the absolute difference between benchmark totals of small areas and estimated marginal totals for the same area. Ideally this measure is close to 0.

$$TAE = \sum_{i}^{n} |T - \hat{T}| \tag{2.42}$$

(9) Standardized absolute error (SAE)

The SAE aims to make the TAE measure comparable between simulation which normally use different samples for the simulation.

$$SAE = \sum_{i}^{n} |T - \hat{T}| \div pop_i \tag{2.43}$$

(10) Percentage absolute error (PAE)

This measure is the same measure as the SAE measure but express its result as a percentage value.

$$PAE = \sum_{i}^{n} |T - \hat{T}| \div pop_i \times 100 \tag{2.44}$$

(11) Modified Z-statistic (Z)

The modified Z-statistic describes the performance of the individual attributes of the population used as constrains in the simulation model, the Z-statistic proposed by (Blalock, 1979) has been used by many authors to perform an internal validation of microsimulation methods (Birkin & Clarke, 1988; Williamson, Birkin & Rees, 1998; Voas & Williamson, 2000).

$$Z = \frac{r - p}{\sqrt{p \times (1 - p) \div \sum T}} \tag{2.45}$$

$$r = \frac{\hat{T}}{\sum T} \tag{2.46}$$

$$p = \frac{T}{\sum T} \tag{2.47}$$

(12) Independent samples t-test

The t-test allows for a comparison of the difference between simulated and expected proportions. This test is useful to identify any statistical difference between the synthetic and real population.

Implemented in: (Edwards & Clarke, 2013).

```
ttest <- t.test(T, hT)
```

(13) Pearson correlation

This test looks for correlation between simulated and expected marginal totals.

```
pearson <- cor(cbind(T, hT), use="complete.obs", method="pearson")
```

(14) Coefficient of determination

This test allows analyzing how well the simulated data fits the expected data.

```
lm.X <- lm(T~hT)
r2 <- summary(lm.X)$r.squared
r2.adj <- summary(lm.X)$adj.r.squared
```

Where:

T	Row vector of benchmark totals
$\hat{T}$	New resulting marginal totals
m	survey sample size

(15a) Root mean squared error (SRMSE)

$$RMSE = \sqrt{\sum \left(\hat{T} - T\right)^2 / N} \tag{2.48}$$

(15b) Standardized root mean squared error (SRMSE)

Source: (Pritchard & Miller, 2012; Farooq, Bierlaire, Hurtubia & Flötteröd, 2013).

$$SRMSE = \frac{\sqrt{\sum \left(\hat{T} - T\right)^2 / N}}{\sum (T) / N} \tag{2.49}$$

Where:

T	Row vector of benchmark totals
$\hat{T}$	New resulting marginal totals
N	Population size

(16) Absolute Standardized Residual Estimate (ASRE)

Source: (Rahman, Harding, Tanton & Liu, 2013).

$$ASRE = \frac{\sum \left| T_i - \hat{T}_i \right|}{\sqrt{\frac{\sum \left(T_i - \hat{T}_i\right)^2}{n}}} \tag{2.50}$$

Where:

T	Row vector of benchmark totals
$\hat{T}$	New resulting marginal totals
n	Number of geographical areas i

(17) Hellinger Distance

Source: (Rao, 1995) and implemented on: (Ma & Srinivasan, 2015)

$$H^2 = \frac{1}{2} \sum \left(\sqrt{T_i \div T} - \sqrt{\hat{T}_i \div \hat{T}} \right) \tag{2.51}$$

Where:

$T = \sum T_i$ and $\hat{T} = \sum \hat{T}_i$

(18) Standard Error around Identity (SEI)

Implemented on: (Ballas, Clarke, Dorling & Rossiter, 2007; Tanton, Vidyattama, Nepal & McNamara, 2011) and described on (Edwards & Tanton, 2013).

$$SEI = 1 - \frac{\sum \left(\hat{T}_i - T_i\right)^2}{\sum \left(T_i - \bar{T}\right)^2} \tag{2.52}$$

Where:

$\bar{T} = \frac{1}{n} \sum_{i=1}^{n} T_i$

(19) Chi squared

$$Chi = \frac{\left(\hat{T}_i - T_i\right)^2}{T_i} \tag{2.53}$$

2.3 Classification of the Building Stock: Building Typologies

2.3.1 The Use of Building Typologies in Energy and Urban Planning

In the framework of this thesis we use and compare and asses five building typologies. All these typologies are constructed following a similar structure. In this section we will make a short description of these typologies and briefly explain how these typologies are used: (a) to estimate energy (mainly heat) demand in urban spaces; and (b) to perform macro analysis of the building stock.

The selected building typologies are following:

1. **Blesl**, a simplified typology developed by Blesl, Kempe, Ohl und Fahl (2007);
2. **IWU-de**, a typology developed for Germany by the IWU[3] institute;
3. **IWU-he**, a typology developed for the federal state of Hessen by the IWU institute;
4. **BSU**, a typology developed for the city of Hamburg by the BSU[4]; and
5. **EcoFYS**, a typology developed for the city of Hamburg by the consultancy firm EcoFYS[5].

These typologies have been developed for different federal states or for the entire federal Republic of Germany. Often an available building typology for a specific region, city or state is a modified version (or calibrated) of a national typology.

In the following section we present and discuss these building typologies. I start with a typology developed at a national level (IWU-de, see sub section 2.3.2), this is probably the most used typology in Germany for the estimation of heat demand of the building stock. The second example is a derivation from the first typology. The German typology developed by the IWU institute is calibrated for the use of this typology to the specific federal state of Hessen (see sub section 2.3.3).

The behavior of these typologies differ depending on the simulation scale and location. With the method developed in the framework of this thesis (see Chapter 4) we aim to show how the different typologies perform in the estimation of heat demand for the city of Hamburg.

It is expected that the typology developed for Germany will perform better than the one calibrated for the building stock of Hessen. We go further with our analysis and analyze two different typologies specially developed for the city of Hamburg, expecting them to perform much better by the analysis of a small urban area of the city. The difference between these typologies is the level of detail of both typologies. While the first typology developed by the Ministry of Urban Development and Environment

[3]IWU Institut Wohnen und Umwelt GmbH `http://www.iwu.de`
[4]BSU (Behörde für Stadtentwicklung und Umwelt) Ministry of Urban Development and Environment`https://www.hamburg.de/bsu/`
[5]EcoFYS`http://www.ecofys.com/`

(BSU) presents a strong simplification of a typology (see Sub-Section 2.3.4), giving only specific heat demand of the buildings based on two parameters:

1. Construction age; and

2. Construction type.

The second typology, developed for the city of Hamburg, delivers a comprehensive technical report (see (Hermelink, Manteuffel, Lindner & John, 2011)) from which many parameters for the building classification can be derived (see Sub-Section 2.3.5).

In order to present a complete overview of the building typologies we make a short review of the building typologies at a European level, driven by the European initiative TABULA (see Sub-Section 2.3.6). This sub section also serves to build a stepping stone towards a systematic approach to building typologies and the implementation of them in energy and urban planning at a European scale. In the final chapter of our thesis we revisit the use of building typologies at a European level to highlight our envisioned next steps.

There are many building typologies for a lot of different regions all over Europe, a good overview of the available building typologies in Europe, driven by the TABULA initiative, is provided in (TABULA Project Team, 2010). In Germany there are many examples of these typologies calibrated for specific regions or cities:

Düsseldorf (ebök, 2005);

Münster (Hildebrandt, Hellmann & Zantner, 2003);

Freistaat Sachsen (ebök, 2000);

Mannheim (ebök/ifeu, 1998); and

Heidelberg (ebök/ifeu, 1996)

Nonetheless, just a few have reached a scientific discussion, finding a way into the scientific literature, few examples of these are: Kragh und Wittchen (2013); M. K. Singh, Mahapatra und Teller (2013); Hrabovszky-Horváth, Pálvölgyi, Csoknyai und Talamon (2013); Caputo et al. (2013). Most of the developed typologies are developed as technical reports or brochures targeting real estate developers or home users, especially in Germany, this makes it difficult to reproduce or use the typologies in a standardized fashion for the estimation of heat demand in urban areas.

In this section we present an overview of different building typologies used in Germany and Europe. Chapter 4 describes our approach to automate the classification of the building stock, introducing the concept of an array filter that will allocate the building types to the building stock in a stochastic manner. In that chapter we also present the results from the postulated filter array and compare the results with monitoring consumption for the same area.

The use of building typologies is commonly used to estimate different retrofit scenarios for single buildings promoting this way energy-efficient measure through the building stock (Kragh & Wittchen, 2013; M. K. Singh et al., 2013; Dascalaki, Droutsa, Balaras & Kontoyiannidis, 2011; TABULA Project Team, 2012a), the use of building typologies finds also use at estimating the mitigation potential and vulnerability of the residential building stock (Hrabovszky-Horváth et al., 2013). Through the literature we also find approaches addressing the issue at an urban scale (Caputo et al., 2013), these approaches make use of a GIS system for the systematic collection of building information and further classification of the building stock. The collection of building data in GIS system presents an interesting tool for the analysis of urban setting rather than the analysis of individual buildings or the analysis of the national building stock at an aggregated level. Nonetheless, a big problem of such an approach is the intensive data recovery needed to fill the building database. The aim of this thesis is to develop and analyze a method that can deal with missing data on existing databases, as for example with the use of a digital cadaster. Other free and open databases as is the use of open street map present an interesting alternative to institutionalized data sources. The main challenges for the use of such open source are: (1) the coverage rate of the building stock; and (2) the missing characteristics of the individual buildings, needed for the classification process (e.g. Construction year and construction type).

The main difference between an analysis of the national building stock, for which the building typologies are developed, and an analysis of new decentralized energy supply systems, which is the scope of our research, relies on scale. While the first analysis focuses exclusively on the aggregated effect of energy efficiency measures of the building stock, the planning of decentralized energy supply needs to:

1. Estimate the heat demand of small urban areas, suitable for small decentralized energy supply systems; and

2. Identify *heat spots* in urban areas in order to allocate retrofit priorities.

For the proper dimensioning of heat supply systems the estimation of heat demand is a key issue, especially at a low aggregation level. The use of building typologies for the simulation of heat demand improves as the numbers of buildings are summed together (Blesl et al., 2007), this is because many key parameters like internal heat gains, internal temperature, ventilation rates, thermal bridges, etc., cannot be recovered for every single building. In order to perform the heat demand computation the authors have to use average values defined in national guidelines. For the construction of the typology the authors classify the buildings by construction epoch and construction type (single family house, terrace house, etc.). This type of classification is a standard in the construction of building typologies (Ebel, 1990; ebök/ifeu, 1996, 1998; ebök, 2000, 2005; Hildebrandt et al., 2003; IWU, 2003; Loga, Diefenbach & Born, 2011; BSU, 2011; Hermelink et al., 2011; Kragh & Wittchen, 2013; M. K. Singh et al., 2013; Hrabovszky-Horváth et al., 2013; Caputo et al., 2013).

An example of such an approach can be seen in Table 2.6. This approach does not use a building stock database for its construction, neither uses real consumption data for its construction. The construction of such a typology is based on average parameters taken from national guideline VDI 3807 (Verein Deutscher Ingenieure, 1994). Not only for heat demand is a simulation at a low aggregation level important. We see a trend towards decentralized heat and electricity production (KEMA, 2012), driven mainly by the expansion of renewable energies. High temporal simulation of energy signatures of small urban areas for on site generation may be a key development towards low energy demand neighborhoods (Koch & Girard, 2013). In order to perform, either a simple energy balance of the individual

Table 2.6: Building matrix with specific building parameters (Blesl et al., 2007)

Baualtersklassen		<1918	1919-1948	1949-1957	1958-1968	1969-1978	1979-1983	1984-1995	1996-2000	2001-2005
EFH	Fl[a]	132.0	220.0	101.0	242.0	158.0	161.0	136.0	134.0	142.0
	WB[b]	22.0	33.0	17.8	36.3	26.1	26.6	21.3	13.6	10.3
	WKZ[c]	167.0	150.0	176.0	150.0	165.0	165.0	156.0	101.0	72.0
RDH	Fl	103.0	103.0	136.0	72.0	97.0	99.0	81.0	128.0	128.0
	WB	15.6	14.5	23.8	11.7	18.6	17.0	10.5	11.4	9.0
	WKZ	152.0	141.0	175.0	162.0	192.0	171.0	129.0	89.0	70.0
KMH	Fl	616.0	349.0	593.0	2 845.0	1 500.0	595.0	1 263.0	351.0	351.0
	WB	110.9	58.5	69.1	543.0	253.7	74.5	128.1	32.9	22.8
	WKZ	180.0	167.0	117.0	191.0	169.0	125.0	101.0	94.0	65.0
GMH	Fl	649.0	1 349.0	1 457.0	3 534.0	3 020.0	595.0	2 075.0	2 075.0	2 075.0
	WB	121.0	249.4	246.9	524.4	438.9	68.9	170.6	151.1	105.5
	WKZ	187.0	185.0	169.0	148.0	145.0	116.0	82.0	73.0	51.0
HH	Fl				10 408.0	18 012.0				
	WB				1 074.0	2 114.0				
	WKZ				103.0	117.0				

(a) (Fl) Heated living space (Nutzwärmefläche) $[m^2]$; (b) (WB) Heat demand (Wärmebedarf) $[MWh/a]$; (c) (WKZ) Specific Heat demand (spez. Wärmebedarfskennzahl) $[kWh/m^2a]$

(EFH) Einfamilienhaus (single family house); (RDH) Reihenhaus (row house); (KMH) Kleines Mehrfamilienhaus (small multi-family house); (GMH) Großes Mehrfamilienhaus (big multi-family house); and (HH) Hochhaus (high-rise building);

buildings or a high temporal thermal simulation model, the input data of the buildings has to be available. Although digital cadaster data offer a vast amount of information of the individual buildings, needed U-values of the building components are not available in the digital cadaster of the city of Hamburg ALKIS[6]. The use of building typologies to fill this information gap may be an interesting option.

2.3.2 IWU-de — a Typology for Germany

The first German typology, although developed under a scope aiming at energy measures for the federal state of Hessen, was developed in (1990) by Ebel. The author was part of the IWU team at that time. This typology has been further developed and improved through the years, the current "base typology" for Germany was developed in (2003) by the IWO institute IWU. The "base typology" defines key parameters needed for the computation of heat demand of the individual types, like average heating space or roof type. The method used for the construction of building typologies developed by (Ebel, 1990) is essentially the same method used on all available building typologies. For a technical description of the approach currently used for the Germany building typology, see (Loga, Diefenbach, Stein

[6]ALKIS (Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV), 2006) `http://www.hmdk.de/trefferanzeige?docuuid=AB8C6B21-BAFF-4230-A686-0C918FEBEE2F`

& Born, 2012). A synthesis of the typology can be seen in Table 7.1 derived from (Loga et al., 2011), this typology is the one used in this analysis.

2.3.3 IWU-he — a Typology for Hessen

Derived from the "base typology" (IWU, 2003), the IWU institute developed a tailored typology for the federal state of Hessen (Born, Diefenbach & Loga, 2003). We present this typology as an example of a derived typology "calibrated" for a specific region and use it as control within my analysis. Because this typology has been specifically developed for a different region than the one we are analyzing, we expect that this typology will perform poorly in comparison to typologies developed for Germany (Loga et al., 2012) (see sub Section 2.3.2) or to typologies developed for the city of Hamburg (BSU, 2011) (see Section 2.3.4); and (Hermelink et al., 2011) (see Section 2.3.5).

The aim of this typology is to determine the potential energy savings of the residential building stock through: (1) retrofit of the thermal envelope of buildings; and (2) a modernization of the heating system (Born et al., 2003). The study by Born et al. analyses typical residential buildings for the federal state of Hessen. As basis for the energy balancing of the building types the authors use the national regulation DIN V 4108–6 (Deutsches Institut für Normung e. V, 2003b) and DIN V 4701–10 (Deutsches Institut für Normung e. V, 2003a). The typology is represented by 31 building types. Table 2.8 describes this typology.

2.3.4 BSU — a Typology for Hamburg

The data for the Hamburger building typology is based on the results from the presented energy passes since 1997 (BSU, 2011). Following the European directive "Energy Performance of Buildings Directive, EPBD" 2002/91/EC (The European Parliament and the Council of the European Union, 16 December 2002) and the subsequent directive 2010/31/EU (The European Parliament and the Council of the European Union, 19 May 2010) all member states are required to implement measures for the reduction of energy in the building stock. As part of the application of this directive, Germany introduced step by step mandatory issue of an "energy pass". Since the introduction of the German energy demand regulation law EnEV in 2007, all new constructions have to issue such an energy pass. Subsequently since 2008 all buildings constructed previous to 1965 had to issue an energy pass, previous to being rented or sold. Since 2009 all buildings have to issue an energy pass in order to sell them or rent them (BMVBS Bundesministerium für Verkehr, Bau und Stadtentwicklung, 2008).

The BSU typology is very simple and does not offer much background information, In the first analysis we only want to compare the performance of different typologies. In order to achieve this we don't need the background information of the typology. In Chapter 4 we present a method to classify the building stock, the performance of this classification method is influenced by the available background information of the building typologies.

Table 2.7: IWU-de building typology matrix for Germany (Loga et al., 2011)

Construction period		<1859	1860-1918	1919-1948	1949-1957	1958-1968	1969-1978	1979-1983	1984-1994	1995-2001	2002-2009
EFH	WKZ[a]	183.0	180.5	164.8	181.3	146.5	155.6	118.4	132.7	110.1	88.8
	FL[b]	199.0	128.9	275.0	101.0	242.0	157.5	196.0	136.6	110.8	133.2
RH	WKZ		153.7	137.1	156.6	106.3	127.9	127.5	98.8	78.1	86.8
	FL		87.2	102.5	136.0	106.7	96.6	98.4	116.0	135.3	138.1
KMH	WKZ	190.1	143.8	168.1	156.2	129.7	134.0	118.3	122.9	92.8	79.9
	FL	615.9	284.0	350.0	574.8	2844.6	426.0	594.5	707.4	759.0	1991.0
GMH	WKZ		127.4	144.4	142.7	131.5	117.9				
	FL		754.0	1349.1	1457.0	3534.0	3020.0				
HH	WKZ					114.1	113.9				
	FL					10408.0	18012.0				

(a) (WKZ) Specific Heat demand (spez. Wärmebedarfskennzahl) (b) (Fl) Heated living space (Nutzwärmefläche) $[m^2]$;; $[kWh/m^2a]$
(EFH) Single family house "Einfamilienhaus"; (RH) Terrace house "Reihenhaus"; (KMH) Apartment house "Mehrfamilienhaus"; (GMH) Large apartment house "Großes Mehrfamilienhaus"; (HH) High-rise "Hochhaus";

Table 2.8: IWU-he building typology matrix for Germany (Born et al., 2003)

Baualtersklassen		< 1918	1919–1948	1949–1957	1958–1968	1969–1978	1979–1983	1984–1994
EFH	WKZ[a]	250/210	194	223	166	182/123	120	140
RDH	WKZ	204	166	163	135	159	129	97
KMH	WKZ	241/180	193	211	168	139	118	122
GMH	WKZ	159	164	173	172	140		
HH	WKZ				119	103		

(a) (WKZ) Specific Heat demand (spez. Wärmebedarfskennzahl) $[kWh/m^2a]$
(EFH), Einfamilienhaus (single family house); (RDH), Reihenhaus (row house); (KMH), Mehrfamilienhaus (multi-family house); (GMH), großes Mehrfamilienhaus (big multi-family house); and (HH), Hochhaus (high-rise building);

The available background information on building typologies is essential for the next steps of this analysis. After a classification of the building stock into building typologies, the individual buildings inherit attributes from the corresponding building type. These attributes are this background information, particularly important are the U-values used to estimate the heat demand of the individual types.

In this thesis the use of this typology contributes to the discussion[7]. This analysis can be seen as the missing documentation of this typology, with this typology we are able to compare two typologies specifically developed for the same urban space. In Chapter 4 we present a comparison of the BSU typology, presented in this Section and the Ecofys typology presented in the next Section 2.3.5, both typologies developed for the city of Hamburg.

The BSU typology is build analog to the rest of the typologies described in this thesis. The buildings are classified by construction epoch and construction type. For each type of the matrix we find 2 values:

1. (WKZ) Specific Heat demand $[kWh/m^2a]$; and
2. (PWKZ) Potential specific Heat demand after a renovation $[kWh/m^2a]$

Due to lack of information for some types no values are presented in the typology, in order to include them in the analysis we perform a simple polynomial interpolation to estimate these values. We used the available values from the same construction type in order to performed the polynomial interpolation. This values are presented in Table 2.9, the interpolated values are marked with an asterisk.

[7] This typology is no longer available online, making it even more important to discuss the role of this typology in the framework of my thesis

Table 2.9: BSU building typology matrix for Hamburg (BSU, 2011)

Construction period		<1918	1919-1948	1949-1958	1959-1968	1969-1978	1979-1983	1984-1994
EFH/DHH	WKZ[a]	260	262	258	262	204	170	126
RDH	WKZ	235	225	238	221	202	*198	*117
KMH	WKZ	194	171	171	156	146	105	85
GMH	WKZ	165	150	157	138	117	112	*88

* Values have been computed using a polynomial interpolation method as values for these types were not available on the original typology.
a) (WKZ) Specific Heat demand (spez. Wärmebedarfskennzahl) [kWh/m^2a]; (EFH) Single family house "Einfamilienhaus"; (RDH) Terrace house "Reihenhaus"; (KMH) Small apartment house "Kleines Mehrfamilienhaus"; and (GMH) Large apartment house "Großes Mehrfamilienhaus"

2.3.5 Ecofys — a Typology for Hamburg

In contrast to the previous typology, the typology presented by Ecofys, delivers detailed information on the individual types as well as good documentation of the applied methodology. In contrast to the BSU typology we use the results, rather than the formal definition of the typology, to perform the analysis (see Chapter 4 for methodological details). In Table 2.10 the postulated typology, with specific heat demand values can be seen. In the subsequent table, Table 2.11 a more detailed description of the typology, describing the number of floors and roof type, is presented. This table represents the result of the analysis performed by Hermelink et al. (2011). We use these results to create an "extended typology". For our analysis we base the typical building typology on construction epoch and construction type and include other parameters like roof type or number of stories for further classification of buildings that do not have either a construction year or a construction type. In Chapter 4 we describe this process in more detail. Because this typology contains more parameters that we can use for the classification of the building stock into building types we expect that this typology (Ecofys) will perform better than the previous typology (BSU).

The detailed description of the building typology, containing the "extra" parameters, described in Table 2.11 shows the percentage of buildings of each type allocated to the specific parameter. In their work the authors (Hermelink et al., 2011) conducted a classification of the building stock for the city of Hamburg.

2.3.6 TABULA — European Building Typologies

In the EU-financed project TABULA[8] 15 European countries have developed building typologies following the structure proposed and implemented by the IWU Institute from Germany, see (TABULA Project Team, 2012b) for an overview of the single country reports involved in the tabular project.

[8](TABULA) typology approach for building stock energy assessment `http://www.building-typology.eu/`

Table 2.10: Building typology for the city of Hamburg developed by EcoFYS showing the heat demand per residential space* in $[kWh/m^2a]$ (Hermelink et al., 2011, pp. 18)

Construction period		<1918	1919-1948	1949-1957	1958-1968	1969-1978	1979-1983	1984-1994	>1995
EFH/DHH	WKZ[a]	223	245	232	221	209	194	138	120
RDH	WKZ	147	154	129	140	126	99	88	78
MFH-E	WKZ	204	191	204	195	191	147	120	97
MFH-G	WKZ	136	145	145	145	115	111	94	91
MFH-W	WKZ	161	159	158	160	145	121	106	92
MFH-H	WKZ				131	132			

a) (WKZ) Specific Heat demand (spez. Wärmebedarfskennzahl) $[kWh/m^2a]$
(EFH) Single family house "Einfamilienhaus"; (RDH) terrace house "Reihenhaus,"; (MFH-E) Single apartment house "Mehrfamilienhaus Einzelhaus"; (MFH-G) Group apartment house "Mehrfamilienhaus Gruppenhaus"; (MFH-W) Building block apartment house "Mehrfamilienhaus Wohnblock"; (MFH-H) High-rise apartment house "Mehrfamilienhaus Hochhaus";

One of the aims of this project is the estimation of national energy potential savings of the residential building stock.

> "One important objective of the set-up of national building typologies is the elaboration of bottom-up models which enable a calculation of the energy consumption of the respective building stocks. A typical application field is the investigation of energy saving potentials for a national or regional building stock as well as the design and evaluation of instruments and political strategies." (TABULA Project Team, 2012a, p.28)

The report presented by (TABULA Project Team, 2012a) points towards the development of national building stock models, the development of a building stock model at a city level is not explicit mentioned in the report. The development of a specific typology for a given city proves to be complicated because of lack of available and homogeneous statistics. For an overview of the different data sources used in the project by the participant institutions see Table 2.12.

For some typologies presented in the TABULA report, a differentiation between climate zones is performed. This is especially important for large countries expanding through different climate zones. Countries expanding through climate zones with a substantial difference in air temperatures will have different building types. Countries differentiating climate zones in the projects are:

1. Denmark (TABULA Project Team, 2012c, p. 49);
2. Greece (TABULA Project Team, 2012c, p. 59);
3. Sweden (TABULA Project Team, 2012c, p. 101); and

Table 2.11: Building typology for the city of Hamburg developed by EcoFYS showing the parameters of each Type, in percentage. (Hermelink et al., 2011, pp. 18)

	Year (BAJ)	Floor (AOG)							Roof (DAF)				
		01-02	03-03	04-05	06-09	10-13	14-15	16-20	sa	m	w	f	so
Freist. EFH / DHH Semidetached / Single family house	<1918	94	6	0	0	0	0	0	51	13	26	5	5
	1919-1948	99	2	0	0	0	0	0	60	5	26	4	5
	1949-1957	99	1	0	0	0	0	0	71	2	19	4	4
	1958-1968	98	2	0	0	0	0	0	65	3	20	9	3
	1969-1978	99	2	0	0	0	0	0	61	3	24	10	3
	1979-1983	99	1	0	0	0	0	0	65	4	22	6	3
	1984-1994	99	1	0	0	0	0	0	47	5	42	3	2
	>1995	98	2	0	0	0	0	0	50	3	36	5	6
Reihenhaus Terrace house	<1918	84	13	3	0	0	0	0	37	24	14	23	2
	1919-1948	95	5	0	0	0	0	0	68	3	19	10	1
	1949-1957	99	1	0	0	0	0	0	85	1	2	11	1
	1958-1968	100	0	0	0	0	0	0	77	1	1	20	1
	1969-1978	97	3	0	0	0	0	0	52	2	1	43	2
	1979-1983	96	4	0	0	0	0	0	79	7	1	10	2
	1984-1994	98	2	0	0	0	0	0	72	12	9	2	5
	>1995	76	23	0	0	0	0	0	50	0	6	24	19
MFH-Einzelhaus Single family house	<1918	49	38	11	1	0	0	0	30	29	25	10	6
	1919-1948	69	26	4	0	0	0	0	24	12	53	7	4
	1949-1957	68	22	10	1	0	0	0	44	8	28	17	3
	1958-1968	68	21	8	3	0	0	0	57	8	13	19	3
	1969-1978	59	32	8	0	0	0	0	44	8	12	34	2
	1979-1983	65	26	8	1	0	0	0	39	24	15	20	2
	1984-1994	72	21	6	1	0	0	0	36	27	24	7	6
	>1995	48	31	18	2	0	0	0	31	6	22	29	12
MFH-Wohnblock Block family house	<1918	6	19	67	8	0	0	0	19	66	2	11	2
	1919-1948	13	26	51	9	0	0	0	29	40	11	18	3
	1949-1957	7	25	63	5	0	0	0	51	20	11	17	1
	1958-1968	10	25	59	7	0	0	0	48	28	5	18	2
	1969-1978	4	22	65	8	0	0	0	34	39	6	19	2
	1979-1983	6	23	63	9	0	0	0	35	38	7	17	3
	1984-1994	7	23	61	8	0	0	0	34	43	5	14	4
	>1995	25	31	34	9	0	0	0	30	18	6	38	8
MFH-Gruppenhaus Group family house	<1918	6	33	55	6	0	0	0	41	38	10	7	4
	1919-1948	8	28	54	10	0	0	0	33	4	37	25	1
	1949-1957	12	23	62	3	0	0	0	62	1	20	17	1
	1958-1968	17	43	38	3	0	0	0	59	2	4	34	1
	1969-1978	12	29	49	9	0	0	0	35	3	11	51	1
	1979-1983	12	30	48	10	0	0	0	44	11	11	33	1
	1984-1994	19	37	40	5	0	0	0	49	18	15	15	3
	>1995	14	33	45	9	0	0	0	29	4	13	45	9
HH*	1958-1968	0	0	0	55	32	5	6	5	3	0	89	3
	1969-1978	0	0	1	39	47	7	6	0	0	0	99	0

* MFH-Hochhaus, massiv. (sa) pitched roof, "Satteldach"; (m) curg roof, "Mansardendach"; (w) hip roof, "Walmdach"; (f) flat roof, "Flachdach"; (so) other, "Sonstiges".

4. Spain (TABULA Project Team, 2012c, p. 115).

The implementation of a regional typology could be derived from this classification. The representation of the Hellenic building stock with help of the developed building typologies in the TABULAR project are presented by Dascalaki et al., in their paper publish on (2011) the authors present an assessment of various energy conservation measures used for the estimation of heat consumption of the residential sector in Greece. Kragh und Wittchen deliver an overview of the Danish building typology in their paper from (2013), This typology is based on the database of the Danish Energy Performance Certificate Scheme. The Danish building typology presents two types of buildings models:

1. Real example models; and
2. Average designed models

The calculation of the heat demand of the models is performed trough an energy balancing method. An example of the use of building typologies in the scientific community is delivered by M. K. Singh et al. in their paper from (2013). The performed analysis by the authors focus on Liege, Belgium. The aim of this work is to apply the developed building typologies to estimate building characteristics and subsequent its heat consumption. The result is used for the identification of possible measures that the federal state can implement in order to efficiently reduce the heat demand of the residential in the specific region.

The use of building typologies to estimate the mitigation potential and vulnerability of the residential sector is addressed by Hrabovszky-Horváth et al. in their paper from (2013), developed for Hungary. An application of Building typologies at an urban scale is performed by Caputo et al. in their paper presented in (2013). The authors make use of a GIS system to collect and classify the building stock in the city of Milan. The aim of their work is to support energy policies at an urban level. The authors validate their model with help of an energy information system (SIRENA), which contains aggregated energy consumption for the Lombardy region. The authors make an important contribution in the analysis of small urban areas and show the potential of such an analysis for the further development of cities evaluating possible expansion areas for district heating and cogeneration systems. Dall'O', Galante und Torri (2012) present an analysis for the same region also focusing on urban scale.

Table 2.12: Used data for the different building typologies of the TABULA project (TABULA Project Team, 2012c)

Country	Data source. As quoted in (TABULA Project Team, 2012c)
Belgium	(1) General Socio-economic Survey performed in 2001 by the National Institute of Statistics NIS (2) Energy Advice Procedure database
Czech Republic	(1) Public database of the Czech Statistical Office (2) National census 2001 (3) Microcensus ENERGO 2004
Denmark	(4) The Danish building stock register (BBR) (5) The building Energy Performance Certification (EMO) database
Germany	(6) Energy certificate data base of the German Energy Agency (dena) (7) “Daten-basis Gebäudebestand" (Diefenbach, Cischinsky & Rodenfels, 2010)
Greece	(1) National Census 1990 and 2000 (2) Hellenic Ministry for the Environment, Physical Planning and Public Works, Directorate Urban Planning & Housing - MEPPPW
Italy	(1) National Institute of Statistics (ISTAT - Report 2004) (2) Centre Economical, Social and Market Surveys in the Building Sector (CRESME) (3) National Energy Agency (ENEA)
Slovenia	(1) Registry of buildings of Slovenia

2.4 Allocating People to the Building Stock in Space

The aim of this thesis is the simulation of heat consumption in space at a low level of aggregation. In order to further explain what this is and why do we need such a vast literature base (presented in the last 4 sections) to achieve this, we need to break up this aim into individual parts:

1. Simulation of heat consumption (see Section 2.1);
2. The need to take demographic characteristics of the population into account (see Sub-Section 2.1.3 for the integration of user behavior into simulation models and Section 2.2 for the generation of a synthetic population);
3. A spatial referenced classified building stock, see Section 2.3 for a classification of the building stock using building typologies; and
4. The allocation of individual families into the classified building stock.

The first point distinguish the simulation of heat consumption from the simulation of heat demand. This small distinction drives the architecture and data structure of the entire model and of this thesis. The differentiation between consumption and demand is normally applied between estimated values (demand) and monitored values (consumption). In this case we talk about the simulation of heat consumption because the developed model doesn't aim to capture an average heat demand of the building but to simulate a possible heat consumption of this building. In order to simulate this consumption we need to consider human behavior in the model.

The explicit consideration on human behavior in a traditional heat demand model presents itself as a big challenge, especially for the residential sector where behavior patterns vary considerably among individuals. For the simulation of heat demand at a building level we could design a stochastic simulation model to consider a wide range of behaviors' influencing heat consumption, but not its driving heat demand pattern. By doing this on small geographical areas we can allocate specific individuals to specific buildings. This advantage allows us to analyze consumption patterns of these small areas and quantify the effect of user behavior at a low aggregation level. The quantification of this effect is important for the proper dimensioning of decentralized supply systems (outside the scope of this thesis).

In the following chapters we present a step by step description and discussion of the individual components of the developed model. The scope of this thesis is divided into two main topics: (1) building simulation; and (2) spatial microsimulation. The first topic deals with the underlying physics for the simulation of energy demand mainly at a building level, while the second topic deals with the development of simulation models at a micro level with a spatial reference. Commonly on the second topic the individual simulation units are either families of individuals. An essential part of the microsimulation models is the generation of a synthetic population for small urban areas that matches the aggregated statistics for the same area. In this thesis we make use of these methods for the generation of synthetic families and in a second step allocate these families to the classified buildings stock. We use all this generated micro-data to feed a thermal simulation model for the estimation of heat demand at a low aggregation level with an explicit consideration of user behavior.

3 Model Architecture and Data Sources

Most of the model is implemented using the python language using some common python libraries, see (Payne, 2010; Oliphant, 2007) for the *numpy* and *scipy* python library used for the scientific computation and (Hunter, 2007; Tosi, 2009) for the *matplotlib* for the used library for plotting and visualization. Most of the data processing performed on the model uses the *Pandas* library (Pandas development team, o. J.).See McKinney (2012) for an overview of the library. Many of the examples presented in this thesis are implemented in the *IPython Notebook* (Pérez & Granger, 2007).For a reference on the IPython interface see (Rossant, 2013).The manipulation of spatial elements as well as most spatial analysis are performed using the shapely library (Gillies, 2013). Most of the spatial elements stored in the PostgreSQL database (see below) are parsed as shapely objects within the model. For the integration of the developed R libraries into the model architecture we make use of the *RPy* library (Moreira & Warnes, 2004). Specially the spatial microsimulation model *GREGWT* and the corresponding library written in the R language are used in a python environment through the use of the RPy library. In addition to these libraries, the relational database management system (RDBMS) *MySQL* is used to store the input data as well as part of the constructed data. See (DuBois, 2007; Oracle, 2011) for more information regarding the database functionality and (Dustman, 2013, 2007; Python Software Foundation, o. J.-a) for the python interface to *MySQL*.

For the storage and process of spatial data the *PostgreSQL* object-relational database is used (see Section 3.2 for data specifications). See (The PostgreSQL Global Development Group, 2012, 2013) for documentation of the *PostgreSQL* database. In addition to the *PostgreSQL* database the *PostGIS* database extender is used to add support for geographic objects. See (PostGIS, 2013) for further information. For the storage and process of spatial data the *PostgreSQL* object-relational database is used (see Section 3.2 for data further information and application possibilities see (Python Software Foundation, o. J.-b; psycopg, 2013). Westra (2010) provides an overview of spatial analysis using the python language.

The model architecture is based on an object oriented programing principle, not only because it is built upon an object oriented programming language (python) but the structure of the model uses all the advantages of an object oriented language, representing so each individual unit as an object described by a language class. Many authors argue in favor of a true object oriented model (Ballas et al., 1999; Miller, Hunt, Abraham & Salvini, 2004; Rahman et al., 2010). Having an object oriented architecture sets the ideal platform for a development of the model towards an agent-based model (ABM). Many authors acknowledge the benefits of an agent-based simulation, a simulation capable of computing the emergence of properties arising from the interaction of single individuals (Birkin & Wu, 2012). We see the combination of this (and other approaches e.g. Cellular Automata) as the direction to go and therefore prepare our model for our future work towards an agent based urban model.

The objects in our model are not only objects interacting with each other but are build hierarchical, where spatial objects are able to contain objects on a smaller geographical aggregation level. That means that the object "Ward" contains "Building Blocks" which contain "Buildings" with have "Dwelling Units" that have "Individuals". This hierarchical architecture is not a simple relationship within the tables stored in the database but the object is created with a node to the corresponding upper and lower levels. This hierarchical architecture gives the developer a lot of flexibility, making it possible to iterate trough all items in a given ward without having to connect them first, either through a spatial join or through an id in a database table.

In the framework of my thesis I tested some of the most common approaches for the generation of synthetic micro data: (1) Iterative Proportional Fitting (IPF); (2) Combinatorial Optimization (CO); and (3) Generalized Regression and Weighting (GREGWT). All of these approaches present some advantage over the others.

The Combinatorial Optimization algorithm, more specifically a Simulated Annealing, based on the guidelines presented by Harland et al. (2012) its particularly interesting for the case of Hamburg because:

1. available data for the city of Hamburg is ideal for a simulated annealing, as I am able to directly construct households containing individuals based on the micro-census data;

2. I can take advantage of an object oriented programming structure, simulating households and individuals as object from the initial state of the simulation.

3. I need individual families (rather that weights) in order to allocate them to the building stock.

The IPF algorithm is a well-studied algorithm not only in the microsimulation community. Implementations of this algorithm are fast and robust. The output of the algorithm are weights of the individual records of the initial survey. This algorithm does not take initial weights into account for the simulation. The use of initial weights can be useful in certain scenarios where the generation of synthetic population needs to be alight to a second aggregate.

GREGWT tries to minimize the weight distance, that is the distance between the initial weights of the original survey and the estimated new weights. This minimization can be useful for certain scenarios, but can also create problems during the simulation. Because of this extra constraint, the GREGWT algorithm can fail to find appropriate estimations in remote areas where the population differs significantly from the survey used in the re-weighting process.

3.1 Data Storage and Management

One of the most time consuming tasks of this thesis has been the data processing and data management. We decide to store and maintain all the data on two different databases: (1) a *MySQL* database dedicated to store all the demographic data and (2) a *PostgreSQL* database dedicated to store all the spatial data.

3.2 Data Storage and Management

database and create a synthetic population. On a second step we populate the buildings store in the *PostgreSQL* database with the created synthetic families. An alternative to this method is discussed on Section 7.2 where we create not only synthetic families but also a synthetic building stock. The advantage of this method is that it allows us to simulate the heat consumption at a micro scale for the entire country at a NUTS-3 level.

Figure 3.1 gives a schematic overview of the simulation architecture. In this figure both databases are represented as well as the different available aggregation levels. The available spatial layers stored in the *postgreSQL* database are used at different stages of the simulation. From aggregated statistics we generate a synthetic population, in theory a synthetic population can be generated at any spatial aggregation level. Limits to which level of aggregation a generation of a synthetic population are attached to the availability of data. Problems with the presented method may also occur at a low aggregation scale. Because the presented method is a reweighting algorithm rather that a synthetic reconstruction mechanism the reweighting of a population at a very low aggregation level may have difficulties achieving convergence because populations at these levels may be specific to the reweighted population. This case is not very common as data at a very low level of aggregation is hard to get or protected through data scrambling mechanisms.

The classification of the building stock can be achieved through the use of building typologies, see Section 2.3 for a description of building typologies and Section 4.2 for the implemented method for the automatic classification of the building stock through the use of building typologies.

The building stock data is retrieved from the city digital cadastre ALKIS and stored in the postgreSQL database. The "raw" data is then processed into hierarchical python objects.

The classified building stock is populated by the synthetic families. The data of each individual building (building geometry and other energy relevant parameters) and the resident living on them (demographic characteristics and specific family schedules) is used as input for the estimation of heat demand. The computation of heat demand can be performed via: (a) a heat balance method (see Section 7.1) or a more detailed thermal simulation model (see Section 5.6).

Table 3.1 lists the used spatial layers in the analysis. Most of these layers are available directly from the digital cadastre except for the layer containing the boundaries of the statistical areas. Most of the data from the digital cadastre is very clean and does not need any special cleaning. The "extra" layer added to the database, containing the boundaries of the statistical areas is not as accurate as the rest of the digital cadastre. In this case problems can arise at the edges of the areas. In order to avoid this, the internal engine of the model always takes the building (and other spatial objects) centroid to perform spatial joins between data layers.

Table 3.1: Main layers / data sources from ALKIS and the constructed spatial joins. This table shows the example of the building (Geb) layer, similar spatial joins are constructed for the other layers and follow a similar syntax

	ID	Layer-de	Layer-en	Layer-ALKIS
a	HH	Hamburg		
b	Bez	Bezirk	Borough	`Bezirke`
c	Std	Stadtteil	District	`Stadtteile_2010`
d	Sg	S. Gebiete	S. Area	`Statistische_Gebiete_utm`
e	BBZ	Baublock	Building Block	`AX75001`
f	USE	Nutzung	Actual use	`SD41000`
g	Gr	Grundstück	Lots	`AX11001`
h	Geb	Gebaude	Buildings	`AX31001`
	BJA		Construction year	`AX31001_BJA`
i	Add	Adresse	Address	`AX12006-UTF`

Probably the most important data layer is `AX32001`, this layer contains the geometrical and attribute data from the building stock. The data containing information about the construction years of the building stock is not stored in the same spatial layer, an attribute join is necessary in order to create a layer with both: construction year and building geometry. The digital cadastre also provides address points, this information is relevant for the allocation of consumption data to the individual buildings. The lots data is used to extend the address information to all the buildings on the same lot.

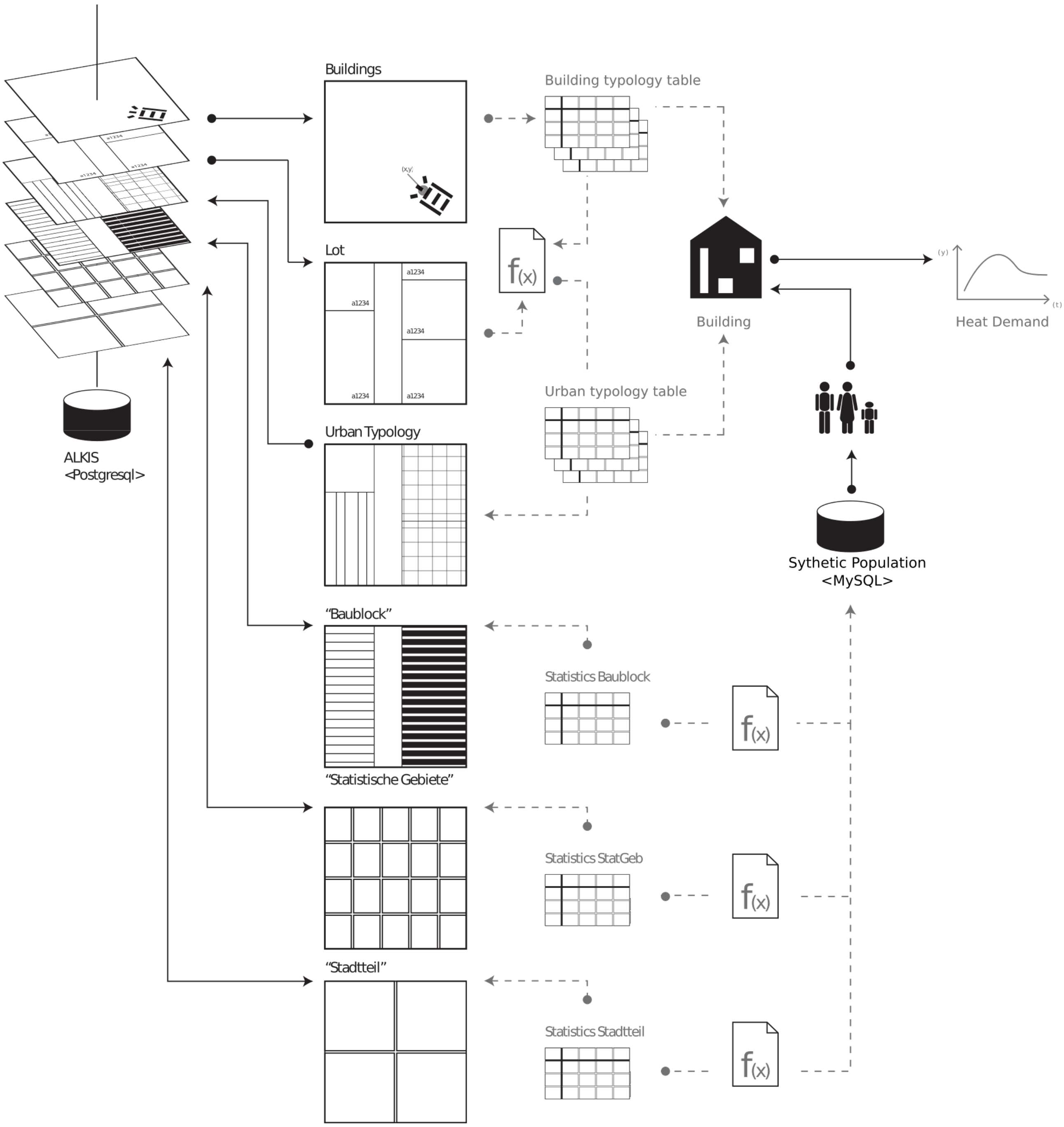


Figure 3.1: Geographical and population databases with main relationships of the simulation model

3.3 Main Data Sources

For the presented method in this thesis we use the following data sets:

1. The 2010 German micro census[1].

2. Census records available at a district level, which correspond to the *Nomenclature of Territorial Units for Statistics*[2,3] **(NUTS 3)** aggregation level.

3. The time-use survey[4].

4. Historic demographic benchmarks for Hamburg districts[5]. This data set contains historical data, describing the following parameters: (1) total population, (2) gender distribution and (3) foreign national share.

5. Demographic projections for Hamburg until year 2030[6]. The available projections for the city of Hamburg are age/gender tables projected from base year 2008 until 2030.

6. The digital cadastre of the city of Hamburg ALKIS[7] for the year 2010. The digital cadastre contains information regarding the building stock of the city. This dataset contains: (1) the geometry of the entire building stock of the city; and (2) some attributes of the individual buildings like: construction year, construction type (single family house, terrace house, etc.), number of stories, etc. We use a pre classified version of the cadastre performed by Muñoz H. und Peters (2014a), describing the buildings as types of a defined building typology.

In order to create synthetic population at a small geographical level we need two base data sets: (1) a micro data set containing individual records and (2) aggregated statistics available at the geographical level we aim to reweight the dataset containing individual records. For the purpose of this simulation we want to create a synthetic population describing the demographic characteristics of the individuals and characteristics of the dwelling units these individuals reside on. In order to generate such a data set we select data describing three different aggregation units: (1) individuals, (2) household/dwelling units and (3) buildings. Table 3.4 list the used variables from the two data sets.

[1] `forschungsdatenzentrum.de/bestand/mikrozensus`
[2] `http://ec.europa.eu/eurostat/web/nuts/overview`
[3] `http://database.espon.eu/db2/jsf/DicoSpatialUnits/DicoSpatialUnits_html/ch01s01.html`
[4] `forschungsdatenzentrum.de/bestand/zeitbudget`
[5] Bevölkerungsentwicklung in Hamburg
`www.statistik-nord.de/daten/bevoelkerung-und-gebiet/bevoelkerungsstand-und-entwicklung`
Years 2011–2013 Update based on Zensus 2011 (Fortschreibung auf Basis des Zensus 2011) Years 2003–92011 Update based on Zensus 1987 (Fortschreibung nach den Ergebnissen der Volkszählung 1987)
[6] Bevölkerungsvorausberechnung für Hamburg
`www.statistik-nord.de/publikationen/publikationen/statistische-berichte/bevoelkerung-und-gebiet/`
[7] ALKIS Amtliches Liegenschaftskatasterinformationssystem

3.3.1 The Micro Census

The micro census contains 489 630 individual entries. Each entry in the data set corresponds to a real person living at the time in Germany. The individual records can be group into families with help on one of its attributes. The data set has a detailed description of each individual, describing the individual with help of 529 attributes. Out of this 529 attributes we use 11 to fetch a representative individual form the time-use survey.

The free available German micro census contains 23, 374 records and corresponds to the 3% of the official micro census data for scientific use containing 1% of the total German population. This data set represents 0.03% of the German population. The file can be downloaded from the *German Scientific Data Center*[8].

Table 3.2 shows the data structure of the micro census. Both R libraries, GREGWT and IPF, accept this data format as input. In the case of the SAS implementation of the GREGWT algorithm this table had to be translated to binary data, because the GREGWT method requires all individual categories listed as columns on the input matrix. The R implementation of the GREGWT algorithm will make this conversion internally.

Table 3.2: Survey data structure

ID	Age	Marital Status	Household Size	Weights
1	65...74	Divorced	1.person	90,173
2	40...49	Married	4.persons	119,987
3	40...49	Married	4.persons	119,987
4	18...24	Single	4.persons	119,987
5	75.and.over	Widowed	4.persons	119,987
⋮	⋮	⋮	⋮	⋮

3.3.2 The Census Data

The synthetic population is represented as the reweighted German micro census, we reweight this survey with help of an R library implementing the GREGWT method (Muñoz H., Vidyattama & Tanton, 2015a). The GREGWT method is classified as a deterministic reweight method (Tanton et al., 2014). Deterministic reweighting methods aim to reweight a survey to match known aggregated values of geographical areas. The size and available data of these geographical areas vary between countries. For the European Union a standard incorporating the different national definitions exist. This is the Nomenclature of Territorial Units for Statistics (NUTS[9]) standard. This nomenclature described four hierarchies: (0) national territories; (1) NUTS–1; (2) NUTS–2; and NUTS–3. A reweighting of a national survey could be implemented at any of the NUTS levels. Depending on the research question a suitable geographical area should be selected. These geographical areas have different names on each

[8] `http://www.forschungsdatenzentrum.de/bestand/mikrozensus/cf/2010/`
[9] `http://ec.europa.eu/eurostat/web/nuts/overview`

country and the authors referee them according to the use case location. These areas are known as: (a) Summary Files in the U.S.; (b) Profile Tables or Basic Summary Tabulations (BSTs) in Canada; (c) and Small Area Statistics in the U.K. (Pritchard & Miller, 2012).

The census data listed on Table 3.4 comes from the census 2011 and can be accessed through the official web portal of the census 2011[10]. All the results from the census can be accessed through the web portal. Unfortunately there is no API to retrieve the data remotely. In order to download the data, the desired table has to be created giving as input the desire parameters and areas to be downloaded, we export this table as a CSV file but the data can be exported into different data types.

Table 3.3: Used benchmarks from the 2011 Census and corresponding micro census attributes

MC Code*	Census Code	Unit**	Description
EF1	/	/	Federal State (NUTS 2)
EF952	/	Person	Weight
EF44	ALTER_01JS	Person	Age (yearly stages)
EF44	ALTER_KURZ	Person	Age (five classes of years)
EF49	FAMSTND_AUSF	Person	Marital status (in detail)
	FAMSTND_KURZ	Person	Marital status
EF809	FAMTYP_LEB	Person	Type of family nucleus (by living arrangement)
EF46	GESCHLECHT	Person	Sex
EF20	HHGROESS_KLASS	Person	Size of private household
EF809	HHTYP_FAM	Person	Type of private household (by family)
EF368	STAATSANGE_KURZ	Person	Citizenship
	RAUMANZAHL	Dwelling	Number of rooms
EF491/95	WOHNEIGENTUM	Dwelling	Ownership of dwelling
EF492	WOHNFLAECHE_20S	Dwelling	Floor area of the dwelling ($20m^2$ intervals)
EF494	BAUJAHR_MZ	Building	Year of construction (micro census classes)
EF495	EIGENTUM	Building	Type of ownership of building
EF570	GEBTYPBAUWEISE	Building	Type of building (construction)
EF496	HEIZTYP	Building	Type of heating
EF635	ZAHLWOHNGN_HHG	Building	Number of dwellings in a building

*Micro Census Code
**Refers only to Census

The structure of the table describes geographical areas by parameters categories. This table represents the total number of individuals corresponding to the parameter in a specific area. In this case the categories of the parameters from both datasets match, this is not the normal case. Many times a time intensive preparation of the input data is needed in order to reweight the sample.

[10] `https://ergebnisse.zensus2011.de/`

Table 3.4: Structure of the census data for the four simulation areas

Area	Area Code	Under.3	3...5	...	Divorced	...	1.person	...
Hamburg	2	47757	45575	...	143354	...	400440	...
Berlin	11	94867	86753	...	333989	...	860542	...
Bremen	4011	12996	12714	...	48267	...	121056	...
Bremerhaven	4012	2643	2822	...	9723	...	26444	...

3.3.3 Time-Use Survey

Similar to the micro census, the time-use survey consists of individual records of people. Each record contains a wide set of attributes, for the method described on this thesis we make use of just a small section of the survey. In order to handle this data set in a more efficient way, we divided the survey into 11 different tables. Two of these tables are used to generate the needed data for the analysis on this thesis. The first table describes general characteristics of the individuals. This table contains 397 attributes and 13 691 records. The second table contains more records because each individual may have recorded a time use journal for more than one day. This table contains 148 attributes, 144 of these attributes represent the location of the individual in a 10 minute interval. The other four attributes are: (1) an internal unique ID;(2) a household ID;(3) the individual ID; and (3) a day ID.

4 The Use of Building Typologies[1]

4.1 Estimating Heat Demand for Small Urban Areas

The planning of cities and the planning of the underlying infrastructure to support the resources demand of the individual buildings is a challenge for both urban planners and infrastructure planners. In this section we discuss the first step into a knowledge base urban planning schema. In order to achieve such a planning schema we need information about the urban areas, here we present an approach to classify the building stock into building types for the estimation of heat demand.

We want to provide a **reproducible** and **flexible** method for the estimation of heat demand of urban areas at a low aggregation level. In this section we describe our first attempt to produce such a method. In order to make the method reproducible we create a github repository containing all the required scripts to reproduce the presented computations and the data used in this analysis. The provided data had to be decoupled from its spatial reference to ensure anonymity. We also provide an Ipython notebook containing step by step explanations on how to reproduce the classification process and estimation of heat demand with the provided scripts. All this data can be found under the link: `https://github.com/emunozh/btyp`.

In order to ensure flexibility we have constructed our set of scripts so that these scripts can be expanded to incorporate new parameters and new functions. The definition of individual typologies are hard coded into individual python files. The addition of new typologies would require to only provide a new python file. In order to facilitate this process we also provide a small script to generate a base typology for the easy generation of new typologies. The syntax used for this file can be compared to a JSON file. A knowledge of the python language is not a requirement for the use of these scripts.

We present a method for the estimation of heat demand of small urban areas at a low aggregation (individual buildings). For this estimation we make use of building typologies which provide us with a specific heat demand for each building type. Because the use of building typologies is central to this endeavor we present a comparison of different building typologies used for the estimation of heat demand in space. In this chapter we compare and asses the performance of these typologies.

We divide this chapter into three main sections: (1) Section 2.3.1 describes the used typologies and the general structure and use of building typologies. This section focuses on German typologies but mentions also the use of building typologies for other European countries; (2) Section 4.2 presents the developed method for the classification of the buildings stock; and (3) Section 4.3 presents and discuss

[1]This chapter is heavily based on: Muñoz Hidalgo, Dochev & Peters (2015) and Muñoz Hidalgo & Peters (2015)

the performance of the individual building typologies presented in Section 2.3.1.

In order to give a first comparison between typologies we plot the heat demand values of the types of the typologies we want to analyze. In Figure 4.1 we can see difference between typologies, we are able to compare the typologies in this way because all of them follow a similar classification structure. All the typologies compared in this section classify the single types by construction epoch and construction type. There are some small differences by the exact construction epoch of some typologies, there are differences between the division years (±1 year) and between the number or epochs. So, for example, in the last epoch not all typologies have a type representing this last construction epoch.

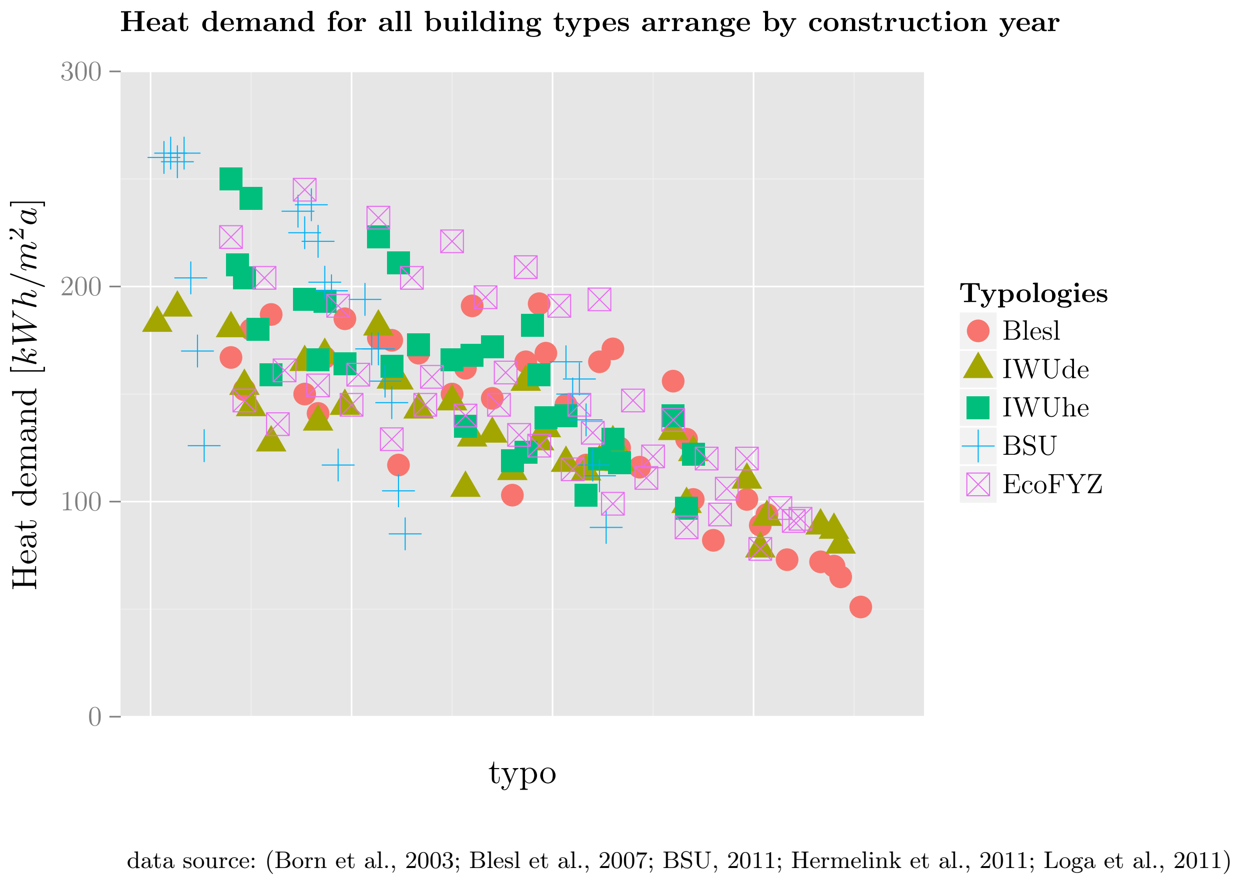


Figure 4.1: Different values for heat demand of building typologies used in Germany. The building types are arrange by construction year and construction type along the X-axis. The Y-axis shows the specific heat demand of the single typologies in [kWh/m^2a]

In Figure 4.1 the types of the different typologies are arranged by construction year (construction epoch), the graph clearly shows the improvements in heat demand over the years, independent of construction type. The figure only contains typologies developed for Germany as are these typologies the ones we want to understand and analyzed.

4.2 Automatic Classification of the Building Stock

In this section we present the developed methodology to classify the building stock into the different types of the presented typologies in the previous section. The buildings from the digital cadaster cannot be directly classified into the building typologies because: (1) the building attributes from the digital cadaster do not correspond to the building attributes of the building typologies; and (2) some attributes of the digital cadaster are not available for the entire building stock. The digital cadaster has a different classification for construction types to that of the buildings from the building typologies. In Sub-Section 4.2.2 we present the constructed method to cope with this problem. Some attributes available in the digital cadaster are not available for the entire building stock, this is specially problematic for buildings without a construction year, because the construction epoch is one of the key parameters to classify the buildings. The proposed algorithm for the classification can cope with unknown parameters, taking a random type from the used typology is case of missing attributes. We present two methods, or two arrays as are the arrays which differentiate the methods: (1) using a binary array; and (2) using a float array.

The aim of our research is to develop a generic method that can be used with any typology on any given building stock at a micro level, that is, each building has a building type and all its attributes. The combination of both data-sets: a building typology and a the geometry of the digital cadastre, create an optimal base for the estimation of heat demand. The building typology provides information about building materials and other characteristics of the individual building that are not present in the digital cadastre. The digital cadastre provides two important characteristics of the building stock: (1) the building geometry; and (2) the building location in space. The first attribute is important for the calculation of building heating space and transmission area of the building envelope, the second parameter is essential for the planning the distribution of heat.

In order to achieve these goals we developed a methodology that is able to cope with data gaps, both in the building stock database and in the building typology itself. The first step is to develop a data structure for the formal definition of the typologies, our computer program has to be able to read and understand the typology. In order to achieve this we have design a simple data structure, that is flexible and can be adapted to a new or to a different typology (see Sub-Section 4.2.2). We call this process "Typology Hard Code" as we translate the information of each individual typology (until now represented in simple tables) into a data structure expressed in the python language, that is, machine readable. We use the python language because the rest of the algorithm is also developed in this language, nonetheless, this data type may as well be defined in other more common data types like JSON[2] or xml[3].

In this section we: (a) describe the data used in the analysis; (b) present the postulated data structure to represent the typologies; and (c) conclude the section with the definition and discussion the filter array.

[2]JavaScript Object Notation (JSON)
[3]Extensible Markup Language (xml)

4.2.1 Data Sources: The Digital Cadastre and Gas Consumption Data

For the analysis performed in this section we make use of two main data sources (excluding the typology data): (1) the digital cadaster system of the city of Hamburg, ALKIS; and (2) monitored gas consumption values provided by the heat provider E.On-Hanse[4]. The data from the digital cadaster is a snapshot of the year 2010. With the available information of the digital cadaster we classify the buildings into types of the building typology under analysis and attribute each building with a specific heat consumption value (the corresponding value of the building type). We then estimate the living space of the buildings for the computation of absolute energy demand of the individual buildings. A detailed description of this process is presented in Section 4.3. The monitored gas consumption has been converted by the provider into heat demand values, presenting them as heat consumption in $[kWh/a]$. The consumption values used in this analysis correspond to the same year of the available digital cadastre 2010.

The original gas consumption dataset contains 308 observations. Each observation represents the gas consumption of a set of buildings. This forces us to analyze the performance of the individual typologies at this aggregation level. We are able to aggregate the estimated heat demand (through the typologies) because we know the address of each building within the buildings sets of the gas consumption data. Out of the original 308 observation we filter some implausible values e.g. the gas consumption was too low to be used as heating, leaving 290 observations in the data set.

The digital cadastre for the city of Hamburg has information on 369,416 individual buildings. We only classify the buildings that are represented in the gas consumption data (5,300 buildings). Although the building typologies are designed for the classification of the residential sector only, we have classified all the buildings including non-residential buildings. In our analysis we distinguish between residential and non-residential buildings for the computation of typology performance.

4.2.2 Typology Hard Code

The data structure presented here expands the building typology beyond the commonly used parameters of construction year and construction type. The definition of a typology under this data structure makes use of 6 parameters: (1) building use; (2) construction type; (3) construction year; (4) living space; (5) number of floors; and (6) roof type. The selection of these parameters arise from the available parameters on the digital cadaster and define parameters in the building typologies. This data structure is open and flexible, meaning that it could be expanded, including different parameters depending on the available information on the analyzed urban area and the available data. The parameters used in our analysis are the combination of available parameters form the digital cadaster and parameters used on the individual typologies. The algorithm processing these parameters will eventually have to be updated, it is therefore imperative to use a transparent and open source algorithm.

[4] `http://www.eon-hanse.com`

In the following sub-sections we describe these parameters, the relevance of the individual parameters and the use of these parameters within the typologies.

Building Use

The first parameter, *building use*, is important in order to define the analysis scope of the typology. With this parameter we can filter buildings from the digital cadaster by use. So for example if we use a typology design for the residential sector we can filter all buildings that have a not residential use. In our specific case the available consumption data, delivered by the gas provider, are aggregated in a way in which it is impossible to filter just the heat consumption from the residential sector. Therefore for our analysis we have to use the same typology for both residential and for non-residential buildings, in this case we simply change the typology code so that no building gets filtered by its use.

Integrating this parameter will allow us to eventually expand the analysis scope, to include an analysis of the non-residential sector. For such an analysis the parameter *building use* will make a first classification of the individual buildings into residential, commercial and other common uses. There are similar approaches, implementing building typologies, for the classification of the tertiary sector. Such an approach is presented by Loga et al. (2011), where an analog typology was developed for the non-residential sector. A different approach is presented by Blesl et al. (2007), in his approach buildings are classified by the required temperature demand for a specific building use and its underlying process. The integration of the last approach will require a redefinition of the source code under our postulated classification model.

Construction Type

The construction type is a very important parameter in the definition of the different buildings typologies and for the classification. The challenge dealing with this parameter, is that the classification of buildings into a construction type is a rather subjective matter. There is not a numerical definition or a systematic behind the construction types. That's why each typology may have a different definition and a different classification of construction types. The second problem is that the digital cadaster for the city of Hamburg has its own classification scheme.

In order to cope with this problem we had to define rules to merge the different construction types from the typology with the construction types from the digital cadaster.

These rules are defined in the form of a simple table (see Table 4.1). Because of the stochastic nature of the algorithm the rules in the table show possible combinations between the construction types, rather than deterministic rules to combine the different construction types. In Table 4.1 the construction types from the different typologies are represented horizontally in the table and the construction types from the digital cadaster, vertically. The X marks a possible combination between this construction types. This method may define that all construction types are possible, for example if the building is classified as of construction type "other" in the digital cadaster this building can be attributed to construction type from the typology. This attribute is also used to filter constructions that have a residential use but are not necessarily heated, at least not with the same intensity. So, for example we filter all the garage floor area from our analysis with this parameter.

Table 4.1: Construction types defined in the digital cadaster and aggregation scheme for typologies

		Building typologies								
		EFH	RDH	KMH	MFH-E	MFH-G	MFH-W	MFH-H	GMH	HH
Digital cadaster	SH	X								
	DSB	X								
	HR	X	X	X	X					
	TH		X							
	FBB			X				X	X	X
	BBCD			X			X		X	
	GH			X		X			X	
	O	X	X	X	X	X	X	X	X	X
	OH	X	X	X	X	X	X	X	X	X
	SG									
	DG									
	CG									

(EFH) Single family house (RDH) Terrace house (KMH) Small apartment house (MFH-E) Single apartment house (MFH-G) Group apartment house (MFH-W) Building block apartment house (MFH-H) High-rise apartment house (GMH) Large apartment house (HH) High-rise (SH) Semi-detached building (DSB) Detached single building (HR) House in a row (TH) Terrace house (FBB) Free-standing building block (BBCD)Building block in closed design (GH) Group building (OH) Open hall (O) Other (SG) Single garage (DG) Double garage (CG) Collective garage

Construction Year

The third parameter is the most common parameter, especially for the German typologies. This is because in Germany the quality of the building envelope has been regulated since the first "Wärmeschutzverordnung" (WSVO) Heat conservation ordinance in 1977. Since its first introduction, the German government has systematically introduced new regulations over the past decades creating substantial gaps in the quality of the building envelope over the years.

Other approaches in Europe developed for estimating heat demand at a neighborhood level do not take the building construction epoch into account but rather population density of the neighborhood (Finney, Chen et al., 2012; Finney, Sharifi et al., 2012; Finney et al., 2013). A classification of the building stock by construction epoch may not be well suited for areas where there is no substantial difference of the thermal properties of the building envelope between epochs.

The defined ranges of construction years in the algorithm are design to be used by any of the typologies analyzed in this section. For example the IWU-de typology has an extra construction period that none of the other building typologies have (see Table 7.1). The defined data structure contains construction periods that satisfy all the construction periods define in the individual typologies. If a single typology contains fewer periods than the ones define in the data structure, the periods will be simply merge together. See next subsection for a more detailed example.

Based on these two parameters (*construction period* and *construction type*), a "base typology" is generated, able to contain all the analyzed typologies. Table 4.2 is a graphical representation of this base typology with all the different typologies filling it. The table also contains the type number for each typology. These numbers are important as these are used within the algorithm to filter the single types of each typology. In the case of the construction year the extreme range of the typology will include all the lower or upper values. So for example in typology (3) IWU-he (highlighted green in the table) construction period [1860–1910] will be translated to construction period [< 1910]. This is defined in the hard code of the typology by defining both periods equally [1860–1910] = [< 1859].

Table 4.2: Base typology for the algorithm and position of the analyzed typologies

		<1860	1860-1918	1919-1948	1949-1958	1959-1968	1969-1978	1979-1983	1984-1994	1995-2000	>2000
(a)	EFH		01	05	09	13	18	23	27	31	35
(c)	RDH		02	06	10	14	19	24	28	32	36
(d)	KMH		03	07	11	15	20	25	29	33	37
(g)	GMH		04	08	12	16	21	26	30	34	38
(j)	HH					17	22				
(a)	EFH	01	03	07	11	15	20	25	28	31	34
(c)	RDH		04	08	12	16	21	26	29	32	35
(d)	KMH	02	05	09	13	17	22	27	30	33	36
(g)	GMH		06	10	14	18	23				
(j)	HH					19	24				
(a)	EFH		01	07	11	15	20	26	29		
(b)	EFH-b		02								
(c)	RDH		03	08	12	16	21	27	30		
(d)	KMH		04	09	13	17	22	28	31		
(e)	KMH-b		05				23				
(g)	GMH		06	10	14	18	24				
(j)	HH					19	25				
(a)	EFH		01	05	09	13	17	21	25		
(c)	RDH		02	06	10	14	18	22	26		
(d)	KMH		03	07	11	15	19	23	27		
(g)	GMH		04	08	12	16	20	24	28		
⋮	⋮										

Table 4.2: (continued)

		<1860	1860-1918	1919-1948	1949-1958	1959-1968	1969-1978	1979-1983	1984-1994	1995-2000	>2000
(a)	EFH		01	06	11	16	22	28	33	38	
(c)	RDH		02	07	12	17	23	29	34	39	
(f)	MFH-E		03	08	13	18	24	30	35	40	
(h)	MFH-G		04	09	14	19	25	31	36	41	
(i)	MFH-W		05	10	15	20	26	32	37	42	
(k)	MFH-H					21	27				

(EFH) Single family house “Einfamilienhaus”; (RDH) Terrace house “Reihenhaus”; (KMH) Small apartment house “Kleines Mehrfamilienhaus”; (MFH-E) Single apartment house “Mehrfamilienhaus Einzelhaus”; (MFH-G) Goroup apartment house “Mehrfamilienhaus Gruppenhaus”; (MFH-W) Building block apartment house “Mehrfamilienhaus Wohnblock”; (MFH-H) High-rise apartment house “Mehrfamilienhaus Hochhaus”; (GMH) Large apartment house “Großes Mehrfamilienhaus”; (HH) High-rise “Hochhaus”. Blesl ; IWU-de ; IWU-he ; BSU ; (5) EcoFYS ;

Floor Space

The floor space is an important parameter used in many of the typologies, this parameter can be used to differentiate between small apartment house (KMH) and big apartment house (GMH). Not all typologies define the average floor area of its types. In Figure 4.2 we plot the floor space used to estimate heat demand of three building typologies for which the floor space is defined (a, d) Blesl, (b, e) IWU-de and (c, f) IWU-he. The figure shows in the first row (a, b and c) the floor space for all types while the second row (d, e and f) shows only types with more than 4,000 m^2. In the figure we also plot the limits line (red), this limits represent the rules for the classification of buildings into building types based on floor space. With these limits a probability for the filter array (see Sub-Section 4.2.3) can be defined. The definition of this probability uses 4 values representing: (1) probable, no building gets filtered out (1.0); (2) less probable, 20% of buildings are not attributed to this typo (0.8); (3) un-probable, 40% of buildings are not attributed to this typo (0.6); and (4) impossible, 100% of buildings are not attributed to this typo (0.0). A summary of these probabilities for the typologies are described in Table 4.3.

The living space is estimated for all the buildings in the digital cadaster. We estimate the floor space as:

$$sqm = groundarea \times stories \times k \tag{4.1}$$

Where; sqm is the living space in m^2; $groundarea$ is the polygon area of the building in the digital cadaster; and k is a constant (0.6), differentiating so construction space (exterior end internal walls)

Table 4.3: Probabilities assign to building types as function of floor space

living space $[m^2]$			EFH (a) EFH-b (b) RDH (c)	KMH (d) KMH-b (e) MFH-E (f)	GMH (g) MFH-G (h) MFH-W (i)	HH (j) MFH-H (k)
$<$	-	400	1.0	0.6	0.0	0.0
400	-	600	0.6	1.0	0.6	0.0
600	-	4 000	0.0	0.6	1.0	0.6
4 000	-	10 000	0.0	0.0	0.6	1.0
$>$	-	10 000	0.0	0.0	0.0	1.0

from living space.

This parameter is only used with the float array filter (see Sub-Section 4.2.3 for a detail description of the array filter) and only for the three typologies (blesl, iwu-de, iwu-he) which define the average living space of its types.

Number of Floors

The number of floors is directly retrieved from the digital cadaster. This parameter is defined in some typologies and can help us to classify the buildings between construction types. This parameter is only used with the float array filter (see Sub-Section 4.2.3 for a detail description of the array filter), as defining this parameter in the binary filter may lead to inconsistencies within the data. E.g.: a building in the digital cadaster may have no possible type because it eliminates all possible types, if the building is of construction type (KMH) and is 10 stories high. In the defined typology, a particular building may be correctly attributed a construction type, but may be filtered out, because of a deterministic definition of a limit in the number of stories category. In the float array this will only decrease the probability of a type being peeked out.

The "rules" for the attribution of probabilities for the individual types, as function of number of stories, works similarly to the attribution of probabilities as a function of the living space (see sub Sub-Section 4.2.2). In Table 4.4 the number of stories for the different types of two typologies (iwu-he and Ecofys) are plotted along with the defined limits needed for the probability definition. The definition of these probabilities can be seen in Table 4.4.

Roof Type

Some typologies define the type of roof of the individual types. Because we can gather this information from the digital cadaster, the parameter can be used to classify the buildings. This parameter is only used with the float array filter (see Sub-Section 4.2.3 for a detail description of the filter array). Figure 4.4 shows the different roof types per construction type for two typologies.

Table 4.4: Probabilities assign to building types as function of number of stories

# **stories**			EFH (a) EFH-b (b) RDH (c)	KMH (d) KMH-b (e) MFH-E (f)	GMH (g) MFH-G (h) MFH-W (i)	HH (j) MFH-H (k)
<	-	4	1.0	0.6	0.0	0.0
4	-	5	0.6	1.0	0.6	0.0
6	-	9	0.0	0.6	1.0	0.6
9	-	15	0.0	0.0	0.6	1.0
>	-	15	0.0	0.0	0.0	1.0

4.2.3 Filter Array

In this subsection we present the functionality of the filter array with a small example, presenting both array types: (1) the binary array; and (2) the float array.

Imagine we define construction period [< 1859] as att_1 and construction period $[1860 - 1918]$ as att_2 and we analyze typology IWU-de (see Table 7.1) and typology BSU (see Table 2.9). Typology IWU-de has 36 types and typology BSU has 28 types. We create two binary vectors, one for each attribute att_1 and att_2, for each typology, represented in Equation 4.2 and 4.3 for the BSU and IWU-de typology. The length of these vectors correspond to the number of types on the corresponding typology.

$$att_{1\ bsu} = [1, 0, 0, \ldots, 0]$$
$$att_{2\ bsu} = [1, 0, 0, \ldots, 0] \tag{4.2}$$
$$att_{1\ iwu-de} = [1, 0, 0, \ldots, 0]$$
$$att_{2\ iwu-de} = [0, 1, 0, \ldots, 0] \tag{4.3}$$

Each element in the vector represents the probability of being of certain building type, given that att_i is true. In the binary vector we only define the probability as yes/no values represented by 1 and 0. In this case, position 1 of vector $att_{i\ bsu}$ corresponds to building type [efh/dhh < 1918], position 2 corresponds to type [efh/dhh 1919–1948] ans so forth (see table 2.9). If we want to give a probability to two different buildings having two different construction years, e.g.: (1) building (a) has a construction year = 1900; and (2) building (b) has a construction year = 1800. In this example all types that do not correspond to the given construction period will have a probability of 0, they will be filtered out. We still have different types that satisfy this restriction. In a second step we apply the same procedure, but using the construction type as filter. After this filter we may still have two options, in this case one of these types will be randomly selected (see Figure 4.5a). We perform this filtering a defined number of times (i.e. Monte Carlo) defining so the possible distribution of building types of that particular building and subsequent, the heat distribution, and therefore the attached uncertainties, of heat supply for small urban areas.

Because some typologies define the average living area and the average number of stories we can use these parameters to further filter possible types for a specific building. In order to do this we apply a *float array*, which defines the probability of being from a certain type as? float[5] values between 1.0 and 0.0 (see Figure 4.5b).

With this method we can define a matrix containing all the filter probabilities for each attribute defined in the typology, if an attribute is not defined in the typology the probabilities for this attribute (vector) will all be equal to 1. The matrix can be defined as in Equation 4.4, and a probability vector for each single building can be extracted, given the individual parameters of the building taken out of the digital cadaster. This vector is used to choose a single type for the typology, the mathematical expression of this vector is written in Equation 4.5 and Equation 4.6. The entire system is graphically described on Figure 4.5.

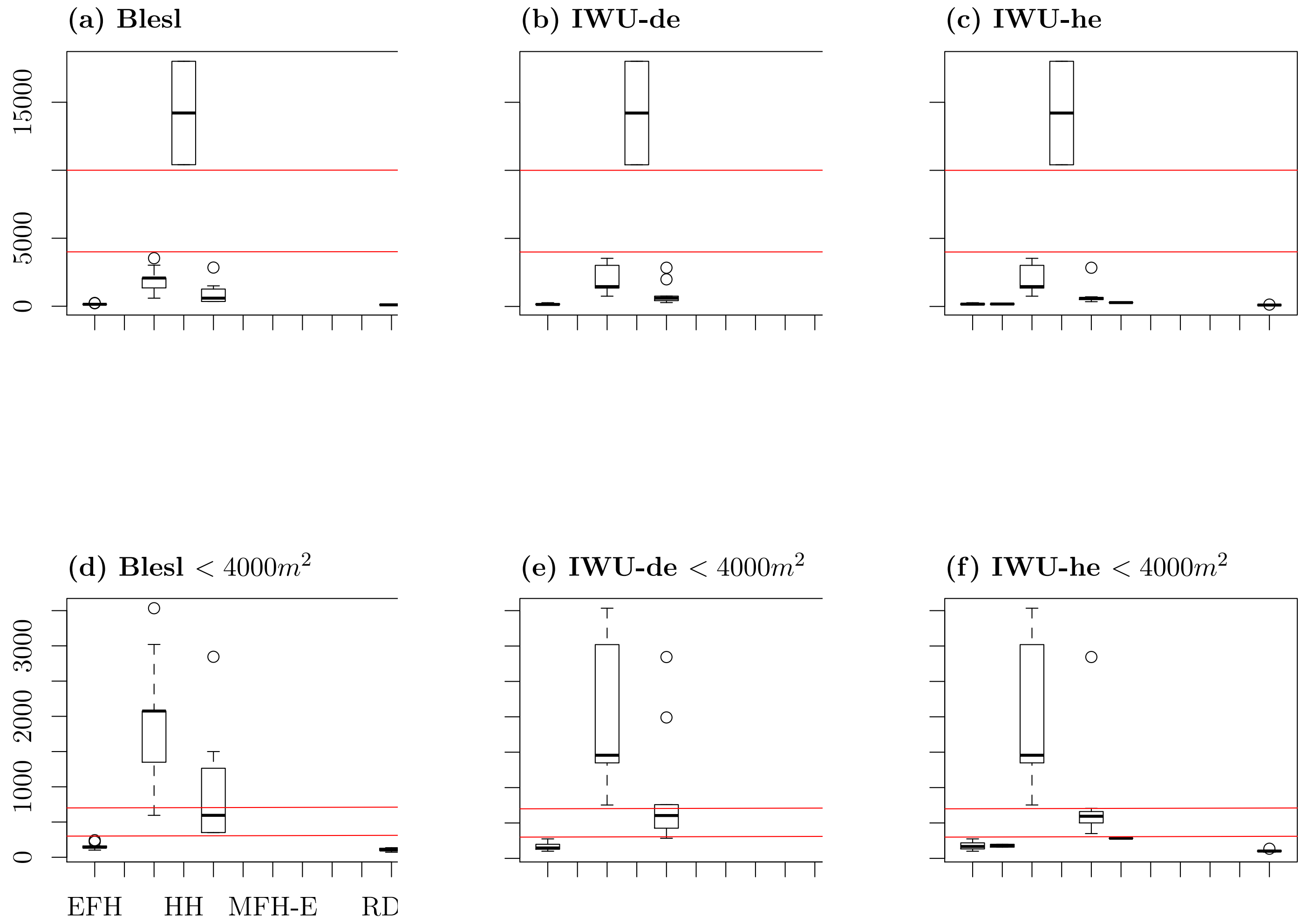
$$att_{m,n} = \begin{array}{c c} & \begin{array}{cccc} typ_1 & typ_2 & \cdots & typ_n \end{array} \\ \begin{array}{c} att_1 \\ att_2 \\ \vdots \\ att_m \end{array} & \begin{pmatrix} att_{1,1} & att_{1,2} & \cdots & att_{1,n} \\ att_{2,1} & att_{2,2} & \cdots & att_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ att_{m,1} & att_{m,2} & \cdots & att_{m,n} \end{pmatrix} \end{array} \tag{4.4}$$

$$min = \begin{pmatrix} min(typ_1) & min(typ_2) & \cdots & min(typ_n) \end{pmatrix} \tag{4.5}$$

$$p = \begin{pmatrix} min_1 & min_2 & \cdots & min_n \end{pmatrix} \div \sum min \tag{4.6}$$

Figure 4.5 shows the two developed arrays, the first line of the arrays represent the possible building types (A, B, C, etc...). Each array has 4 attributes (Att_i) that will filter the typologies depending on the building characteristics. The first example (a) representing the use of a binary array, in which attributes filter certain typologies deterministically. In this example the building X has attributed Att_1 and therefore the probability of building X of being of type A is 0. and the second array (b) representing the use of a float array, in which attributes filter the typologies in a probabilistic way. In this example the building X has attribute Att_1, the probability of building X of being of type A remains 0, but the probability of this building of being of a different type is also defined in this array.

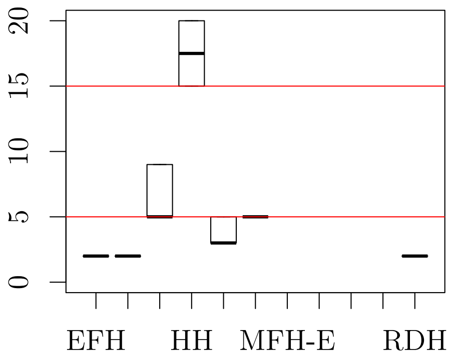
[5] float as in number definition in the python language, see `http://docs.python.org/2/library/stdtypes.html`



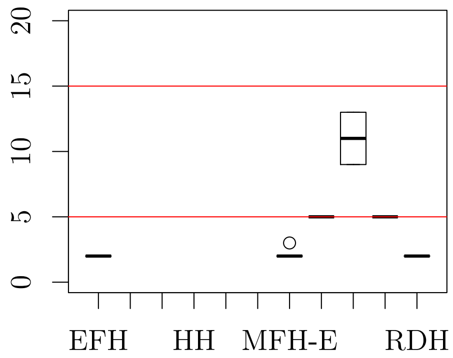
data source: (a-d) (Blesl et al., 2007) (b-e) (Loga et al., 2011) (c-f) (Born et al., 2003)

Figure 4.2: Different values for living space of building typologies used in Germany. The building types are arrange by construction type in alphabetical order along the X-axis. The Y-axis shows the living space of the single typologies in $[m^2]$. EFH; EFH-b; GMH; HH; KMH; KMH-b; MFH-E; MFH-G; MFH-H; MFH-W; RDH

(a) IWU-he



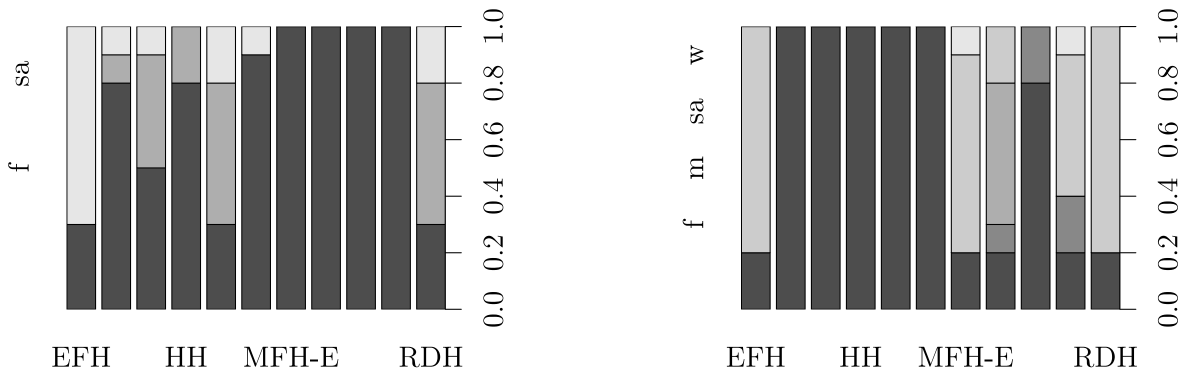
(b) EcoFYS



data source: (Born et al., 2003; Hermelink et al., 2011)

Figure 4.3: Different values for number of floors of building typologies used in Germany. The building types are arrange by construction type in alphabetical order along the X-axis. The Y-axis shows the number of stories of the individual typologies. EFH; EFH-b; GMH; HH; KMH; KMH-b; MFH-E; MFH-G; MFH-H; MFH-W; RDH

Roof types for construction types of two typologies

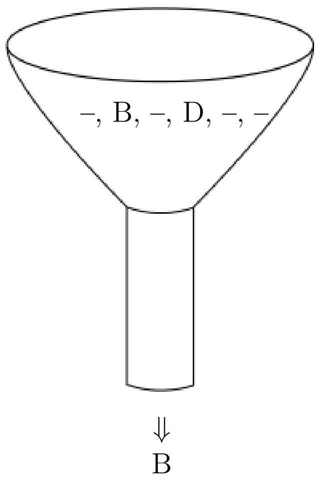


data source: (Born et al., 2003; Hermelink et al., 2011)

Figure 4.4: Different values for roof types of building typologies used in Germany. The building types are arrange construction type in alphabetical order along the X-axis. The Y-axis shows the roof types of the single typologies. EFH; EFH-b; GMH; HH; KMH; KMH-b; MFH-E; MFH-E; MFH-G; MFH-H; MFH-W; RDH. (sa) pitched roof, “Satteldach”; (m) curg roof, “Mansardendach”; (w) hip roof, “Walmdach”; (f) flat roof, “Flachdach”; (so) other, “Sonstiges”.

(a) Filter using a binary array

	A	B	C	D	E	F
	⇓	⇓	⇓	⇓	⇓	⇓
Att_1	0	1	1	1	1	1
Att_2	1	1	0	1	0	1
Att_3	1	1	1	1	1	1
Att_4	1	1	1	1	1	0
min	0	1	0	1	0	0
p	0	.5	0	.5	0	0



(b) Filter using a float array

	A	B	C	D	E	F
	⇓	⇓	⇓	⇓	⇓	⇓
Att_1	0	1	2	3	4	4
Att_2	4	4	0	4	0	4
Att_3	4	4	4	4	4	4
Att_4	4	4	3	2	1	0
min	0	1	0	2	0	0
p	0	.3	0	.6	0	0

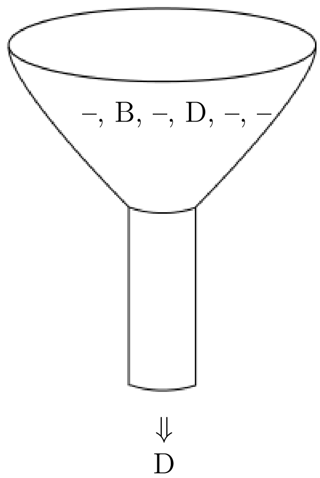


Figure 4.5: Example of the two developed arrays

4.3 Heat Demand and Heat Consumption

In this section we present: (1) the performance of the analyzed typologies; (2) a simple visualization of the result in space; (3) a discussion regarding the performance of buildings typologies for the analysis of heat demand in space; and (4) present some possible implementation of the presented method and future path in our analysis. We present a comparison of the typologies and analyze their performance in a small urban area in the city of Hamburg, Germany. We use this urban area because we have access to gas consumption data for this area, supplied as part of a monitoring program, making it possible to analyze the performance of the different typologies for the estimation of heat demand.

The monitoring data is suboptimal for this analysis but is the only data we have access to for an urban area for which we also have data regarding the characteristics of the building stock. The monitoring data consists of gas consumption data given as consumed volume of gas, this has been transformed to heat consumption data. The major problem of this data set for the analysis of the performance of building typologies for the estimation of heat demand is the aggregation format. The data set records monitoring points, each point represents a group of buildings, the building groups contain both residential and non-residential buildings. The building typologies analyzed in this section are calibrated to estimate the residential part of the building stock only. Because we know which building is in each group, we can at least know what percent of the heated area corresponds to the residential sector. This makes it possible to weight the observation points by heated residential floor area.

4.3.1 Performance of the Individual Typologies

In order to quantify the performance of the individual typologies we calculate the relative difference between consumption and demand for each single monitoring point, defined as Wd. The weighted mean of the single monitored points serves as a parameter to define the performance of the typologies and compare them, see Equation 4.7. The mean is weighted by the share of residential floor space per aggregation level (monitoring point), define by r_i. The share of residential floor space is simple calculated as the percentage of residential floor space form the total estimated floor space in that monitoring point. So for example, if a given monitoring point has only non-residential buildings $r_i = 0$ the difference between consumption and demand for that given point will not be considered for the weighted mean.

This method performs relatively well for this particular urban area, but might not be the best approach for other urban areas with a larger and more energy intensive non-residential building stock. The floor space of buildings is good indicator for heat demand of the residential sector, but might not be adequate as an indicator of gas consumption for the non-residential sector. A small business might be using large quantities of gas for more than just heating in this case the weighting by floor space would not be enough for the performance analysis of building typologies. On this specific area the non-residential buildings do not affect the overall performance of the individual typologies, Figure 4.8 shows the result for monitoring points which contain exclusively residential buildings and the resulting weighting mean. Although the figures are different the performance of the individual typologies follows the same order as with the sample containing non-residential buildings.

$$Wd = 1 \div n \sum_{i}^{n} C_i \div D_i \times R_i \tag{4.7}$$

Where C represent the monitored gas consumption for measure point i, D represents the average (1000 iterations) heat demand aggregated to point i and R represents the share of residential floor space in point i

We present four results from the comparison as: (1) the relative difference between demand and consumption weighted by share of residential floor area (see Table 4.5); (2) the difference of the sum of demand and consumption for the entire analyzed area (see Table 4.6); (3) the percentage of measuring points for which the developed algorithm over estimated heat demand (see Table 4.7); and (4) deviation of a weighted Z-statistic (see Table 4.8). In all four tables we present the results for both types of arrays **(B) binary** and **(F) float**, for three limit values l. These limit values define the maximum value taken into account for the analysis. These limit allows us to differentiate the performance of the typology for different consumption levels.

Table 4.5: Average relative difference between demand and consumption for three different limit point l

	$l = 1e+07$		$l = 1e+06$		$l = 1e+05$	
	B	F	B	F	B	F
Blesl	1.09	1.09	0.98	0.99	0.53	0.55
IWUde	1.12	1.12	1.01	1.02	0.54	0.55
IWUhe	0.95	0.95	0.85	0.86	0.46	0.47
BSU	0.94	0.94	0.81	0.81	0.42	0.42
EcoFYS	0.98	0.98	0.86	0.86	0.43	0.43

Table 4.6: Difference between monitored gas consumption and estimated heat demand in GWh

	$l = 1e+07$		$l = 1e+06$		$l = 1e+05$	
	B	F	B	F	B	F
Blesl	73.64	99.62	77.68	86.61	7.74	7.93
IWUde	56.57	83.83	71.85	81.71	7.28	7.35
IWUhe	60.81	108.23	83.35	98.04	8.18	8.15
BSU	172.30	172.88	125.76	126.48	8.74	8.73
EcoFYS	129.04	129.33	105.24	105.73	8.73	8.72

In addition to both tables we present the comparison between consumption and demand for all five typologies in form of tree graphs: (1) Figure 4.6 shows the comparison limited to values below $1e+07$; (2) Figure 4.7 shows the comparison limited to values below $1e+06$; and (3) Figure 4.8 shows the comparison between estimated heat demand and monitored gas consumption for all monitored point with exclusively residential buildings limited to values below $1e+06$. The first observation in these

Table 4.7: Percentage of measuring point for with the algorithm over estimates the heat demand

	$l = 1e + 07$		$l = 1e + 06$		$l = 1e + 05$	
	B	F	B	F	B	F
Blesl	.52	.53	.60	.58	.84	.84
IWUde	.47	.47	.54	.54	.84	.84
IWUhe	.59	.59	.68	.67	.88	.84
BSU	.69	.69	.76	.76	.92	.92
EcoFYS	.65	.65	.74	.74	.92	.92

Table 4.8: Weighted sum of Z-statistic deviation of the single typologies

	$l = 1e + 07$		$l = 1e + 06$		$l = 1e + 05$	
	B	F	B	F	B	F
Blesl	-76.46	-78.95	-70.08	-76.94	- 825.76	- 708.57
IWUde	-72.68	-73.73	-67.95	-73.07	-1983.75	- 1376.57
IWUhe	-82.53	-89.89	-76.14	-84.06	-3534.62	-12094.36
BSU	-79.41	-78.93	-72.56	-71.83	-1973.63	- 2022.98
EcoFYS	-75.54	-74.98	-71.30	-70.59	-5726.72	- 4507.83

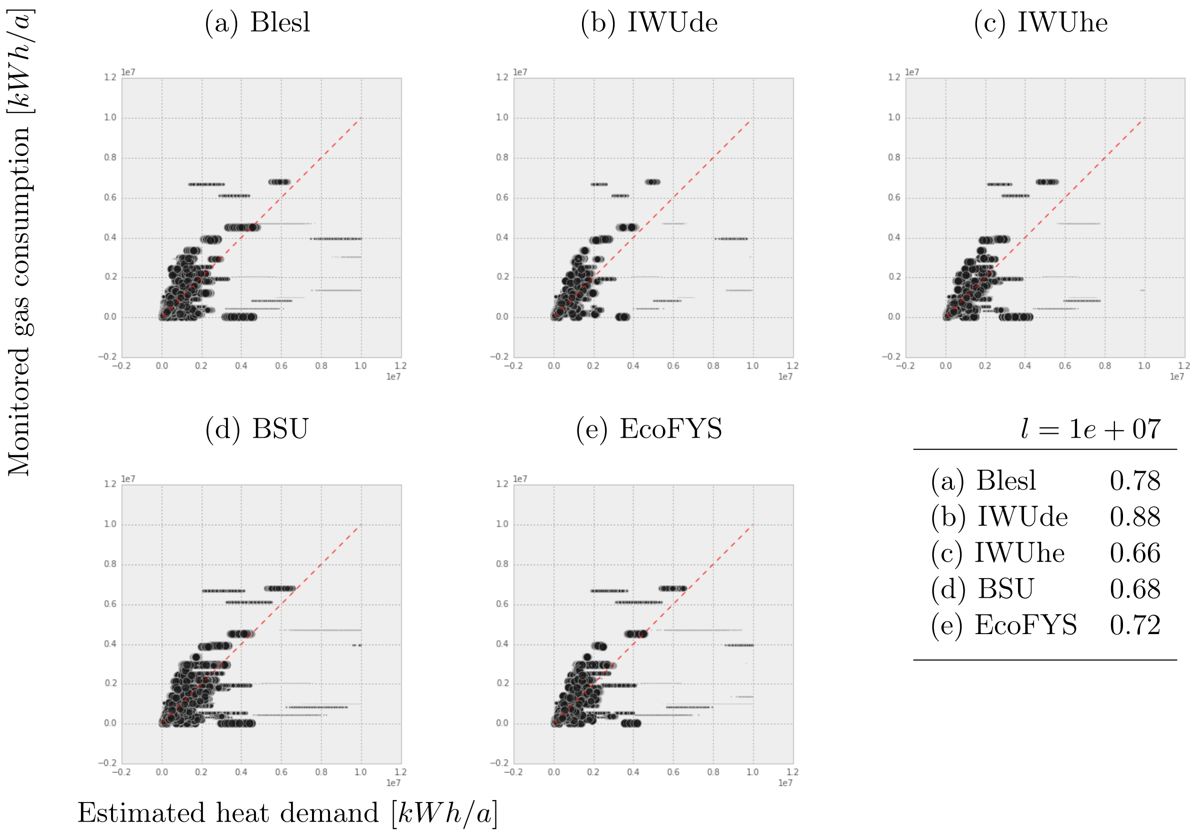
plots is a clear over estimation of the heat demand, This overestimation disappears when filtering monitoring points with non-residential buildings on them. This overestimation may relay not on the typologies itself but on the estimation of floor space of the individual buildings (see Equation 4.1) as the total heat demand is computed by multiplying the specific heat demand of the assigned type by the estimated living space. Another alternative explanation could be achieved by looking at the climate conditions at the specific monitored year. The monitored year may have been an especially warm year, making the heat consumption lower than the average, for which the typologies are designed. As the aim of this section is to compare the performance between typologies, the overall performance of the typologies plays a secondary role in this analysis.

$$Z = 1 \div n \sum_{i}^{n} \frac{d_i - c_i}{\sqrt{\left\| \frac{c_i - (1 - c_i)}{\sum D_i \times r_i} \right\|}} \tag{4.8}$$

$$c = \frac{C_i}{\sum_i C_i} \tag{4.9}$$

$$d = \frac{D_i \times r_i}{\sum_i (D_i \times r_i)} \tag{4.10}$$

Where C_i represents the monitored gas consumption for measured point i, D_i the estimated heat demand aggregated to measure point i and r_i the share of non-residential floor space in measure point i



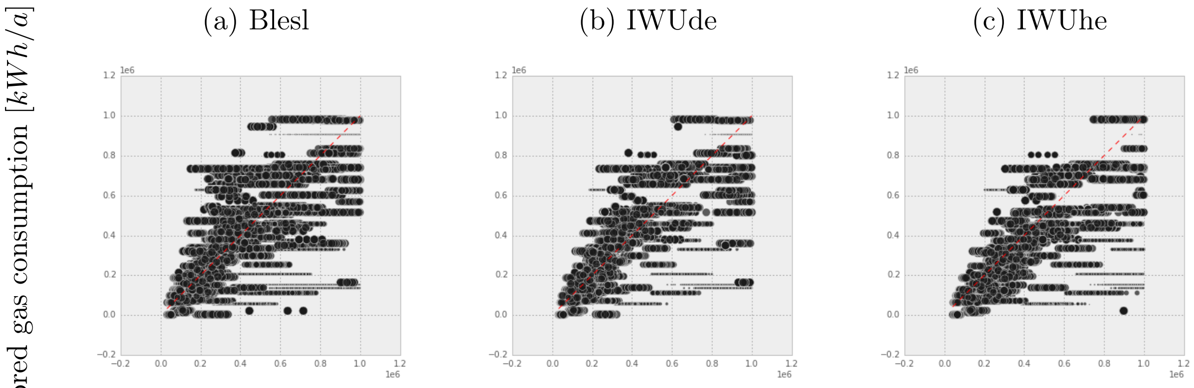
	$l = 1e + 07$
(a) Blesl	0.78
(b) IWUde	0.88
(c) IWUhe	0.66
(d) BSU	0.68
(e) EcoFYS	0.72

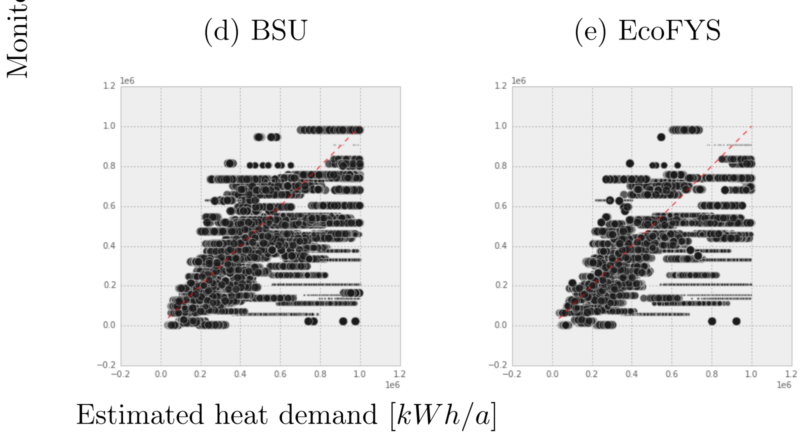
Figure 4.6: Shows the monitored consumption of the measuring points and the computed demand aggregated at the same level (gas consumption measuring point) using the float array, For values lower than $1e + 07$

The result from this analysis shows little variation between typologies. This is very interesting because the underlying data and methodology for the construction of the typologies differs between typologies. While the typology from (Blesl, 2002) does not use any consumption values for the construction of the building typology the IWUde typology (IWU, 2003) used data from the entire country. The performance of the EcoFYS typology does not perform significantly better than the other typologies but uses Hamburg specific data for the construction of the typology.

The result from our first analysis, comparing the performance of the typologies by the weighted mean difference between estimated heat demand and monitored gas consumption, show little variation between the typologies. The use of the float array has a small effect on the typologies that provide more background information, needed for the classification with the float array. In the case of the BSU typology, that provides little background information, the use of the float typology has almost no impact on the performance of the typology.

As expected both typologies developed for the city of Hamburg (BSU and EcoFYS) achieve the best performance. This good performance applies only for consumption levels lover than $1e + 05kWh/a$, for the higher consumption levels the best performing typology is the IWU-he typology.





	$l = 1e + 06$
(a) Blesl	0.86
(b) IWUde	0.87
(c) IWUhe	0.74
(d) BSU	0.74
(e) EcoFYS	0.77

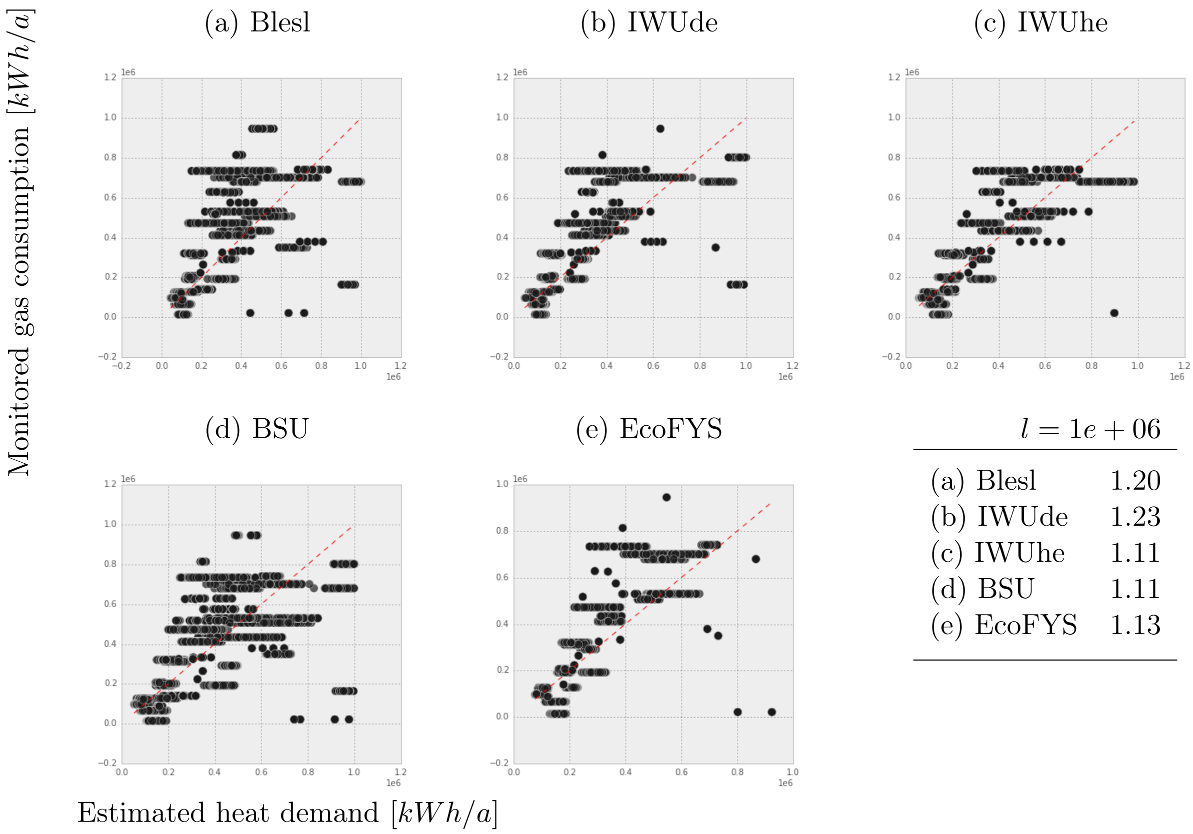
Figure 4.7: Shows the monitored consumption of the measuring points and the computed demand aggregated at the same level (gas consumption measuring point) using the float array, For values lower than $1e + 06$

For the absolute difference between consumption and demand we observe the same pattern, for low consumption levels the typologies constructed for Hamburg outperform the ones constructed for Germany. At a higher consumption level the IWU-he typology performs best.

The percentage of estimation points that were overestimated follows a different pattern. Typology IWU-de has the lowest share of overestimated points among all typologies. The share of overestimated points does not vary much with a change in consumption level for this typology. The share of overestimated points can be explained by the non-residential sector. The share of overestimated points reduced drastically by analyzing only monitoring points with only residential buildings. For monitoring points with only residential sector we see the opposite pattern, the building typologies underestimate heat demand.

The Z-statistic gives us a sense of the typology performance at an individual level rather that the overall performance of the typology. The performance at an individual level show that the typologies constructed for Germany outperformed the typologies design for Hamburg.

We identify five points of action in order to increase the performance of building typologies for the estimation of heat demand at low aggregation level:



$l = 1e + 06$	
(a) Blesl	1.20
(b) IWUde	1.23
(c) IWUhe	1.11
(d) BSU	1.11
(e) EcoFYS	1.13

Figure 4.8: Shows the monitored consumption of the measuring points and the computed demand aggregated at the same level (gas consumption measuring point) using the float array, For values lower than $1e + 06$ and measuring point with only residential buildings.

1. The number of types is not enough to distinguish relevant building characteristics at a low level of aggregation, we need a more precise classification of the building stock.

2. The typologies do not consider user influence on the consumption of heat demand, we argue for the integration of demographic characteristics of the occupants for a specific modeling of user behavior in the computation of heat demand. See (Muñoz H. & Peters, 2014b) for the allocation of families to the building stock and (Muñoz H., 2014) for the integration of user behavior in urban heat demand models.

3. The used data for the construction of the typologies is rarely monitored heat consumption but estimated values. We pledge the community to for a transparent and open share of monitored data for the construction of better building typologies.

4. There is much room for improvement for the classification of the building stock. We see this endeavor as part of our ongoing research. Available analysis tools can be used for a better classification of the building stock, an improvement of the classification can improve the performance of these building typologies.

5. The building typologies are design for the estimation of yearly heat demand. In order to estimate heat demand at a lower temporal resolution the user has to perform their own heat balance for each individual building. In this case the use of the building types is not the primary data set

but the underlying values of the building components (U-values, share of glassing, etc.). We need a typology design for this specific cases.

For the proper classification of the building stock plays a significant role in the accurate estimation of heat demand in small urban areas. The use of a digital cadastre for this classification presents itself as a great data source, to achieve a good classification, unfortunately this data source alone is not enough for a proper classification of the building stock.

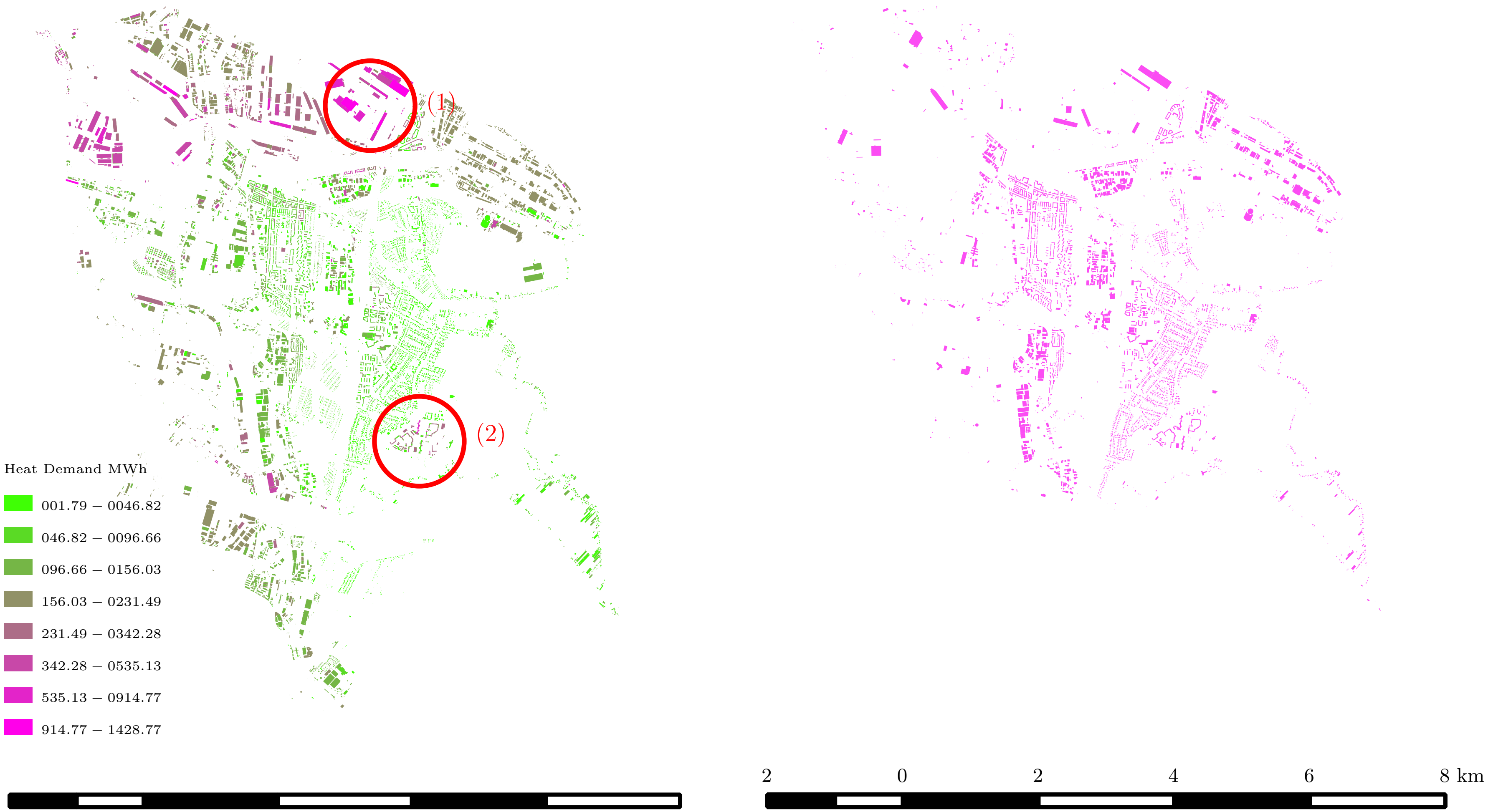
We want to develop methods to enrich the digital cadastre previous to the classification of the building stock. We identify two key parameters for the classification of the buildings stock and for the estimation of heat demand: (1) construction year; and (2) floor space. The proper estimation of construction year would reduce the uncertainty of the classification procedure. The second parameter, floor space, is extremely important for the estimation of heat demand because building typologies provide a specific value for heat demand in $[kWh/m^2a]$ we multiply this value by the "heated space" to estimate absolute heat demand in $[kWh/a]$.

The resulting temporal resolution achieve with this method is not enough to perform an appropriate heat planning distribution strategy. None of these typologies are constructed for the estimation of heat demand a lower resolution level. There is a need for another type of typologies addressing this issue.

4.3.2 Visualizing the Result in Space

The visualization of the computed result in space may help us understand the problems of the developed algorithm. For this we present two maps (see Figure 4.9): (1) showing the estimated absolute heat demand for the single buildings using the EcoFYS typology; and (2) showing all the buildings connected to the gas grid for which we have the monitored consumption values at an aggregated level. See Section 4.2.1 for details on data sources.

We have identified two important observations in these maps: (1) an over estimated heat demand for non-residential buildings, agglomerated in the north part of the map; and (2) a possible identification of "heat spots" in urban areas. These two observations are particularly interesting. The first may explain an overestimation of the heat demand for some monitored points and the second one presents an interesting application of this method. We briefly discuss the first observation in the next subsection (Sub-Section 8.1) making the argument for the need to expand this analysis to the non-residential sector. In the subsequent section (Sub-Section 8.2) we discuss the further implementation of the analysis and discuss further paths to expand this method.



(a) Map showing the estimated heat demand in MWh using typology EcoFYS. The red circles mark: (1) an over estimation of heat demand for non-residential buildings; and (2) identification of a (residential) heat spot in the urban area. **(b)** Map showing the buildings connected to the gas grid (pink)

Figure 4.9: Maps showing the estimated heat demand and the monitored gas consumption

5 Integrating User Behavior Into Thermal Simulation Models[1]

5.1 Using Family Specific Occupancy Rates to Simulate Heat Consumption

A significant variation between estimated heat demand and monitored heat consumption among residential buildings with similar physical characteristics has lead the scientific community to the conclusion that this variation has to be explained trough occupant behavior (Guerra Santin et al., 2009; Haldi & Robinson, 2011; Durand-Daubin et al., 2013; D'Oca, Fabi, Corgnati & Andersen, 2014). The aim of this section is to present a strategy to use time-use data to enrich the population census, different families taken from the census are allocated to a generic building with a simple geometry for the estimation of heat demand. The estimation of heat demand is performed with the well-established building simulation software EnergyPlus[2]. The generated occupancy rates are parsed as input files to the building simulation model in form of a CSV file. This file contains 52560 (144×365) record points, 144 points for each day, because the data from the time-use dataset describes events in a 10 minute interval. For the integration of building occupants at an urban scale we propose to allocate synthetic families, with their corresponding time budget, into the digital cadastre. The biggest challenge in these endeavors is to develop robust models able to create plausible scenarios for this merge. First steps in this direction have been taken (Muñoz H. & Peters, 2014b). This section marks an important step towards this goal.

5.2 The User Influence on Heat Consumption

The influence of the user on heat demand is a topic of much debate among the community. There is little empirical data available at: (1) the required resolution, needed to identify difference between behavioral patterns during the day; and (2) required amount of data, needed to control for all other variables influencing heat demand (Zhun Yu et al., 2011). An important factor for the simulation of energy demand seems to be occupant presence (Guerra Santin et al., 2009). The later authors observe this effect in their work, (Page, Robinson & Scartezzini, 2007) developed an interesting method for the integration of this parameter into simulation models generating stochastic occupation patters using Markov chains. The integration of user behavior into building simulation models can be achieved, difficulties arise by the validation of such simulations as empirical data at the needed temporal and disaggregation resolution is difficult to recover. The integration of behavior models into well-established

[1]This chapter is heavily based on: Muñoz Hidalgo (2014)
[2]`http://apps1.eere.energy.gov/buildings/energyplus/`

energy demand models have presented into two forms: (a) probabilistic and (b) deterministic approaches, the latter being the “native” method for common simulation models (D’Oca et al., 2014). A good example of this distinction can be found in the simulation of window opening (Borgeson & Brager, 2008), the integration of user behavior models (e.g.: natural ventilation rate through opening windows) into energy demand models has posed a challenge for models, because resulting models from empirical observation deliver probabilistic models which are hard to integrate into the “classic” deterministic architecture of conventional building simulation models. (Bourgeois, 2005) develop the model SHOCC (Sub-Hourly Occupancy Control) which can be directly used with the common thermal simulation model ESP-r.

5.3 Allocating Occupancy Patterns to Individuals

For each individual described in the micro census a representative record from the time-use dataset is selected. In table 5.1 the available attributes present in both datasets: (a) the micro-census; and (b) the time-use data set are listed by relevance. The order of the attributes is important for the selection of a matching record in the time-time use dataset because of two reasons: (1) the number of records in the time-use data is smaller than of the micro census and; (2) the attributes classes are not equal for both data sets. It is possible that the developed algorithm can’t find a matching using all the attributes as query constraints, in such a case the algorithm drops the last constraint and performs the query with one constrain less, this procedure is performed until at least one matching record is recovered from the database. In the case than more than one journal is recovered an average is constructed upon all retrieved records. The recovered data contains information regarding the location of the individual in a 10 minute step for workdays and weekends (day types). This information is simplified to two one dimensional vectors for each day type. Each vector of size 144 represents the probability of being at home.

Table 5.1: Attributes used to allocate occupancy schedules to individuals from the census

#	Time-use ID	Microcenssus ID	Attribute name
1.	P27X	EF131	Working hours
2.	P2610	EF171	Work-at-home
3.	VO_TE_N	EF129	Full-time or part-time job
4.	P281	EF149	Work on Saturday
5.	P282	EF150	Work on Sunday
6.	PH01B2X	EF44	Age
7.	PH01E	EF49	Family status
8.	BERUF_N	EF117	Occupational status
9.	P11	EF287	Attending school
10.	PH01C	EF46	Gender
11.	P29X	EF436	Net-income

5.4 Generating a Schedule File for the Simulation

Based on the matched profile from the time-use data a file for the entire simulation year is generated. Two files are given as input for this computation describing the probability of being home at: (1) weekdays and (2) weekends. Based on these files the occupation patterns for the entire year are generated. Figure 5.1 shows this probability for the average individual and for two sample households, see below for details on the definition of households. The file containing the presence of the single individual for the entire year in a 10 minute interval is generated based on the retrieved information from the time-use data set. This data-set delivers the probability of the specific individual of being home at a particular hour of the day. This probability is used as base to generate the schedule of the individual. In order to avoid fluctuations during time period without a clear probability (50%) a transition probability is added to the equation, labeled $trans$ in equation 5.1 and equation 5.2. Because out of the micro census we can group the individuals into families, the generated schedule used as input for the thermal simulation model is the average occupation of the building for the number of persons living in the house.

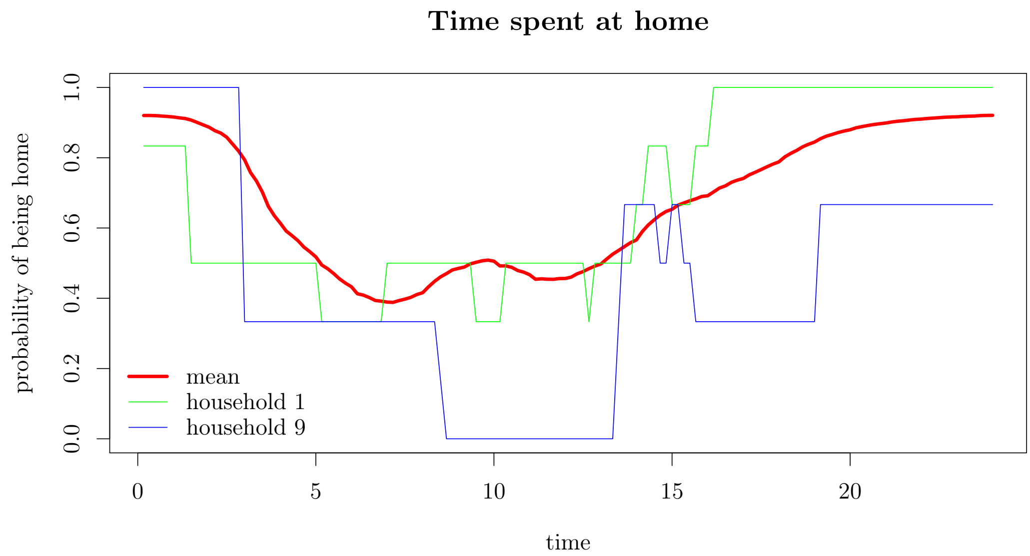


Figure 5.1: Average probability of being at home and two probability curves for two individual households

$$B_t(i,dt) = \begin{cases} 0 & \text{if } rnd > P_t(i,dt) + trans(t-1) \\ 1 & \text{if } rnd < P_t(i,dt) + trans(t-1) \end{cases} \tag{5.1}$$

$$trans(t) = \begin{cases} +0.4 & \text{if } B_t = 1 \\ -0.4 & \text{if } B_t = 0 \end{cases} \tag{5.2}$$

Where:

B_t	Determinant of being home at time step t
$trans$	Transition probability
P_t	Probability of being home at time step t
i	Unique individual from the micro census
dt	day type (Workday/Weekend)

5.5 Building Geometry Used in the Simulation

For the purpose of this section we defined a simple building geometry, A box of 6 $\times$ 8 meters with a ceiling height of 2.7 meters. A double glazed window was located on each side of the building. We aim to expand this approach by defining the building characteristics, including geometry and heat transmission coefficient of components through the use of the available digital cadastre for the city of Hamburg. Using the digital cadastre to acquire geometrical information of the individual buildings is not the only advantage of the digital cadastre, because the buildings are geo-referenced an allocation of representative synthetic families to the building stock is possible (Muñoz H. & Peters, 2014b). A detail description of the building stock geometry and schedules of its occupants is the development goal we am working on. The presented contribution represents an important step to achieve this goal.

5.6 Estimating Heat Demand with Family Specific Occupational Schedules

For the simulation of heat demand the energy simulation software EnergyPlus, was used. The `*.idf` input file used for the simulation is dynamically generated with the new data. For now, only two variables of the input file are modified for the different simulations: (1) the generated schedule for the specific family, third variable in listing 5.6; and (2) the number of occupants, or in this case household size, listed as the fifth variable on listing 5.6. In order to show the function of this method we have only used the generated schedule to control for internal gains, this method can be expanded to control other relevant parts of the simulation e.g.: ventilation rates, appliance used, lighting or temperature set points. For the defined building we run 100 simulations with the first 100 families in the micro census. This type of simulation could also been achieved with the use a stochastic model. The aim of the presented method is not to simulate possible spreads of heat demands for a single building but aims to simulate building agglomerations while taking into account the specific socio demographic characteristics of its residents. The efforts presented in this section make a significant contribution toward the development of urban heat demand models which take the user behavior into account? The resulting heat demand estimation from the simulation is depicted in figure 5.2. There is a significant variation in the consumption of heat for the different family types. We expect to see a similar variation at an urban scale. Such a heat demand model may prove interesting for the proper dimensioning of decentralized heat supply systems.

```
Schedule:File,
    OCCUP,                      !- Name
    Fraction,                   !- Schedule Type Limits Name
    shedules\shedule.csv,       !- File Name
    1,                          !- Column Number
    0,                          !- Rows to Skip at Top
    8760,                       !- Number of Hours of Data
    Comma,                      !- Column Separator
    no,                         !- Interpolation
    10;                         !- Minutes per item
```

Definition in the .idf input file for the schedule controlling the presence of occupants during the simulation

```
People,
    SPACE1-1 People 1,       !- Name
    ZONE ONE,                !- Zone or ZoneList Name
    OCCUP,                   !- Number of People Schedule Name
    people,                  !- Number of People Calculation Method
    3,                       !- Number of People
    ,                        !- People per Zone Floor Area
    ,                        !- Zone Floor Area per Person
    0.3,                     !- Fraction Radiant
    0.55,                    !- Sensible Heat Fraction
    ACTIVITY_SCH;            !- Activity Level Schedule Name
```

Base code used in the input .idf file for the simulation of internal heat gains

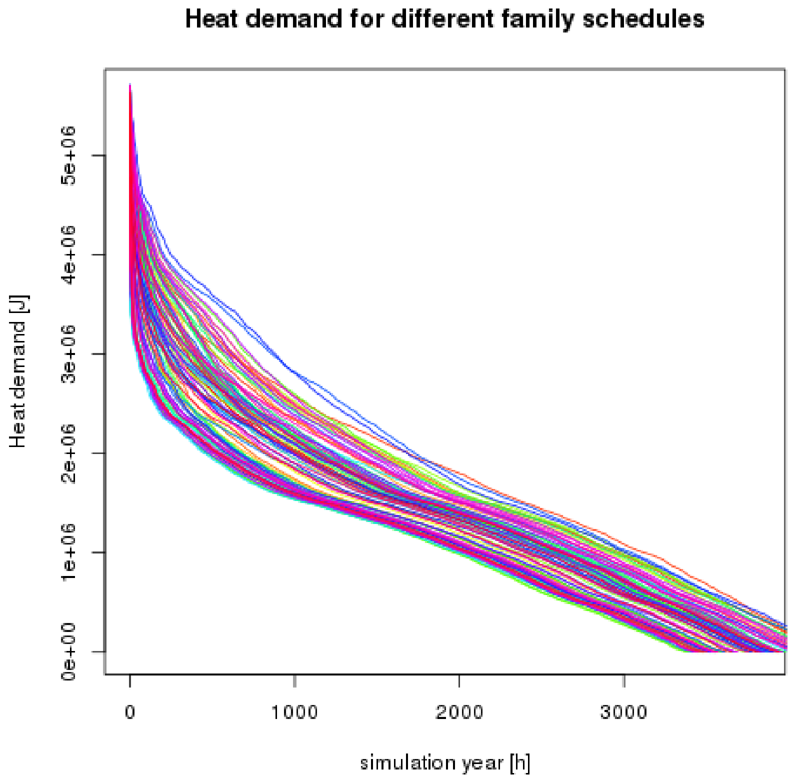


Figure 5.2: Simulated heat demand for 100 different occupation schedules of 100 different families occupying the generic building

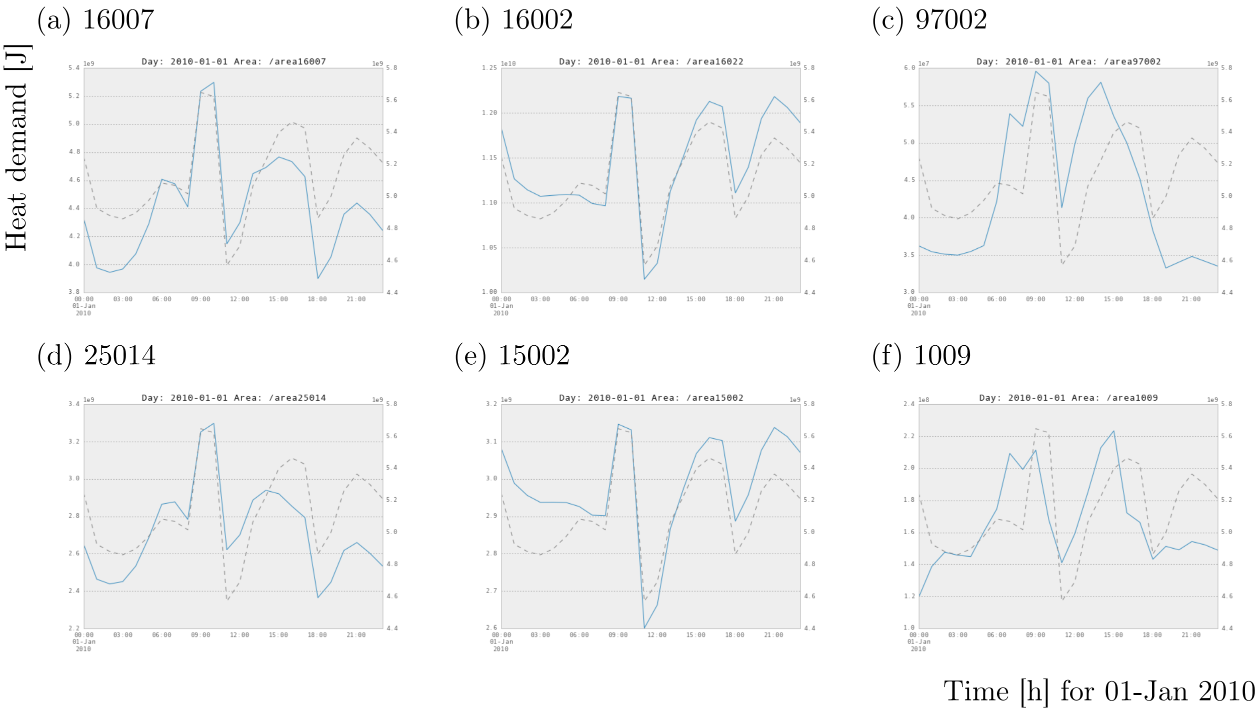


Figure 5.3: Heat demand profiles for selected small areas

Table 5.2: Socio-demographic data from selected small areas

STATGEB	A>65	Migration	Density	H-Size	1P-H	Kids	#
16007	0.0633	36	0.0066	1.7	64.10	18.60	3332
25014	0.0871	13	0.0061	1.4	74.30	8.60	1803
16002	0.0000	100	45.9286	1.0	100.00	0.00	14
15002	0.1059	31	2.3529	1.4	69.40	5.60	170
97002	0.0211	42	2.1789	1.3	82.40	3.90	95
1009	0.1189	17	0.0492	1.2	80.30	2.30	244

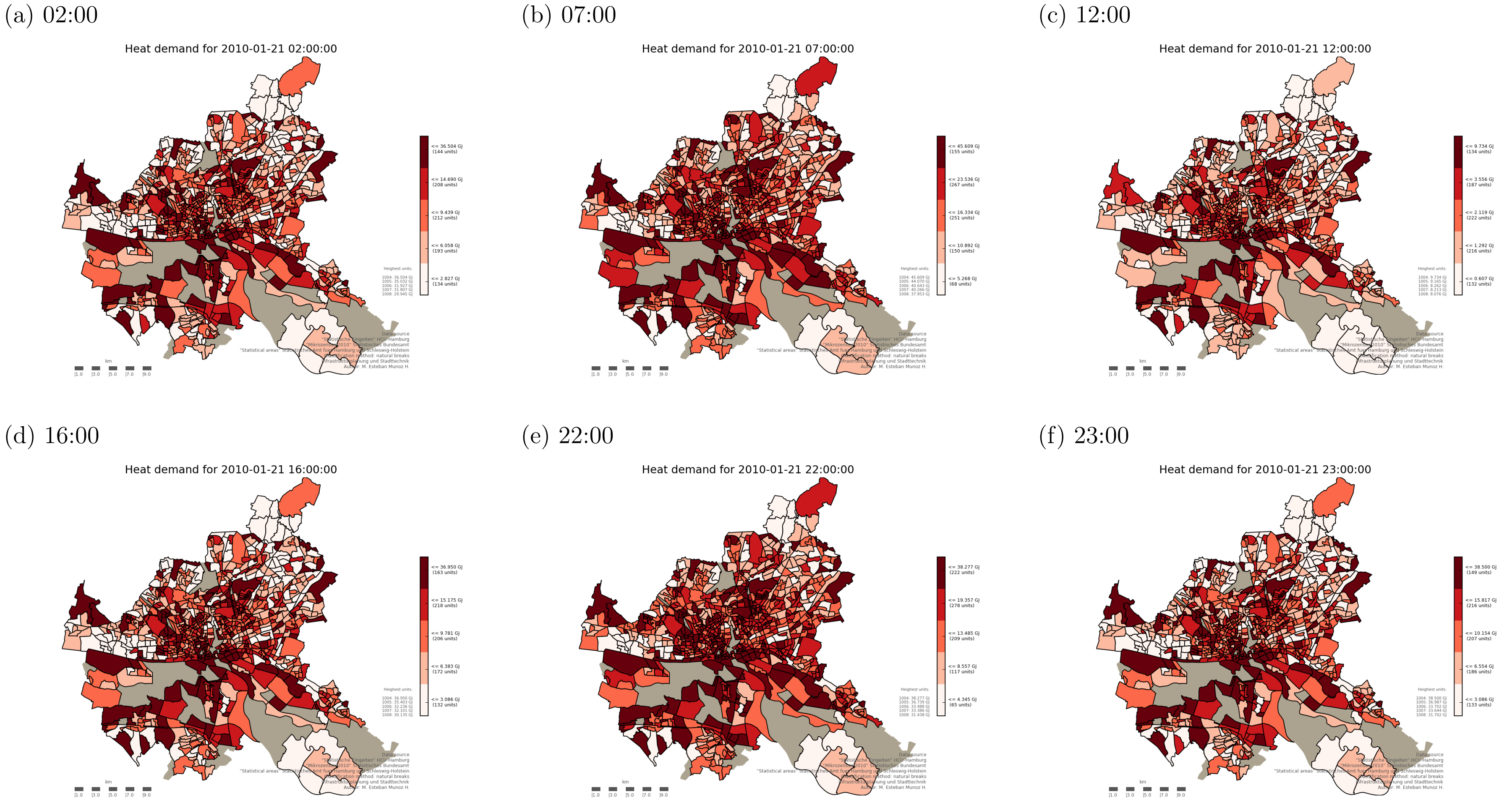
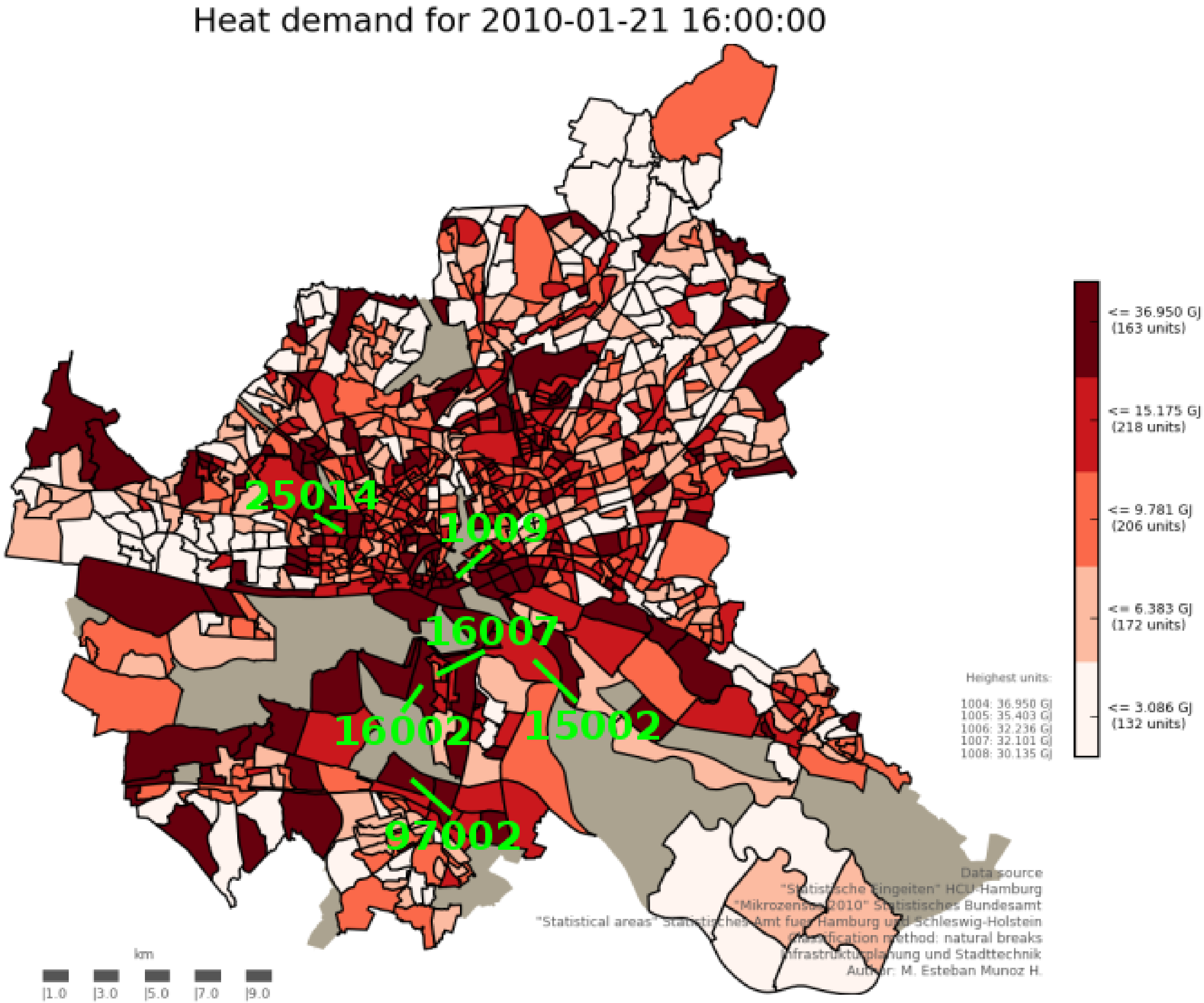


Figure 5.4: Heat demand maps for different hours of the 21 of January 2010


Heat demand for 2010-01-21 16:00:00
25014
1009
16007
16002
15002
97002
<= 36.950 GJ (163 units)
<= 15.175 GJ (218 units)
<= 9.781 GJ (206 units)
<= 6.383 GJ (172 units)
<= 3.086 GJ (132 units)
Heighest units:
1004: 36.950 GJ
1005: 35.403 GJ
1006: 32.236 GJ
1007: 32.101 GJ
1008: 30.135 GJ
Data source
"Statistische Eingeiten" HCU-Hamburg
"Mikrozensus 2010" Statistisches Bundesamt
"Statistical areas" Statistisches Amt fuer Hamburg und Schleswig-Holstein
Classification method: natural breaks
Infrastrukturplanung und Stadttechnik
Author: M. Esteban Munoz H.
km
|1.0
|3.0
|5.0
|7.0
|9.0

6 Using Spatial Microsimulation to Populate the Building Stock[1]

6.1 Generating Synthetic Populations for Microsimulation Models

Microsimulation, introduced by Orcutt in 1957 is a widely used method in social sciences, the generation of a synthetic population is the first step of most microsimulation models. The methods used for the generation of synthetic population constitute the scope of this section. Synthetic populations are used as input for simulation models of a diverse set of disciplines, among them: health (Brown & Harding, 2002); transport (Farooq et al., 2013); water (Williamson, Clarke & McDonald, 1996; Williamson, Mitchell & McDonald, 2002); energy (Chingcuanco & Miller, 2012; Muñoz H. & Peters, 2014b). A synthetic population is also used on many urban simulation models, e.g.: ILUTE[2], MATSim[3], UrbanSim[4]. O'Donoghue et al. (2014) provides an overview of applications of microsimulation models, the used methods to generate synthetic populations and a description of validation techniques.

Deterministic reweighting algorithms are not the only method for the generation of synthetic populations. An alternative to these algorithms is a combinatorial optimization (CO) algorithm (Williamson et al., 1998; Voas & Williamson, 2000), the advantage of a combinatorial optimization algorithm is that this algorithms produce integer weights. Integer weights are needed for the representation of individuals within a simulation framework, this is a requirement for many microsimulation models, especially agent-based models. The performance of deterministic reweighting algorithms declines with a post-processing of weights for the generation of integer weights (Pritchard & Miller, 2012). Newer approaches to the generation of synthetic populations exist, Farooq et al. (2013) propose the use of a Markov Chain Monte Carlo (MCMC) simulation based approach for the generation of synthetic populations, with the presented method the authors were able to achieve a better performance than by using an IPF algorithm. Depending on the simulation model and the model scope, a synthetic population with integer weight may not be required.

There are many reviews of methods for the generation of synthetic populations. Harland et al. (2012) make a comparison between three methods: (1) a deterministic reweighting algorithm (IPF); (2) a stochastic synthetic reconstruction method; and (3) a stochastic Combinatorial Optimization (CO) method (simulating annealing). The authors compare the simulation results by means of the Total Absolute Error (TAE). The results from the CO algorithm achieves the best results out of these three

[1]Sections of this chapter are based on: Muñoz H. & Peters (2014) Muñoz H. Tanton & Vidyattama (2015) and Muñoz H, Vidyattama & Tanton (2015b)
[2]Integrated Land Use, Transportation, Environment (ILUTE) Modeling System `http://www.civ.utoronto.ca/sect/traeng/ilute/ilute_microsimulation.htm`
[3]Multi-Agent Transport Simulation `http://www.matsim.org/`
[4]Simulation system for supporting planning and analysis of urban development `http://www.urbansim.org`

algorithms. Hermes und Poulsen (2012) deliver a review of spatial microsimulation method, the authors also speak in favor of a CO algorithm, the analysis does not perform any test of the method but based this observation on literature review only. Müller, Axhausen, Axhausen und Axhausen (2010) discuss in their paper different implementations of the IPF algorithm. Rahman et al. (2013) presents a comparison on microsimulation method and a detailed description of the GREGWT algorithm, the paper also discusses the use of a "new" microsimulation method implementing Bayesian statistics for the generation of synthetic populations. Tanton (2014) presents a comprehensive review of microsimulation methods and its applications. Williamson (2013) presents a comparison of two methods to generate a synthetic population: (1) a Synthetic Reconstruction (SR), and (2) a CO method. The author concludes that the CO method outperforms the SR method. Tanton et al. (2014) compare the performance of a CO and a GREGWT algorithm, the performed analysis shows that the CO outperforms the GREGWT method.

6.2 Deterministic Reweighting Algorithms

The following section aims to: (1) compare different methods commonly used in the microsimulation community to generate geographical allocated synthetic populations; (2) asses the performance of our new R package implementing the GREGWT method; and (3) discuss the role of the initial weights of the individual record survey.

In order to assess the performance of the new R library implementing the GREGWT algorithm we perform the same reweighting for four small areas using both implementations of the GREGWT algorithm and a reweighting using an R implementation of the IPF method. We compare the results using:

1. An implementation of the **GREGWT** algorithm in the **SAS** language (Bell, 2000);
2. An implementation of the **GREGWT** algorithm in the **R** language (Muñoz H., Vidyattama & Tanton, 2015a).
3. An implementation of the **IPF** algorithm in the **R** language (Blocker, 2013).

For each simulation we use the same input data (see Section 3.3). All the data used in this analysis is freely available and can be directly downloaded from the corresponding German statistical offices. The R implementation of the GREGWT algorithm allows us to specify the reference category with the `prepareData` function. We make use of these feature in order to select the category resulting in the minimal error within the reweighting process. This process is described in Section 6.3.

The comparison of the different algorithms and libraries is discussed on Section 6.4. First we compare the resulting new weights from each simulation and each small area. In this section we also compare the internal errors of the different simulations by means of the Total Absolute Error (TAE) and Percentage Specific Absolute Error ($PSAE$).

Under Section 6.5 we compare the weight distance for each simulation area and each algorithm implementation. The weight distance is defined as the *Chi-Squared-Distance*. In Section 6.6 we present a summary of the simulation results and discuss the role of initial weights in the reweighting process. The section concludes with a discussion of further developments and implementations of the reweighting algorithms (see Section 8.5).

6.2.1 GREGWT

The GREGWT method is part of the model SpartialMSM developed by the National Center for Social and Economic Modeling (NATSEM) (Tanton, 2007). The GREGWT algorithm used by NATSEM is a SAS macros developed by the Australian Bureau of Statistics (ABS) (Bell, 2000). The method for reweighting the survey sample is based on method number 5 of A. Singh und Mohl (1996). Tanton et al. (2011) make a detailed description of the algorithm. The mathematical description of the GREGWT algorithm presented below is taken from Rahman et al. (2010).

Aim of the GREGWT algorithm is to find a set of new weights w that can be used to match a survey X to a set of given benchmarks T so that $T = \sum w_j X_j$ (e.g. small area aggregates) by minimizing the difference between new weights w and the sample design weights d from the survey. For the distance D between design and estimated weights the GREGWT algorithm makes used of the truncated Chi-Squared-Distance function, represented in Equation 7.4.

$$D = \frac{1}{2} \sum_j \frac{(w_j - d_j)^2}{d_j} \tag{6.1}$$

The equation needed to minimize the weight distance constraint to some given marginal totals of a geographical area T can be expressed as the Lagrangian function of the chi-squared function, as follows:

$$L = \frac{1}{2} \sum_j \frac{(w_j - d_j)^2}{d_j} + \sum_k \lambda_k \left(T_k - \sum_j w_{j,k} X_{j,k} \right) \tag{6.2}$$

By differentiating (7.4) with respect to w_j and applying the first order condition, we have:

$$\frac{\delta L}{\delta w_j} = \left(\frac{w_j - d_j}{d_j} \right) - \sum_j \lambda_j X_j = 0 \tag{6.3}$$

With this equation we can formulate an equation for the new weights. Where $X_j^{'} = \sum \lambda_k X_{j,k}$.

$$w_j = d_j + d_j X_j^{'} \tag{6.4}$$

The new weights computed by the GREGWT algorithm are float values. Without any restrictions the algorithm will produce negative weights, both implementations of the algorithm introduce boundary constraints as a user input. The user is able to define an upper and lower bound, if the algorithm computes weights outside these bounds the weights will be truncated to the corresponding bounds. In this case the algorithm will iterate with the new computed weights until a predefined convergence parameter is met or there is no improvement in the iteration.

In this section we make use of two implementations of the algorithm. The first implementation is a SAS implementation of the algorithm developed and maintained by the Australian Bureau of Statistics (Bell, 2000). The second implementation of the algorithm — developed to replicate the results from the SAS algorithm — is an R implementation of the procedure. The development of this library has been performed by the same author of this thesis (Muñoz H., Vidyattama & Tanton, 2015a). A development version of the library is available on a github repository under: `github.com/emunozh/GREGWT`. This library is still under development, this section aims to present the first results of our development efforts.

6.2.2 IPF

The IPF has a long tradition within the microsimulation community (Birkin & Clarke, 1989, 1988). This method is used not only among the spatial-microsimulation community but is also implemented as the first step to generate synthetic populations of many agent-based models.

The IPF algorithm computes the new weight in an iterative manner. The algorithm estimates a new weight for each small area benchmark in sequence. The new weights are estimated as the share between survey marginal totals T and the observed geographical areas marginal totals $T(Obs)$ (see Equation 6.5). The initial design weights of the survey d are also taken into account, although many implementations of the IPF method will just set all initial weights to 1.

$$w_t = d_t \times \frac{T_t(Obs)}{T_t} \tag{6.5}$$

The new weights are computed for each constraint sequentially, for each constraint the new weights are computed as the ratio between the observed marginal totals (benchmarks) and the marginal totals on the original survey. The order of the input benchmarks has an impact on the simulation results (Lovelace, Birkin, Ballas & van Leeuwen, 2015).

The R implementation of the IPF method is a package called ipfp(Blocker, 2013), the package is a wrapper of a C implementation of the IPF algorithm. We use all the defaults of the package to run the simulation. These are: (1) a maximal iteration number of $1e3$ and (2) a tolerance level define as `.Machine$double.eps`, in our case this value is equal to $2.22e-16$. Because the main computation of the IPF is performed by the C code, the computation is extremely fast. Both implementations of the GREGWT algorithm are implemented natively on the SAS and R languages correspondingly. The R implementation of the GREGWT is almost 100 times faster than the IPF library.

6.3 Selection of Reference Categories for the Simulation

As described before the R implementation of the GREGWT algorithm uses a generalized regression for the computation of new weights, for this reason the input matrix cannot contain correlated variables and therefore we need to define a reference category. The R implementation of the GREGWT algorithm will use the function `model.matrix` to automatically define the reference categories. The use of this default option is suboptimal because this function will reduce the number of observations in case of missing values, this is not always desire for the reweighting of a survey. We aim to expand the R library in future releases in order to provide a more robust default function for the selection of reference categories. There are many factors to consider for the selection of the reference category. Because the algorithm does not benchmark the weights to the reference category this category will allocate the biggest error (see Section 6.4). The reference category should be selected depending on the research question, for this analysis we have selected the "optimal" (In terms of lowest $PSAE$ value) category. In order to identify this optimal category we reweight to every possible combination of reference categories, the result of this iterative simulation if plotted in Figure 6.5. The selected reference categories are specific to the four simulation areas. The figure shows the sorted $PSAE$ values for each combination of reference category, the values are read independently, for each area the "optimal" reference categories are different. The SAS implementation of the GREGWT algorithm will drop the last category in case of a correlation in the input matrix.

The implemented process to find the "optimal" combination of reference categories was possible in this case because the presented use case contains only 3 benchmarks with a total of 23 categories and 4 simulation areas. This computational expensive process is not viable for a real world simulation. We do not prioritize a development of a method able to identify the "optimal" reference categories, because this decision should be based on the specific research question rather than on the internal performance of the algorithm.

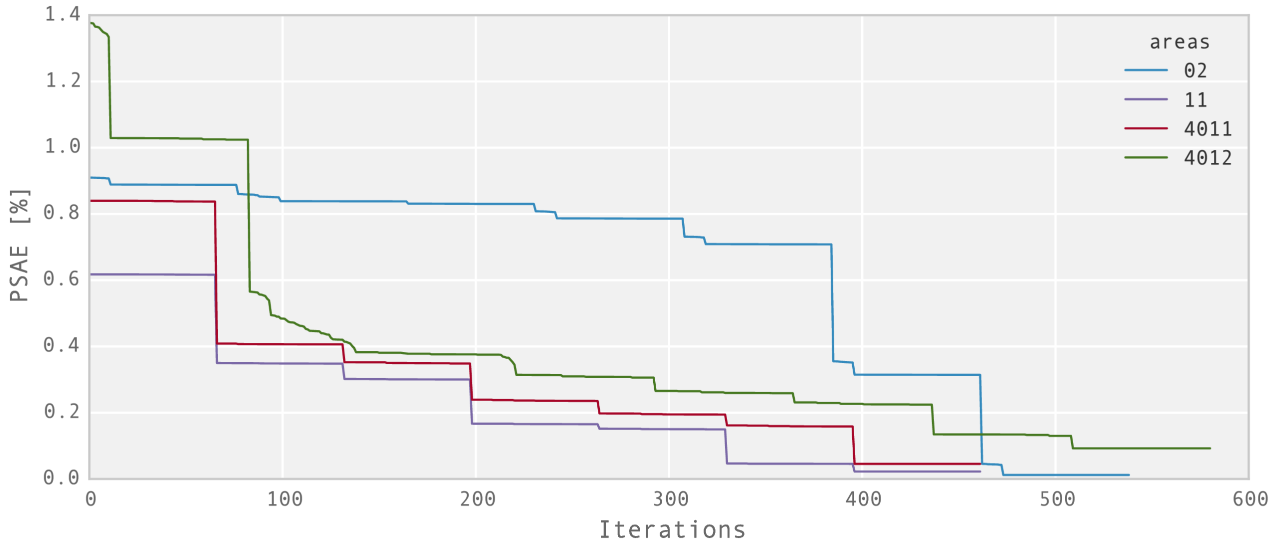


Figure 6.1: Performance of the algorithm based on the selected reference category (sorted at each area individually)

6.4 Internal Error of the Simulations

In the following section we compare the simulation results from the three reweighting algorithms. We compare two measures of the simulation results: (1) the distance between initial and estimated new weights and (2) the internal error of the simulation measured by means of the Total Absolute Error (TAE) and the Percentage Specific Absolute Error ($PSAE$).

6.4.1 Weight Distance

The weight distance is a useful measure to visualize and quantify the algorithm output. Figure 6.2 shows a weight comparison between the three reweighting algorithms, as expected both implementations of the GREGWT algorithm delivers almost the same weights. In order to see the small difference in the estimated weights we show in Table 6.1 shows the estimated new weights for the first five records of the input survey. In order to quantify the difference between the algorithms we define an absolute weight distance (see Equation 6.6). The computed absolute weight distances WD between the algorithms are presented on Tables 6.2.

$$WD_j = \frac{\left| w^R_{i,j} - w^{SAS}_{i,j} \right|}{pop_j} \div n \tag{6.6}$$

Where the weight distance WD is calculated as the mean absolute difference between estimated weights. In this case we compare the estimated weights from the R implementation of the GREGWT algorithm with the SAS implementation of the same algorithm. In order to compare this result between geographical areas we calculate the specific weights by dividing them by the population size of geographical area j. The results are listed on Table 6.1.

Table 6.1: Merged data from the different simulation methods showing estimated weight for the first 5 records in the survey

Algorithm	#	02	11	4011	4012
R	1	80.046	194.129	27.633	5.945
SAS	1	80.046	193.978	27.521	5.924
IPF	1	79.434	197.644	27.752	6.078
R	2	59.225	85.885	19.430	3.432
SAS	2	59.225	85.854	19.456	3.428
IPF	2	60.707	92.272	19.682	3.550
R	3	59.225	85.885	19.430	3.432
SAS	3	59.225	85.854	19.456	3.428
IPF	3	60.707	92.272	19.682	3.551
R	4	76.039	129.754	29.137	4.350
SAS	4	76.039	129.819	28.987	4.359
IPF	4	74.670	125.042	28.379	4.297
R	5	70.084	92.604	32.682	5.273
SAS	5	70.084	93.172	32.313	5.350
IPF	5	69.924	100.678	32.264	4.994
⋮	⋮	⋮	⋮	⋮	

Although the GRWGWT algorithm implements a deterministic procedure, the R implementation of the algorithm was not able to exactly reproduce the resulting weights from the SAS implementation. We attribute this small discrepancy between the implementations to the internal mechanics and selected reference category on both implementations rather than discrepancies in the core GREGWT algorithm. We were able to exactly reproduce the results presented by Rahman et al. (2010), in their paper the authors present a simple example and implement the GREGWT algorithm to reweight a theoretical survey. In order to perform a "real" reweight of a survey — as opposed to a simplified hypothetical example — there are many factors that can influence the results. In this case both GREGWT implementations need to perform some preprocessing of the data in order to feed these data to the core GREGWT algorithm. These small manipulations of the data are: (1) not well documented and therefore hard to reproduce; (2) many of these manipulations are performed using language internal functions, we did not investigate the exact procedure behind these functions and (3) the selection of the reference category within the GREGWT algorithm has an impact on the algorithm results.

6.4.2 Total Absolute Error (TAE) and Percentage Absolute Error ($PSAE$)

This measure is commonly used for the internal validation of spatial microsimulation models and has an extensive use in the spatial microsimulation community (Burden & Steel, 2015; B. Anderson, 2013;

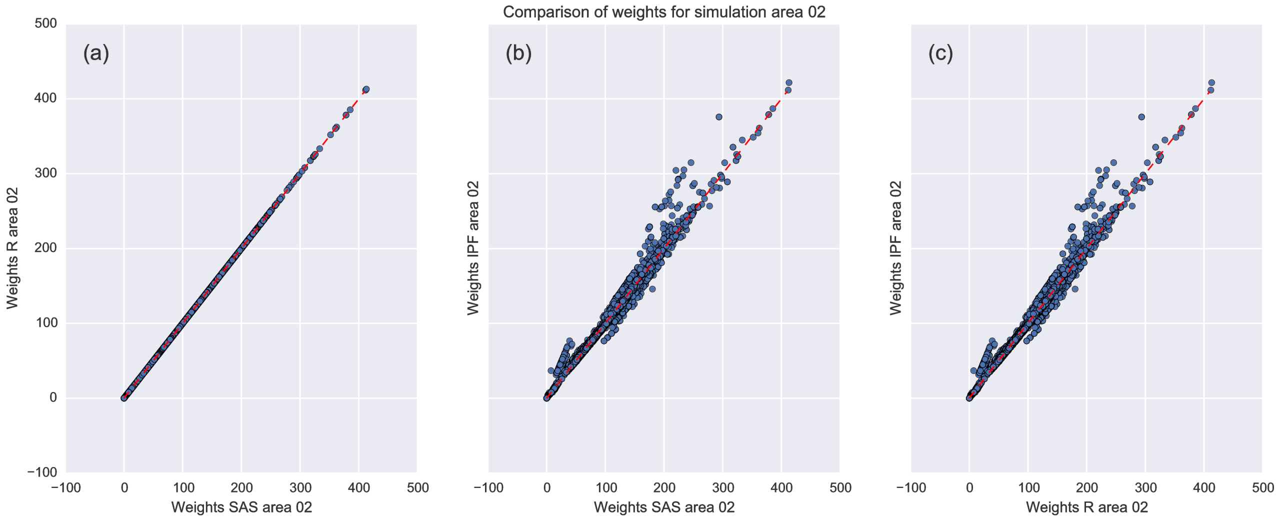


Figure 6.2: Comparison between estimated weights of the GREGWT **SAS** and **R** implementations and **IPF** for simulation area 02.

Table 6.2: Absolute weight difference between implementations

	IPF — SAS	R — SAS
Area 02	1.771e-02	8.048e-12
Area 11	2.941e-02	8.850e-04
Area 4011	2.002e-02	3.309e-03
Area 4012	1.589e-02	2.795e-03

Edwards & Tanton, 2013; Harland et al., 2012; Huang & Williamson, 2001; Tanton & Vidyattama, 2010). The total absolute error measures the absolute difference between used benchmark totals T of small areas and estimated marginal totals $\hat{T}$ for the same area. Ideally this measure is close to 0. The mathematical expression of the TAE measure is presented in Equation 6.7.

$$TAE = \sum_{i}^{n} |T - \hat{T}| \tag{6.7}$$

The comparison between observed marginal sums (benchmarks) and simulated marginal sums show a very good performance of all three implementations. Figure 6.3 plots the observed marginal sums in the y-axis and the different estimations of the algorithm implementations in the x-axis. The use of the TAE and the $PSAE$ measures to validate spatial microsimulation models give us an idea of how the algorithms are performing. With these measures we can verify that the algorithms are in fact behaving in the way they supposed to. Nonetheless, an internal validation of spatial microsimulation models cannot replace an external validation of simulation results. An external validation of spatial microsimulation models is normally not possible at a micro level but can be validated at an aggregated level.

The error of the reweighting process is minimal, on Table 6.3 we can see the difference in TAE between implementations. The difference between the two GREGWT implementations differ only on simulation area 4011. This difference is attributed to the selected reference category by the two implementations. The TAE of the IPF is slightly higher than both GREGWT implementations, except for geographical area 4011. A summary comparing all implementations and simulation areas in terms of the TAE is depicted on Figure 6.4. This figure clearly shows the disproportional error in simulation area 4011 of the SAS implementation of the GREGWT algorithm.

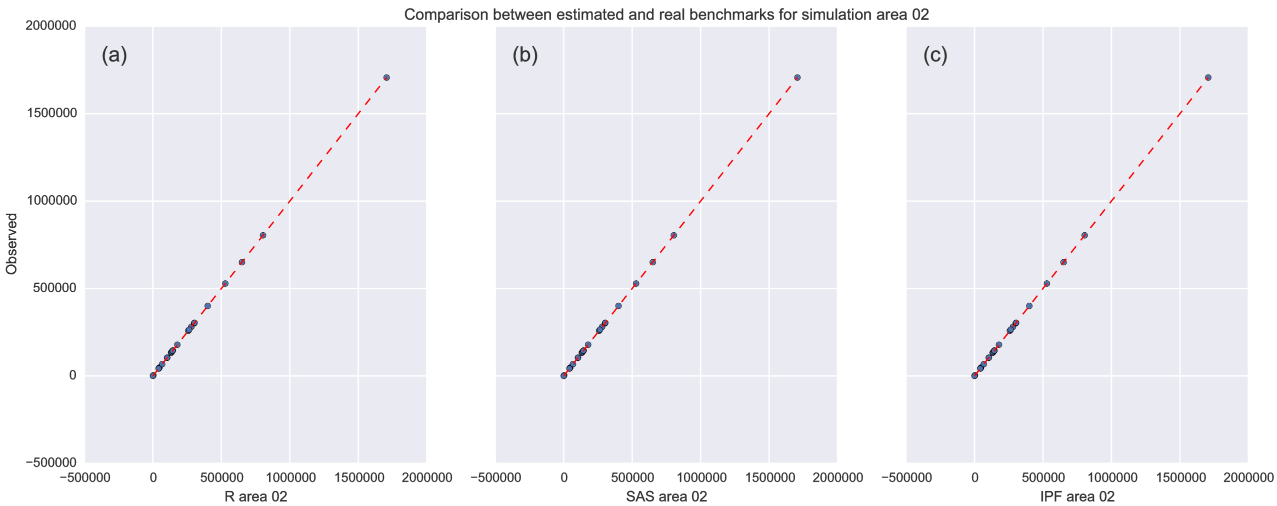


Figure 6.3: Comparison between simulated marginal sums and census constrains for: (a) the **R** implementation of the GREGWT algorithm; (b) the **SAS** macros; and (c) **IPF**.

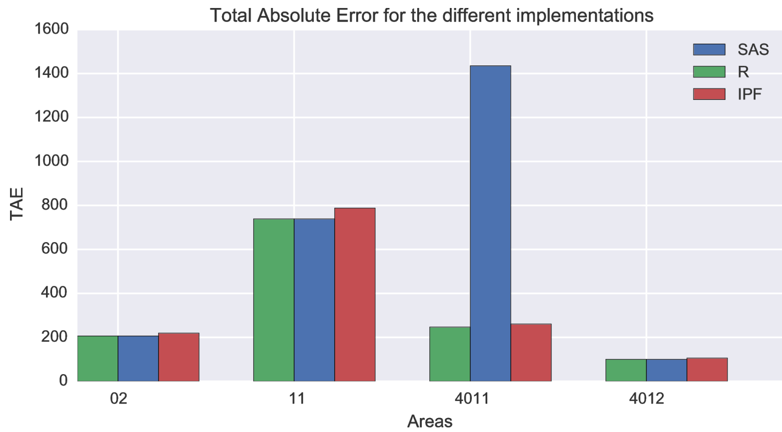


Figure 6.4: Total Absolute Error for all simulation methods

The *Standardized Absolute Error* (SAE) aims to make the TAE measure comparable between simulations which normally use different samples for the simulation and between geographical areas with

Table 6.3: TAE, $PSAE$ and Chi-Square-Distance for all areas and all simulation methods

Method	Area	TAE	PSAE	Chi Sqrt Diss
R	02	206.00	0.01	$2.26E+05$
SAS	02	206.00	0.01	$2.26E+05$
IPF	02	219.43	0.01	$2.28E+05$
R	11	739.00	0.02	$7.36E+05$
SAS	11	739.00	0.02	$7.36E+05$
IPF	11	788.03	0.02	$7.40E+05$
R	4011	247.00	0.05	$6.36E+04$
SAS	4011	1435.34	0.26	$6.35E+04$
IPF	4011	260.79	0.05	$6.39E+04$
R	4012	100.00	0.09	$1.21E+04$
SAS	4012	100.00	0.09	$1.21E+04$
IPF	4012	105.65	0.10	$1.21E+04$

different population size. The $PSAE$ measure is the same measure as the SAE but expresses its result as a percentage value. The $PSAE$ is calculated as the TAE divided by the geographical area population size and is expressed as percentage. The mathematical expression of this relationship is expressed on Equation 6.8.

$$PSAE = \sum_{i}^{n} |T - \hat{T}| \div pop_i \times 100 \tag{6.8}$$

The resulting values from the $PSAE$ are listed on Table 6.3. The error is minimal for all implementation and all geographical areas, the highest $PASA$ value (0.26%) is for geographical area 4011 with the GREGWT SAS implementation. The error for these geographical areas is so small that can be attributed to missing information on the aggregated values of the individual geographical areas. All the algorithms perform extremely good because we use a limited number of benchmarks and performed the simulation for just four geographical areas, we expect this value to increase as we add more constrains to the model and perform the simulation for a larger number of geographical areas.

With this measure we can compare the performance of the algorithm for different geographical areas and for each benchmark independently. The $PSAE$ by benchmark and geographical is plotted in Figure 6.5. The analysis of the $PSAE$ allows us to compare the results from the different implementations at a finer scale. In this figure we can see how the error is distributed among the defined benchmarks. We can see the differences between the GREGWT and IPF implementations. While the GREGWT tends to allocate most of the error to the reference category, the IPF implementation distributes it among all categories of each benchmark. The exception to this observation is area 4011.

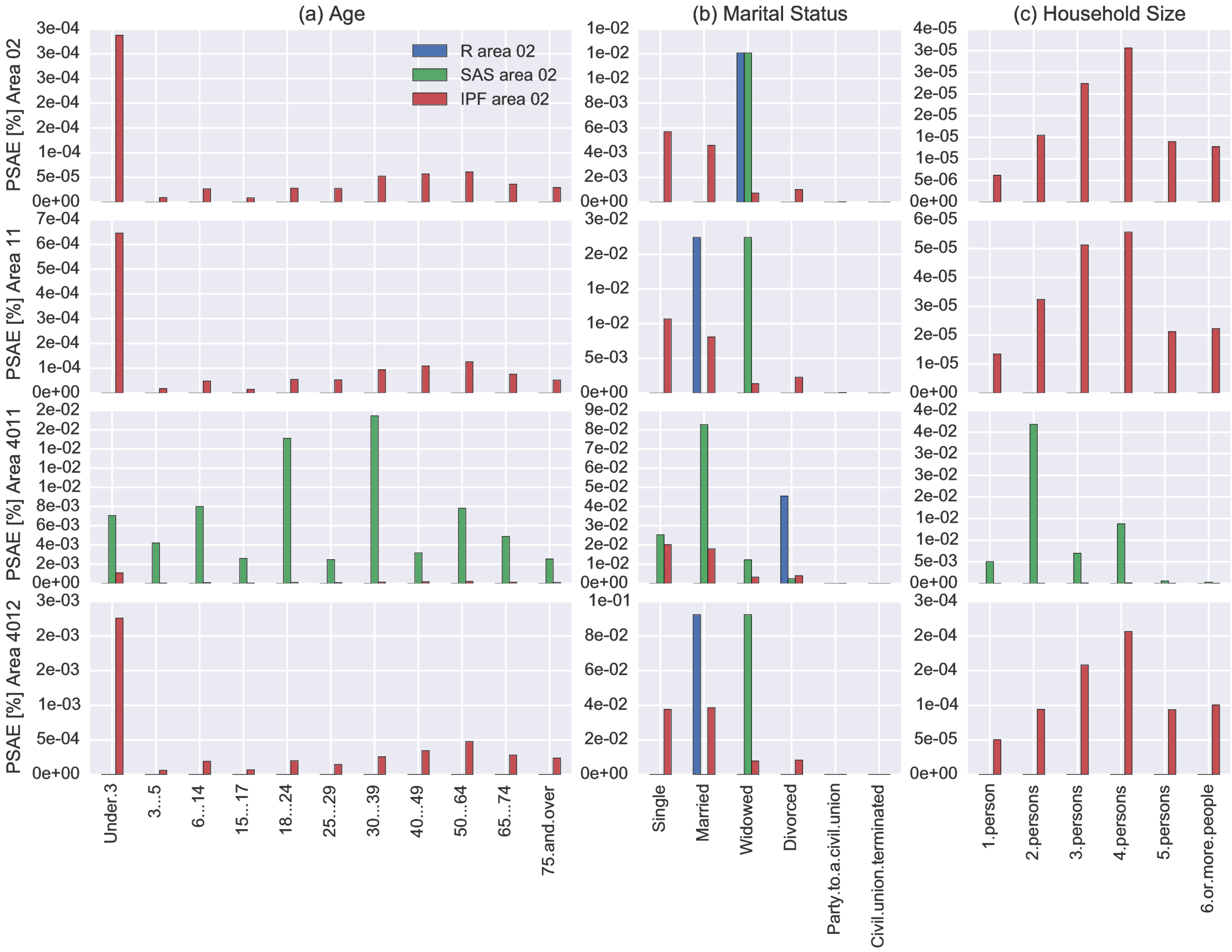


Figure 6.5: Percentage specific error for all categories by area and implementation

6.5 Chi-Squared-Distance Between Weights

The chi squared distance is used in this analysis to measure the distance between estimated new weights w and initial weights d. The chi squared distance is calculated as in Equation 6.9. In this section we compare the difference in chi squared distance between the three implementations. A summary of this measure is listed on Table 6.3 for all implementations and simulation areas. On this table we see that the values of the chi squared distance is very similar for both GREGWT implementations. In this case the chi square distance of the reweighted survey with the SAS implementation for simulation area 4011 shows the lowest figure from all other implementations.

It is important not only to assess the performance of a reweighting algorithm by means of the TAE but to observed the movement of initial weights. One of the biggest advantages of reweighting a survey for the generation of spatially allocated micro-data is that we maintain all the information of the survey even if we only benchmark this reweighting process to a few variables of the survey. The initial weights define the distribution of the individuals represented in the survey, this reweighting algorithms takes an "extra" parameter as a reweighting constrain: weight distance (the IPF algorithm is also constrained to weight distance, see next section for details). This means that the initial weights are very important within the reweighting process. In most cases we have a design weight attached to the input survey that we can use and trust. On some other case we do not have an input weight and we need to be aware of the induced error by defining a uniform distributed weight (e.g. all weights set to 1). Vidyattama, Tanton und Biddle (2015) present an example of a synthetic input survey for which all input weights are unknown. The authors generate a synthetic input survey using a probability table to ensure the distribution of the data is similar enough to the aggregate known distribution, then the authors set a uniform distributed weight matching the total population size of the corresponding aggregation unit.

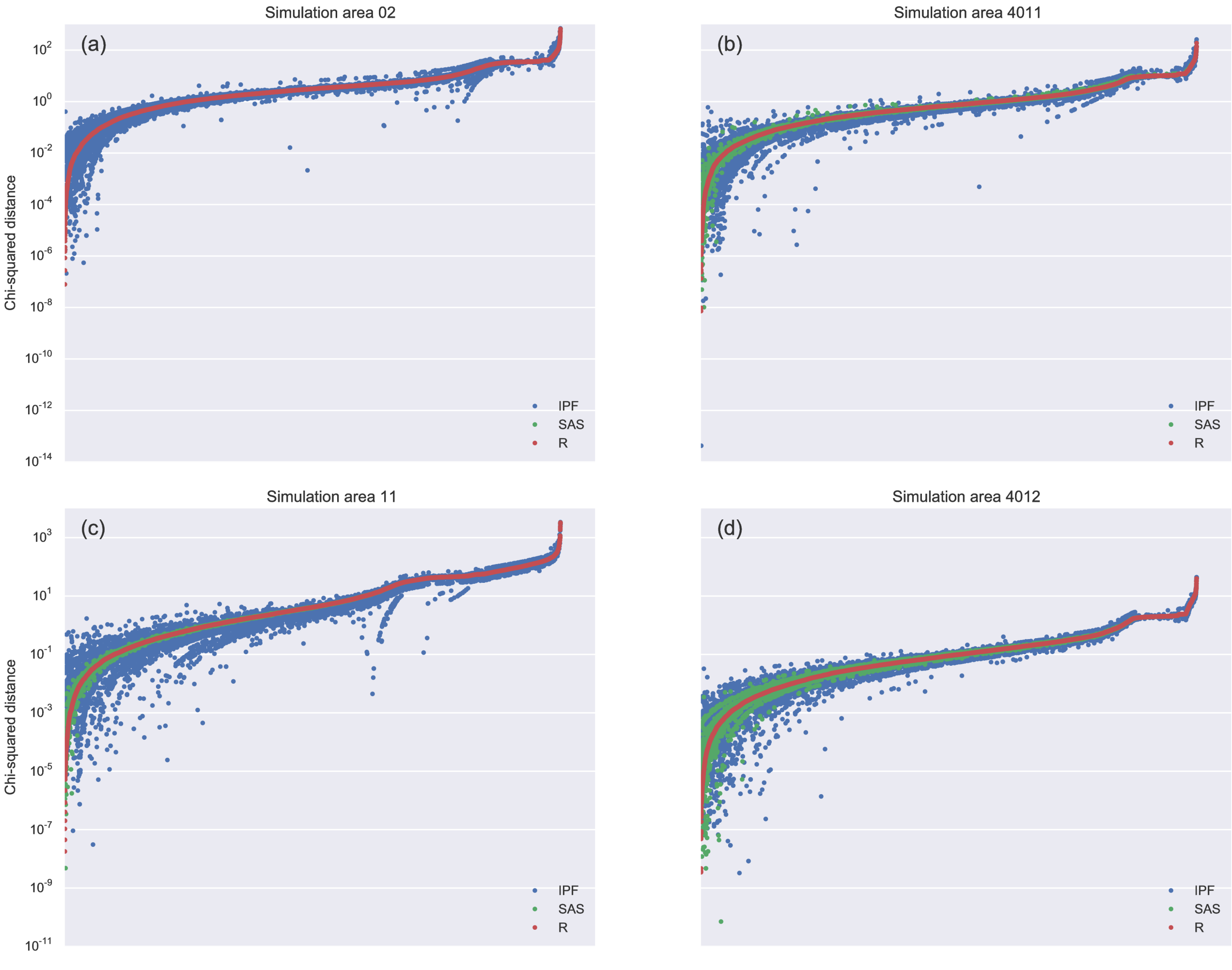


Figure 6.6: Chi-squared distance

All three implementations analyze in this section present a similar value for the chi squared distance. The distance for the individual records of the survey are plotted on Figure 6.6. The plot shows the Chi squared distance of all implementations sorted by the values of the R implementation of the GREGWT algorithm. As expected, the difference between both GREGWT implementations is minimal. The values for the IPF implementation follow the same pattern as the GREGWT implementations but are not as close as both GREGWT implementations are.

$$Chi_i = \sum_{j}^{m} \frac{(w_j - d_j)^2}{2d_j} \tag{6.9}$$

6.6 Manipulating the Initial Weights

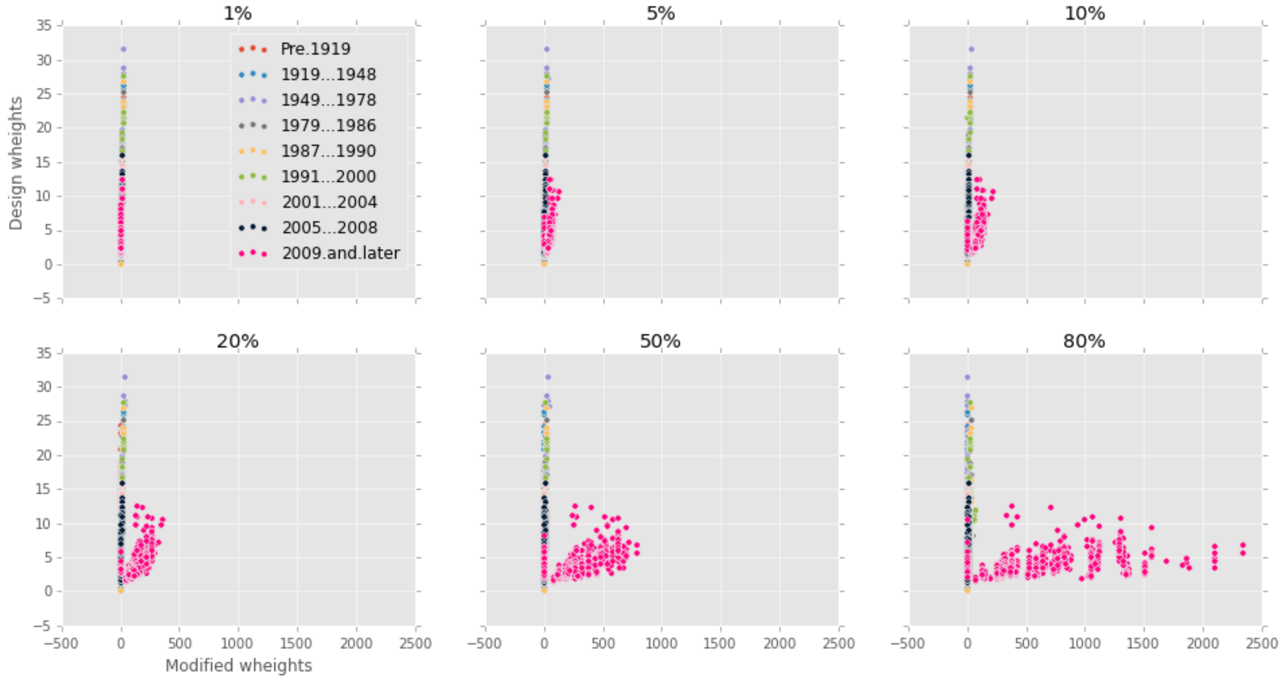
Because the GREGWT algorithm aims to minimize the chi square distance between, we can use this as a simulation tool. In the following section we present an application of this in a simple spatial microsimulation and discuss other possibilities for the application of this method in the context of the design of spatial microsimulation models.

For the purpose of this section we manipulate the weight distribution of the initial survey based on the construction year of the buildings on which individuals represented in the survey reside. This manipulation of the initial weights represents a “retrofit” of the building stock by re allocating weights to buildings with a newer construction year, and there for a better energy standard. The example is a modification of the spatial microsimulation performed by Muñoz H., Vidyattama und Tanton (2015b).

Lovelace et al. (2015) shows that a manipulation of initial weights has a negligible impact on TAE using an IPF algorithm. We make a similar observation for TAE values using the IPF and the GREGWT R implementations. In this case the TAE values do not vary much, $2.95e-08$ and $3.95-06$ for the GREGWT and the IPF implementations correspondingly for a manipulation of initial weights assuming an 80% retrofit of the entire building stock (See Table 6.4). A variation of the TAE does not mean that there is no variation on the resulting weights. This variation of the weights is depicted on Figure 6.7 for the GREGWT (a) and IPF (b) implementations. These plots show that the modifications made on the input weights of the survey have a direct effect on the resulting weights by maintaining an almost constant TAE value. The change on the resulting weights is almost a factor 10 lower for the IPF algorithm. These means that the IPF algorithm is less sensitive to a modification of the input weights. Depending on the model design a low sensitivity to input weights might be desired. This could be the case for which design weights are not available. In the other hand if we want to explicitly use the input weights as part of the model design we might prefer an algorithm with a higher sensitivity to modifications to input weights. Therefore we recommend taking a conscious decision regarding the defined initial weights of the input survey. In the case of unavailable input weights we need to understand that the implicit decision of a uniform distributed weight has to be assumed. For many cases the assumption of a uniform distribution might be the correct one, but we might find better distributions more accurately representing the population of a particular geographical area.

In the case of the GREGWT algorithm, the input weights have a direct impact on the resulting weights. The GREGWT algorithm does not only reweight a survey to meet given area benchmarks but also minimized the distance between initial and estimated weights (see Equation 7.4). We see this as an opportunity to control a simulation process by manipulating the initial weights of the input survey. In the following section we present a small use case of this technique. In this example we change the initial weights based on the construction year of buildings, aiming to simulate retrofits on the building stock.

(a)



(b)

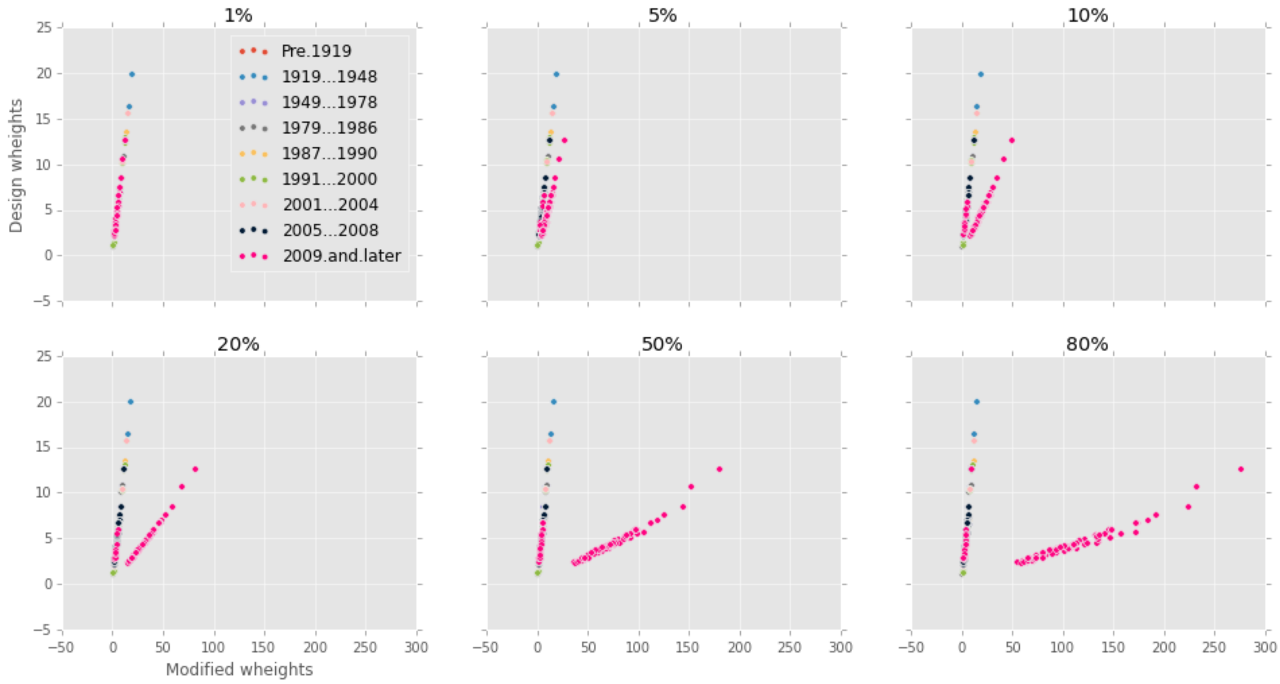


Figure 6.7: Modified Weights and reweighted with (a) GREGWT and (b) IPF

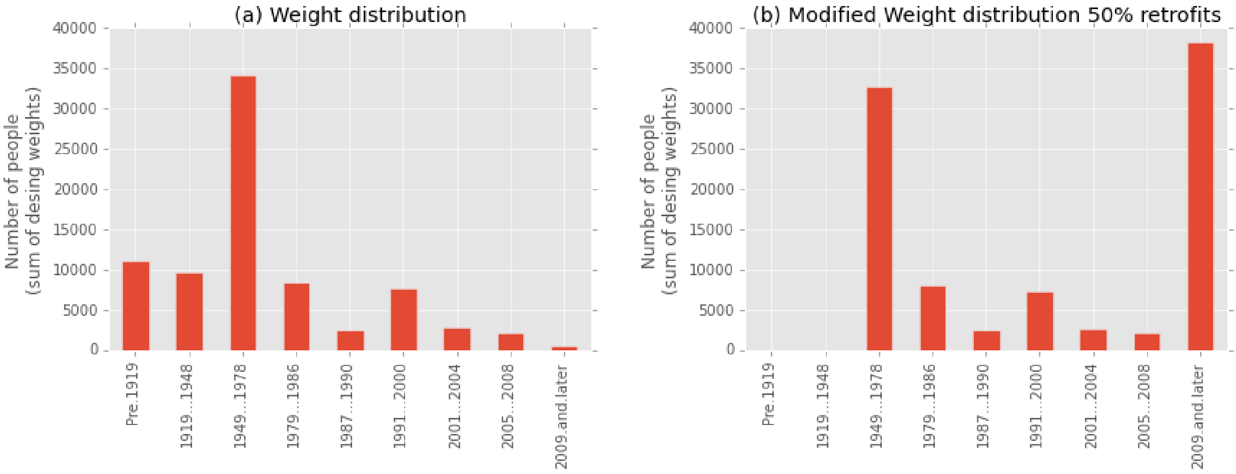


Figure 6.8: Number of people living in buildings by construction year for: (a) survey design weights and (b) manipulated weight representing a 50% retrofit of the building stock

The input survey data, describing individuals, contains also some information describing the residence of the individual, construction year of the building in which the individual resides in one of these parameters. Muñoz H., Vidyattama und Tanton (2015b) use these parameters to simulate retrofits of the building stock by benchmarking the survey to construction year, the retrofit rates are estimated at a geographical area level. This projection at a low aggregation level is not always accurate. An alternative to this method it a manipulation of the initial weights, in this case we won't benchmark the survey to construction year but will modify the initial weights. The way in which we modify the initial weight is by changing the weight distribution. Figure 6.8 compares the weight distribution by construction year, Figure 6.8a shows the distribution of the design weights, in this distribution we can clearly see the post war reconstruction period that dominates the building stock in Germany. In this example we want to represent a 50% retrofit of the building stock, a simple way to model this is to move the weight of the older buildings into the newest category. Figure 6.8b shows the new distribution. Of course, more accurate and elaborated manipulations of the initial weights are possible with this method.

With this new modified weight we rerun the reweighting procedure and as expected, the GREGWT algorithm respects the modification made to the initial weights. Table 6.4 lists the TAE for the different modifications of the weights (expressed as percentage of retrofitted building stock), the difference from the initial TAE (00%, survey design weights), and the Chi squared distance between: (a) survey design weights and estimated new weights and (b) modified survey weights and estimated new weight. From this table we can see that the variation in TAE is minimal while the Chi square distance changes, the change on both Chi square distances is bigger with the GREGWT implementation than with the IPF one. It is important to notice that we do not benchmark the survey to construction year in this example. The manipulation of the initial weights presents an alternative method to constrain the reweighting process to known or projected values at a more aggregated level. The advantage of constraining the model at a different aggregation level can be useful in many model designs scenarios, for example: (1) we may have important data available at a higher aggregation level, in this case we can modify the initial weight to match this value and subsequently reweight the survey to the underlying geographical areas, Vidyattama et al. (2015) uses a similar method for the creation of a synthetic unit record survey of indigenous population in Australia; or (2) we can modify, as briefly presented in this

Table 6.4: TAE, Difference from initial TAE and Chi squared distance for different modifications of the initial weights d and modified initial weight md reweighted with: (a) GREGWT and (b) IPF

	00%	**01%**	**05%**	**10%**	**20%**	**50%**	**80%**
(a) GREGWT							
TAE	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$
Diff	$0.00e+00$	$8.07e-08$	$9.46e-08$	$7.49e-08$	$9.31e-08$	$5.69e-08$	$9.26e-08$
Chi d	$1.60e+07$	$1.61e+07$	$1.27e+08$	$6.96e+08$	$4.15e+09$	$4.73e+10$	$4.88e+11$
Chi md	$1.60e+07$	$1.61e+07$	$1.18e+08$	$6.39e+08$	$3.83e+09$	$4.33e+10$	$4.61e+11$
(b) IPF							
TAE	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$	$3.79e+04$
Diff	$0.00e+00$	$0.00e+00$	$2.03e-06$	$1.60e-05$	$5.90e-06$	$3.12e-06$	$3.95e-06$
Chi d	$1.40e+07$	$1.40e+07$	$1.44e+07$	$1.79e+07$	$3.54e+07$	$3.14e+08$	$1.09e+09$
Chi md	$1.40e+07$	$1.40e+07$	$1.41e+07$	$1.61e+07$	$2.64e+07$	$1.85e+08$	$6.10e+08$

section, the input weight to project the survey matching target distributions define by a postulated scenario or target values define on policies.

6.7 Projecting the Synthetic Population Into the Future

In this section we analyze a plausible scenario for many developed countries, this is: a parallel increase in the energy efficiency of the building stock and an ageing of population. The consequences of this development for a secure heat supply of the residential sector is the focus of this section.

The influence of users on domestic heat demand has been identified as an important factor to reduce the gap between estimated heat demand and consumed heat demand (D'Oca et al., 2014; Durand-Daubin et al., 2013; Haldi & Robinson, 2011; Guerra Santin et al., 2009). In a scenario with a large share of energy efficient buildings and an old population the incorporation of residents influence on energy demand models could be essential for: (a) securing heat supply of the residential sector, and (b) achieving an optimal heat supply in this sector.

In order to assess these consequences we: (1) project demographic benchmarks at a district level fitted to national statistics; and (2) project the characteristics of the building stock at the same aggregation level. We use these benchmarks to reweight the German micro census survey for each district in the city and each simulation year.

The method presented here makes use of microsimulation techniques developed for the reweighting of surveys to small areas, see (Tanton, 2014) for an overview of these techniques and their applications.

Microsimulation is a simulation working at a micro scale (individuals, households, firms, buildings). This would include tax transfer, transport and land use models. In this section we make use of two methods used by the spatial microsimulation community: (1) a survey reweighting for the creation of synthetic populations allocated to geographical areas, this method is the starting point of many spatial microsimulation models (a reweighing of a population survey is not the only method for the generation of synthetic populations); and (2) the reweighting of the same survey to projected statistics of the same geographical areas. These methods allow us to generate a synthetic population for each simulation year. The use of these methods to estimate energy demand is not very common among the microsimulation community, some examples of this application are (Muñoz H., 2014; Muñoz H. & Peters, 2014b).

Static spatial microsimulation models have also been used to project micro-data, benchmarking the survey to projections at a small area level (Vidyattama & Tanton, 2013).

The section is organized as follows: (1) we present the data used for this analysis; (2) we discuss the used methods; and (3) discuss the results and future improvements of the developed method.

6.8 Data Extrapolation

In order to project the population growth at district level we: (1) extrapolate the growth rate based on historical data for the corresponding district; and (2) fit this extrapolation to the national population projections.

For the extrapolation of historical values we use a linear function. We created a sample with eight observations (2003–2010), being all the available records online at this aggregation level.

In a second step we fit these extrapolations to the available national projections.

$$G(y) = P_y \div P_{y-1}^{1/\Delta y} \tag{6.10}$$

$$E(y, s) = a_s y + b_s \tag{6.11}$$

Where $G(y)$ is the growth rate as a function of year y, P_y is the projected population in year y, $E(y, s)$ is the extrapolated growth rate as a function of year y and district s and a_s and b_s are the resulting coefficients from the extrapolation function as a function of the historical data for district s.

$$G_y = \sum_{s=1}^{n} E_{s,y} \tag{6.12}$$

$$D(y) \sim \mathcal{N}(G_y, sd) \tag{6.13}$$

$$sort(E_y) = sort(D_y); \text{by } s \tag{6.14}$$

Where $\mathcal{N}(G_s, sd)$ is a set of random numbers with a normal distribution with mean G_y and standard deviation sd (default set at $1e-3$).

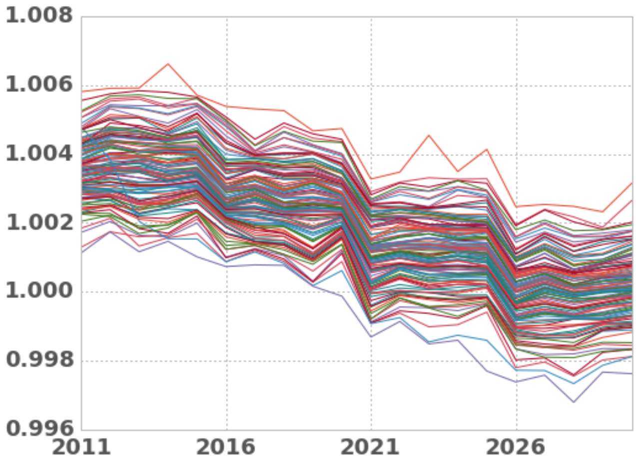


Figure 6.9: Extrapolated and fitted growth rates for all districts

Figure 6.9 shows the fitted extrapolated growth rates for all districts in the city of Hamburg. These growth rates are capped in order to meet the aggregated projections of the city.

The available projections for the city contain information about age, gender and foreign national share, but historical data from individual districts contain only information about gender and foreign national share. At this time we do not make any assumptions regarding different changes in age distribution for the individual district, but simply inflate these values to match the city official projections.

6.9 Synthetic Building Stock

Because we use the micro census to create a synthetic population for each district of the city, constrained by aggregated statistics, we decided to define a synthetic building stock based on the micro census. We define a synthetic building for each individual in the micro census. We use the following available parameters to generate the synthetic buildings: (1) household size, (2) construction year, (3) dwelling unit size and (4) number of dwelling units in the building.

With these variables we create a micro building stock that in its aggregation represents the "real" building stock of each district. The advantages of describing the building stock in this fashion are many. This simplification allows us to perform a relatively simple reweighting of the micro census, making the method transferable to many cities around the world, for German cities we only have to change the aggregated district statistics, a simulation at a national level is also possible.

6.10 Estimating Heat Demand

For the estimation of heat demand we use a simple heat demand balance method (Muñoz H., 2015) developed in the R language. This library is a simplification of the German DIN V 18599 standard.

The library does only compute heat gain and losses. The difference of these values are interpreted as the heat demand. Factors like heat recovery systems of mechanical ventilations or the efficiency of the heat supply infrastructure are not considered within the library.

We vary the following input parameters for the estimation of heat demand:

- Geometry of the buildings, expressed as: (a) length, (b) width and (c) height. See previous Section "Creating a Synthetic Building Stock"
- Heat transmission coefficients of building components, U-values of: (a) roof, (b) walls and (c) widows. See next subsection "Building typology"
- Ratio of glazing surface. See next subsection "Building typology"
- User influenced parameters: (a) Internal heat gains Q_i, (b) Internal temperature set point T_i and (c) Air exchange rate n. See subsection "Estimation with user influence" and Table 6.5.

Because the focus of this model is primarily on method we decided to use a heat balance model instead of a thermal simulation model like EnergyPlus or ESP-r because of: (a) the heat balance model requires less input parameters and less complicated input parameters; and (b) the computation time of the heat balance model is less that of the thermal simulation model. The use of a thermal simulation model is possible (see (Muñoz H., 2014)), the advantage of using a thermal simulation model is its ability to take occupational schedules as input.

Table 6.5: User influenced variables used in the model as function of working hours (WH)

WH	$n[h^{-1}]$	$T_i[C^\circ]$	$Q_i[W/m^2]$
≤ 1	0.7	22	7
≤ 4	0.6	21	6
≤ 8	0.5	20	5
≤ 9	0.4	19	4
> 9	0.3	18	3

6.10.1 Building Typology

For the estimation of heat demand we make use of the well-established building typology from the IWU institute (Loga et al., 2011). We use this typology to define heat transmission coefficients and glazing ratio of the building stock. The same typology is used by Muñoz H. und Peters (2014b) to classify the building by building type. With the heat transmission coefficients and glazing ratio from the building typology we compute two absolute heat demand values for each synthetic building on the survey: (1) only taking into account characteristics of the building stock — building geometry and heat transmission coefficients — variables influenced by the user are maintained constant for all buildings; and (2) including demographic characteristics, by manipulating variables influenced by the user as function of the working hours of each individual in the sample.

Because of the nature of the model the synthetic population is generated for each simulation year, benchmarked to aggregated statistics of both: the population (demographic parameters); and characteristics of the building stock (capture by the distribution of building types). A retrofit rate of the building stock modifies the distribution of building types at an aggregated level, a 2% retrofit rate will pick 2% of the buildings of each geographical area and attribute them to the new energy efficient building types. See section "Benchmarking population" for a more detailed description of this computation step.

6.10.2 Estimation With User Influence

The variables used to simulate the user influence on heat demand are listed in Table 6.5. These variables are modified as a function of the occupant working hours. The variation in these parameters has not been empirically validated. The aim of this modification is to represent a hypothetical change in heat consumption based on demographic characteristics. We use the variable *Working Hours* as a proxy to induce this influence based on empirical analysis of other authors. A brief literature review by Muñoz H. und Peters (2015a) concludes that occupancy rates of users seem to be the determinant of user influence in heat demand in the residential sector.

One of the attributes of the population survey is "working hours". We use this attribute in order to define: (a) ventilation rates; (b) internal temperatures; and (c) internal heat gains. These three parameters are given as input to the heat demand model.

There exist approaches to simulate occupancy rates in a stochastic fashion (Page et al., 2008; Hoes, Hensen, Loomans, de Vries & Bourgeois, 2009). We aim to contribute to these efforts by simulating the influence of a specific group of users on heat demand at a district level.

Figure 6.10 shows the relative difference between both estimations of heat demand: (1) taking user influence into account; and (2) maintaining variables influenced by the user constant. As expected, the influence of users is higher on more efficient buildings. Building type 18 (large multifamily houses of construction period 1958–1968) shows a lower variation and lower mean than the rest of the building typologies, this effect is caused by the low number of individuals from the micro census attributed to this type.

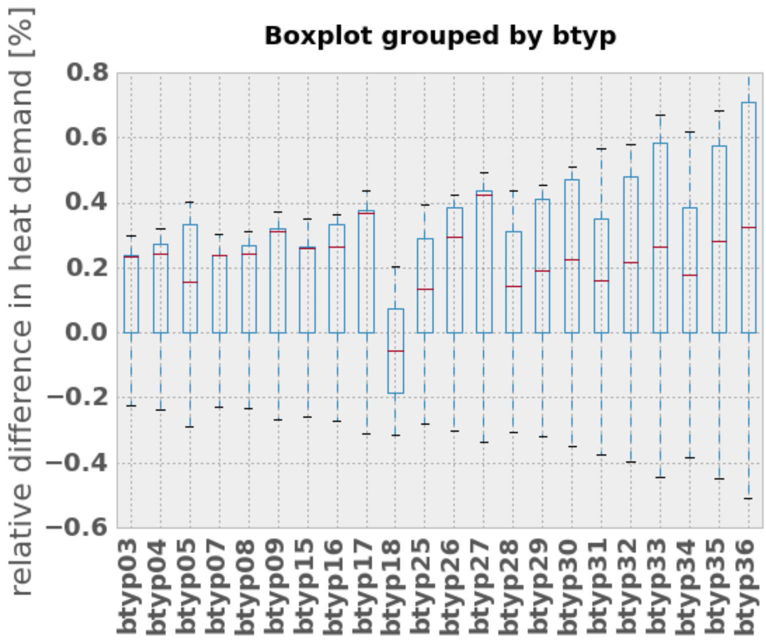


Figure 6.10: Difference between estimated heat demand with and without demographic parameters, grouped by building type.

An analogue comparison of relative heat difference is presented in Figure 6.11, in this plot the difference is grouped by age class. This plot shows a higher variation in heat demand for ages older than 15. We can see that the mean of both age classes, 21–45 and 46–65 are exactly on the zero line (no difference), these are the values used in the models that do not take user influence into account.

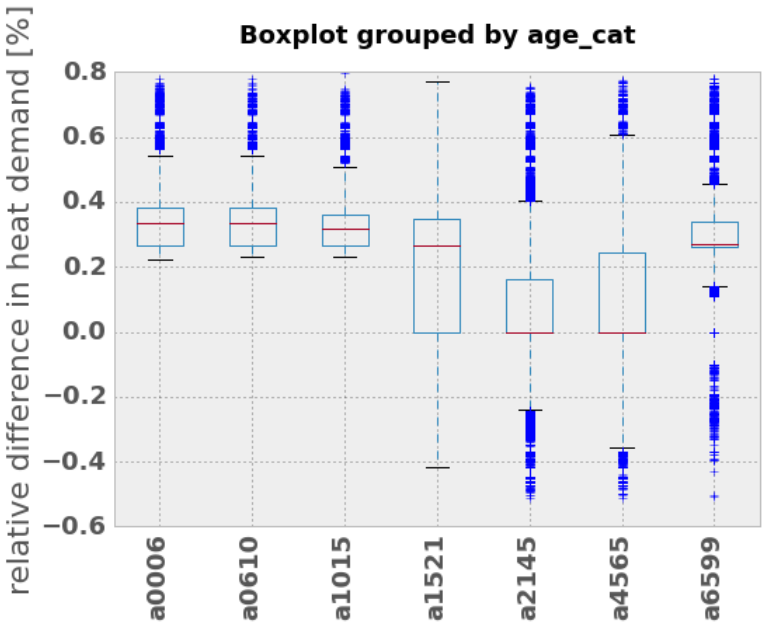


Figure 6.11: Difference between estimated heat demand with and without demographic parameters, grouped by age class of user.

6.11 Benchmarking Population

In order to assess the impact of: (1) an ageing population; and (2) a rising of efficiency rates of the building stock, we reweight the survey to demographic and building stock benchmarks projected until 2030 in a five year step at a district level.

For the reweighting of the survey, containing heat demand values for each individual in the sample, we use an implementation of the GREGWT algorithm in the R language (Muñoz H., Vidyattama & Tanton, 2015a). This method is used in the spatial microsimulation community to generate synthetic populations of small areas, Tanton et al. (2011) deliver a detail description of the algorithm and its application. The GREGWT algorithm is based on "method 5" from A. Singh und Mohl (1996). For a technical description of the algorithm see (Bell, 2000) and for applications of it see (Tanton & Vidyattama, 2010; Tanton, Harding & McNamara, 2013).

6.12 Simulation Scenarios

For the development of simulation scenarios we define six different scenarios divided into two groups: (1) base scenarios, and (2) Different retrofit rates.

1. Base scenarios

Scenario 1 Base scenario.
The building stock benchmarks are constant through the simulation. The distribution of building types is constant.

Scenario 2 New buildings scenario.
New buildings added to the building stock (driven by population growth) are attributed to the newest building type.

2. Different retrofit rates

- Average user.
  User parameters (n, T_i and Q_i) are constant for all buildings and all simulation years.

  Scenario 3 0.5% retrofit rate

  Scenario 4 1.0% retrofit rate

  Scenario 5 1.5% retrofit rate

  Scenario 6 2.0% retrofit rate

- User influence.
  User parameters (n, T_i and Q_i) are defined based on the working hours of the building residents.

  Scenario 3-User 0.5% retrofit rate

  Scenario 4-User 1.0% retrofit rate

  Scenario 5-User 1.5% retrofit rate

  Scenario 6-User 2.0% retrofit rate

6.12.1 Base Scenarios

For this analysis we define six different scenarios. The first “base” scenario only considers changes in the population, maintaining the state of the building stock constant. This “base” scenario is not very realistic as all new families introduced into the district will be attributed a building with the characteristics of the district. A more realistic “base” scenario is therefore introduced, we call this scenario “new buildings”. In this scenario each new family introduced to the district is attributed a new building with energy efficiency standards corresponding to the last four types of the building typology.

6.12.2 Different Retrofit Rates

The following scenarios define different retrofit rates for the entire city. In these scenarios we take a weighted sample of the entire building stock and "retrofit" these buildings to new construction standards defined by taking one of the last four building types of the building typology.

In order to pick the buildings to be retrofitted we take a sample with a given probability based on the building construction year, which is embedded in the building typology.

This probability is expressed with help of an exponential function in Equation 6.15 where $p(b)$ is the sampling probability as a function of building type. This function is used to assign probabilities to the survey, V contains these probabilities (Equation6.16). We select m number of buildings from the sample with the estimated probability (Equation 6.17). Set S contains the m buildings that will get retrofit in the specific simulation step.

$$p(b) = e^{1/b} \tag{6.15}$$

$$V = p(B_i); \text{for } i = 1 \text{ to } l \tag{6.16}$$

$$S_k = v_i \quad \begin{array}{l} ; \in V - S \\ ; \text{while } k < m \\ ; \text{if } v_i \geq rand\left(e^{1/\max(b)}, e^{1/\min(b)}\right) \end{array} \tag{6.17}$$

$$m(r, y) = B \times (1 + r)^{y-2010} - B \tag{6.18}$$

We define four annual retrofit rates, which define our scenarios: (1) 0.5%, (2) 1.0%, (3) 1.5%, and (4) 2.0%. For each one of these scenarios we run two simulations: (1) taking the occupant influence into account ("user") and (2) using the same "average" occupant for all buildings. The number of buildings selected for retrofitting at a specific at a specific simulation year is defined by Equation 6.18.

6.13 Results

The simulation results show: (1) the impact of the different retrofit scenarios on total heat demand for the city of Hamburg, see normal lines on Figure 6.12; and (2) the difference between the model that takes user influence into account and the one that maintains this variation constant, see dotted line in Figure 6.12. A map showing this difference for simulation year 2030 is depicted on Figure 6.13 for all districts in Hamburg. The map shows the simulated difference between the "user" scenario (taking occupant influence into account) and the "average user" scenario. The difference is expressed as heat density for simulation year 2030 with a 2% retrofit rate. We see a concentration in the city center

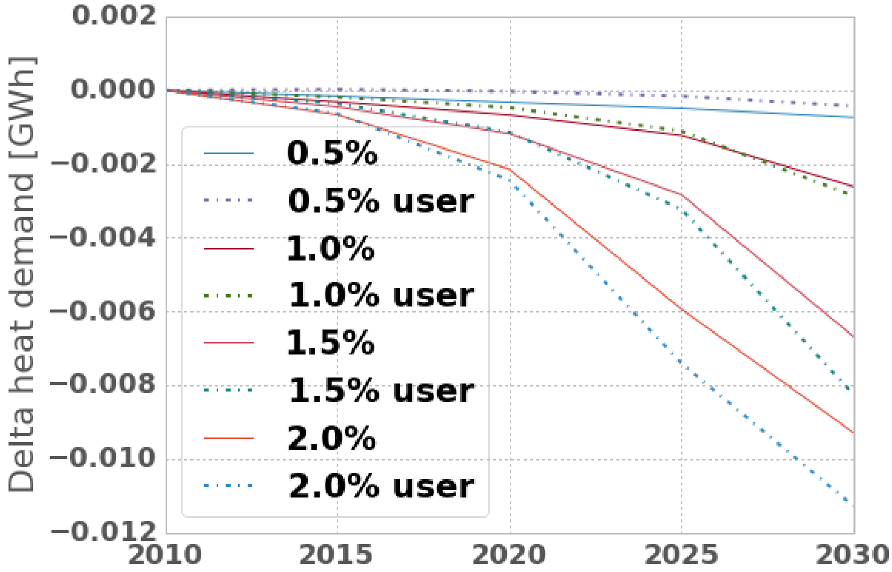


Figure 6.12: Estimated development for all four retrofit scenarios in heat demand for the city of Hamburg expressed as difference from base scenario, simulations including user influence are depicted with dashed lines

where most of the old buildings are located.

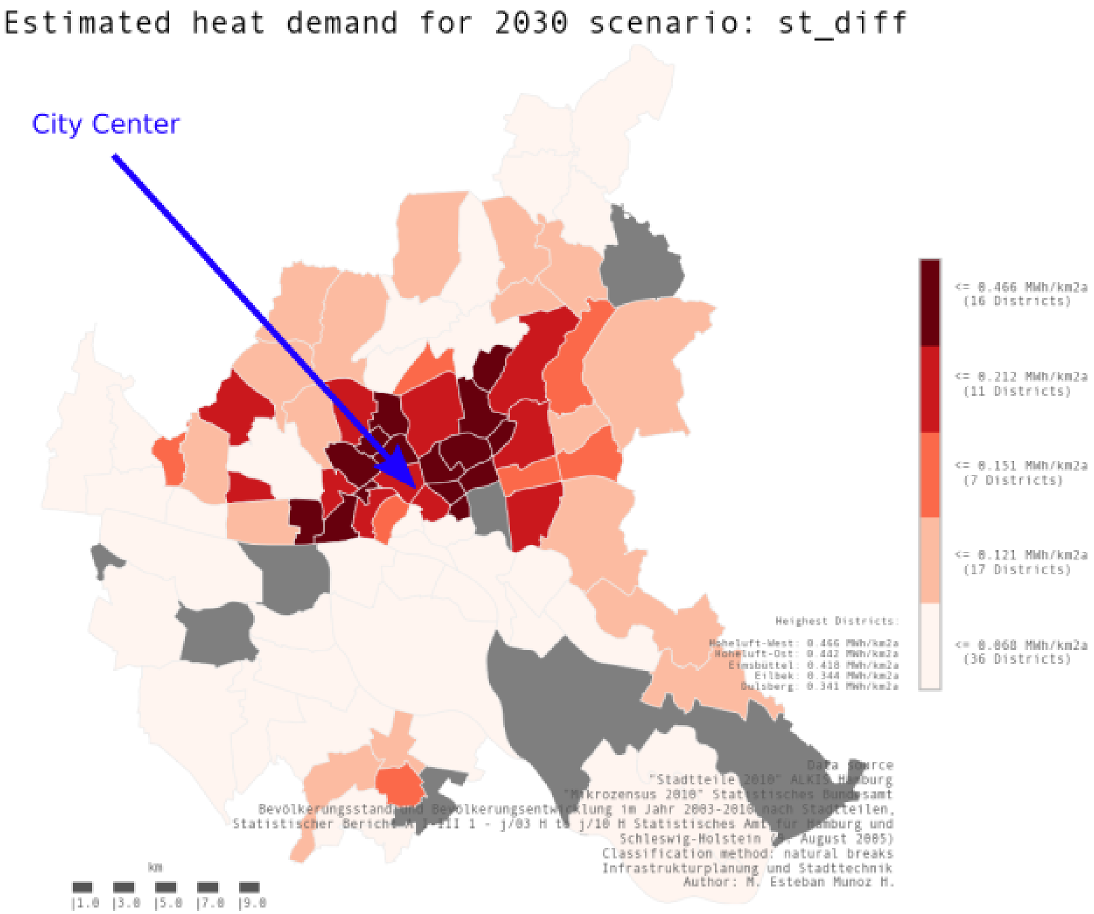


Figure 6.13: Difference in heat density $[MWh/km^2a]$ between model including and not including user influence in the computation for all districts.

An interesting observation in the results is that the difference in heat demand (absolute difference from base scenario) is bigger (more net savings) when we include user behavior in the model. This observation does not occur on all scenarios. On scenario "0.5% retrofits" we observe the expected behavior, this is a higher heat demand under explicit consideration of user behavior. On Scenario "1.0% retrofits" we see a breakpoint on simulation year 2025. For all the other scenarios this breakpoint occur in 2020 and 2015 for scenario "1.5% and 2.0% retrofits" respectively.

We observe this effect because the absolute influence of user behavior is higher (more kWh) on older buildings while the relative user influence (higher %) is higher on energy efficient buildings. Figure 4 shows the net energy savings for the different scenarios. For scenarios including user behavior (dotted lines), the line represents the simulated heat demand of corresponding scenario minus the simulated heat demand of the "base" scenario (no retrofits) also taking user influence into account. For the base scenario, the absolute user influence on heat demand is higher than the influence on simulation scenarios with more energy efficient buildings. This result on higher net savings for scenarios with user behavior. It must be noted that the relative influence of user behavior increases as does the energy efficiency of the building. This is important for the dimensioning of heat supply systems. This effect is also represented in the map plotted in Figure 5, we see a concentration of the highest difference in the city center where most of the old buildings are located.

7 Estimating Heat Demand with a Heat Balance Method[1]

7.1 Heat Balance Method

Following equations show a simple procedure to estimate the heat demand of a building (see equations 7.9 and 7.1.4). As input parameter for such calculation a table containing the solar radiation on the vertical plane for the city of Hamburg is used (needed for solar gains). This table contains values for the monthly average solar radiation at different orientations. The second input needed in order to perform this calculation is the geometry of the building. Furthermore the computation of final and primary heat demand, are presented in equations 7.1.5 and 7.1.5.

Specific values corresponding to the building types, needed for this calculation, are calculated based on a building stock analysis for the city of Hamburg. This analysis calculates the average floor space and average height for the specific building types define in the process. U-values for the building types are estimated based on literature review.

The performed calculation represents a simplified variation from the DIN V 18599 (2011) standard calculation method. This method is widely used for the estimation of heat demand in buildings.

7.1.1 Heat Gains Qg

$$Qg_{(m)} = 0.024 \times (Ss_{(m)} + Si) \times t_{(m)} \tag{7.1}$$

Where: m Month of the year.

$0.024\ kWh$	$= 1Wd$
Ss	average monthly solar heat flow $[W]$
Si	heat flow by internal heat sources $[W]$
t	number of the days in a particular month $[d/m]$

The heat gains, as well as the rest of the computations, are performed on a monthly basis. In Equation 7.1 the heat gains (both solar and from internal heat sources) are computed for each month of the

[1]Sections of this chapter are based on: Muñoz H., Seller, & Peters (2015)

year.

Internal heat gains Si

$$Si = qi \times An \tag{7.2}$$

Where: qi	Average value for internal heat emissions.
An	Heated area.

The internal heat gains are computed with help of a factor qi. This factor is constant through all computed months. The used value is to be understood as an average value for the whole year. The factor represents the heat emissions of a household per m^2 living space.

Monthly solar heat flows Ss

$$Ss_{(m,j)} = I_{(m,j)} \times A_{(j)} \tag{7.3}$$

Where: m	Month of the year.
Ss	average monthly solar heat flow $[W]$
I	average monthly solar intensity of radiation $[W/m^2]$
A	actual collector surface $[m^2]$
j	orientation (direction and down-grade to vertical)

The monthly heat gains are computed for each month based on the preloaded table of solar emissions for the city of Hamburg. The solar gains are computed for each window based on its orientation.

7.1.2 Heat Losses Ql

$$Ql_{(m)} = 0.024 \times H \times \left(Ti - Te_{(m)}\right) \times t_{(m)} \tag{7.4}$$

Where: m	Month of the year.
$0.024\ kWh$	$= 1\ Wd$
H	Specific total heat loss $[W/K]$
$Ti - Te$	Difference between internal and ambient temperature $[K]$
t	Number of the days in a particular month $[d/M]$

As a contra part to the heat gains the heat losses are computed. The heat losses of the building are computed for each month, as the transmission of heat directly depends on the temperature difference between the inside and the outside of the building. Here we define a constant operative internal temperature, define by the corresponding family class. The outside temperature is taken from the previously loaded climate date for the city of Hamburg. In order to compute the heat losses we first have to compute the specific total heat loss H (see equation 7.9) compose of: Ventilation losses Hv define in equation 7.5, a correction factor for thermal bridges Hwb define in equation 7.6 and transmission losses Ht define in equation 7.7.

Ventilation losses Hv

$$Hv = AirCRate \times V \times plCpl; \tag{7.5}$$

Where: AirCRate	air change rate $[h-1]$	
V	volume of air in heated building	$= 0.8 \times Ve$
Ve	building volume	
plCpl	heat storage capacity of air	$= 0.34[Wh/(m^3K)]$

The ventilation losses depend directly on the air exchange rate $AirCRate$. This value varies according to the defined family class. The ventilation losses are computed only taking into account natural ventilation and therefor no heat recovery measures are taken into account. This simplification is needed as the aim of this computation is to compare plCpl heat storage capacity of air $= 0.34[Wh/(m^3K)]$ An extension of this computation will have to take place in order to incorporate such a technology from special interest in low energy consumption buildings.

Correction value for thermal bridges Hwb

$$Hwb = Uwb \times A \tag{7.6}$$

Where: Uwb	correction value for thermal bridges $[W/m^2K]$
A	total heat transmitting building envelope $[m^2]$

The correction for thermal bridges is computed with help of a correction factor define in DIN V 18559. This factor is multiplied by the heat transmitting envelope of the building, this fact makes the building very reactive to the well-known A/V relationship influencing heat consumption.

Transmission losses Ht

$$Ht = \sum_{i=1}^{n} \left(U_{(i)} \times A_{(i)} \right) \tag{7.7}$$

Where: n	Number of building components encounter with ambient air
$U_{(i)}$	Heat transfer coefficient in countercurrent with ambient air $[W/(m^2K)]$
$A_{(i)}$	Analogical to building part surface $[m^2]$

For every building envelope component (walls, roof, slabs) the transmission losses are computed. The heat transfer coefficient for the building components are defined trough the building typologies.

Specific total heat loss H

$$H = Hv + Ht + Hwb \tag{7.8}$$

7.1.3 Monthly Heat Demand Qh

$$Qh_{(m)} = Ql_{(m)} - \eta \times Qg_{(m)} \tag{7.9}$$

Where: m	Month of the year.
η	Factor for heat gains
Qg	Heat gains
Ql	Heat losses

The heat supply is computed using the previously computed heat gains and heat losses. The heat demand is the heat needed to maintain the operative temperature and cover the heat losses. A fraction of all the computed heat gains are subtracted from the heat losses, this fraction is the usable share of the total heat gains. The fraction is computed with help of the *eta* (η) Factor.

Factor η

The factor η is computed as:

$$\eta = \begin{cases} a_{(m)} \div \left(a_{(m)} + 1\right) & \text{for } y = 1 \\ \left(1 - y_{(m)}^{a_{(m)}}\right) \div \left(1 - y_{(m)}^{a_{(m)}+1}\right) & \text{for } y \neq 1 \end{cases} \tag{7.10}$$

Where: m	Month of the year.
a	Numerical parameter considering thermal inertia of building.
y	heat-gain-loss relation.

Numerical parameter $a_{(m)}$

$$a_{(m)} = 1 + StorageCapacity \times Ve \div H_{(m)} \div 16 \quad (7.11)$$

Where: m	Month of the year.
H	Specific total heat loss (see equation 7.8)
Ve	Heated volume of building
a$_{(m)} = 1 + StorageCapacity \times Ve \div H_{(m)} \div 16$	

Heat-gain-loss relation.

$$y_{(m)} = Qg_{(m)} \div Ql_{(m)}$$

StorageCapacity$Storage capacity of building$

Where: Qg	Heat gains (see equation 7.1)
Ql	Heat losses (see equation 7.4)

7.1.4 Specific Heat Demand Qhs $[kWh/m^2a]$ and Specific Transmission Losses Hts $[W/m^2K]$

$$Qhs = Qh/An \quad (7.12)$$

$$Hts = Ht/An \quad (7.12)$$

7.1.5 Final Heat Demand Q_{eh} $[kWh/a]$ and Primary Heat Demand Q_{ph} $[kWh/a]$

$$Q_{ph} = eP \times Q_h \quad (7.12)$$

$$Q_{eh} = Q_{ph}/fP \quad (7.12)$$

7.1.6 Specific Final Heat Demand Q_{ehs} $[kWh/m^2a]$ and Specific Primary Heat Demand Q_{phs} $[kWh/m^2a]$

$$Q_{phs} = Q_{ph}/A_n \tag{7.12}$$

$$Q_{ehs} = Q_{eh}/A_n \tag{7.12}$$

7.2 Creating a Synthetic Building Stock

The development of a synthetic city representing the urban fabric of urban agglomerations has proved helpful for the development of urban models, the focus of this endeavor has been the use of remote sensing data or other type of image and laser data available at a low aggregation level in order to generate urban structures. Laycock und Day (2003) presents an overview of used data sources and methods for the generation of urban structures. Parish und Müller (2001) used a procedural approach based on L-systems to generate urban structures. The authors use different maps as input for the generation of the geometrical representation of the city. Our aim is to expand these methods by creating appropriate data regarding the building stock and the population living and working on this building stock that can be used as input to these models for the generation of synthetic cities. Kang, Ma, Tong und Liu (2012) create a set of synthetic cities in order to assess the relationship between urban morphologies and human mobility. Kii, Akimoto und Doi (2014) develop a simple synthetic city for the assessment of urban transport policies, the authors generate the synthetic city and use this data as input for a land use model, subsequently the authors apply the postulated policies to the model in order to assess them. The authors argue in favor of including user behavior in urban models. Farber, Neutens, Miller und Li (2013) generate eighty different synthetic cities in order to analyze the Social Interaction Potential of these environments based on commuting patterns and land use distribution. Bagchi, Sprintson und Singh (2013) use a synthetic city for the simulation of fire dispersion on an electrical distribution grid, in this case the authors represent the building stock and an electrical grid. Mei et al. (2015) develop a synthetic city in order to study the diffusion of infectious diseases. The authors use the synthetic city in order to understand the outbreak of influenza in dense populated urban areas in China. In this case the authors do not require a detail description of the building stock geometry but only the building use. Stötzer, Hauer, Richter und Styczynski (2015) define a simple synthetic city for the estimation of the potential load shift of commercial and residential electricity demand. Because the authors are only interested in electric consumption, they do not create a geometrical representation of the building stock not do they have any type of geo-reference in the synthetic city. In this chapter the authors define the electric consumption as a function of household size. Many of the examples presented above could be implemented on a more realistic environment describing the characteristics of the building stock and the population living on them.

The development of a robust method for the creation of synthetic cities representing specific urban agglomerations with known population aggregates constitutes the scope of this chapter. This chapter presents first results from this endeavor. It is shown the developed method for the representation of a synthetic city, describing individuals and the characteristics of their households and the building they reside on. Here we do not discuss in detail the pursued method for the representation of geo-referenced geometrical objects. We present some discussion regarding the different alternatives to generate a geometrical representation of the synthetic city extending the methods described in this chapter.

The method used in this analysis is a spatial microsimulation method. Microsimulation, introduced by Orcutt (1957) is a commonly used method among social scientist used to simulate a large range of social phenomena at a micro-level. The first step of this method is normally the generation of a synthetic population representing the population under analysis. The spatial microsimulation methodology extends this concept by allocating estimated synthetic populations to geographical areas (Clarke & Holm, 1987). This simulation method is applied by many disciplines. Brown und Harding (2002) and Smith, Pearce und Harland (2011) argue the use of this type of models for the analysis of health systems, the modeling of resources consumption is addressed by Williamson et al. (1996, 2002) for the estimation of water consumption and Chingcuanco und Miller (2012) and Muñoz H. und Peters (2014b) for the estimation of energy demand. Many transport models use this approach for the generation of synthetic populations (Farooq et al., 2013). For overview of spatial microsimulation models, its applications and methods see (Tanton, 2014; O'Donoghue et al., 2014). In this chapter we make use of the GREGWT algorithm to create a synthetic population. We use the available R library (Muñoz H., Vidyattama & Tanton, 2015a), implementing the GREGWT algorithm, originally developed by the Australian Bureau of Statistics (ABS) (Bell, 2000). This algorithm is used by the National Center for Social and Economic Modeling (NATSEM) on their spatial microsimulation model spatialMSM (Tanton, 2007).

The presented chapter is structured as follows: Section 3.3 presents and describes the used data for the analysis and the undertaken steps in preparing the data for the simulation, the next section, Section 7.3, describes the computation method to estimate heat demand for each individual on the micro census and the assumptions made for this computation. Section 7.4 describes the method used for the reweighting of the micro census with the computed heat demand values. We benchmark the survey to benchmarks describing individuals, dwelling units and buildings. This process is described under Section 7.5. The results from the simulation are described and discussed under Section 7.6. On Section 7.7 we highlight the benefits and shortcomings of the developed method and propose extensions to this method in order to address some shortcomings of the method.

7.3 Heat Demand

In order to estimate heat demand we estimate it for each individual in the micro census. For the estimation of heat demand of each individual we need to consider the characteristics of the building stock. In order to take these characteristics into account we make use of a building typology (Diefenbach, Cischinsky, Rodenfels & Clausnitzer, 2010; Loga et al., 2011). We classify each individual to one of the 36 types. When aggregating these values we need to take into account the household size and divide the estimated heat demand by this value. This is necessary because the computed heat demand is the estimated heat demand per dwelling unit squared meter. An alternative to this would be to divide the dwelling unit area by household size.

In order to classify the micro census into the building types we use three parameters from the micro census: (1) building construction year, (2) dwelling unit size, and (3) number of dwelling units per building.

The main parameter used for the classification is the building construction year. We use the number of dwelling units to differentiate between single-family and multi-family houses. Finally, the dwelling unit size multiply by the number of dwelling units is used to distinguish between small multi-family housed, large multi-family houses and high rise buildings. The specific heat values used in this analysis are listed in Table 7.1.

Table 7.1: IWU-de building typology matrix for Germany

	< 1859	1860–1918	1919–1948	1949–1957	1958–1968	1969–1978	1979–1983	1984–1994	1995–2001	2002–2009
EFH[a]	183	180	164	181	146	155	118	132	110	88
RH		153	137	156	106	127	127	98	78	86
KMH	190	143	168	156	129	134	118	122	92	79
GMH		127	144	142	131	117				
HH					114	113				

source: (Loga et al., 2011) (a) Specific Heat demand (spez. Wärmebedarfskennzahl) $[kWh/m^2a]$
(EFH) Single family house “Einfamilienhaus”; (RH) Terrace house “Reihenhaus”; (KMH) Apartment house “Mehrfamilienhaus”; (GMH) Large apartment house “Großes Mehrfamilienhaus”; (HH) High-rise “Hochhaus”;

The advantage of using a building typology for the estimation of heat demand is that we don’t need to take assumptions about the building geometry. This is because the numbers listed under the building typology represent specific heat demand values (per square meter). In order to compute the absolute heat demand we simply multiply this value by the building size, in square meters. In order to explicitly account for building geometry we need to allocate individuals to a digital cadastre describing the building properties, see (Muñoz H. & Peters, 2014b).

The disadvantages of using the digital cadastre for the computation of heat demand are: (1) the digital cadastre of other type of building information data are not as homogeneous as demographic data and building typologies, there are many building typologies available for Europe (Caputo et al., 2013; Hrabovszky-Horváth et al., 2013; Kragh & Wittchen, 2013; M. K. Singh et al., 2013; Dall’O’ et al., 2012; Dascalaki et al., 2011; Balaras et al., 2007), as is demographic data. (2) the complexity of data representing the building stock makes it difficult to make projections into the future, Muñoz H., Vidyattama und Tanton (2015b) present an application of a synthetic building stock projected into the future.

7.4 Simulation: Using GREGWT to Reweight the Micro-census

For this analysis we use an implementation of the GREGWT method in the R language (Muñoz H., Vidyattama & Tanton, 2015a). The GREGWT method is used in the SpatialMSM model of the

National Center for Social and Economic Modeling (NATSEM) (Tanton, 2007). This method was developed by the Australian Bureau of Statistics (ABS) (Bell, 2000). The aim of the GREGWT is to reweight a survey implementing method number 5 from A. Singh und Mohl (1996). Tanton et al. (2011) makes a detail description of the algorithm and its applications. The mathematical description of the GREGWT algorithm presented below is taken from Rahman et al. (2010) and the algorithm description from Muñoz H., Tanton und Vidattama (2015).

Aim of the GREGWT algorithm is to find a set of new weights w that can be used to match a survey X to a set of given benchmarks T so that $T = \sum w_j X_j$ (e.g. small area aggregates) by minimizing the weight difference between these new weights w and the sample design weights d from the survey. For the distance D between design and estimated weights the GREGWT algorithm makes used of the truncated Chi-Squared distance function, represented in Equation 7.4.

$$D = \frac{1}{2} \sum_j \frac{(w_j - d_j)^2}{d_j} \tag{7.12}$$

The equation needed to minimize the weight distance constraint to some given marginal totals of a geographical area (T) can be expressed as the Lagrangian function of the Chi-Squared function, as follows:

$$L = \frac{1}{2} \sum_j \frac{(w_j - d_j)^2}{d_j} + \sum_k \lambda_k \left(T_k - \sum_j w_{j,k} X_{j,k} \right) \tag{7.12}$$

By differentiating (7.4) with respect to w_j and applying the first order condition, we have:

$$\frac{\delta L}{\delta w_j} = \left(\frac{w_j - d_j}{d_j} \right) - \sum_j \lambda_j X_j = 0 \tag{7.12}$$

With this equation we can formulate an equation for the new weights. Where $X_j^{'} = \sum \lambda_k X_{j,k}$.

$$w_j = d_j + d_j X_j^{'} \tag{7.12}$$

The new weights computed by the GREGWT algorithm are float values. Without any restrictions the algorithm will produce negative weights, both implementations of the algorithm introduce boundaries constrains as user input. The user can define an upper and lower bound, if the algorithm computes weights outside these bounds the weights will be truncated to the corresponding bounds. In this case the algorithm will iterate with the new computed weights until a predefine convergence parameter is met or there is no improvement in the iteration.

The implementation of the GREGWT algorithm in the R language add an extra calibration method for the new estimated weights. This last process makes sure that the sum of the new weights is equal to the total population of the specific area. Equation 7.4 shows the alignment used to calibrate the resulting weights.

$$wo_i = \frac{p \times w_i}{\sum w_i} \tag{7.12}$$

Where wo are the new calibrated weights, p is the population total and w are the computed new weights from the GREGWT algorithm.

7.5 Benchmarking to Different Aggregation Units

The defined benchmarks for the reweighting of the survey are aggregated by different units: (1) Individuals, (2) Families, and (3) Buildings (see Table 3.4 and Section 3.3). The used R implementation of GREGWT algorithm is able to perform integrated reweights. This is important for maintaining unit aggregations, e.g. to maintain the family structure given by the survey. Nonetheless, this does not allow us to benchmark to different aggregation units. Attempts to address this issue exist in the literature, Guo und Bhat (2007) benchmark an initial survey to household characteristics and fit the result to individual benchmarks in the integerization of the weights via a Monte Carlo process. This method allows to "benchmark" to both aggregation units. The disadvantage of this problem relies on the integerization process, this process is only required for the construction of agents at the cost of a decline of the algorithm precision and performance (computational time). For the presented simulation we do not require individual agents and therefor any form of integerization would be contra productive for the final result. For a description of other similar approaches see Pritchard und Miller (2012) and Ma und Srinivasan (2015). Pritchard und Miller (2012) propose a "Conditional Monte Carlo Synthesis Procedure" to fit the synthetic population to both: household and individual benchmarks. Again this procedure assumes an integerization of the survey. As describe above we aim to develop a method that does not require an integerization step in order to avoid: (a) a decline in the performance of the algorithm, (b) an explosion of computational time, and (c) introduction of a stochastic element to a deterministic approach. Ma und Srinivasan (2015) developed another method to create a synthetic population: "fitness-based synthesis" (FBS). The method presented by Ma.2015 proposed the computation of two fitness values expressing the adding and subtracting probability of individuals from the random selected population from the reweighted population survey.

For our simulation we make use of a simpler method. Because we don't need an integerized survey, we reweight the original survey to three benchmark groups (aggregation units): (1) individuals, (2) families and (3) buildings. For each of these groups there is available data at a NUTS 3 level.

The GREGWT algorithm need to transform the input matrix X (micro census) to a binary matrix, each 1 on the matrix corresponds to an individual. This works well if we benchmark the survey to aggregates counting individuals (e.g. number of individuals on age category 18 to 20) but fails if we try to benchmark to an aggregate, counting dwelling units (e.g. number of dwelling units with floor space of 60 m^2). We need to control for this difference. With the available information of the input survey we can manipulate the binary matrix in order to make the benchmarking to different aggregation units possible. For variables benchmarked to dwelling units we divide them by household size, an individual with household size 3 will get a 1/3 instead of 1 in the X input matrix. For all variables counting buildings we divide the values by the household size and number of dwelling units, the same individual living in a building with 6 dwelling units will have a value of 1/3/6. This calibration is formalized in Equation 7.5 and 7.5, where HH is household size and DU is number of dwelling units.

$$X_i^{du} = X_i \div HH_i \tag{7.12}$$

$$X_i^{bu} = X_i \div HH_i \div DU_i \tag{7.12}$$

With this simple modification of the survey we are able to benchmark the input survey to three aggregation levels. In theory this method allows us to benchmark to any number of aggregation units. The used library implements three aggregation levels: (1) individuals, (2) households; and (3) buildings.

7.6 Results: German Heat Demand

In this section we present the main results from the performed spatial microsimulation. The results show the estimated heat demand for the German residential sector. First we present an internal validation of the spatial microsimulation model and the resulting heat demand. The results are aggregated back to the geographical areas in order to visualize them.

In order to internally validate the model we compare the results at the NUTS 3 level. We compare the output results with the benchmarks used in the reweighting process. For the comparison we make use of the Total Absolute Error TAE and the Percentage Absolute Error $PTAE$ as measures of the model internal error. The TAE is the absolute difference between the simulated $\hat{T}$ and observed T benchmarks, the $PTAE$ is an extension of the TAE measure. The $PTAE$ divides the computed TAE by the total population pop of the geographical area i. The mathematical expressions of both measures are expressed below.

$$TAE_i = \sum_i \left| T_i - \hat{T}_i \right| \tag{7.12}$$

$$PTAE_i = TAE_i \div pop_i \times 100 \tag{7.12}$$

The resulting $PTAE$ values show a very low miss-allocation of individuals at this aggregation level. There are only four areas with a $PTAE$ value higher than 0.2% and 52 areas with a value higher than 0.1%. Figure 7.1 shows: (a) the distribution of the $PTAE$ values for all simulation areas; and (b) a

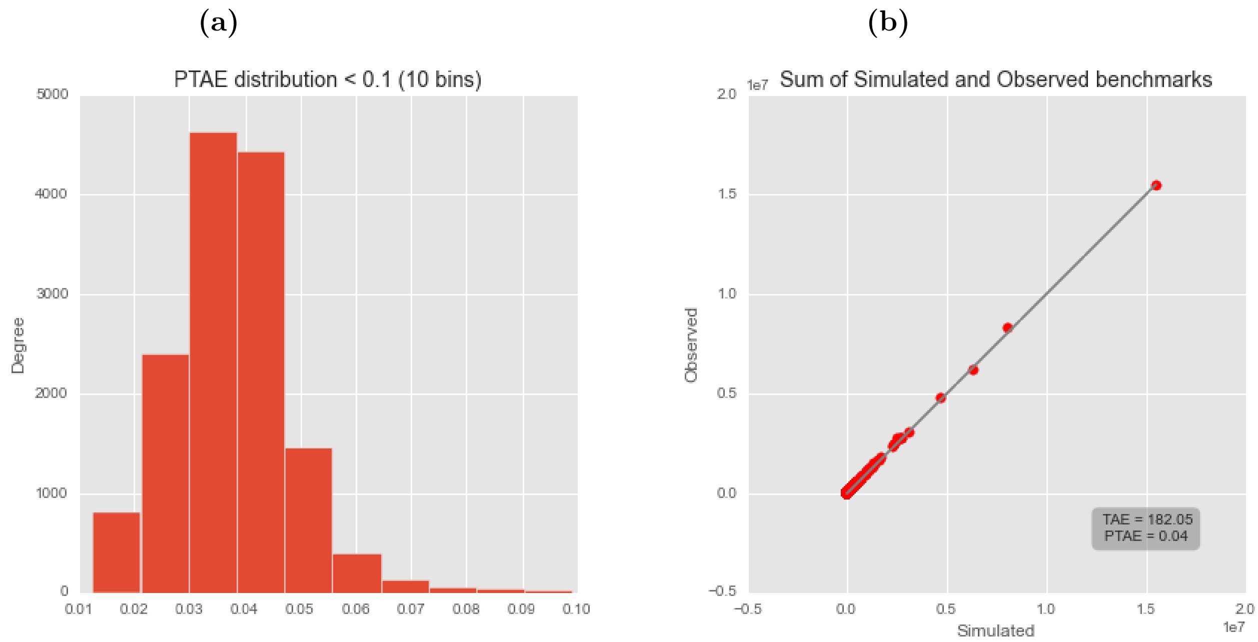


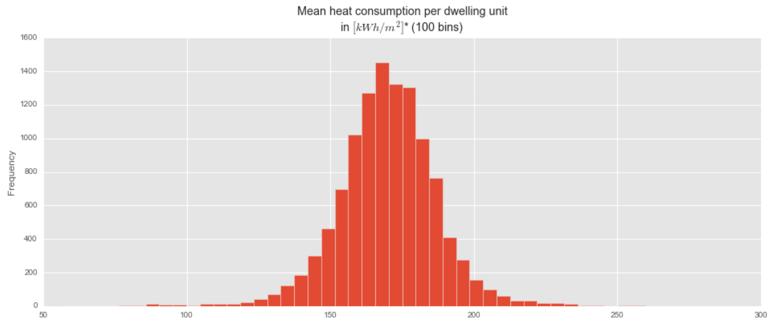
Figure 7.1: Distribution of $PTAE$ and difference between estimated and simulated population

scatter plot comparing the observed and simulated marginal sums.

The Internal validation of the model shows extremely good results, these good results do not reflect on the accuracy of the estimated heat demand. An external validation of the model is necessary in order to validate the estimation of heat demand. This chapter shows that the use of a spatial microsimulation model for the estimation of heat demand at low aggregation levels is possible, the underlying model used for the estimation of heat demand has not yet been validated, still the estimation is able to show plausible differences in heat demand between small geographical areas. This method can be applied at a city level, differentiating the heat demand at a neighborhood level.

An external validation of the model is not possible because available energy consumption data is not available at this disaggregation level. Available energy consumption data exist at a higher aggregation level. A problem with this data is that this data does not differentiate between consumption sectors neither between primary and end energy consumption. Our model simulates residential end heat demand. There is also a difference between heat demand and heat consumption, our model does not consider efficiency rates of heat supply infrastructure. Further steps are planned to make an external validation possible. The first step we envision towards a validation of estimated heat demand is the integration of the non-residential sector to the model. With an estimate of the non-residential sector we might be able to validate the sum of residential and non-residential heat demand at a higher aggregation level. We also plan to include efficiency rates of the underlying heat supply infrastructure as well as the energy carrier used to supply the heat.

In order to show the results in a more meaningful way we divide the estimated total heat demand by the number of dwelling units in each geographical area and by a constant (60) representing the



* for a 60 m^2 dwelling unit

Figure 7.2: Mean heat consumption per dwelling unit

average dwelling unit size in m^2 for Germany[2]. The resulting specific heat demand allows us to assess the performance of the model. The distribution of this value for all simulation areas is plotted on Figure 7.2, the values and distribution shown in this plot are consistent to known consumption values for the German building stock.

We present a figure showing the estimated heat demand at a NUTS-3 level. The first map (Figure 7.3) shows the distribution of specific heat demand for the entire country. The second map shows the detail of four specific areas of Germany. In addition to the estimated heat demand we plot urban agglomerations retrieved from Landsat images, we make use of the GADM dataset containing this information (Hijmans et al., 2014).

On the first map (see Figure 7.3) we can identify a differentiation between West- and East-Germany. The specific heat consumption is higher on East-Germany. The specific heat consumption on East-Germany is higher because the building stock on this part of Germany is older. This regional differentiation on heat demand shows that the presented model is sensitive to regional characteristics of the building stock. The model does not take other regional characteristics into account like temperature, solar radiation or heat supply infrastructure. There is a big difference between East and West regarding the heat supply infrastructure. The share of households connected to a district heating network in East-Germany is much higher than in West-Germany.

The second map (Figure 7.4), showing a detail of the distribution of heat demand for the two largest cities in Germany: (a) Berlin and its surroundings, (b) Hamburg and its surroundings and two urban agglomerations: (c) the south-east part of North Rhine-Westphalia “Nordrhein Westfalen” and (d) Thuringia “Thüringen”, a federal state of East-Germany. In the case of the German cities, Berlin and Hamburg, the official NUTS 3 level corresponds is equivalent to the NUTS 2 and NUTS 1 level. Both areas have a disproportional large population size compare to all the other geographical areas. Available statistics at a lower aggregation level are available for both cities and a second reweighting algorithm could be perform for these two areas separately. In both cases we see that the areas corresponding to the cities have a higher specific heat demand and the peripheries show a lower specific heat demand,

[2] 71,5 m^2 (2013 west Germany) & 63.4 m^2 (2013 east Germany) from: Income and Consumption survey “Einkommens- und Verbrauchsstichprobe” (EVS) as quoted in: “Haushalte zur Miete und im Wohneigentum nach Anteilen und Wohnfläche in den Gebietsständen am 1.1.” destatis.de

we find an explanation for this differentiation in the construction year of the corresponding building stock. We expect to have an old building stock within the historical urban fabric of the cities, this part of the city will be well within the corresponding geographical area. Both cities have outgrowth the boundaries of these geographical areas. The development of urban settlements at the periphery of the city, constructed in more recent years, are constrained to newer building codes regulating the heat transmission losses of the building shell and therefore consuming less heat demand.

We see a similar effect on the large urban agglomeration of North Rhine-Westphalia (see Figure 7.4c), the high values of heat demand correspond to the original settlements in the region. These settlements have the oldest building stock in the region. Large parts of these settlements are under heritage protection, making a retrofit of the buildings extremely expensive and difficult. We can't see this differentiation on the East-German urban agglomeration in the Thuringia state (see Figure 7.3d). The urban agglomeration is not a big as in North Rhine-Westphalia and the urban and population growth hasn't been as big ans in other German regions. In the case of North Rhine-Westphalia we see the opposite phenomena in which more rural areas in the periphery have higher consumption values.

The results presented in this chapter show that an estimation of heat demand at this level of aggregation with relative little input data is possible. Future developments of the models aim to integrate projections of the heat demand consumption at the same area of aggregation. A theoretical background to perform a projection of heat demand with a spatial microsimulation model exist (Muñoz H., Vidyattama & Tanton, 2015b). Muñoz H., Vidyattama und Tanton (2015b) project the heat demand at a low aggregation level for the city of Hamburg using a spatial microsimulation model. Recent developments show an alternative method to project heat demand by simulating retrofits through the manipulation of the input weights of the micro census (Muñoz H., Tanton & Vidattama, 2015), this method present a clear advantage while simulating and projecting heat demand for the entire country at a low level of aggregation because we do not have to make assumptions for the projection of retrofit levels at a low level of aggregation. A possible implementation of this method is to align the initial weight distribution to proposed policies implemented at a federal level. In this scenario the aggregated retrofit rates are defined at a federal level and the allocation of these retrofits could be simulated as function of demographic characteristics at a lower geographical level.

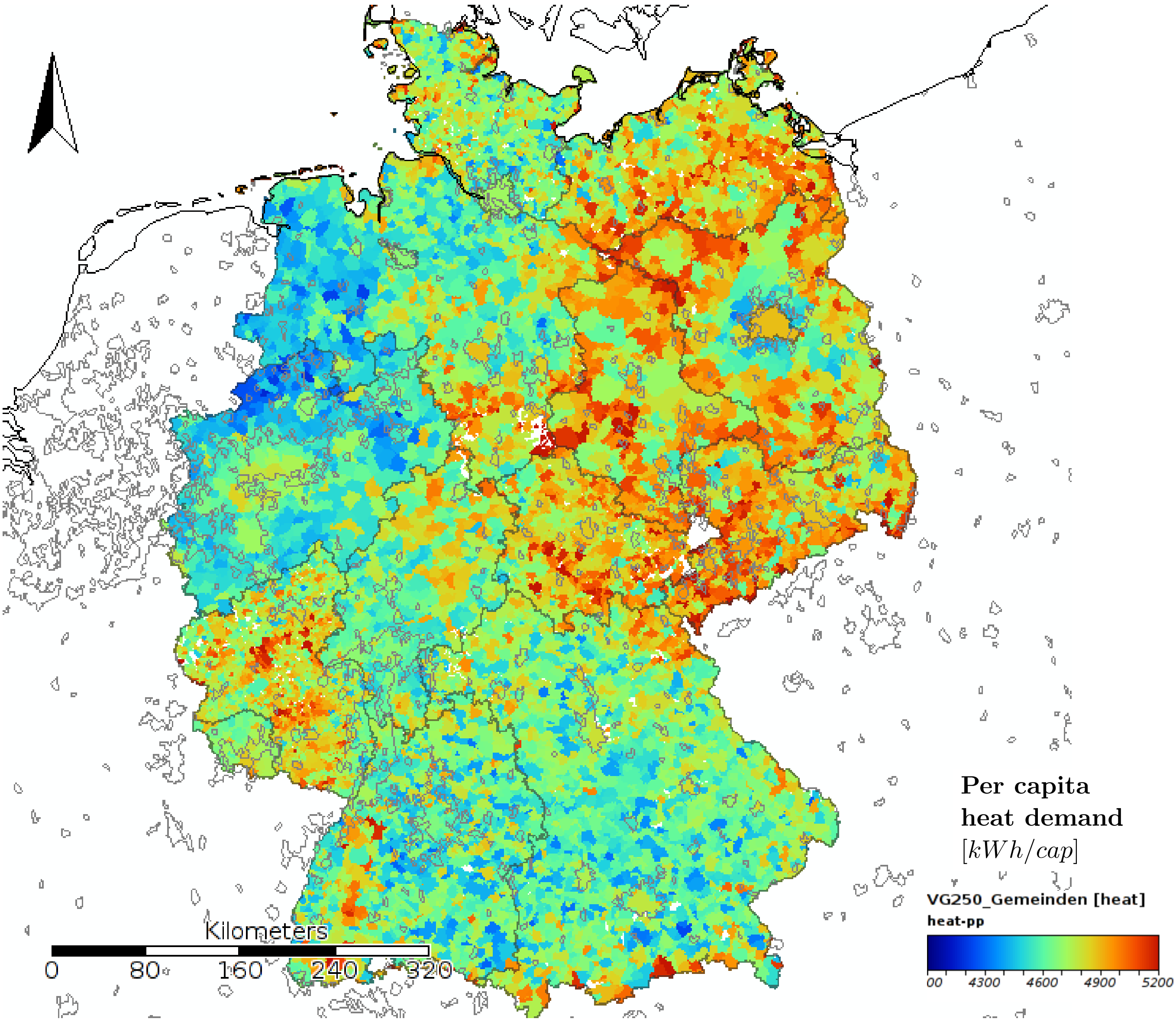


Figure 7.3: Simulated heat demand for Germany at a NUTS-3 level as per capita heat demand in kWh/cap

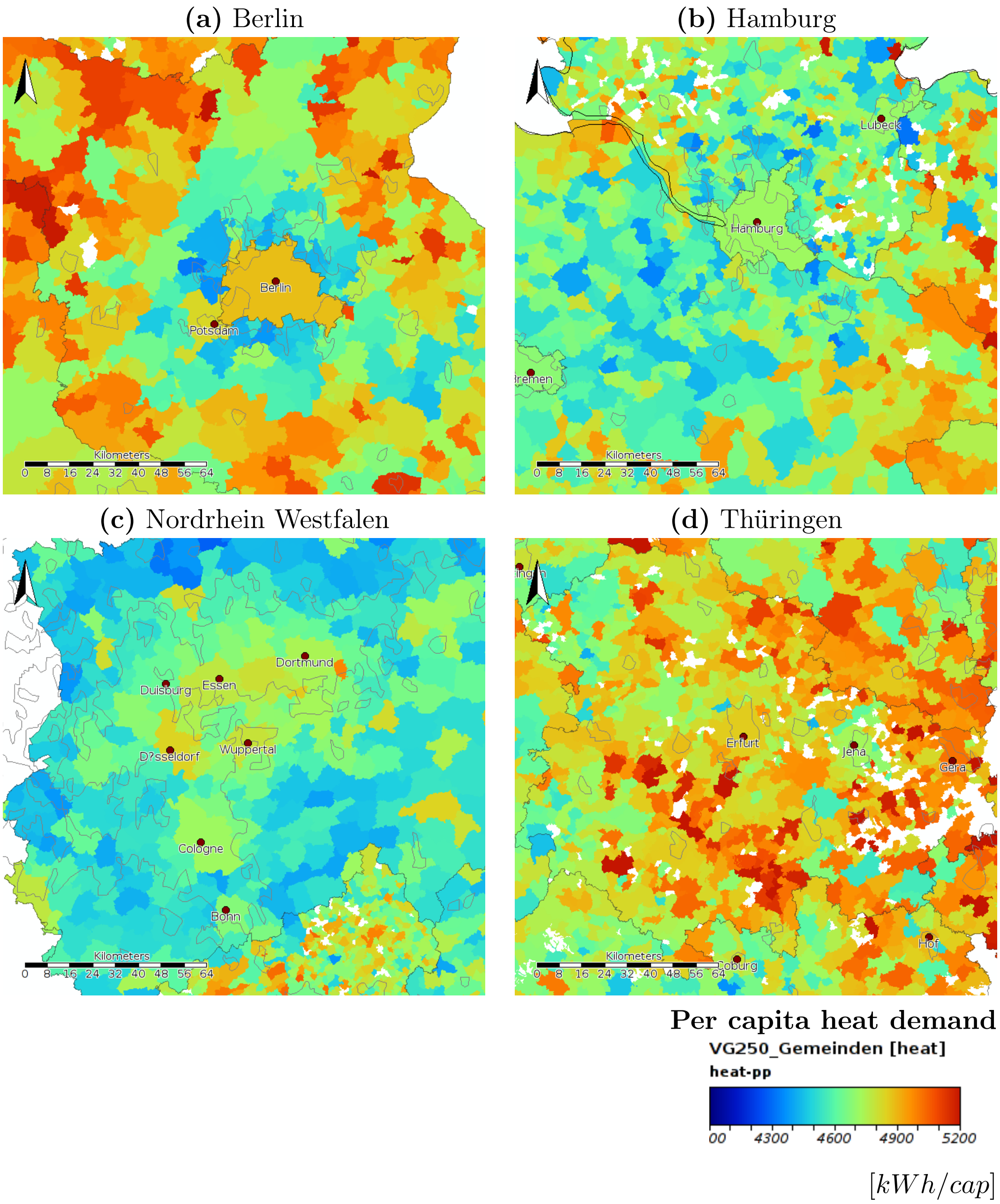


Figure 7.4: Simulated heat demand for selected regions of Germany

7.7 Next Steps: Integrated Reweighting

The next envisioned step is to group the synthetic individuals into: (a) households and (b) into buildings. The advantage of grouping individuals into single buildings are: (1) the ability to represent energy consumption at a building level and (2) the possibility to geo-reference these buildings at a finer spatial resolution.

One of the applications for this synthetic data is the dimensioning and planning of decentralized energy supply systems. In order to make this data usable for this application we need to represent energy demand at a finer spatial resolution. We may achieve this by obtaining aggregated benchmarks at a finer spatial resolution, this may not be possible due to data protection concerns. For this reason we want to develop new method to further disaggregate heat consumption values. For this disaggregation we see two possible paths, the first step for both paths is the grouping of individual data into buildings.

The first approach makes use of the digital cadastre. The problem of this method is that a digital cadastre may not always be able for the desire area, this is especially true for rural areas. The second approach makes use of satellite images for the computation of spatial probability distributions of relevant building parameters (e.g. density, construction year, etc.) and stochastically allocated buildings to build up areas based on the computed spatial probabilities. The advantage of this method is that satellites are available for the entire world, making the transferability of the method higher.

With a more efficient building stock in place the role of user-behavior will take a dominant place in the estimation of heat demand (Hong, D'Oca, Turner & Taylor-Lange, 2015). The presented method already solves the biggest problem of urban models aimed at the estimation of residential heat demand with an explicit consideration of human behavior, this is the allocation of families to the building stock. Muñoz H. und Peters (2014b) use a spatial microsimulation model to describe households and the building characteristics they reside on for the estimation of heat demand varying user related parameters on the model according to the demographics of the user. A more elaborated approach is presented by Muñoz H. (2014), in this paper the author enriches the German micro census with a time-use survey for the generation of household schedules, these schedules are use as input in a thermal simulation model.

8 Conclusions and Outlook

In this section we discuss the use of building typologies at a European and global level as well a short discussion on the use of building components as an alternative to building typologies. We argue the need to work on method rather than on specific national building typologies with focus on available data of individual countries. We also make a small excursion to the use of urban typologies for the estimation of heat demand and highlight the benefits and problems of using urban typologies rather that building typologies. We highlight the need to include the non-residential sector in this type of analysis.

A reflection on the use of time-use data for the description of human activity on urban environments is presented. We argue in favor of an activity-based model for the simulation of urban processes. The generation of rich micro data-sets can be used for many type of urban simulations.

We present a brief discussion on the use of demographic data for the projection of urban environments. The decision of which building is to be retrofitted within the projected model is based on this thesis only on building characteristics. Such a decision should be simulated based not only on characteristics of the building stock but on the characteristics of the population living on them.

An important contribution of this thesis to the spatial microsimulation community is the performed analysis of the initial weights of the sample survey. We argue that the manipulation of the initial survey sample weights can be used as a modeling tool. The projection of the building stock within a spatial microsimulation model needs to take assumptions at a small area level. Projected data at this level of aggregation is normally not available and assumptions are hard to make. We see the manipulation of initial weights as an alternative to this method for the projection of synthetic populations into the future.

8.1 The Need to Include Non Residential Buildings in the Analysis Scope

Many of the developed typologies have extended their approach to the non-residential sector (Loga et al., 2011; Hermelink et al., 2011). An estimation of the industrial sector will differ too much from the typology approach in order to be include into this methodology. A different approach is presented by (Blesl et al., 2007), in which the authors attempt to classify the non-residential sector by the required temperature of the processes within the building. An automatic classification of buildings by the driving needed temperature may be possible thanks to the comprehensive classification of building use in the digital cadastre (231 classes). A parallel comprehensive classification of building uses and their corresponding energy demand was developed by (Zeine, Gebhardt, Bockting, Mantai & Wei, 2007). The connection of these both data sets may be an interesting option for the estimation of nonresidential buildings.

Another interesting alternative, especially for an urban setting without a comprehensive digital cadastre, can be the use of urban typologies. An example for such a typology is the one developed by Hegger et al. (2014), which divides the urban territory into areas with specific heat demand and potential for renewable energy sources. The heat demand is based upon the character of the urban fabric, for example predominantly single family houses, terrace houses, prefabricated blocks etc. The potential for renewable energy is derived again by the character of the areas with the amount of non-built up areas being a signal for the potential of geothermal and biomass energy and the typical roof form (slope, aspect), for the given area, –a signal for solar energy potential. In this typology, however, the heat demand of the non-residential buildings is taken down to the building level, being analyzed as "single-elements" rather than urban areas (with one urban area exception, in which the demand of the nonresidential sector is calculated per area of the urban type). Since almost the entire heat demand of these urban areas does come from the buildings (and the nonresidential sector is actually tackled at the building level), it becomes arguable if the switching from buildings to areas presents a benefit to the analysis. It can be argued that it is useful for areas without a digital cadastre, however the lack of such a cadastre may render the assignment of the urban areas into types impossible or at least with low precision. Therefore, simply methodologically, estimating the heat demand of urban areas rather than buildings may not bring benefits precisely because areas do not demand any heat, apart from the buildings on them (there are marginal exceptions, street lightning demands electrical power for example). Looking at the supply of heat however, the urban area typology approach presents a more solid case. Generally, the supply of a building is external to it, thus exploring the areas outside of the buildings as well, makes more sense. Therefore, if the aim of the analysis relies not on the single buildings but on entire urban areas taking into account not only the single heat demand of the buildings but the underlying infrastructure supporting the area, a more holistic approach, using areas rather than buildings, may be useful. Roth und Häubi (1980) developed an urban typology design for the dimension of district heating systems. In Germany and Switzerland, the use of urban typologies for the description of the building stock and the underlying infrastructure has a long tradition. For the use of urban typologies in relationship with: (1) heat supply, see (Roth & Häubi, 1980; Sieverts und Volwahsen, Roth & Volwahsen, 1980; Blesl, 2002; Jentsch et al., 2008); (2) for the estimation of the potential of solar power generation, see (Everding, 2004); (3) a general description of the potential for renewable energies in open spaces, see (Genske, Jödecke & Ruff, 2009); (4) transportation volume estimation and transportation patterns, see (Marconi, 2006; Krug, 2006); and (5) for a description of general infrastructure networks see (Buchert, 2004; Ecoplan & Schweiz. Bundesamt für Raumentwicklung., 2000; Einig, 2006; Erhorn-Kluttig, 2011).

8.2 Next Steps: Expanding the Method

Having access to a detailed digital cadaster of the city offers a vast set of possibilities for the simulation of heat demand of urban areas. A simple balance of heat demand can be performed without much effort and with a relative low input of data. Much of the needed data for an energy balance of a single building can be recover through the digital cadaster. The missing data for such a computation are the U-values of the individual building components. The implemented building typologies are not constructed for this purpose. There is an alternative to the use of building typologies, this is the use of a building component typology. For Germany there is a regional material catalog developed by the Center for Environmental Friendly Construction[1] (Klauß, Kirchhof & Gissel, 2009). With help of this catalog we can simulate heat consumption based on building components rather than on building typologies. First attempts have been performed (Muñoz H., 2016). The simulation at such level of detail may foster further insights on different topics needed for a holistic understanding of urban systems. This method could be applied for the simulation of material flows in urban systems as well as an estimation of retrofit cost and the construction and demolition waste arising from such retrofits.

8.3 Extending the Use of Occupation Schedules

In the presented approach the generated occupation schedule distinguish only between workdays and weekends, this method can be expanded to distinguish occupational patterns between seasons. This method can also be expanded to other type of simulation models. The simulation of electricity demand can directly profit from this approach. The available data in the time-use survey delivers not only information about the location of single individuals over time but describes the specific activity as well as appliances used in the household. We expect that this type of information, in combination with a spatial microsimulation (that is the selection of representative individuals from the micro census for specific urban areas) make a contribution towards the reduction of the variation between simulated heat demand and monitored heat consumption. When occupants have control over the system, they can adapt, and this can lead to a significant reduction of energy (Bourgeois, 2005; Toftum, Andersen & Jensen, 2009; Fabi, Andersen, Corgnati & Olesen, 2012), especially on new or retrofitted buildings where heat transmission losses through the wall and through ventilation are rather low. We need models able to capture this influence for the proper design of future district heating supply systems.

8.4 Expanding the Use of Demographic Data Onto Urban Models

This thesis shows the application of a microsimulation model used for the estimation of heat consumption with an explicit consideration of user influence. We project this estimation into the future considering: (a) the aging population of the city of Hamburg, Germany and (b) ambitious scenarios to retrofit a substantial share of the buildings stock. The estimate indicates that the biggest changes could be concentrated in the city center where most of the old buildings are located. Nevertheless, this estimation can still be improved by integrating further characteristics of both the population and the building stock into the model. For future analysis we want to take other demographic parameters to sample the building stock and include new building typologies in order to simulate retrofit cycles.

Further applications of this method may integrate more elaborated user behavior models developed by the building simulation community. The developed model architecture also can accommodate the use of weather files projected under different climate change scenarios. A dataset describing the building stock and the individuals living on it can be used for many analyses. The impact of fluctuating temperatures has a health consequence, especially on the elderly population (Shi, Kloog, Zanobetti, Liu

[1](Zentrum für Umweltbewusstes Bauen e.V.) `http://www.zub-kassel.de/`

& Schwartz, 2015). Models projecting this type of effects on the population need to take the build environment and possible developments of the build environment into account.

The presented method shows a model to project the building stock and the population living on it. The developed scenarios can growth in complexity as the model matures. We aim to include more parameters (and the corresponding algorithms) driving: (a) the selection of buildings to be retrofitted and (b) the interaction between buildings and residents.

8.5 Using the Initial Weights of Population as a Modeling Tool

The presented thesis makes a comparison of two well established methods for the reweighting of population surveys to geographical areas. This comparison is based on the results of three reweighting processes using different implementations of the algorithms: (1) an implementation in the R language of the GREGWT algorithm, developed by author of this thesis (Muñoz H., Vidyattama & Tanton, 2015a); (2) an implementation in the SAS language of the GREGWT algorithm (Bell, 2000), developed by the Australian Bureau of Statistics; and (3) an implementation into the R language of the IPF algorithm (Blocker, 2013). The results from these three implementations are compared in terms of the weight distance and in terms of total absolute error (TAE). Results from the reweighting process show that all three implementations deliver robust results, the resulting weights from the GREGWT and IPF implementations are very similar. The resulting weights from both GREGWT implementations are almost identical, we attribute the difference in weights between the GREGWT implementations to the internal architecture of both implementations and the selected reference categories used on the two implementations.

The achieved percentage specific absolute error ($PSAE$) is extremely low for all three implementations, the highest $PSAE$ value is for geographical area 4011 reweighted with the GREGWT SAS implementations. These values are still extremely low with a 0.26%. We attribute this larger error to the selection of reference categories within the SAS implementation.

In this thesis we also discuss the role of initial weights and discuss the advantages of having a reweighting algorithm sensitive to a modification of the initial weights. The manipulation of the initial weights of the input population survey offers an extra tool to the user for the design of simulation models. In this thesis we discuss the use of this tool within static projections, the manipulation of initial weights can be used to make projections of the population without the need to project the desired benchmark at a low aggregation level. Form a model design a projection at a low level aggregation might be difficult or undesired, in these cases a manipulation of the weight distribution might be a better solution. It is important to mention that a manipulation of the initial weights will only be effective if the reweight is not performed on the same variable used to manipulate the weight distribution.

Bibliography

Anderson, B. (2013). Estimating small-area income deprivation: An iterative proportional fitting approach. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6, S. 49–67). Springer.

Anderson, J. G. & Kushman, J. E. (1987). A model of household heating demand: Home production with satiety and an endowment. *Journal of consumer affairs*, *21* (1), 1–20. doi: 10.1111/j.1745-6606.1987.tb00185.x

Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland (AdV). (2006). *Dokumentation zur modellierung der geoinformationen des amtlichen vermessungswesens (geoinfodok): Kapitel 7 alkis-katalogwerk. abschnitt 7.1 alkis-objektartenkatalog: Version 5.1.*

Bagchi, A., Sprintson, A. & Singh, C. (2013). Modeling the impact of fire spread on an electrical distribution network. *Electric Power Systems Research*, *100*, 15 - 24. doi: 10.1016/j.epsr.2013.01.009

Baker, P., Blundell, R. & Micklewright, J. (1989). Modelling household energy expenditures using micro-data. *The Economic Journal*, *99*, 720–738. doi: 10.2307/2233767

Balaras, C. A., Gaglia, A. G., Georgopoulou, E., Mirasgedis, S., Sarafidis, Y. & Lalas, D. P. (2007). European residential buildings and empirical assessment of the hellenic building stock, energy consumption, emissions and potential energy savings. *Building and Environment*, *42* (3), 1298–1314. doi: 10.1016/j.buildenv.2005.11.001

Ballas, D., Clarke, G., Dorling, D. & Rossiter, D. (2007). Using simbritain to model the geographical impact of national government policies. *Geographical Analysis*, *39* (1), 44–77. doi: 10.1111/j.1538-4632.2006.00695.x

Ballas, D., Clarke, G. & Turton, I. (1999). Exploring microsimulation methodologies for the estimation of household attributes. In *4th international conference on geocomputation.* Mary Washington College and Virginia and USA and.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G. & Dorling, D. (2005). *Geography matters: Simulating the local impacts of national social policies.* University of Leeds. Zugriff auf `http://www.dannydorling.org/wp-content/files/dannydorling_publication_id1233.pdf`

Battiston, S., Farmer, J. D., Flache, A., Garlaschelli, D., Haldane, A. G., Heesterbeek, H., ... Scheffer, M. (2016). Complexity theory and financial regulation. *Science*, *351* (6275), 818–819. doi: 10.1126/science.aad0299

Beckman, R. J., Baggerly, K. A. & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30* (6), 415 - 429. doi: http://dx.doi.org/10.1016/0965-8564(96)00004-3

Bell, P. (2000). GREGWT and TABLE macros — User guide [Software-Handbuch]. Australian Bureau of Statistics (ABS), Canberra.

Birkin, M. & Clarke, M. (1988). Synthesis – a synthetic spatial information system for urban and regional analysis: methods and examples. *Environment and Planning A*, *20* (12), 1645–1671. doi: 10.1068/a201645

Birkin, M. & Clarke, M. (1989). The generation of individual and household incomes at the small area level using synthesis. *Regional Studies*, *23* (6), 535-548. doi: 10.1080/00343408912331345702

Birkin, M. & Wu, B. (2012). A review of microsimulation and hybrid agent-based approaches. In A. J. Heppenstall, M. Batty, A. T. Crooks & L. M. See (Hrsg.), *Agent-based models of geographical systems.* Dordrecht: Springer Netherlands.

Blalock, H. M. (1979). *Social statistics* (Rev. 2d ed Aufl.). New York: McGraw-Hill.

Blesl, M. (2002). *Räumlich hoch aufgelöste Modellierung leitungsgebundener Energieversorgungssysteme zur Deckung des Niedertemperaturwärmebedarfs* (Dissertation, Univ, Stuttgart and Stuttgart and Stuttgart). Zugriff auf `http://elib.uni-stuttgart.de/opus/volltexte/2002/1193/pdf/FB92.pdf`

Blesl, M., Kempe, S., Ohl, M. & Fahl, U. (2007). *Wärmeatlas Baden-Wüttemberg: Erstellung eines Leitfadens und Umsetzung für Modellregionen: Institut für Energiewirtschaft und Rationelle Energieanwendung. Universität Stuttgart.* Zugriff auf `http://elib.uni-stuttgart.de/opus/volltexte/2009/4840/`

Blocker, A. W. (2013). ipfp: Fast implementation of the iterative proportional fitting procedure in c [Software-Handbuch]. CRAN. Zugriff auf `https://cran.r-project.org/web/packages/ipfp/`

BMVBS Bundesministerium für Verkehr, Bau und Stadtentwicklung. (2008). *Energieausweis für gebäude – nach energieeinsparverordnung (enev 2009).* Berlin.

Bohi, D. R. (op. 1981). *Analyzing demand behavior: A study of energy elasticities.* Baltimore: Johns Hopkins University Press.

Boonekamp, P. G. (2007). Price elasticities, policy measures and actual developments in household energy consumption – a bottom up analysis for the netherlands. *Energy Economics*, *29*, 133–157. doi: 10.1016/j.eneco.2005.09.010

Borgeson, S. & Brager, G. (2008). *Occupant control of windows: Accounting for human behavior in building simulation.* Zugriff am 16.08.2012 auf `http://www.cbe.berkeley.edu/research/pdf_files/Borgeson2008-OperableWindowSimulation.pdf`

Born, R., Diefenbach, N. & Loga, T. (2003). *Energieeinsparung durch verbesserung des wärmeschutzes und modernisierung der heizungsanlage für 31 musterhäuser der gebäudetypologie: Studie im auftrag des impulsprogramms hessen: Endbericht.* Darmstadt. Zugriff auf `http://www.iwu.de/fileadmin/user_upload/dateien/energie/klima_altbau/GebTyp_Impulsprogramm_Hessen_22_01_2003.pdf`

Bourgeois, D. J. (2005). *Detailed occupancy prediction, occupancy-sensing control and advanced behavioural modelling within whole-building energy simulation* (Unveröffentlichte Dissertation). Université Laval, Québec.

Brown, L. & Harding, A. (2002). Social modelling and public policy: Application of microsimulation modelling in australia. *Journal of Artificial Societies and Social Simulation*, *5* (4). Zugriff auf `http://jasss.soc.surrey.ac.uk/5/4/6.html`

BSU. (2011). *Hamburger gebäudetypologie: Verbrauchswerte und einsparpotentiale.* Zugriff am 07.05.2011 auf `http://klima.hamburg.de/energiediagnose/1998140/gebauedetypologie.html`

Buchert, M. (2004). *Nachhaltiges bauen und wohnen in deutschland: Stoffflussbezogene bausteine für ein nationales konzept der nachhaltigen entwicklung ; verknüpfung des bereiches bauen und wohnen mit dem komplementären bereich "Öffentliche infrastruktur" ; [endbericht] ; forschungsbericht 29892303/02* (Bde. 04,01). Berlin: Umweltbundesamt.

Burden, S. & Steel, D. (2015). Constraint choice for spatial microsimulation. *Population, Space and Place*, n/a–n/a. doi: 10.1002/psp.1942

Caputo, P., Costa, G. & Ferrari, S. (2013). A supporting method for defining energy strategies in the building sector at urban scale. *Energy Policy*, *55*, 261–270. doi: 10.1016/j.enpol.2012.12.006

Chingcuanco, F. & Miller, E. J. (2012). A microsimulation model of urban energy use: Modelling residential space heating demand in ilute. *Computers, Environment and Urban Systems*, *36* (2), 186–194. doi: 10.1016/j.compenvurbsys.2011.11.005

Clarke, M. & Holm, E. (1987). Microsimulation methods in spatial analysis and planning. *Geografiska Annaler. Series B, Human Geography*, *69* (2), pp. 145-164. doi: 10.2307/490448

Cuijpers, C. (1995). A joint model of household space heat production and consumption: Empirical evidence from a belgian micro-data survey. Washington D.C.: 18th IAEE International Conference.

Dall'O', G., Galante, A. & Torri, M. (2012). A methodology for the energy performance classification of residential building stock on an urban scale. *Energy and Buildings*, *48*, 211–219. doi: 10.1016/j.enbuild.2012.01.034

Dascalaki, E. G., Droutsa, K. G., Balaras, C. A. & Kontoyiannidis, S. (2011). Building typologies as a tool for assessing the energy performance of residential buildings – a case study for the hellenic building stock. *Energy and Buildings*, *43* (12), 3400–3409. doi: 10.1016/j.enbuild.2011.09.002

Deutsches Institut für Normung e. V. (2003a). *Energetische Bewertung heiz- und raumlufttechnischer Anlagen - Teil 10: Heizung, Trinkwassererwärmung, Lüftung* (Bd. ICS 91.140.10; 91.140.30; 91.140.65) (Nr. 4701-10). Berlin: Beuth.

Deutsches Institut für Normung e. V. (2003b). *Wärmeschutz und Energie-Einsparung in Gebäuden - Teil 6: Berechnung des Jahresheizwärme- und des Jahresheizenergiebedarfs* (Bd. ICS 91.120.10) (Nr. 4108-6). Berlin: Beuth.

Deutsches Institut für Normung e. V. (2011). *Din V 18599: Energetische Bewertung von Gebäuden – Berechnung des Nutz-, End- und Primärenergiebedarfs für Heizung, Kühlung, Lüftung, Trinkwarmwasser und Beleuchtung: Berechnung des Nutz-, End- und Primärenergiebedarfs für Heizung, Kühlung, Lüftung, Trinkwarmwasser und Beleuchtung Sonderdruck 2012* (4. Aufl., Bd. 91.120.10; 91.140.01) (Nr. V 18599). Berlin: Beuth.

Diefenbach, N., Cischinsky, H. & Rodenfels, M. (2010). *Summary of the research project data base of the building stock – data survey of the state and the trends of energy saving measures in the german residential building stock".*

Diefenbach, N., Cischinsky, H., Rodenfels, M. & Clausnitzer, K.-D. (2010). *Datenbasis gebäudebestand: Datenerhebung zur energetischen qualität und zu den modernisierungstrends im deutschen wohngebäudebestand* (1. Aufl.). Darmstadt: Institut Wohnen und Umwelt (IWU) and Bremer Energie Institut (BEI).

D'Oca, S., Fabi, V., Corgnati, S. P. & Andersen, R. K. (2014). Effect of thermostat and window opening occupant behavior models on energy use in homes. *Building Simulation*. doi: 10.1007/s12273-014-0191-6

Douthitt, R. A. (1986). The demand for residential space and water heating fuel by energy conserving households. *Journal of consumer affairs*, *20* (2), 231–248. doi: 10.1111/j.1745-6606.1986.tb00380.x

Douthitt, R. A. (1989). An economic-analysis of the demand for residential space heating fuel in canada. *Energy*, *14* (4), 187–197. doi: 10.1016/0360-5442(89)90062-5

DuBois, P. (2007). *Mysql cookbook* (2ND ED Aufl.). England: O'REILLY & ASSOCIATES INC (CA).

Durand-Daubin, M., Huebner, G. M., McMichael, M., Shipworth, D., Shipworth, M. & Summerfield, A. (2013). Heating patterns in English homes: Comparing results from a national survey against common model assumptions. *Building and Environment, Volume 70*, *70*.

Dustman, A. (2007). *Mysqldb user's guide: a python interface for mysql.* Zugriff am 19.09.2013 auf `http://mysql-python.sourceforge.net/MySQLdb.html`

Dustman, A. (2013). *Mysqldb1: Github, inc. repository.* Zugriff am 19.09.2013 auf `https://github.com/farcepest/MySQLdb1`

Ebel, W. (1990). *Energiesparpotentiale im gebäudebestand.* Darmstadt: Inst. Wohnen und Umwelt.
ebök. (2000). *Einführung eines energiepasses für den freistaat sachsen. gebäudetypologie für sachsen. endbericht: In auftrag des sächsischen staatsministeriums für umwelt und landwirtschaft (smul).* Tübingen.
ebök. (2005). *Gebäudetypologie für die stadt düsseldorf: Endbericht: Umweltamt landeshauptstadt düsseldorf.* Zugriff am 17.04.2012 auf http://www.eboek.de/downloads/gebaeudetypologie.pdf
ebök/ifeu. (1996). *Heidelberger gebäudetypologie: Tübingen / heidelberg.* Tübingen/Heidelberg. Zugriff am 08.10.2013 auf http://www.ifeu.de/bildungundinformation/pdf/Heidelberg_Gebaeudetypologie_1997.pdf
ebök/ifeu. (1998). *Mannheim waermefibel. der ratgeber zur wärmedämmung und energieeinsparung der ratgeber zur wärm und energieeinsparung: Gebäudetypologie für die stadt mannheim: Tübingen / heidelberg.* Tübingen / Heidelberg. Zugriff am 08.10.2013 auf http://www.ifeu.de/energie/pdf/Mannheim_Waermefibel_1999.pdf
Ecoplan & Schweiz. Bundesamt für Raumentwicklung. (2000). *Siedlungsentwicklung und infrastrukturkosten: Schlussbericht.* Bern: Ecoplan.
Edwards, K. L. & Clarke, G. (2013). Simobesity: Combinatorial optimisation (deterministic) model. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6, S. 69–85). Springer.
Edwards, K. L. & Tanton, R. (2013). Validation of spatial microsimulation models. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6, S. 249–258). Springer.
Einig, K. (2006). *Siedlungsentwicklung und infrastrukturfolgekosten: Bilanzierung und strategieentwicklung ; endbericht ; ein projekt des forschungsprogramms -aufbau ost- des bundesministeriums für verkehr, bau und stadtentwicklung (bmvbs) und des bundesamtes für bauwesen und raumordnung (bbr).* Bonn: BBR. Zugriff auf http://www.bbr.bund.de/cln_005/nn_21272/DE/Veroeffentlichungen/BBR-Online/2002-2006/DL__ON032006,templateId=raw,property=publicationFile.pdf/DL_ON032006.pdf
Erhorn-Kluttig, H. (2011). *Energetische Quartiersplanung: Methoden - Technologien - Praxisbeispiele.* Stuttgart: Fraunhofer-IRB-Verl.
Everding, D. D. (2004). *Leitbilder und Potenziale eines solaren Städtebaus.* Zugriff auf http://www.ecofys.com/files/files/zusammenfassung_000.pdf
Fabi, V., Andersen, R. V., Corgnati, S. & Olesen, B. W. (2012). Occupants' window opening behaviour: A literature review of factors influencing occupant behaviour and models. *Building and Environment*, *58* (0), 188–198. doi: 10.1016/j.buildenv.2012.07.009
Farber, S., Neutens, T., Miller, H. J. & Li, X. (2013). The social interaction potential of metropolitan regions: A time-geographic measurement approach using joint accessibility. *Annals of the Association of American Geographers*, *103* (3), 483-504. doi: 10.1080/00045608.2012.689238
Farooq, B., Bierlaire, M., Hurtubia, R. & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58* (0), 243 - 263. doi: 10.1016/j.trb.2013.09.012
Finney, K. N., Chen, Q., Sharifi, V. N., Swithenbank, J., Nolan, A., White, S. & Ogden, S. (2012). Developments to an existing city-wide district energy network: Part ii – analysis of environmental and economic impacts. *Energy Conversion and Management*, *62* (0), 176–184. doi: 10.1016/j.enconman.2012.03.005
Finney, K. N., Sharifi, V. N., Swithenbank, J., Nolan, A., White, S. & Ogden, S. (2012). Developments to an existing city-wide district energy network – part i: Identification of potential expansions using heat mapping. *Energy Conversion and Management*, *62* (0), 165–175. doi: 10.1016/j.enconman.2012.03.006

Finney, K. N., Zhou, J., Chen, Q., Zhang, X., Chan, C., Sharifi, V. N., ... Bradford, R. (2013). Modelling and mapping sustainable heating for cities. *Applied Thermal Engineering*, *53* (2), 246 - 255. doi: http://dx.doi.org/10.1016/j.applthermaleng.2012.04.009

Genske, D. D., Jödecke, T. & Ruff, A. (2009). *Nutzung städtischer freiflächen für erneuerbare energien.* Zugriff am 26.04.2012 auf `http://www.bbsr.bund.de/BBSR/DE/Veroeffentlichungen/BMVBS/Sonderveroeffentlichungen/2009/DL_NutzungFreiflaechen.pdf`

Gillies, S. (2013). *The shapely user manual — shapely 1.2 and 1.3 documentation.* Zugriff am 2015-11-16 auf `http://toblerity.org/shapely/manual.html`

Guerra Santin, O., Itard, L. & Visscher, H. (2009). The effect of occupancy and building characteristics on energy use for space and water heating in dutch residential stock. *Energy and Buildings*, *41* (11), 1223–1232. doi: 10.1016/j.enbuild.2009.07.002

Guo, J. & Bhat, C. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, *2014*, 92-101. doi: 10.3141/2014-12

Haldi, F. & Robinson, D. (2011). The impact of occupants' behaviour on building energy demand. *Journal of Building Performance Simulation*, *4* (4), 323–338. doi: 10.1080/19401493.2011.558213

Harland, K., Heppenstall, A., Smith, D. & Birkin, M. (2012). Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. *Journal of Artificial Societies and Social Simulation*, *15* (1), 1. Zugriff auf `http://jasss.soc.surrey.ac.uk/15/1/1.html`

Hegger, M., Dettmar, J., Drebes, C., Greiner, M., Kern, T., Meinberg, T., ... Wagner, C. (2014). *Energetische Stadtraumtypen.* Stuttgart: Fraunhofer IRB Verlag.

Hensen, J. & Lamberts, R. (2011). *Building performance simulation for design and operation.* Abingdon and Oxon and and New York and NY: Spon Press.

Hermelink, A., Manteuffel, B. v., Lindner, S. & John, A. (2011). *Flächendeckende erhebung des energetischen zustandes des hamburger gebäudebestandes: Kurzdarstellung der arbeitsschritte.*

Hermes, K. & Poulsen, M. (2012). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, *36* (4), 281–290. doi: 10.1016/j.compenvurbsys.2012.03.005

Hijmans, R., Kapoor, J., Wieczorek, J., Garcia, N., Maunahan, A., Rala, A. & Mandel, A. (2014). *Gadm database of global administrative areas, version 2.0. gadm.* Zugriff auf `http://www.gadm.org/`

Hildebrandt, O., Hellmann, R. & Zantner, M. (2003). *Evaluation des förderprogramms zur altbausanierung in der stadt münster: Gebäudetypologie: Im auftrag der: Stadt münster, amt für grünflächen und umweltschutz klenko (koordinierungsstelle klima & energie); gefördert durch das land nordrhein-westfalen im rahmen der landesinitiative zukunftsenergien.* Zugriff am 08.10.2013 auf `http://www.muenster.de/stadt/klima/pdf/EB_Anhang2Gebaeudetypologie.pdf`

Hoes, P., Hensen, J., Loomans, M., de Vries, B. & Bourgeois, D. (2009). User behavior in whole building simulation. *Energy and Buildings*, *41* (3), 295–302. doi: 10.1016/j.enbuild.2008.09.008

Hong, T., D'Oca, S., Turner, W. J. & Taylor-Lange, S. C. (2015). An ontology to represent energy-related occupant behavior in buildings. part i: Introduction to the {DNAs} framework. *Building and Environment*, *92*, 764 - 777. doi: 10.1016/j.buildenv.2015.02.019

Hrabovszky-Horváth, S., Pálvölgyi, T., Csoknyai, T. & Talamon, A. (2013). Generalized residential building typology for urban climate change mitigation and adaptation strategies: The case of Hungary. *Energy and Buildings*, *62*, 475–485. doi: 10.1016/j.enbuild.2013.03.011

Huang, Z. & Williamson, P. (2001). *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata: Working paper 2001/2.* Zugriff am 23.09.2013 auf `http://pcwww.liv.ac.uk/~william/microdata/workingpapers/hw_wp_2001_2.pdf`

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9* (3), 90–95. doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2007.55

IWU. (2003). *Deutsche gebäudetypologie: Systematik und datensätze.* Zugriff auf `http://www.iwu.de/fileadmin/user_upload/dateien/energie/klima_altbau/Gebaeudetypologie_Deutschland.pdf`

Jentsch, A., Bohn, K., Pohlig, A., Dötsch, C., Richter, S. & Manderfeld, M. (2008). *Leitungsgebundene wärmeversorgung im ländlichen raum: Handbuch zur entscheidungsunterstützung - fernwärme in der fläche.* Zugriff auf `http://publica.fraunhofer.de/eprints/urn:nbn:de:0011-n-1136251.pdf`

Kah, O., Peper, S., Ebel, W., Kaufmann, B., Feist, W. & Zeno, B. (2010). *Untersuchung zum außenluftwechsel und zur luftqualität in sanierten wohnungen mit konventioneller fensterlüftung und mit kontrollierter lüftung: Forschungsvorhaben im rahmen der nationalen beteiligung an der arbeitsgruppe 'advanced housing renovation with solar & conservation' der internationalen energie agentur, iea: Iea shc task 37.*

Kang, C., Ma, X., Tong, D. & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, *391* (4), 1702–1717. doi: 10.1016/j.physa.2011.11.005

Kavgic, M., Mavrogianni, A., Mumovic, D., Summerfield, A., Stevanovic, Z. & Djurovic-Petrovic, M. (2010). A review of bottom-up building stock models for energy consumption in the residential sector. *Building and Environment*, *45* (7), 1683–1697. doi: 10.1016/j.buildenv.2010.01.021

Keirstead, J., Jennings, M. & Sivakumar, A. (2012). A review of urban energy system models: Approaches, challenges and opportunities. *Renewable and Sustainable Energy Reviews*, *16* (6), 3847–3866. doi: 10.1016/j.rser.2012.02.047

KEMA. (2012). *Anpassungs-und investitionserfordernisse der informations- und kommunikationstechnologie zur entwicklung eines dezentralen energiesystems (smart grid): Endbericht – kurzfassung: Proprietary 9011-664.* Zugriff am 08.10.2013 auf `http://www.e-energy.de/images/2012.05_VKU_Kurzstudie_V1.0__final.pdf`

Kii, M., Akimoto, K. & Doi, K. (2014). Measuring the impact of urban policies on transportation energy saving using a land use-transport model. *{IATSS} Research*, *37* (2), 98 - 109. doi: 10.1016/j.iatssr.2014.03.002

Klauß, S., Kirchhof, W. & Gissel, J. (2009). *Katalog regionaltypischer Materialien im Gebäudebestand mit Bezug auf die Baualtersklasse und Ableitung typischer Bauteilaufbauten: 2., berichtigte version.* Zentrum für Umweltbewusstes Bauen e.V. Universität Kassel.

Koch, A. & Girard, S. (2013). Urban neighbourhoods – an intermediate scale for the assessment of energy performance of buildings. In *eceee 2013 summer study* (S. 1377–1385).

Kragh, J. & Wittchen, K. (2013). Development of two danish building typologies for residential buildings. *Energy and Buildings*, *68, Part A*, 79 - 86. doi: 10.1016/j.enbuild.2013.04.028

Krug, H. (2006). *Räumliche wahlmöglichkeiten als effizienzkriterium für siedlung und verkehr* (Dissertation, Univ, Kassel). Zugriff auf `http://kobra.bibliothek.uni-kassel.de/bitstream/urn:nbn:de:hebis:34-2006110115361/3/Krug%2BR%C3%A4umliche%2BWahlm%C3%B6glichkeiten.pdf`

Laycock, R. G. & Day, A. M. (2003). Rapid generation of urban models. *Computers & Graphics*, *27* (3), 423–433. doi: 10.1016/S0097-8493(03)00037-2

Loga, T., Diefenbach, N. & Born, R. (2011). *Deutsche Gebäudetypologie: Beispielhafte Maßnahmen zur Verbesserung der Energieeffizienz von typischen Wohngebäuden.*

Loga, T., Diefenbach, N., Stein, B. & Born, R. (2012). *Tabula – scientific report germany –: Further development of the german residential building typology.* Zugriff am 09.10.2013 auf `http://www.building-typology.eu/downloads/public/docs/scientific/DE_TABULA_ScientificReport_IWU.pdf`

Lovelace, R. & Ballas, D. (2013). Truncate, replicate, sample: A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41* (0), 1–11. doi: 10.1016/j.compenvurbsys.2013.03.004

Lovelace, R., Birkin, M., Ballas, D. & van Leeuwen, E. (2015). Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. *Journal of Artificial Societies and Social Simulation*, *18* (2), 21. Zugriff auf `http://jasss.soc.surrey.ac.uk/18/2/21.html`

Lutzenheiser, L. (2002). *An exploratory analysis of residential electricity conservation survey and billing data.* Zugriff auf `http://www.energy.ca.gov/reports/2002-05-20_400-02-006F.PDF`

Ma, L. & Srinivasan, S. (2015). Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*, *30* (2), 135–150. doi: 10.1111/mice.12085

Manfren, M., Caputo, P. & Costa, G. (2011). Paradigm shift in urban energy systems through distributed generation: Methods and models. *Applied Energy*, *88* (4), 1032–1048. doi: 10.1016/j.apenergy.2010.10.018

Marconi, D. S. A. S. H. H. R. C. P. B. M. (2006). *Raumstruktur und mobilität von personen. argumentarium. unterstützung nachhaltiger mobilitätsstile durch raumplanung. ergebnisse einer sonderauswertung des mikrozensus zum verkehrsverhalten 2000.*

Mavrogianni, A., Davies, M., Kolokotroni, M. & Hamilton, I. (2009). A gis-based bottom-up space heating demand model of the london domestic stock. In *11th conference of international building performance simulation association* (S. 1061–1067). Glasgow. Zugriff am 20.11.2012 auf `http://www.ibpsa.org/proceedings/BS2009/BS09_1061_1067.pdf`

McKinney, W. (2012). *Python for data analysis* (Bd. 1). O'Reilly Media. (Despite the explosive growth of data in industry after industry, learning and accessing data analysis tools has remained a challenge. This pragmatic guide demonstrates the nuts and bolts of manipulating, processing, cleaning, and crunching data with Python. It also serves as a modern introduction to scientific computing in Python for data-intensive applications.)

Mei, S., Chen, B., Zhu, Y., Lees, M., Boukhanovsky, A. & Sloot, P. (2015). Simulating city-level airborne infectious diseases. *Computers, Environment and Urban Systems*, *51*, 97 - 105. doi: 10.1016/j.compenvurbsys.2014.12.002

Micklewright, J. (1989). Towards a household model of uk domestic energy demand. *Energy Policy*, *17* (3), 264–276. doi: 10.1016/0301-4215(89)90051-7

Miller, E. J., Hunt, J. D., Abraham, J. E. & Salvini, P. A. (2004). Microsimulating urban systems. *Computers, Environment and Urban Systems*, *28* (1-2), 9–44. doi: 10.1016/s0198-9715(02)00044-3

Moreira, W. & Warnes, G. R. (2004). *Rpy reference manual: (version 0.3.3).*

Müller, K., Axhausen, K. W., Axhausen, K. W. & Axhausen, K. W. (2010). *Population synthesis for microsimulation: State of the art.* ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

Muñoz H., M. E. (2014). A microsimulation approach to generate occupancy rates of small urban areas. In *2nd Asia Conference on International Building Performance Simulation Associatio, IBPSA Japan.* Nagoya University, Japan. Zugriff auf `http://ibpsa.org/proceedings/asim2014/140_AsimC3-28-370.pdf`

Muñoz H., M. E. (2015). *heat: R package to estimate heat demand of residential buildings. github.com/emunozh/heat.* Zugriff auf `github.com/emunozh/heat`

Muñoz H., M. E. (2016). Construction of building typologies from a regional material catalog: Assessment of urban heat demand and the environmental impact of retrofit policies. *Management of Environmental Quality - An international journal*, *27* (6). doi: 10.1108/MEQ-05-2015-0096

Muñoz H., M. E., Dochev, I. & Peters, I. (2015). Allocating heat consumption of residential buildings in space with help of a filter array to determinate the building type in a digital cadaster.

Muñoz H., M. E., Dochev, I., Seller, H. & Peters, I. (2016). Constructing a synthetic city for estimating spatially disaggregated heat demand. *International Journal of Microsimulation*, *–in Press–*.

Muñoz H., M. E. & Peters, I. (2014a). *Allocating heat consumption of residential buildings in space with help of a filter array to determinate the building type in a digital cadaster: Working paper, Technical Urban Infrastructure Systems Group, HafenCity University (HCU).*

Muñoz H., M. E. & Peters, I. (2014b). Constructing an urban microsimulation model to assess the influence of demographics on heat consumption. *International Journal of Microsimulation*, *7* (1), 127-157. Zugriff auf http://www.microsimulation.org/IJM/V7_1/6_IJM_7_1_Munoz_Peters.pdf

Muñoz H., M. E. & Peters, I. (2015a). Berücksichtigung des Nutzers in der Schätzung der Wärmennachfrage von Gebäuden. In IBA Hamburg GmbH, Umweltbundesamt & TU Darmstadt (Hrsg.), *Energieatlas Werkbericht 1* (S. 224). JOVIS Verlag.

Muñoz H., M. E. & Peters, I. (2015b). The use of diffusion algorithms for the estimation of construction year within urban models: Working paper, Technical Urban Infrastructure Systems Group, HafenCity University (HCU).

Muñoz H., M. E., Tanton, R. & Vidattama, Y. (2015). A comparison of the GREGWT and IPF methods for the re-weighting of surveys. In *5th world congress of the international microsimulation association (IMA).*

Muñoz H., M. E., Vidyattama, Y. & Tanton, R. (2015a). *GREGWT: an implementation of the GREGWT algorithm in R. github.com/emunozh/gregwt.* Zugriff auf github.com/emunozh/GREGWT

Muñoz H., M. E., Vidyattama, Y. & Tanton, R. (2015b). The influence of an ageing population and an efficient building stock on heat consumption patterns. In *14th international conference of the international building performance simulation association (IBPSA).* Zugriff auf http://www.ibpsa.org/proceedings/BS2015/p2917.pdf

Nouidui, T. S., Wetter, M. & Zuo, W. (2012). Validation of the window model of the modelica buildings library. In *Proc. of the 5th simbuild conference.* Berkeley. Zugriff am 21.08.2012 auf http://simulationresearch.lbl.gov/wetter/download/2012-simBuild-windowValidation.pdf

O'Donoghue, C., Morrissey, K. & Lennon, J. (2014). Spatial microsimulation modelling: a review of applications and methodological choices. *International Journal of Microsimulation*, *7* (1), 26-75.

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science & Engineering*, *9* (3), 10–20. Zugriff auf http://link.aip.org/link/?CSX/9/10/1

Oracle. (2011). *Mysql 5.5 reference manual: Including mysql cluster ndb 7.2 reference guide.*

Orcutt, G. H. (1957). A new type of socio-economic system. *Review of Economics and Statistics*, *39* (2), 116–123.

Page, J., Robinson, D., Morel, N. & Scartezzini, J.-L. (2008). A generalised stochastic model for the simulation of occupant presence. *Energy and Buildings*, *40* (2), 83–98. doi: 10.1016/j.enbuild.2007.01.018

Page, J., Robinson, D. & Scartezzini, J.-L. (2007). Stochastic simulation of occupant presence and behaviour in buildings. In *10th conference of international building performance simulation association* (S. 757–764). Zugriff am 21.11.2012 auf http://www.ibpsa.org/proceedings/BS2007/p385_final.pdf

Pandas development team. (o. J.). *Python data analysis library — pandas: Python data analysis library.* Zugriff am 2015-11-16 auf http://pandas.pydata.org/

Parish, Y. I. H. & Müller, P. (2001). Procedural modeling of cities. In *in proceedings of the 28th annual conference on computer graphics and interactive techniques* (S. 301–308). Press.

Payne, J. (2010). *Beginning python: Using python 2.6 and python 3.1.* Indianapolis and IN: John Wiley & Sons Inc.

Pérez, F. & Granger, B. E. (2007, Mai). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, *9* (3), 21–29. doi: 10.1109/MCSE.2007.53

Peters, I., Brassel, K.-H. & Spörri, C. (2002). A microsimulation model for assessing urine flows in urban wastewater management. In A. E. Rizzoli & A. J. Jakeman (Hrsg.), *Integrated assessment and decision support. proceedings of the first biennial meeting of the international environmental modelling and software society.* Zugriff auf http://www.iemss.org/iemss2002/proceedings/pdf/volume%20due/25.pdf

PostGIS. (2013). *Postgis 2.0 manual - postgis-2.0.pdf: Svn revision (11136).* Zugriff am 04.03.2013 auf http://www.postgis.net/stuff/postgis-2.0.pdf

Pritchard, D. & Miller, E. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39* (3), 685-704. doi: 10.1007/s11116-011-9367-4

psycopg. (2013). *psycopg2: Github, inc. repository.* Zugriff am 19.09.2013 auf https://github.com/psycopg/psycopg2/

Python Software Foundation. (o. J.-a). *Mysql-python 1.2.4: Python package index.* Zugriff am 19.09.2013 auf https://pypi.python.org/pypi/MySQL-python/1.2.4

Python Software Foundation. (o. J.-b). *psycopg2 2.5.1 : Python package index.* Zugriff am 19.09.2013 auf https://pypi.python.org/pypi/psycopg2/2.5.1

Rahman, A., Harding, A., Tanton, R. & Liu, S. (2013). Simulating the characteristics of populations at the small area level: New validation techniques for a spatial microsimulation model in australia. *Computational Statistics & Data Analysis*, *57* (1), 149–165. doi: 10.1016/j.csda.2012.06.018

Rahman, A., Harding, A., Tanton, R. & Shuangzhe, L. (2010). Methodological issues in spatial microsimulation modelling for small area estimation. *International Journal of Microsimulation*, *3* (2), 3–22.

Rao, C. R. (1995). The use of hellinger distance in graphical displays of contingency table data. *New trends in probability and statistics*, *3*, 143–161.

Rehdanz, K. (2007). Determinants of residential space heating expenditures in germany. *Energy Economics*, *29* (2), 167–182. doi: 10.1016/j.eneco.2006.04.002

Rossant, C. (2013). *Learning IPython for interactive computing and data visualization : learn IPython for interactive python programming, high performance numerical computing, and data visualization.* Birmingham [u.a.]: Packt Publ.

Roth, U. & Häubi, F. (1980). *Wechselwirkungen zwischen der siedlungsstruktur und wärmeversorgungssystemen : Forschungsprojekt bmbau rs ii 4-704102-77.10(1980).* Bonn.

Sanquist, T. F., Orr, H., Shui, B. & Bittner, A. C. (2011). Lifestyle factors in u.s. residential electricity consumption. *Energy Policy* (0). doi: 10.1016/j.enpol.2011.11.092

Schenk, N. J., Moll, H. C. & Schoot Uiterkamp, A. J. (2007). Meso-level analysis, the missing link in energy strategies. *Energy Policy*, *35* (3), 1505–1516. doi: 10.1016/j.enpol.2006.04.013

Schmid, T. & Münnich, R. (2014). Spatial robust small area estimation. *Statistical Papers*, *55* (3), 653-670. doi: 10.1007/s00362-013-0517-y

Scott, A. (1980). The economics of house heating. *Energy Economics*, *2* (3), 130–141. doi: 10.1016/0140-9883(80)90024-9

Shi, L., Kloog, I., Zanobetti, A., Liu, P. & Schwartz, J. D. (2015, #jul#). Impacts of temperature and its variability on mortality in new england. *Nature Clim. Change*, *advance online publication*, –. doi: 10.1038/nclimate2704

Sieverts und Volwahsen, Roth, U. & Volwahsen, A. (1980). *Rationelle energieverwendung im rahmen der kommunalen entwicklungsplanung* (Bd. Heft Nr. 083). Bonn: Der Bundesminister für Raumordnung Bauwesen und Städtebau.

Singh, A. & Mohl, C. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, *22*, 107–115.

Singh, M. K., Mahapatra, S. & Teller, J. (2013). An analysis on energy efficiency initiatives in the building stock of Liege, Belgium. *Energy Policy*, *62*, 729–741. doi: 10.1016/j.enpol.2013.07.138

Smith, D. M., Pearce, J. R. & Harland, K. (2011). Can a deterministic spatial microsimulation model provide reliable small-area estimates of health behaviours? an example of smoking prevalence in new zealand. *Health & Place*, *17* (2), 618 - 624. (Geographies of Care) doi: 10.1016/j.healthplace.2011.01.001

Stötzer, M., Hauer, I., Richter, M. & Styczynski, Z. A. (2015). Potential of demand side integration to maximize use of renewable energy sources in germany. *Applied Energy*, *146*, 344 - 352. doi: 10.1016/j.apenergy.2015.02.015

Swan, L. G. & Ugursal, V. I. (2009). Modeling of end-use energy consumption in the residential sector: A review of modeling techniques. *Renewable and Sustainable Energy Reviews*, *13*, 1819–1835. doi: 10.1016/j.rser.2008.09.033

TABULA Project Team. (2010). *Use of building typologies for energy performance assessment of national building stocks: Existent experiences in european countries and common approach: – first tabula synthesis report –.* Zugriff am 08.10.2013 auf http://www.building-typology.eu/downloads/public/docs/report/TABULA_SR1.pdf

TABULA Project Team. (2012a). *Application of building typologies for modelling the energy balance of the residential building stock: – tabula thematic report nr. 2 –.* Zugriff am 07.10.2013 auf http://www.building-typology.eu/downloads/public/docs/report/TABULA_TR2_D8_NationalEnergyBalances.pdf

TABULA Project Team. (2012b). *Typology approach for building stock energy assessment: Main results of the tabula project: - final project report -.* Zugriff am 07.10.2013 auf http://www.building-typology.eu/downloads/public/docs/report/TABULA_FinalReport.pdf

TABULA Project Team. (2012c). *Typology approach for building stock energy assessment: Main results of the tabula project: – final project report: Appendix volume –.* Zugriff am 06.10.2013 auf http://www.building-typology.eu/downloads/public/docs/report/TABULA_FinalReport_AppendixVolume.pdf

Tanton, R. (2007). spatialMSM: The Australian spatial microsimulation model. In *The 1st general conference of the international microsimulation association.*

Tanton, R. (2014). A review of spatial microsimulation methods. *International Journal of Microsimulation*, *7* (1), 4–25.

Tanton, R. & Edwards, K. L. (2013). Introduction to spatial microsimulation: History, methods and applications. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6, S. 3 – 9). Springer.

Tanton, R., Harding, A. & McNamara, J. (2013). Spatial microsimulation using a generalised regression model. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6, S. 87–103). Springer.

Tanton, R. & Vidyattama, Y. (2010). Pushing it to the edge: Extending generalised regression as a spatial microsimulation method. *International Journal of Microsimulation*, *3* (2), 23–33.

Tanton, R., Vidyattama, Y., Nepal, B. & McNamara, J. (2011). Small area estimation using a reweighting algorithm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174* (4), 931–951. doi: 10.1111/j.1467-985X.2011.00690.x

Tanton, R., Williamson, P. & Harding, A. (2014). Comparing two methods of reweighting a survey file to small area data: generalised regression and combinatorial optimisation. *International Journal of Microsimulation*, *7* (1), 76-99.

The European Parliament and the Council of the European Union. (16 December 2002). *Directive 2002/91/ec on the energy performance of buildings: Directive 2002/91/ec* (Bd. L 1) (Nr. 65).

The European Parliament and the Council of the European Union. (19 May 2010). *Directive 2010/31/eu on the energy performance of buildings: Directive 2010/31/eu* (Bd. L 153) (Nr. 13).

The PostgreSQL Global Development Group. (2012). *Postgresql 9.2 documentation.* Zugriff am 25.01.2013 auf `http://www.postgresql.org/files/documentation/pdf/9.2/postgresql-9.2-A4.pdf`

The PostgreSQL Global Development Group. (2013). *Postgresql: About.* Zugriff am 19.09.2013 auf `http://www.postgresql.org/about/`

Toftum, J., Andersen, R. V. & Jensen, K. L. (2009). Occupant performance and building energy consumption with different philosophies of determining acceptable thermal conditions. *Building and Environment*, *44* (10), 2009–2016.

Tosi, S. (2009). *Matplotlib for python developers.* BIRMINGHAM - MUMBAI.

Van Raaij, W. & Verhallen, T. M. (1983a). A behavioral model of residential energy use. *Journal of Economic Psychology*, *3* (1), 39–63. doi: 10.1016/0167-4870(83)90057-0

Van Raaij, W. & Verhallen, T. M. (1983b). Patterns of residential energy behavior. *Journal of Economic Psychology*, *4* (1–2), 85–106. doi: 10.1016/0167-4870(83)90047-8

Verein Deutscher Ingenieure. (1994). *Energieverbrauchskennwerte für gebäude; grundlagen: Richtlinien energieverbrauchskennwerte für gebäude* (Nr. 3807). Berlin: Beuth.

Vidyattama, Y. & Tanton, R. (2013). Projections using a static spatial microsimulation model. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6, S. 87–103). Springer.

Vidyattama, Y., Tanton, R. & Biddle, N. (2015). Estimating small-area indigenous cultural participation from synthetic survey data. *Environment and Planning A*, *47* (5), 1211-1228. doi: 10.1177/0308518X15592314

Voas, D. & Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, *6* (5), 349–366. doi: 10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5

Westra, E. (2010). *Python geospatial development: Build a complete and sofisticated mapping application from scratch using python tools for gis development.* Olton and Birmingham: Packt Pub.

Williamson, P. (2013). An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation. In R. Tanton & K. Edwards (Hrsg.), *Spatial microsimulation: A reference guide for users* (Bd. 6). Springer.

Williamson, P., Birkin, M. & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment and Planning A*, *30* (5), 785–816. doi: 10.1068/a300785

Williamson, P., Clarke, G. P. & McDonald, A. T. (1996). Estimating small area demands for water with the use of microsimulation. In G. P. Clarke (Hrsg.), *Microsimulation for urban and regional policy analysis* (Bd. 6, S. 117–148). London: Pion. doi: 10.1111/j.1747-6593.1997.tb00114.x

Williamson, P., Mitchell, G. & McDonald, A. T. (2002). Domestic water demand forecasting: A static microsimulation approach. *Water and Environment Journal*, *16* (4), 243–248. doi: 10.1111/j.1747-6593.2002.tb00410.x

Wirl, F. (1987). Thermal comfort, energy conservation and fuel substitution: An economic-engineering approach. *Energy Syst. Policy*, *11* (4), 311–318.

Yamasaki, E. & Tominaga, N. (1997). Evolution of an aging society and effect on residential energy demand. *Energy Policy*, *25(11)*, 903–912. doi: doi:10.1016/S0301-4215(97)00040-2

Ye, X., Konduri, K., Pendyala, R. M., Sana, B. & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *88th annual meeting of the transportation research board, washington, dc.*

Zeine, C., Gebhardt, M., Bockting, B., Mantai, A. & Wei, J. P. (Hrsg.). (2007). *Verbrauchskennwerte 2005: Energie- und wasserverbrauchskennwerte in der bundesrepublik deutschland: Forschungsbericht der ages gmbh münster.* Zugriff am 28.10.2013 auf `www.ages-gmbh.de/images/downloads_von_der_homepage/kennwerte/kw2005_inhalt_und_methode.pdf`

Zhun Yu, Benjamin C.M. Fung, Fariborz Haghighat, Hiroshi Yoshino & Edward Morofsky. (2011). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, *43* (6), 1409–1417. doi: 10.1016/j.enbuild.2011.02.002