Bernhard Luger, BSc

Generation of a Synthetic Population for MATSim Models Using Multidimensional Iterative Proportional Fitting and Discrete Choice Models

Master's Thesis

Graz University of Technology

Institute of Software Technology (IST)

Supervisors: Univ.-Prof. Dipl-Ing. Dr.techn. Franz Wotawa
Dipl-Ing. Dr.techn. Birgit Hofer

Graz, August 2017

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, ________________________ ____________________________

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, ________________________ ____________________________

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Abstract

Multi-Agent Transport Simulations (MATSim) offer the possibility to examine the mobility behavior of millions of agents on a high level of detail. Each agent has a data-sheet with its sociodemographic data (e.g. age, sex, employment status) which can be extended by multiple further attributes. However acquiring this data on an agent-level can be tedious, because such kind of information is often missing due to legal restrictions as well as by sparse micro census data in general. Because of legal restrictions such data is often provided only in aggregated form. Thus when it comes to deducing single individuals from such data, there arise certain caveats for processing these data accordingly. Alternatively, to gain detailed data on an agent-level there can be conducted household surveys, which are nevertheless expensive and time-consuming to perform.

Therefore this master's thesis introduces a design which addresses the problem of generating a synthetic agent population with sociodemographic data for each agent by integrating different data sources. In case of generating a synthetic population for the 21^{st} district of Vienna, survey data of the European Union Statistics on Income and Living Conditions (EU-SILC), discrete choice models and public accessible cross-classification tables for socioeconomics and demographics (e.g. sex by age at region) provided by Statistics Austria were used to reassemble the real-world population.

In order to achieve this goal, cross-classification tables are processed to a set of margin constraints and get organized in a data structure for efficiently generating templates. These templates are subsequently used in an iterative procedure of assigning extrapolation factors to satisfy all given margin constraints best. This design copes with shortcomings related to scalability which typically arise when deploying iterative proportional fitting (IPF)-based population generation designs. Finally attributes, which can not be

deduced from Statistics Austria in first place (income per capita and car ownership per household), are estimated using discrete choice models. The result is a synthetic population on an agent-level for the 21^{st} district of Vienna. Liable to costs, spatial data from Statistics Austria was acquired in a grid of 250x250 meters with counts on selected person attribute combinations (sex by age and sex by employment status). This grid information was taken into consideration when finally assigning the population to spatial entities.

Deducing single entities from cross-classification tables may not only be an important topic in the field of transport simulations. Moreover it could also be auspicious in the field of socioeconomic studies when it is necessary to conclude from aggregated to detailed information.

This proposed design for generating a synthetic population allows introducing further information sources leading to improved results (i.e. the synthetic population can be extended by additionally attributes) if the acquired information fits the given context.

Kurzfassung

Multi-Agenten Transport Simulationen (MATSim) bieten die Möglichkeit, das Mobilitätsverhalten von mehreren Millionen Agenten in einem hohen Detailgrad zu untersuchen. Jeder Agent besitzt eine Liste von soziodemographischen Attributen (z.B. Alter, Geschlecht, Erwerbsstatus), welche um viele weitere Attribute ergänzt werden kann. Hinsichtlich gesetzlicher Bestimmungen und fehlender Verfügbarkeit von Mikrozensus Daten ist es schwierig, Informationen zu diesen Attributen auf Einzelpersonenebene zu erhalten. Diese Informationen liegen aufgrund von gesetzlichen Beschränkungen zumeist nur in aggregierter Form vor. Für das Ableiten von Einzelpersonen aus aggregierten Daten ist eine adäquate Informationsverarbeitung notwendig. Alternativ können auch Haushaltsbefragungen durchgeführt werden, um Informationen zu Einzelpersonen zu erhalten, jedoch sind diese teuer und zeitaufwendig in der Durchführung.

Diese Masterarbeit beschäftigt sich mit der Aufgabenstellung zur Generierung einer synthetischen Agentenbevölkerung mit soziodemografischen Attributen für jeden der Agenten. Für diese Umsetzung wurden verschiedene Datenquellen verarbeitet. Für die Erzeugung einer synthetischen Bevölkerung für den 21. Wiener Gemeindebezirk wurden Befragungsdaten der Europäischen Gemeinschaftsstatistik über Einkommen und Lebensbedingungen, diskrete Wahlmodelle und frei zugängliche Kreuzklassifikationstabellen für soziodemographische Daten (z.B. Geschlecht je Alter in einer bestimmten Region) verwendet, um die Bevölkerung realitätsnah nachzubilden.

Für die Lösung dieser Aufgabenstellung wurden Kreuzklassifikationstabellen zu einer Menge von Randsummen verarbeitet und in einer Datenstruktur organisiert, die es ermöglicht Agentenvorlagen zu erzeugen. Diesen Vorlagen wurden iterativ Hochrechnungsfaktoren zugewiesen, um alle Randsummen schrittweise zu erfüllen. Dieses Vorgehen bietet die Möglichkeit

mit Unzulänglichkeiten hinsichtlich der Skalierbarkeit, die typischerweise bei der Erzeugung synthetischer Bevölkerungen per Randsummenausgleich auftreten, umzugehen. Attribute, welche nicht von den Informationen in Kreuzklassifikationstabellen abgeleitet werden können (Pro-Kopf-Einkommen und Fahrzeugbesitz im Haushalt), werden mithilfe von diskreten Wahlmodellen geschätzt. Das Ergebnis ist eine synthetische Bevölkerung auf Einzelpersonenebene für den 21. Wiener Gemeindebezirk. Darüber hinaus wurden räumliche Informationen bezüglich der Anzahl gewisser Kombinationen von Soziodemographien (Geschlecht je Alter und Geschlecht je Erwerbsstatus) in einem Raster von 250x250 Metern von Statistik Austria erworben. Mithilfe dieser Daten wurde schlussendlich die erzeugte synthetische Bevölkerung auch räumlichen Zonen zugewiesen.

Das Ableiten von Einzelpersonen aus aggregierten Informationen stellt bislang ein interessantes Thema im Bereich der Verkehrsplanung dar. Zukünftig könnte dieser Ansatz auch im Gebiet der Sozialwissenschaften Anwendung finden, da detaillierte Informationen zu soziodemographischen Attributen von Einzelpersonen benötigt werden und diese zumeist aggregiert vorliegen.

Dieser Ansatz für die Generierung einer synthetischen Bevölkerung erlaubt es, weitere Informationsquellen einfließen zu lassen (z.B. die Erweiterung der synthetischen Bevölkerung um zusätzliche Attribute), um verbesserte Resultate zu erhalten. Hierfür ist zu beachten, dass die zusätzlichen Daten dem bisher gegebenen Kontext entsprechen.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank all my colleagues from the IKK Kaufmann-Kriebernegg ZT- GmbH company who have provided help and feedback during the course of my master's thesis. I especially wish to thank my advisors, Franz Wotawa and Birgit Hofer, for their attention to my questions and for setting up a pleasant working atmosphere. Special thanks go to Birgit Hofer especially for correcting draft versions of this thesis and providing me with very valuable feedback for writing further works.

Special mention goes to Jacqueline Aspäck, Florian Koppelhuber, Damian Ferk and Georg Kriebernegg for their help and input on the topic of generating synthetic populations for transportation planning simulations.

Last but not least, without the support and understanding of my soon to be wife Nina Krofika, this thesis would not have been possible.

Bernhard Luger
Graz, Austria, August 2017

1. Introduction

As past has shown, the need for more fine-grained transport simulations has become more important in recent times. Due to more powerful computational systems as well as additional research in the fields of transportation planning and computer science, it is nowadays possible to simulate the mobility patterns of each person within a simulation concurrently. Simulation tools which simulate the real-world mobility behavior on such a fine level of detail are called microsimulation tools. One of those is the open source Multi-Agent Transport Simulation (MATSim) model. Internationally MATSim models have become more popular in the field of transport simulation and traffic models. MATSim offers a microscopic transport simulation environment, which is important to provide with fine-grained information on mobility behavior as well as on sociodemographic data of the agents.

In general there are three different types of transportation simulation environments. There are macroscopic, mesoscopic and microscopic simulation environments.

Macroscopic models take a look at traffic flows at a much lower level of detail. Therefore such models can on the one hand easier be maintained but on the other hand also more impractical for the purpose of investigating changes in the context of microsimulations (e.g. introducing and evaluating further modes of transport like bikesharing, carsharing, autonomous vehicles, taxis). Mesoscopic models are some kind of mixture of macro- and microscopic models. Macro- and mesoscopic models use representative person groups (allocation of the population to predefined sociodemographic groups) to depict different sociodemographic specific mobility patterns.

In contrast to these models, a MATSim model handles each agent on its own, whereas an agent can in fact have a detailed and potential arbitrarily expandable character table. Within a microsimulation like MATSim, it is, in comparison to macroscopic and mesoscopic transport simulations, thus even more of a concern to provide sound and detailed input data. Microsimulation environments, such as MATSim, allow more conclusive analysis of traffic-flow behavior depending on sociodemographic characteristics.

As the input data on which a MATSim model is based upon becomes more detailed, the results from this MATSim model for e.g. site evaluation, traffic analysis and predicting potential impacts of future infrastructure planning projects, also become more conclusive. Thus providing as fine-grained information as possible is vital. Therefore it is an important task to generate a synthetic population which depicts the real-world population as good as possible.

Related work, which aimed on creating a near real-world population, are often based on iterative proportional fitting (IPF) designs [1] [15] [14] [26] [7] with additional Monte-Carlo-Sampling [26] [15] strategies in order to deduce population information to a single entity level. Moeckel, Spiekermann & Wegener [14] generated a synthetic population for Switzerland, where they allocated sociodemographic data to persons as well as the geographic positions to households (HH) by deploying Geographic Information System (GIS) techniques. Müller & Axhausen [15] introduced a two-step IPF approach to generate the population of Switzerland, by introducing fitting of the person as well as of the household domain. The IPF-based designs proposed by Zhu & Ferreira [26] and Adiga et al. [1] were quite similar to the one introduced by Moeckel, Spiekermann & Wegener. Farooq et al. [7] proposed an approach for generating a synthetic population by deploying a simulation based Markov-Chain-Monte-Carlo (MCMC) approach and processing information of cross-classification tables specific to a certain spatial region.

These introduced approaches are based on marginal sum information of population attributes, such as the total count of female persons in a certain region. These sums are processed and subsequently allocated to person templates. The basic idea behind these articles is to conclude from marginal distributions of attributes to a joint distribution and eventually to deduce single person entities from the resulting joint distribution. A further approach on how to deal with synthetic population generation, was proposed by Anderson et al. [2], who applied the concept of discrete choice models for deducing information on household compositions in Switzerland. This article advertises the application of discrete choice models in the context of synthetic population generation.

The aim of this master's thesis is to generate a synthetic population for the 21^{st} district of Vienna, *Floridsdorf*, relying on free statistical input data from Statistics Austria as well as on non-public available data of the European Union Statistics on Income and Living Conditions (EU-SILC) household survey [21].

In this survey, which was conducted in the year 2014, approximately 13,000 Austrian citizen took part. The statistics data, given as cross-classification tables, are processed in a data structure to identify possible templates (e.g. eliminate five year old parents). The resulting templates are fit to the counts given in the cross-classification tables, i.e. the templates are assigned extrapolation factors. Subsequently, using discrete choice models and rule matching, the resulting population is allocated to families and moreover to households. Finally the income per capita and the car ownership of each household are estimated by considering parameters deduced from the EU-SILC household survey through deploying discrete choice models. To conclude this work, the households and their inhabitants are allocated to spatial regions by considering spatial information on "sex by age" and "sex by employment status" given in a grid of 250x250 meters provided by Statistics Austria.

Chapter 2 introduces related work for synthetic population generation, which we took in consideration for defining our own design for population synthesis. The approaches, we eventually took in consideration, were based upon cross-classification tables [7] for deducing a baseline population, discrete choice models for refining the population [7] and finally GIS techniques for allocating the population to spatial entities [14].

Chapter 3 shows the technical background of this work. It is presented how MATSim models are defined. Moreover it is highlighted their need for sound input data. Furthermore the input data of Statistics Austria are presented. We processed these data in a tree structure and encoded the resulting leafs (i.e. templates) using a custom 64 bit-mask. Subsequently multinomial logit models are introduced and there is a description how discrete choice models can be based upon those. Eventually parameters which can be deduced by applying Maximum Likelihood Estimation (using BIOGEME [4]) on a discrete choice model are discussed by considering a small example.

Chapter 4 shows the two steps of the technical realization of our design, on how to decompose information of Statistics Austria data to an individual level. The first part deals with the task of setting up a proper tree structure which is needed to create all possible templates, whereas each level of the tree corresponds to another characteristic of the person, family or household. Thus there is the need for coping with the size of this tree, as the characteristics count increases, so does the tree. In the second part the templates get assigned extrapolation factors according to the margin constraints given in the cross-classification tables, therefore resulting in a set of weighted person, family and household templates. Eventually using discrete choice models the population is extended by the characteristics of income per capita and car ownership per household. Finally we conclude our population by allocating the population to grid elements.

Chapter 5 presents the results we gain by performing the design introduced in the previous chapter. Information on the generated population with respect to certain characteristics as well as on the runtime for creating the synthetic population of the 21^{st} district of Vienna are presented .

Chapter 6 discusses the resulting synthetic population and evaluates the introduced design. Limitations as well as strengths are highlighted.

Finally Chapter 7 concludes this work and presents possible future work in the field of synthetic population generation as well as further possible fields of application for populations created by our introduced design.

2. Related Work

This chapter presents a review of already existing approaches for generating synthetic populations with respect to depicting the real-world population best. In this field of research there is a wide range of articles which deploy iterative proportional fitting (IPF) approaches for on the one hand generating populations [1] [15] [14] [26] and on the other hand for benchmarking their own proposed approaches [7]. In this chapter we put six articles on the slate, which we used for analysis of the deployed methods. The following sections will elaborate on the deployed techniques presented in the articles. The different deployed methods result in synthetic populations with different scopes of application. In general the created synthetic populations differ with respect to the spatial level, activity profile of the single persons, household composition, additional household characteristics and moreover household dwelling types.

This chapter is organized by the deployed methods given in the articles. Firstly there are presented approaches which deployed IPF-based approaches and the outcomes of those. Secondly approaches which used discrete choice models for deducing additional information are shown. The additional information were not present in first place when generating the baseline population. Thirdly a work which deployed Markov- Chain-Monte-Carlo (MCMC) Simulation and Gibbs sampling is shown. Subsequently there is introduced an approach, where statistics on trip and travel behavior are assigned to the population s.t. the population also depicts the real-world population with respect to the mobility behavior. Finally approaches which deployed Geographic Information System (GIS) techniques for allocating the population to spatial entities, are stated.

Martial Status (MS)	Count
Single	802,056
Married	669,429
Divorced	109,352
Widower	180,901

Gender (G)	Count
Male	847,483
Female	914,255

Table 2.1.: Margin sums of characteristics "Martial Status" and "Gender" of Vienna citizen.

2.1. Iterative Proportional Fitting

IPF is a commonly used technique, when it comes to population synthesis [5]. The idea of this method is to deduce single person entities with characteristics by processing marginal sums (see Table 2.1). These marginal sums hold information on the totals of selected population characteristics, such as counts of female and male persons as well as counts of single, married, divorced and widowed persons within a certain spatial region. Considering these values, we want to deduce single person entities with characteristics of "martial status" and "sex". Initially a N-dimensional matrix, which is referred to as contingency table (CT), is set up and (in the most trivial case) filled with values equally one. Whereas N corresponds to the number of characteristics we want to take in consideration for our synthetic population. By filling the CT with values unequal to one, it is possible to take into account correlation between certain characteristics (e.g. younger people tend to be more often single). Iteratively the values in the matrix are adjusted to meet the marginal sums (see Table 2.2) present in the column and row. In case the input marginal sums are deduced from the real census data, the IPF procedure will almost perfectly meet the marginal sums in a reasonable number of iterations by design. Due to the computational, as well as the memory overhead, which arises when deploying this method by introducing a further dimension for each additional characteristic we want to synthesize, this approach is limited. Nevertheless to use this approach efficiently, there is the need for deploying heuristics for mitigating the computational overhead.

The study of Moeckel, Spiekermann & Wegener [14] introduced an IPF-based approach for generating a population with characteristics for the domains "sex", age", "religion" and "work location". The resulting population was

MS\G		Male	Female
	Σ	847,483	914,255
Single	802,056	1	1
Married	669,429	1	1
Divorced	109,352	1	1
Widower	180,901	1	1

MS\G		Male	Female
	Σ	847,483	914,255
Single	**802,056**	401,028	401,028
Married	**669,429**	334,714	334,714
Divorced	**109,352**	54,676	54,676
Widower	**180,901**	90,450	90,450

MS\G		Male	Female
	Σ	**847,483**	**914,255**
Single	802,056	385,828	416,227
Married	669,429	322,028	347,400
Divorced	109,352	52,603	56,748
Widower	180,901	87,022	93,878

MS\G		Male	Female
	Σ	**847,483**	**914,255**
Single	**802,056**	385,828	416,227
Married	**669,429**	322,028	347,400
Divorced	**109,352**	52,603	56,748
Widower	**180,901**	87,022	93,878

Table 2.2.: Initial setup of the IPF procedure. Subsequently cell values are alternating adjusted to satisfy row and column margin values as good as possible. After a few iterations the sums given in the cells will converge to the values of the margin sums.

subsequently allocated to households with the aid of Monte-Carlo-Sampling. In the process of allocating, there were also assigned household characteristics (like the number of cars and the household income). Monte-Carlo-Sampling is limited by the possibility of determining relationships among person characteristics as well as among household's characteristics.

Within the scope of the article of Adiga et al. [1], there was created a population by deploying IPF techniques. There was deduced a population with individual and household level characteristics such as "age", "sex", "martial status", "household size" and "location". For modeling correlations between certain characteristics, like "age" and "martial status" (which do in fact highly correlate) there was used a 5% sample of the actual population micro census data to seed the matrix accordingly.

Furthermore Zhu & Ferreira [26] applied an IPF procedure for population synthesis. This procedure differs from the typical IPF because there has been implemented a two-stage IPF. Within the first procedure the population was generated for a certain spatial area, similar to the other IPF-based approaches presented in related work. In the second IPF procedure, the population was assigned to a more disaggregated spatial level. Thus there were processed building and parcel data, the population was also fit to those in an iterative proportional fitting procedure.

Müller & Axhausen [15] proposed a design for performing IPF on two domains. Alternately the household and person domain were fitted to the margin sums of certain sociodemographic characteristics. By switching between the domains an entropy-optimizing fitting step is performed. Hierarchical IPF (HIPF), as this design is called was applied in order to generate a synthetic population for Switzerland, based upon a 5% sample of the Swiss census. The results showed, that the population which was gained by this approach, approximately met the quality and performance of former approaches, that also deployed iterative proportional updating and entropy optimization for population synthesis. A great caveat of this approach was when applying the household expansion factors, the individual person domain expansion factors, were mitigated. The reason for this is, as from the household domain of view, the persons within each household are treated the same without considering the already allocated expansion factors.

For benchmarking their presented approach, Farooq et al. [7] generated a synthetic population based upon IPF, in order for evaluating his newly introduced MCMC-based approach. The actual approach introduced by Farooq et al. for population synthesis is introduced in more detail in Section 2.3. The generated population differed in four characteristics ("age", "sex", "household size" and "level of education"). The two generated populations, IPF- and MCMC-based, were compared to the actual micro census data of the Swiss population by considering the joint distributions of the four characteristics present in the generated populations.

2.2. Discrete Choice Models

In the identified literature there was evident a more common use of discrete choice models (DCM) for population synthesis [2] [7] in the recent past. Actually this technique was often deployed to add further characteristics to baseline populations. As a result of deploying DCMs, the population can become more detailed, as certain characteristics can be estimated by using DCMs, without having to deal with missing statistical data on these characteristics. The idea is to use survey data to estimated parameters which describe the process of deciding between certain alternatives (i.e. characteristics) on an individual level. These kind of models are often applied in the field of transportation research for e.g. mode and location choice [11]. A new field of application for DCMs is to apply these models for population synthesis. In fact decission makers may not decide between certain kind of transportation modes (as it is the case in mode choice models), but may "decide" between different kind of sociodemographic attributes. Defining the alternatives is a huge part of the modeling process. It is possible to define the model to depict the decision making process between person's income alternatives (ranging form low to high income levels) but also to decide if certain households may have a car available for the inhabitants.

Farooq et al. [7] deployed this method for deducing car ownership, income level, dwelling type and household education level of the population in order to gain a sound synthetic population. These characteristics were deduced one after another, as these do highly depend on each other. That means car ownership is highly depended on the income level of the population. The dwelling type is also highly depending on car ownership and income.

Anderson et al. [2] proposed an approach where they performed matching of household positions to a population with the aid of DCMs. In general synthetic populations do often lack information of dependencies between the individual persons. For instance such information is important with respect to the mobility behavior of the population. Anderson et al. [2] deployed discrete choice models for estimating the weights of a bipartite graph which corresponds to the problem of assigning persons best to household positions. Finally this association problem was solved by applying the Hungarian algorithm [9] which solves association problems s.t. the overall utility is optimized (i.e. the overall utility of the bipartite graph is maxed out). In the scope of their work, they assigned the head and spouse position within the households and concluded, that their approach may be extended in future work. Beside allocating head and spouse position, they proposed to also match positions such as children positions. Nevertheless due to the application of the Hungarian algorithm, his approach scaled poorly.

2.3. Markov-Chain-Monte-Carlo Simulation

Another approach to deduce single entity persons, recovered from the analyzed articles, was to apply a Markov-Chain-Monte-Carlo (MCMC) simulation based approach and Gibbs sampling for synthesis. In this approach partial joint distributions of person characteristics are processed. That means joint distributions of sets of characteristics are organized in a Markov-Chain, where transition probabilities correspond to the probabilities evident in the partly joint distribution samples. As the set of joint distributions gets larger, also the quality of different sampled populations gets closer to the real population. An element of the set of joint distributions can be the characteristics of "sex", "age" and "profession" conditioned upon a further characteristic, e.g. "martial status". By deploying Markov-Chains for population synthesis we also can take in consideration previous states of the chain (memory) and therefore create persons depending on previous states (i.e. characteristics) by randomly drawing persons from this chain. As a matter of this design the resulting population will approximately meet the marginal sums (as it was the case with IPF), but it also will closely meet the joint distribution of each of the input joint distribution statements.

Farooq et al. [7] introduced a MCMC simulation based approach for generating a synthetic population with four characteristics ("age", "sex", "household size" and "education level"). The aim of their work was to address the problem of having limited data input. Based upon cross-classification tables there were formulated constraints for the amount of each combination of a set of categories, i.e. each attribute of a category was conditioned upon the remaining ones. These constraints were recovered from the Swiss census, which also were used for evaluating the synthetic population with respect to the actual population on an individual level. Subsequently these constraints were processed by deploying Gibbs sampling for calculating the joint distributions of the set of attributes and their categories. Results of a generated population were presented, whereas the population differed in four attributes: age (eight categories), sex (two categories), household size (six categories) and education level (four categories). For evaluating the proposed approach, there was also generated a population which was derived from the marginal counts of the four attributes of a sample of the census, by deploying an IPF procedure. Considering the real-world population and the outcomes of the IPF-based synthetic population, Farooq et al. [7] concluded good quality of their proposed design. Subsequently the given conditionals were depleted, i.e. there was removed information related to "sex" within the attributes and population synthesis was redone. Tough information was missing, the resulting joint distribution of the generated population was marginally influenced by this lack of information and still greatly outperformed the population generated using the IPF-based approach. The results indicated there was a good quality of the resulting population, that furthermore almost met the real population of Switzerland, by checking the population against the actual microcensus data with respect to these four characteristics.

2.4. Geographic Information System

The preceding methods can be used for creating synthetic populations without any spatial reference. To deal with this limitation, approaches have proposed to deploy Geographic Information System (GIS) techniques for allocating the synthetic population a spatial entitiy. Layers with information on land-use and population density information have been intersected and thus created cells with different levels of attractiveness for assigning persons. By considering the different levels of attractiveness, single person entities can be assigned to spatial entities. The final synthetic population addresses the problem of being assigned to a spatial level, therefore allowing for more precisely depicting reality.

After generating households and assigning persons to those, Moeckel, Spiekermann & Wegener [14] allocated them to geographic coordinates. Therefore spatial data with information on population-density and land-use within zones was disaggregated. They deployed raster disaggregation [17], resulting in a raster representation of the investigated area. Subsequently households were randomly allocated to the cells whereas the number of persons within the cell (i.e. weight) was taken in consideration. Due to processing land-use as well as population-density data it was possible to further disaggregate the population specific information in the zone. Following this approach synthetic populations of Netanya (City) as well as of Dortmund were generated by Moeckel, Spiekermann & Wegener [14].

2.5. Statistics on Trip Chains

By applying the previous techniques it is possible to achieve a synthetic population which reassembles the real-world population with respect to sociodemographic characteristics and spatial information. For elaborating the population there are missing activities which persons may carry out. People carry out certain actives a day (e.g. “working”, “shopping”). In order to add these activities to daily plans of the population, there was presented a work, which deduced activities for the synthetic population by processing statistics on trip chains.

Adiga et al. [1] introduced an approach on how to assign activities and activity locations to the generated population. Each individual was assigned a set of activities, which were recovered from activity time-use survey data. Within the scope of this work, activity locations were finally spatially referenced by applying a gravity model, land-use data and further sources.

2.6. Summary

The approach of Farooq et al. [7] for generating a baseline population indicated good quality by processing cross-classification tables and applying discrete choice models for extending a given population by additional characteristics which were not given in first place. As it was shown by Anderson et al. [2] it is possible to deduce household compositions by deploying DCMs. A more auspicious application was made by Farooq et al. [7] where they assigned income, car ownership and education level by estimating parameters. These parameters were further used for adding the characteristics to the generated baseline population. Therefore for adding characteristics to populations which did lack these so far, it is advised to estimated parameters to add certain characteristics to the persons and households, as it was done by Farooq et al. [7]. The work of Anderson et al. [2] also gives an idea, on how to further deploy such models, but due to the fitting problem (described in Section 2.2) it turned out, this approach has poor scalability. Given a baseline population generated by deploying MCMC and having extended this population by optional characteristics by using DCMs, it is necessary to allocate these populations to a more exact spatial region. As Moeckel, Spiekermann & Wegener [14] could deduce more precise spatial information, it is advised to follow their approach but to extend it by more and more layers s.t. the population can be allocated more exactly.

The related work emphasizes the application of cross-classification tables for generating the baseline population, discrete choice models for deducing additional information (such as car ownership and income per capita) and GIS techniques for allocating the population a spatial context. None of the given approaches generated a synthetic population where all of these methods have been applied concurrently.

3. Technical Background

This chapter introduces the technical background of our work. Firstly we introduce MATSim models and their need for sound and detailed input data. Secondly the data sources, we base our design upon, are highlighted. Thirdly we introduce the data structure, which is applied for efficiently deducing templates. Subsequently we show the procedure of extrapolating our templates by deploying an algorithm for iteratively assigning extrapolation factors to those. Finally there are introduced discrete choice models for deducing further attributes and a procedure for allocating persons, families and households to spatial regions and moreover to geographical street addresses.

3.1. Multi-Agent Transport Simulation

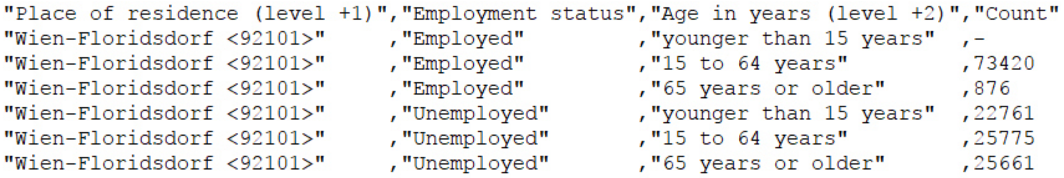
MATSim models can depict the mobility behavior of millions of agents on a high level of detail. As there are simulated persons on an individual level, it is possible to examine those also with respect to their socioeconomic profile. For bringing the simulation one step closer to reality, MATSim can deal with persons allocated to families and households. This offers the possibility for more conclusive evaluations of dependencies between family members with respect to their mobility behavior. In older versions of MATSim, such households were solely identified by the persons who are living in this household and the household income. The persons themselves differed in the characteristics of “age”, “sex”, “car ownership” and “employment status” [10]. Due to the limited characteristics of the persons and households, the evaluations, the models provided, also were not too conclusive. To overcome this shortcoming, in more recent versions of MATSim, there are introduced object attribute files, which allow for adding arbitrary information on persons, households, vehicles, network and anything related to this microscopic simulation environment. Nevertheless handling these additional information accordingly and depicting the impacts, needs to be implemented on its own and is not part of the out of the box functionality of MATSim.

3.2. Data Sources

Our design for generating a synthetic population is based upon four different types of data. Some of these data are not openly available, thus may lead to costs when considering for population synthesis. Firstly we introduce openly available cross-classification tables of the *Abgestimmte Erwerbsstatistik 2013* of Statistics Austria, which are processed for generating the baseline population [18] [19] [20]. Secondly we take a look at the EU-SILC household survey [21] conducted in 2014 with information on the living conditions of approximately 13,000 Austrian citizen. We used that to estimate additional attributes of the baseline population (e.g. income per capita) by deploying discrete choice models. Thirdly we introduce spatial layers with information on citizen characteristics in a grid of 250x250 meters ("sex" by "age" of the year 2015 and "sex" by "employment status" of the year 2013), which were taken from Statistics Austria [18]. Finally we introduce coordinates of all street addresses in Vienna (so-called address-coordinates - ACDs) with information on citizen count per address of the year 2015. This information was provided by a municipal department of the city of Vienna. The following subsections introduce these four data sources.

3.2.1. Statistics Austria Cross-Classification Tables

Statistics Austria provides information of certain statistics in an aggregated form. This information is available through the Statcube portal [18]. One of the provided statistics is the so called *Abgestimmte Erwerbsstatistik 2013* for persons [18], families [19] and households [20]. This statistics gives information on person's sociodemographic profile, such as job status, commuter information and the combination of arbitrary further attributes within a spatial context. Due to privacy concern reasons it is only permitted to query aggregated information of up to five dimensions. This means the user is for example free to query all the persons which do have the following characteristics: "Age is < 20, Job Status is employed, Hometown is Vienna 21^{st} district", and so on. A CSV export of a query, which holds information of three dimensions (i.e. characteristics) is depicted in Figure 3.1. Besides information on persons, the statistics also provides information on households and families in the same data format, s.t. it is processable in the same fashion as it is the case with person entities. The characteristics we processed in the scope of this work for persons, families and households are given in Appendix A, B and C. The counts given in the cross-classification tables are processed line by line as extrapolation factors, which are iteratively assigned to templates such that the citizenship of 2013 in *Floridsdorf* is generated.



```
"Place of residence (level +1)","Employment status","Age in years (level +2)","Count"
"Wien-Floridsdorf <92101>"     ,"Employed"         ,"younger than 15 years"  ,-
"Wien-Floridsdorf <92101>"     ,"Employed"         ,"15 to 64 years"         ,73420
"Wien-Floridsdorf <92101>"     ,"Employed"         ,"65 years or older"      ,876
"Wien-Floridsdorf <92101>"     ,"Unemployed"       ,"younger than 15 years"  ,22761
"Wien-Floridsdorf <92101>"     ,"Unemployed"       ,"15 to 64 years"         ,25775
"Wien-Floridsdorf <92101>"     ,"Unemployed"       ,"65 years or older"      ,25661
```

Figure 3.1.: Input csv data with sociodemographics of "Age-Employment Status".

These CSV exports are used for generating the synthetic population, stated in the subsequent chapters. In contrast to usual margin information, this type of information has a great advantage because it can take dependencies of up to four other margin information in consideration within a certain region (usually district or municipality). Therefore the connection of age of a person, its employment status, sex and field of education can be considered in the generation process. Virtually every category of the statistics can be queried and used in the synthetic population generation process.

Using this information, it is possible to create a set of conditionals, which should be, in the best case, full-conditional. That means, each of the sociodemographic attributes is conditioned upon each other sociodemographic attribute. This allows to deduce real-world populations with respect to the constraints given in the cross-classification tables and thus gain a population which fits the real-world population accordingly.

3.2.2. European Union Statistics on Income and Living Conditions (EU-SILC) Household Survey

This household survey was conducted in 2014 [21]. The participants were interrogated about their living conditions. The resulting survey data was assigned scaling factors, s.t. it represents the total of the Austrian citizenship. In the scope of this master's thesis, we used the given survey data and discrete choice models for estimating additional information for the generated population, which could not be deduced in the baseline creation process in first place. The data of this survey is accessible by researchers only and needs to be requested from Statistics Austria [21].

3.2.3. Sociodemographic Information Grid

For allocating the population to a more conclusive spatial region, we processed layers with information on certain demographics and combinations of demographics, as presented in Figure 3.2. This figure, for example, shows the counts of male persons younger than two years within some grid elements of 250x250 meters. There are more dense populated grid elements with respect to this sociodemographic pattern, but also less dense populated grid elements are given. According to the information given in the grid, the population was allocated s.t. the counts, given in the grids, were fit best with respect to the provided counts of demographic attributes. Liable to costs, these kind of information were acquired from Statistics Austria [18], whereas the costs highly depended on the actual type of data we wanted to process for generating a synthetic population. For our design we processed layers of the years 2014 ("Age by Sex") and 2015 ("Employment Status by Sex").

3.2.4. Coordinates of Street Addresses in Vienna (Address-coordinates - ACDs)

ACDs are the most detailed spatial unit, we processed in the context of synthetic population generation, as presented in Figure 3.3. This figure

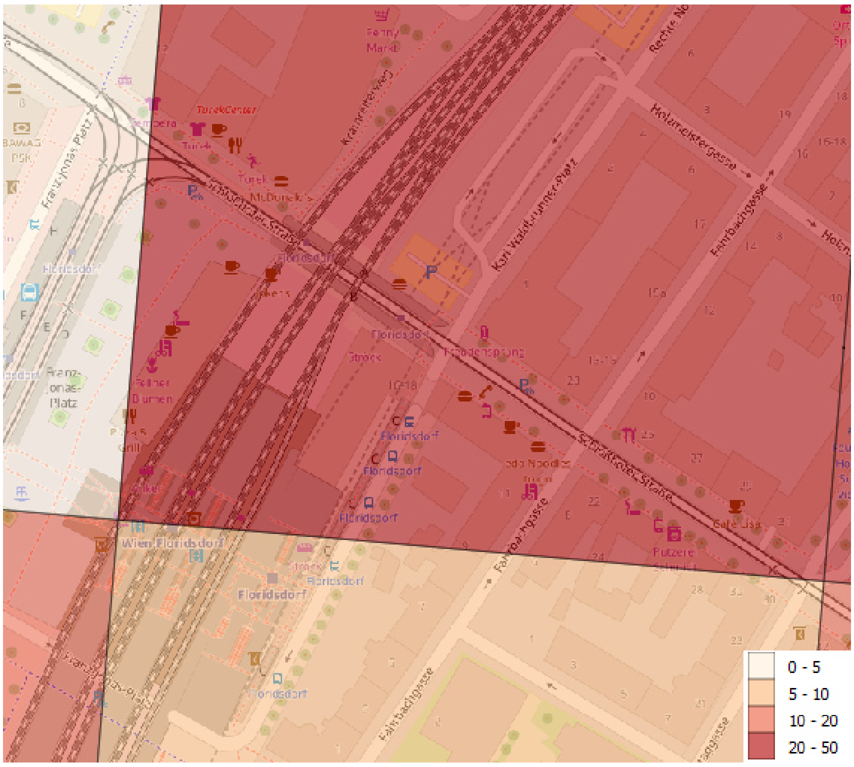


Figure 3.2.: Example of information given in a sociodemographic grid. In this figure, there are highlighted grid elements, with respect to the population-density of male persons younger than two years.

shows the spatial reference of the street-addresses and the markers are colored with respect to the inhabitants count given per ACD. Liable to the overlaying grid, the persons were matched to the grid and subsequently assigned to single ACDs with respect to the counts given in the grid as well as regarding the counts of inhabitants per ACD. Thus resulting in a population of the 21^{st} district of Vienna, *Floridsdorf*. This data source was acquired from a municipal department of the city of Vienna (*Magistratsabteilung 18 - Stadtentwicklung und Stadtplanung*) and depicts the population of 2015.

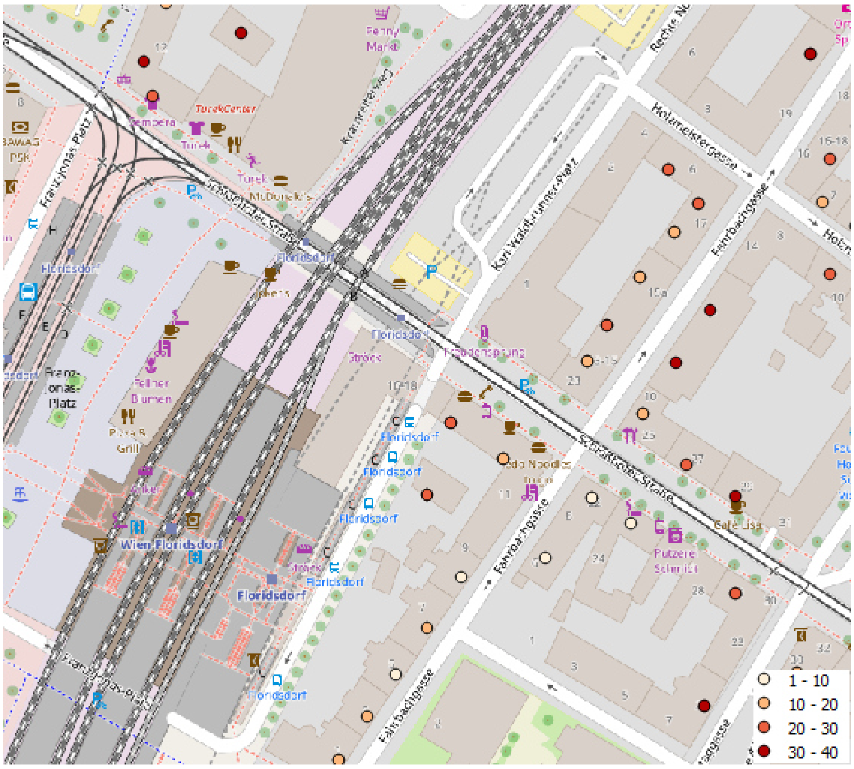


Figure 3.3.: Coordinates of street addresses (ACDs) with inhabitants count, provided by the municipal department of the city of Vienna (*Magistratsabteilung 18 - Stadtentwicklung und Stadtplanung*). Each marker corresponds to a street address and is colored with respect to the inhabitants count.

3.3. Data Structure

For efficiently deducing and extrapolating templates, there is the need for an appropriate data structure. The data structure is responsible for generating templates with a set of possible combination of all sociodemographics. Infeasible templates need to be neglected in the generation process, as the computational and also the memory overhead may become too big. Infeasible templates are templates which hold a combination of sociodemographics that are, according to the information present in the cross-classification tables, occurring zero times. Examples of infeasible templates are given in Figure 3.10. Template 1 and template 3 are marked to be infeasible. This is due to the fact, that persons in Austria which are younger than 15 years are not allowed to be employed (specific for both genders). This information is given in the input CSV file in Figure 3.1, where it is stated, there are no employed persons younger than 15 years.

In the following subsections we demonstrate all the parts of our data structure. Firstly we show the hierarchical tree structure, which we implemented for deducing all possible sets of characteristics. These characteristics correspond to actual possible templates of persons, families and households. Secondly we illustrate the concept of wildcards in our tree structure, which we used for efficiently allocating extrapolation factors, given in the input data, to all affected templates. Finally we address the problem of encoding the generated templates, such that we can subsequently efficiently recover templates by their code. We need this to properly assign all templates to the corresponding margin constraints, which holds counts on sets of characteristics as given in the input cross-classification tables. As soon as the templates are allocated to the appropriate margin constraints, the templates iteratively get assigned extrapolation factors.

3.3.1. Hierarchical Tree

For deducing possible templates for subsequently extrapolating those with multiple extrapolation factors, we introduce a hierarchical tree data structure. The purpose of this tree is to generate all possible combinations of characteristics (i.e. templates). Each level of the tree corresponds to a characteristics, such as “sex”, “age” and “employment status”. Each of the characteristics has a list of attributes. The characteristics of “sex” may be “male” and “female”, for “age” there may be “< 15 years”, “$15 - 64$ years” and “≥ 65 years” and for “employment status” there may be “employed” and “unemployed”. To indicate the count of attributes per characteristics we illustrate the characteristics as “sex” (two categories), “age” (three categories) and “employment status” (two categories). With each additional level, respectively characteristic, the tree (i.e. the count of leafs) grows exponentially. As this induces memory as well as runtime issues, the tree structure is responsible for generating feasible templates only. This is done by truncating tree branches by eliminating infeasible branches as soon as possible. Algorithm 1 demonstrates how all possible templates for extrapolating are deduced.

```
Data: Cross-Classification Tables
Result: List of Weighted Feasible Templates

List<MargCons> marginConsList = parseAllMarginConstraints();
List<Template> listOfTemplates;

// Identify all sociodemographics/attributes
Map<String,List<String>> socMap;
socMap = parseInputFiles();

// For each sociodemographics, add a “wc” entry
addWcEntryToAll(socMap);

// Initialize EncDecoder
Encoder enc = new Encoder(socMap);

// Identify and encode all infeasible combinations of attributes
List<Long> infList = enc.encodeInfeasiableTemplates();

// GeneratePermutations creates all possible combinations
GeneratePermutations(socMap, enc, listOfTemplates, infList);

Extrapolator ext = new Extrapolator(marginConsList,listOfTemplates);
ext.allocateTemplatesToConstraintsByEncoding(enc);
ext.performExtrapolation();
```

Algorithm 1: Algorithm for Deducing feasible Templates for Extrapolating

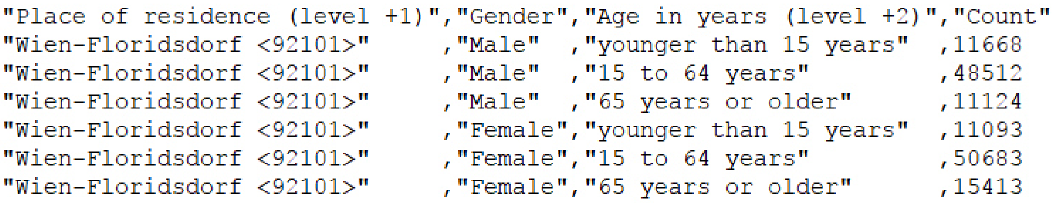
For setting up a comprehensive tree with the introduced Algorithm 1 which generates all of characteristics, the following tasks are explained in more detail.

- Identify all sociodemographics (i.e. levels of the tree) and their attributes (e.g. sex (two categories), employment status (two categories), age(three categories)) in the input CSV files.

- Identify all combinations of characteristics, which occur zero times (i.e. identify all infeasible sociodemographic combinations)
- Introduce an additional attribute in each each sociodemographic category, the so-called wildcard entry.
- Generate all possible combinations of demographics level by level by considering infeasible characteristics combinations
  - For each level check if the current partial set of sociodemographics is feasible according to the identified templates.
  - As soon as a full set of feasible sociodemographics is identified encode this template (i.e. the leaf of the tree) using a custom encoding introduced in Section 3.3.3.
- Keep in mind: Due to checking sets of characteristics in the generation process level by level, we can eliminate infeasible templates as soon as possible, thus slimming the set of possible templates and also the computational time.
- For each margin constraint on sets of characteristics, we assign templates accordingly.
- Using the iterative procedure introduced in Section 3.4, we iteratively assign extrapolation factors. A small example of iteratively assigning extrapolation factors is given in Table 3.1 based upon the input data given in Figures 3.4, 3.6 and 3.8.

Each of the resulting tree leafs corresponds to a possible template, with a value for each characteristics. The idea of this data structure is to eliminate branches of the tree as soon as possible in the generation process. Eliminating infeasible templates is done by exploiting the encoding of the templates, where it is checked if the encoding contains the pattern of an infeasible combination of characteristics.

For synthetic population generation, aggregated Statistic Austria information (cross-classification tables) for persons [18], families [19] and households [20] are processed in trees, where each level corresponds to a characteristics of the population. The design allows to introduce further levels (i.e. characteristics), as a more detailed population are required. Examples of simple trees are given in Figures 3.5, 3.7 and 3.9. The corresponding Statcube input CSV files are given in Figures 3.4, 3.6 and 3.8. These trees and CSV files give an example of a simple abstraction based on the Statistics Austria data. These trees can be generated from the corresponding input cross-classification tables without having to deploy the introduced hierarchical data structure. Nevertheless if we want to create persons with all of the three given sociodemographics, we need to “merge” those trees. Furthermore we need to deduce templates and subsequently extrapolate those. The corresponding merged tree is given in Figure 3.10. The leafs, which are marked as in the color red, are those we neglect in the procedure of generating templates, as those are certainly not present in the final generated population. Figure 3.11 shows the merged tree with additional wildcard entries. The entries are necessary, such that we can also apply the encoding for margin constraints.



```
"Place of residence (level +1)","Gender","Age in years (level +2)","Count"
"Wien-Floridsdorf <92101>"     ,"Male"  ,"younger than 15 years"  ,11668
"Wien-Floridsdorf <92101>"     ,"Male"  ,"15 to 64 years"         ,48512
"Wien-Floridsdorf <92101>"     ,"Male"  ,"65 years or older"      ,11124
"Wien-Floridsdorf <92101>"     ,"Female","younger than 15 years"  ,11093
"Wien-Floridsdorf <92101>"     ,"Female","15 to 64 years"         ,50683
"Wien-Floridsdorf <92101>"     ,"Female","65 years or older"      ,15413
```

Figure 3.4.: Input csv data with sociodemographics of “Age-Sex”.

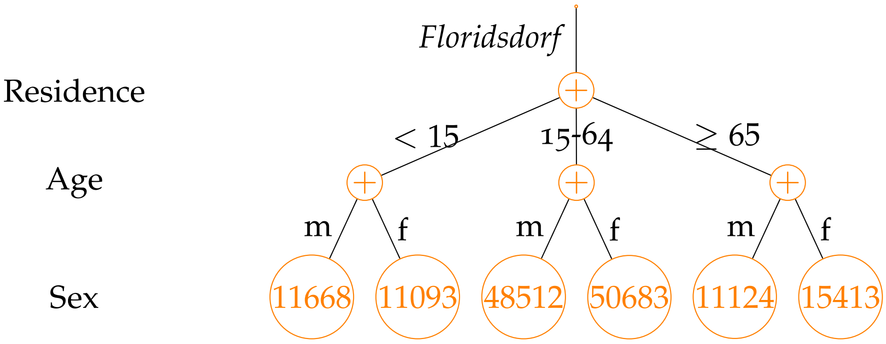
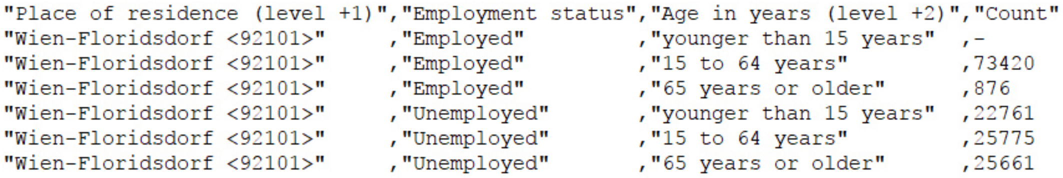


Figure 3.5.: Tree structure of input csv data “Age-Sex”.

3.3.2. Wildcards in Tree Structure

Each template has a complete set of values for their characteristics. Anyway the input margin constraints given in the cross-classification tables lack



```
"Place of residence (level +1)","Employment status","Age in years (level +2)","Count"
"Wien-Floridsdorf <92101>"     ,"Employed"         ,"younger than 15 years"  ,-
"Wien-Floridsdorf <92101>"     ,"Employed"         ,"15 to 64 years"         ,73420
"Wien-Floridsdorf <92101>"     ,"Employed"         ,"65 years or older"      ,876
"Wien-Floridsdorf <92101>"     ,"Unemployed"       ,"younger than 15 years"  ,22761
"Wien-Floridsdorf <92101>"     ,"Unemployed"       ,"15 to 64 years"         ,25775
"Wien-Floridsdorf <92101>"     ,"Unemployed"       ,"65 years or older"      ,25661
```

Figure 3.6.: Input csv data with sociodemographics of "Age-Employment Status".

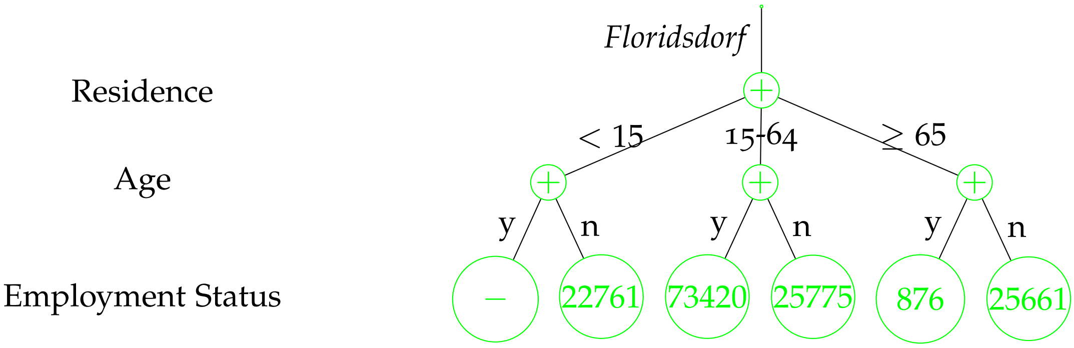
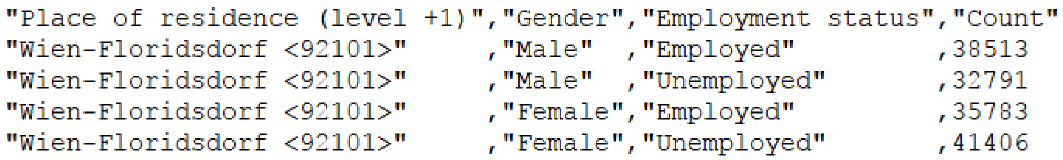


Figure 3.7.: Tree structure of input csv data "Age-Employment Status".

this completeness. Nevertheless to be able to encode the margin constraints accordingly, we introduce a wildcard "wc" entry. The "wc" entry indicates that this margin constraint affects all the outcomes of this characteristics the same.

Each attribute set of the sociodemographics is extended by a so-called wildcard entry ("wc"). Thus each sociodemographic has at least a set of two members. As mentioned before, Statistics Austria forbids to query combinations of more than five sociodemographics. Thus we actually can not deal with a synthetic population with more than five characteristics in first place. To overcome this limitation, we introduce wildcard members for each of the sociodemographic categories. This means each template, on a certain branch, is assumed to be affected by the same extrapolation factors, liable to the input margin constraints on sets of sociodemographic combinations. To indicate the impact on the tree creation, there is given an example in Figure 3.11. The wildcard member is always assumed to be the first of the members and is thus encoded in a special manner, as shown in Section 3.3.3. Due to the fact that each set within the input cross-classification



```
"Place of residence (level +1)","Gender","Employment status","Count"
"Wien-Floridsdorf <92101>"     ,"Male"  ,"Employed"         ,38513
"Wien-Floridsdorf <92101>"     ,"Male"  ,"Unemployed"       ,32791
"Wien-Floridsdorf <92101>"     ,"Female","Employed"         ,35783
"Wien-Floridsdorf <92101>"     ,"Female","Unemployed"       ,41406
```

Figure 3.8.: Input csv data with sociodemographics of “Sex-Employment Status”.

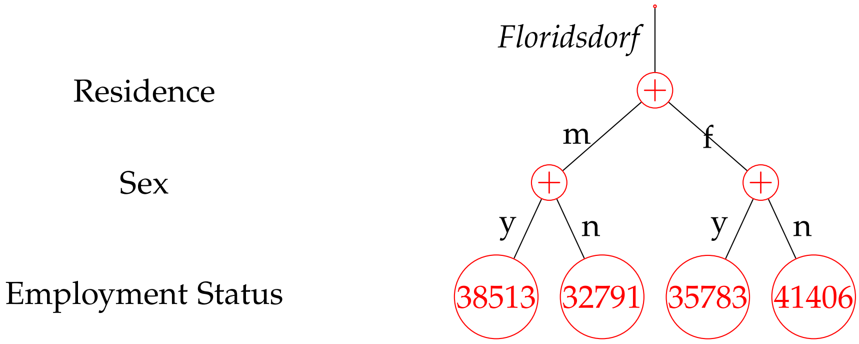


Figure 3.9.: Tree structure of input csv data “Sex-Employment Status”.

tables is limited to five dimensions, the wildcard entry allows to encode the input margin constraints given in the cross-classification tables. Tough these margin constraints do in fact lack information for all combinations of characteristics. That means if we do have a margin constraint on the characteristics of “sex by age”, tough we have templates with characteristics on “sex by age by employment status”, we can encode the margin constraint as “sex by age by all kinds of employment status” in a specific manner. In the template generation process we dismiss all templates which contain a wildcard entry, as we are only interested in templates with values for each characteristics.

3.3.3. Encoding/Decoding of Templates

Each template which is created using the data structure introduced in Section 3.3.1 is assigned a unique identifier. This identifier is a 64-bit Integer value. Intrinsically this value holds all sociodemographic information on the current template. The encoding is introduced to allow the creation

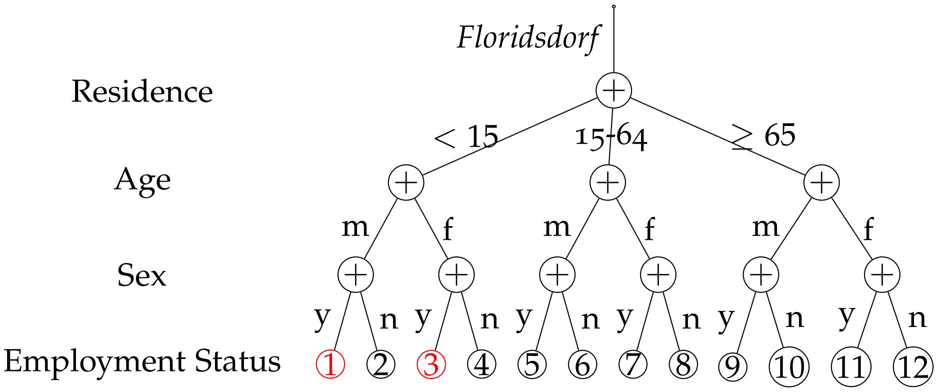


Figure 3.10.: Complete tree with neglectable templates.

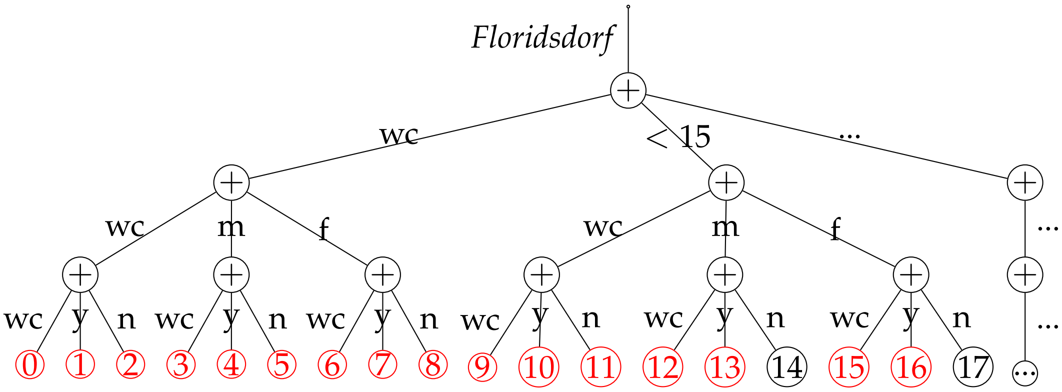


Figure 3.11.: Complete tree with neglectable templates. The additional attributes for each category "wc" which allows for adding wildcards in tree structure and moreover in the encoding scheme for subsequently extrapolation of the templates.

of templates in less time and make it easier to allocate the templates to appropriate margin constraints for extrapolating. Each of the characteristics of the input data is encoded following equation 3.1. The sociodemographic categories have a limited set of attributes. Depending on the amount of the attributes and considering the hierarchical order, which we assume to be arbitrary for now, we can calculate a unique identifier for each composition of characteristics of templates. An important part takes the wildcard stated in the preceding section. Because it was not possible to query more than five dimensions at once at Statistics Austria Statcube portal, it is necessary to add wildcard entries. Therefore we introduced the data structure for efficiently generating templates, which we subsequently use for iterative extrapolation. A tree which implements wildcard entries is given in Figure 3.11. The encoding of each leaf and node of the tree can be calculated by applying Formula 3.1, whereas nodes need to be encoded by assuming the lower levels of sociodemographics as wildcard entries.

The encoding is also important when it comes to the runtime of the design. Because we need to allocate the templates to the appropriate margin constraints in the extrapolation step of the design. This procedure highly depends on the Set specific operation "retainAll", which basically performs an inner join on two sets of templates. This join operation is significantly faster if there is given a custom implementation of the hash method based on the encoding.

A simple encoding of a person template is given in Figure 3.12. The decoding of this template is given in Figure 3.13.

Step(age): 9	wc									< 15								
Step(sex): 3	wc			m			f			wc			m			f		
Step(emp): 1	wc	y	n	wc	y	n	wc	y	n	wc	y	n	wc	y	n	wc	y	n
Code(x,y,z)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

Figure 3.12.: Encoding scheme with custom masking by applying Formula 3.1.

$$\begin{aligned}Code('' < 15'', ''m'', ''n'') =\\ Index('' < 15'') * Step(age) +\\ Index(''m'') * Step(sex) +\\ Index(''n'') * Step(emp) = 14\end{aligned} \tag{3.1}$$

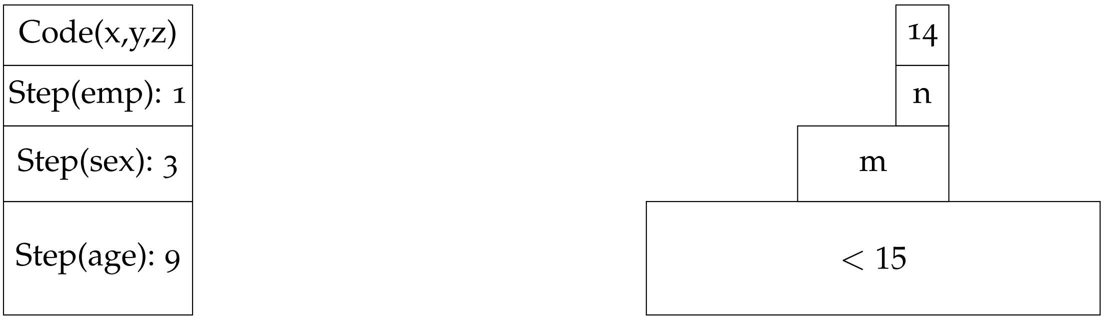


Figure 3.13.: Decoding scheme by applying Formula 3.2.

$$\begin{aligned}age.get(14/Step(age)) = '' < 15''\\ sex.get((14\%Step(age))/Step(sex)) = ''m''\\ emp.get(((14\%Step(age))\%Step(sex))/Step(emp)) = ''n''\end{aligned} \tag{3.2}$$

3.4. Iterative Extrapolation of Templates

Templates, which are generated using the methods and data presented in the precedent sections, are processed such that each template is assigned a set of extrapolation factors. These extrapolation factors represent the weight which is approximately given in the area on a set of appropriate sociodemographic attributes. For extrapolating these templates with the values given in the input CSV files, on the one hand there is the need to assign one extrapolation factor to several templates and on the other hand to assign multiple extrapolation factors to one template, as shown in Table 3.1. In this table there is highlighted how the extrapolation of the templates, derived from the tree given in Figure 3.10, is done for ten iterations. For each template there are shown the affected margin constraints of the partial trees, given in Figures 3.5, 3.7 and 3.9 according to the appropriate color scheme. For handling this issue, we used a method for extrapolating introduced by Otterstätter in his PhD thesis [16]. He extrapolated trip chains by multiple factors such as multiple stationary counting stations. We used a simplified version, where we neglected time-dependencies and replaced trip-chains by templates. The simplified version of this algorithm is presented in this section. The method iteratively weights a list of objects (i.e templates) according to their extrapolation factors.

This procedure iteratively assigns extrapolation factors on sets of templates. Therefore we gain a set of templates with extrapolation factors which meet the input counts given in the the cross-classification better and better by each iteration.

$$c(i,x,j) = \frac{q_Z(j)}{q_T(i,j)} \tag{3.3}$$

$$C(i,x) = \frac{1}{n(x)} * \sum_{j=1}^{n(x)} c(i,x,j) \tag{3.4}$$

$$q_T(i,j) = \sum_{x=1}^{N} w(i,x) \tag{3.5}$$

#	$SF_{i=0}$	Scaling Factors			SF_{i+1}	SF_{i+10}
1	1	A=11,668	**G=0**	M=38,513	6,223	150
2	1	A=11,668	H=22,761	N=32,791	9,381	11,397
3	1	B=11,093	**G=0**	O=35,783	5,824	83
4	1	B=11,093	H=22,761	P=41,406	10,243	11,361
5	1	C=48,512	I=73,420	M=38,513	24,601	37,704
6	1	C=48,512	J=22,775	N=32,791	15,524	10,478
7	1	D=50,683	I=73,420	O=35,783	24,659	35,816
8	1	D=50,683	J=22,775	P=41,406	16,843	13,683
9	1	E=11,124	K=876	M=38,513	6,279	534
10	1	E=11,124	L=25,661	N=32,791	9,774	10,562
11	1	F=15,413	K=876	O=35,783	6,690	557
12	1	F=15,413	L=25,661	P=41,406	11,446	15,162

Table 3.1.: Scaling factors per template for ten iterations by applying method introduced in Section 3.4. The color indicates the origin of the input data, taken from the trees given Figures 3.5, 3.7 and 3.9.

$$W(i,x) = c(i-1,x) * w(i-1,x) \tag{3.6}$$

Subsequently the Equations 3.3, 3.4, 4.2, 3.6 are described in detail.

- $c(i,x,j)$
  - Correction factor for template x in iteration i on margin constraint j.
- $q_Z(j)$
  - Target Value for margin constraint j.
- $q_T(i,j)$.
  - Sum of all weights on margin constraints j at iteration i.
- $C(i,x)$
  - Average correction factor at iteration i for template x.
- $n(x)$
  - Count of margin constraints, which the template x is affected by.

- N
  - Count of all given templates.
- $w(i, x)$
  - Weight of template x at iteration i.

3.5. Discrete Choice Models

Discrete choice models are a widely used method for modeling decisions on an individual level. They can cope with information of the decision making person itself, the decision situation and in the context of synthetic population generation also influencing characteristics of e.g. the household which this decision maker is part of. The decision maker can choose from a fixed set of alternatives, the so-called choice set. In order to describe the decision making process, an appropriate discrete choice model can be defined. The given choice set does have to meet the following criteria according to Train [23]:

- The decision maker chooses only one alternative from the choice set.
- The choice set must be exhaustive and all possible alternatives must be included. The decision maker necessarily chooses one of the given alternatives.
- The number of alternatives must be finite.

The main idea behind the concept of discrete choice models is that the decision maker is assumed to choose the alternative which grants him the highest utility. Therefore it is evident to define a utility function for each alternative, which defines the utility of an alternative depending on observed parameters. Train [23] defines utility as a constructed measure of well-being, which has no scale.

The utility $\mathrm{U}_{\mathrm{alt}}$ of each alternative for a specify decision maker is given by its utility function, as given in Equation 3.7.

$$U_{\text{alt}} = V_{\text{alt}} + \epsilon_{\text{alt}} \tag{3.7}$$

The utility function is composed of the deterministic utility V_{alt} as well as of the stochastic error term ϵ_{alt}. V_{alt} holds all variables which are used in the regression equation, to explain the utility of this alternative. The error term ϵ_{alt} describes all the bias that can not be explained by the variables in the deterministic utility in first place. We highlight the concept of discrete choice models by defining a simple model, which we use for estimating the impact of characteristics of "age" and "sex" on "deciding" among different household sizes. Considering our example model for estimating household sizes, we subsequently define deterministic utility functions of a discrete choice model, where the decision maker may choose among three different alternatives (i.e. household sizes the decision maker may choose to live in). The decision, in the context of our model, is assumed to be depending on "sex" and "age" of the person, which chooses among those alternatives. For demonstration purpose there is given an example where the decision maker may choose among three different household sizes to live in. This model describes where people tend to live, with respect to the size of their household, solely by the influencing characteristics of "sex" and "age". Certainly these two influencing characteristics may not allow to reassemble the complex decision making process behind peoples preferences, on where they prefer to live. Deducing the decision making process to parameters, greatly neglects individual preferences of persons, such as a potentially anti-child attitudes. For the sake of simplicity household sizes larger than size three are neglected in the scope of this simple example model.

$$V_{\mathrm{HH1}} = ASC_{\mathrm{HH1}} + Age * \beta_{\mathrm{Age}} + Sex * \beta_{\mathrm{Sex}} \tag{3.8}$$

$$V_{\mathrm{HH2}} = ASC_{\mathrm{HH2}} + Age * \beta_{\mathrm{Age}} + Sex * \beta_{\mathrm{Sex}} \tag{3.9}$$

$$V_{\mathrm{HH3}} = ASC_{\mathrm{HH3}} + Age * \beta_{\mathrm{Age}} + Sex * \beta_{\mathrm{Sex}} \tag{3.10}$$

The type of the discrete choice model which is deployed by defining the above example (by Equations 3.8, 3.9 and 3.10) is the so called multinomial logit model [8]. This model can deal with more than two different alternatives to choose among and defines the stochastic error term ϵ_{alt} to be Gumbel distributed [23]. The reason for introducing ϵ_{alt} is to make decisions stochastic, rather than choosing an alternative with the highest utility, although another alternative might have almost the same utility. Therefore the probabilities to choose alternatives among are depending on the difference between the alternative specific utilities, rather than the overall magnitude of those. The probability within a multinomial logit model is defined as given in Equation 3.11. Besides multinomial logit models there are also further models, which can be applied. However in the context of this work, we chose the multinomial regression model. For interested readers we recommend further literature [3], [23], [12] where other different models are described.

$$P_{\mathrm{alt}} = \frac{e^{V_{\mathrm{alt}}}}{\sum\limits_{i=1}^{n} e^{V_{\mathrm{i}}}} \tag{3.11}$$

Given the fact that each decision maker is assumed to have chosen the alternative which grants him the highest utility and considering Equations 3.8, 3.9 and 3.10, for each alternative and Equation 3.11, it can be estimated the influence of each β (age and sex) parameter and alternative specific constant (ASC) in the utility functions on the input values. These equations and an input data sample of EU-SILC or similar household survey data can be used for estimating the magnitudes of the parameters in the utility functions and therefore for further generation of unknown information in the process of creating a synthetic population. This method will be applied to our synthetic population, which is deduced from Statistics Austria [18] data, in order to add information on incomes per capita and car ownership per household, as this information is not present in the cross-classification tables of Statistics Austria in first place. The defined set of equations is not deterministic solvable, but the unknown parameters may be estimated with BIOGEME [4]. This open source software can perform a maximum likelihood estimation of parametric models. We deployed BIOGEME to estimated and evaluate values for our discrete choice models, which depict the values given in data samples of the EU-SILC household survey.

As this system of inequalities is usually not deterministic solvable, we apply the software BIOGEME for performing a Maximum Likelihood Estimation in order to find very likely values for the parameters. By using the discrete choice models, we address the problem of revealing connections of sociodemographic data in order to deduce further sociodemographic data.

The example of estimating household sizes regarding "age" and "sex", results in the estimated parameters shown in Table 3.2. First of all it is evident to check the level of significance, which is given by evaluating the t-stat and the p-value. As the p-value equals zero, there is certainly a dependency on age when it comes to decision making among household sizes. The t-value emphasizes the impact of the parameters. Especially the age parameter (specific to choosing households of size two) has a huge influence on the decision making process according to the t-stat value. For a more detailed explanation of the t-stat and p-value we recommend further literature [13] [25]. The parameter for "sex" shows that households for females are both approximately same likely for two and three persons in the household. Indeed according to the estimation, women mostly tend to live on their own (least negative SEX_HH value for households of size one). Furthermore the age parameter shows that younger people are more likely to live in bigger households. This certainly corresponds to real-world circumstances, as usually infants and kids can not live on their own in Austria.

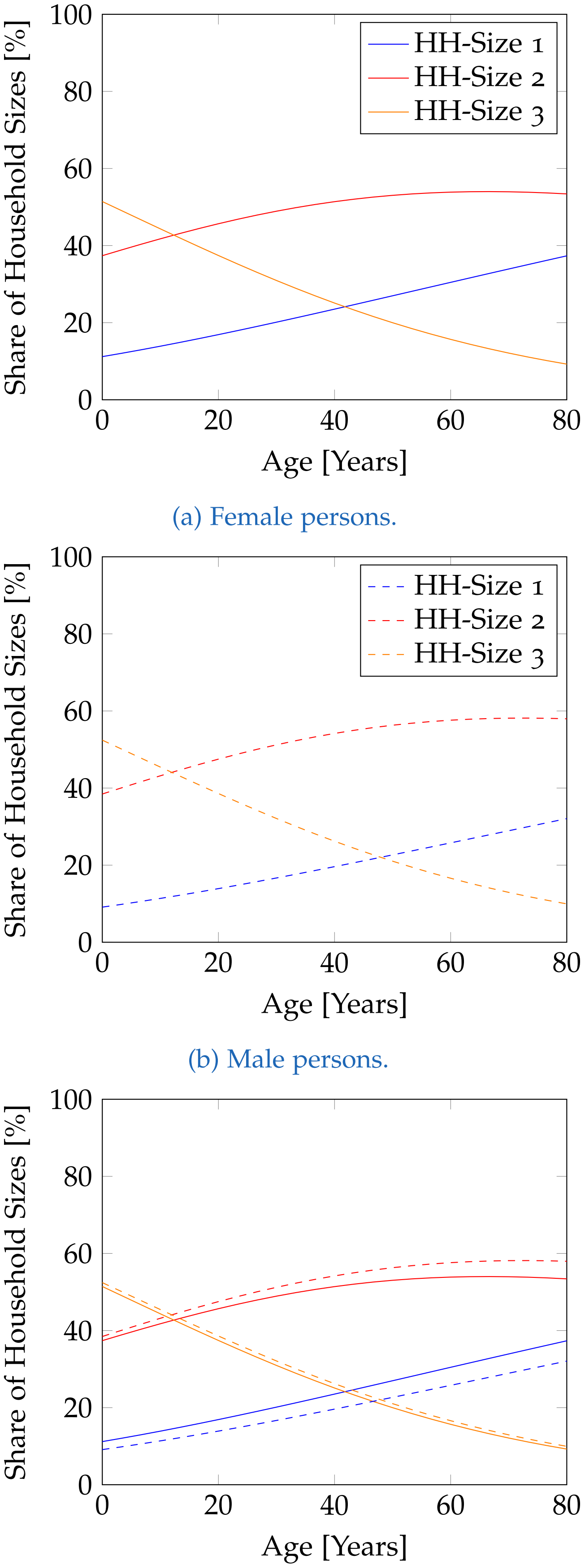
The parameters correspond to our a-priori assumption with respect to signedness of the parameters as well as the ordering of the magnitudes of the parameters. The outcome of the estimated parameters is shown in Figures 3.14a, 3.14b and 3.14, where it is evident that the parameters for "age" and "sex" have indeed a huge impact on choosing a certain household size. In contrast to male persons it is more likely for female persons to live alone as they become older. Male persons are living in households of two persons more often. For both genders it is equally likely, especially at young ages, to life in households of size three. In general households with the size of one or two persons become more evident when people get older.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	*t*-stat	*p*-value
1	AGE_HH2	-0.0106	0.00170	-6.21	0.00
2	AGE_HH3	-0.0365	0.00221	-16.52	0.00
3	ASC_HH2	1.44	0.106	13.53	0.00
4	ASC_HH3	1.75	0.126	13.91	0.00
5	SEX_HH2	-0.234	0.0542	-4.32	0.00
6	SEX_HH3	-0.225	0.0733	-3.07	0.00

Table 3.2.: Estimated parameters for the household size model for deciding among household sizes with respect to age and gender based upon household survey data of EU-SILC. The household of size one is fixed and therefore has a fixed utility of zero.

3.6. Spatial Allocation Procedure

The procedure of allocating the population to spatial regions is done by randomly choosing a grid element. Each grid element holds a count of persons, which live there with respect to certain characteristics. For each household there is randomly chosen a grid element and subsequently it is determined if the household and inhabitants fit the possible sociodemographic profile of the grid element. If the household fits the grid in this manner, the household is allocated. In addition the grid, more precisely the demographics count of the persons attribute combination, is decremented by the sociodemographic profiles of the household members. In case the household does not fit the grid, there is chosen another grid and the procedure is repeated. If there is not found any grid to place the household in 100 tries, the household is not allocated at all. This procedure for allocating households is highly related to the procedure introduced by Spiekermann & Wegener [17], where there are placed persons with respect to the overall population density given within a spatial area.



(a) Female persons.

(b) Male persons.

Figure 3.14.: Household sizes for female (solid lines) and male (dashed lines) persons.

4. Technical Realisation

In this chapter it is shown how the introduced data sources, data structure, extrapolation algorithm and finally the method for assigning the population to spatial coordinates (introduced in Chapter 3) are applied for creating an actual synthetic population of the 21^{st} district of Vienna, *Floridsdorf*. The flowchart in Figure 4.1 gives an overview on the developed design, ranging from the creation of the baseline population, allocation of the population to families and households, refining the population and finally allocating the created households to a spatial level.

Firstly we demonstrate the creation process of the baseline population where there are deduced templates, which are iterative extrapolated and sampled such that we get person, family and household objects. The person objects are allocated to families and to households. Secondly we refine the given population with information on income per capita and car ownership per household. Finally we allocate the households and as a result also the persons and families to a spatial region to get the synthetic population of the 21^{st} district of Vienna, *Floridsdorf*.

4.1. Generation of the Baseline Population

For generating the baseline population of the 21^{st} district of Vienna, we apply the input data and process it in the data structure as described in Chapter 3. The resulting templates, i.e. the feasible templates, for persons, families and households are assigned extrapolation factors, which correspond to the counts in the cross-classification tables provided by Statistics Austria [18] [19] [20].

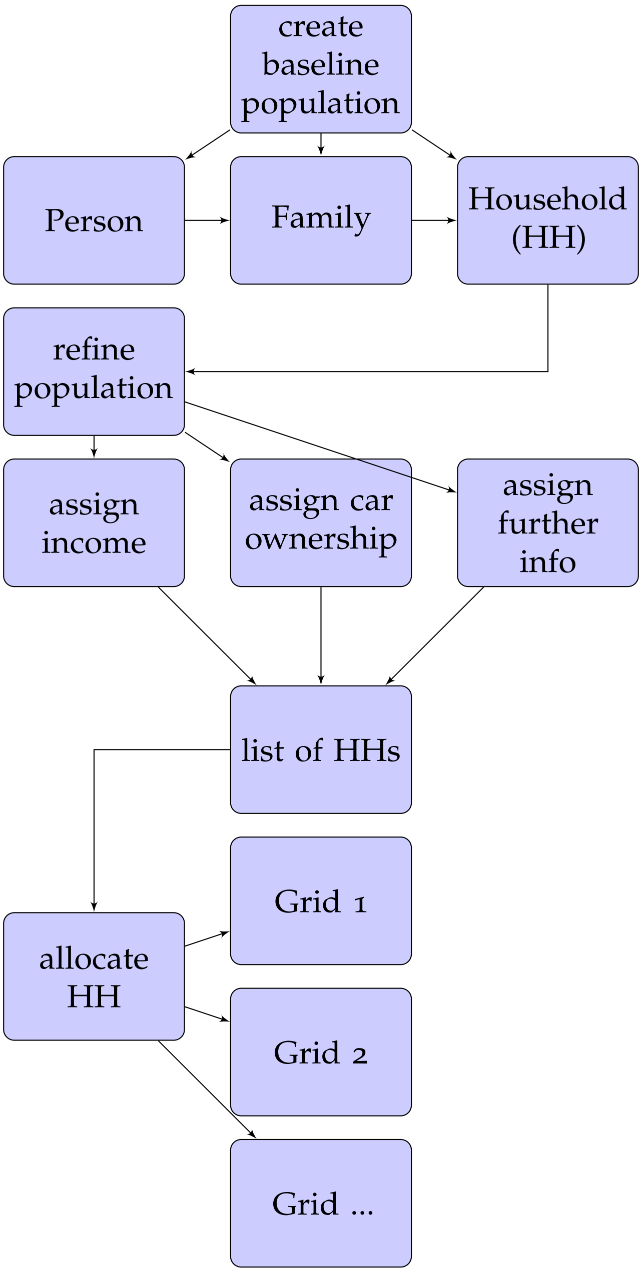


Figure 4.1.: Flowchart for generating synthetic populations.

4.1.1. Deduce all Possible Templates from the Data Structure

The data structure is responsible for creating all possible combinations of sets of sociodemographics. This is important because the amount of templates would increase exponentially regarding the sociodemographics counts, we want to implement in our population. For each additional sociodemographic we introduce, the template count would be multiplied by the number of attributes of the additionally introduced sociodemographic category. This would lead to a fast growing number of templates, which is clearly out of a efficiently processable scope.

Parsing Sets of Sociodemographics

For initializing the data structure accordingly, we parse all the characteristics and attributes presented in the input CSV files. For generating all possible sets of characteristics, i.e. templates, it is mandatory to create all permutations level by level. Each time another level is created, it is checked, if the template is feasible so far. If the template is infeasible, like the combination of characteristics is not possible so far, the template is dismissed. To identify such infeasible combinations, it is mandatory to parse the input files and save those combinations, which occur zero times in the real-world population according to the input files. Such kind of characteristics sets are often related to underage persons, as those are usually not employed or married, so that we can dismiss all the subsequent templates afflicted by this combination of sociodemographics. As the pool of characteristics is extended, this procedure is inevitable as the inflicted problem, i.e. the vast amount of templates, causes serious memory and runtime issues.

Hinting missing Cross-Classification Combinations

As mentioned in the preceding section, it is important to provide the design with information on infeasible combinations of sociodemographics. This information is vital such that the generation process of templates is speed-up, by efficiently dismissing irrelevant templates, as soon as possible. Therefore we keep the set of feasible templates as small as possible.

Once the input data is processed and parsed, all permutations of characteristics and attributes given in the input CSV files are evaluated and potential missing combinations are printed. The implementation is designed to work with cross-classification tables, with information on counts of sets of characteristics given in the real population. Within these tables there are conditioned a maximum of five characteristics upon each other. Nevertheless for creating a synthetic population with more than five characteristics, it is advised to provide the design with a list of cross-classification tables, where all characteristics are conditioned upon all permutations of size four of the other characteristics. Once this is achieved, the input data is said to be full-conditioned and will lead to best possible results in the process of population synthesis (without introducing additional third party data sources). However the design generates a population if there are not given full-conditioned input data, but the runtime for generation as well as the resulting population may lead to unsatisfiable outcomes. Thus this design hints all the combinations of characteristics. This on the one hand improves the results of the population generation and on the other hand significantly speeds-up the generation, as the runtime highly depends on the amount of templates, which need to be processed in the iterative extrapolation step.

Allocating Templates to Extrapolation Factors

Each of the input cross-classification tables holds counts on sets of characteristics, which we refer to as margin constraints. For iteratively assigning extrapolation factors, it is necessary to allocate all templates to the appropriate margin constraints. Each template is affected by a set of margin constraints and in turn each margin constraint has a set of templates. The count given in the margin constraint must be met by the sum of extrapolation factors of the templates in its scope. Once all the templates are assigned, the method shown in Section 3.4 is applied for iteratively assigning extrapolation factors, such that the margin constraints are met best.

For allocating templates to appropriate margin constraints, we start with the master set of templates, where all template objects are contained. For each characteristics (at maximum five characteristics) of the margin constraint we fetch the subset (by deploying “retainAll” method call) on the master set and put it in a map with the encoding illustrated in the preceding chapter. For efficiently generating subsets, we introduced a custom hash representation of the template objects in form of the encoding, such that the operation could be significantly accelerated. Subsequently the recovered subset is used for recovering another subset of the next characteristic of the margin constraint. This procedure is repeated as often as the amount of characteristics in the margin constraints (at max five times). Once there is recovered the final set of templates, which is affected by this margin constraint, we allocate this set of templates to this margin constraint. For the next margin constraint, there is a look-up in the map first, to possible recover partly sets of demographics, for parts of the set of demographics, to significantly speed-up the allocation process. This map serves as a cache, which holds all the recent fetched sets of characteristics combinations (also partly ones).

4.1.2. Assign Extrapolation Factors to Templates

Once the templates are allocated to all appropriate margin constraints, the algorithm of iteratively assigning extrapolation factors takes place. As it is shown in Figure 4.2 the sum of deviations, i.e. the sum of differences of all margin constraints to the sum of extrapolation factors of the afflicted templates, decreases significantly in the first 100 iterations and seems to converge to a minimum sum of deviation at value zero as the algorithm performs more and more iterations.

The *SumOfDeviation* per iteration is defined by:

$$SumOfDeviation(i) = \sum_{j=1}^{J} Abs(q_{\mathrm{T}}(i,j) - q_{\mathrm{Z}}(j)) \tag{4.1}$$

$$q_{\mathrm{T}}(i,j) = \sum_{x=1}^{N} w(i,x) \tag{4.2}$$

- SumOfDeviation(i)
  - Overall Sum of Deviation for all Margin Constraints at Iteration i.
- J
  - Count of all given Margin Constraints.
- $q_T(i, j)$.
  - Sum of all weights on margin constraints j at iteration i.
- $q_Z(j)$
  - Target Value for margin constraint j.
- N
  - Count of all given Templates.
- $w(i, x)$
  - Weight of template x at iteration i.

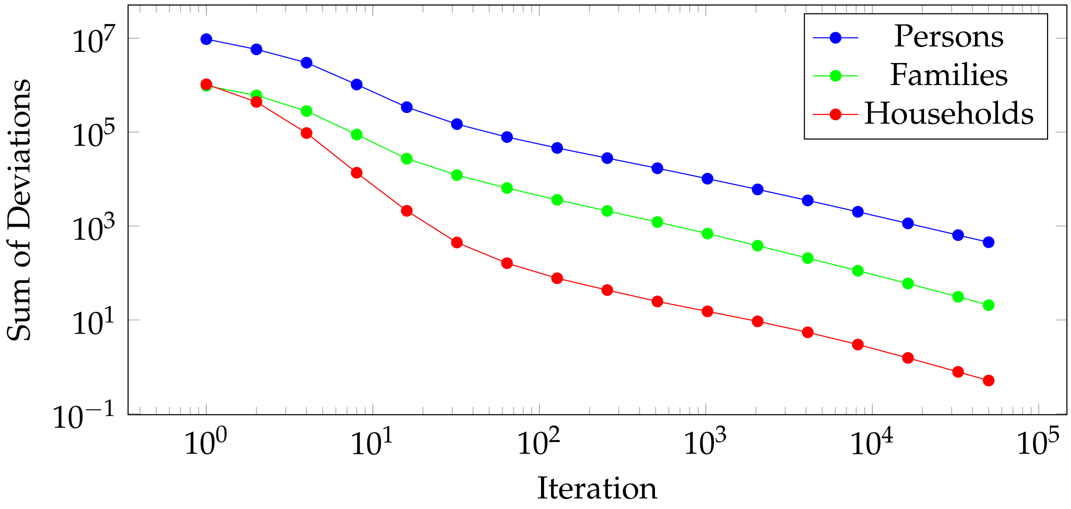


Figure 4.2.: Sum of deviations of cross-classification table counts to person (blue), family (green) and household (red) templates extrapolation factors over 50,000 iterations for person templates with 9 characteristics, family with 8 characteristics and households with 7 characteristics. Sum of deviation ranges from approx. 10,000,000, 1,000,000, 1,000,000 in the beginning to approx. 450, 20, 0.5 after 50,000 iterations.

4.1.3. Sampling Templates and Creating Entities

Once the overall sum of deviation has become reasonable small, there are deduced single entities of persons, families and households. For each margin constraint we recovered the set of affected templates and sampled those with respect to their extrapolation factors. The extrapolation factor is a floating point value. Henceforth we split this value in the integer and the float part. We created as many entities, with the same set of characteristics, as given in the integer value. The float part of the value, which ranges from zero to almost one, was handled as a probability for creating an entity. Due to the randomness of this sampling strategy, the resulting sets of persons, families and households differ in each population synthesis turn.

4.1.4. Assign Persons to Families and Households

To bring the resulting synthetic population closer to the real-world, there have been allocated persons to families and subsequently to households. The entities of persons, families and households share certain characteristics, which need to fit, such that those can be said to belong together.

Each entity has characteristics, which may be used for recovering the relationship to other entities. In appendices A, B and C are presented all the characteristics of the domains of persons, families and households, which allow for defining rules to allocate persons to families and subsequently to households. For example, persons can only live in families which do have the same type of family and in turn families do have to fit the type of the household. This enables to restrict the pools of persons, families and households to potentially allocate. Another restriction is that the household size of the person has to exactly meet the household size present in the household entity. There is the possibility to further refine the allocation rules and to adjust the given ones, in case the results are too weak. The design allows for adding further rules, which might also be probability based, such that persons my be allocated more precisely when processing further statistical information on household compositions.

4.2. Refine Population by Discrete Choice Models

In this section we show how we add further attributes to our baseline population, which we could not deduce in first place due to missing appropriate statistical input data. We use discrete choice models for estimating additional information on income (specific to persons) and car ownership (specific to households). We define models for income per capita and car ownership per household, where we evaluate the impact of certain characteristics on those. The characteristics of the income model are presented in Table 4.1. Table 4.4 respectively presents the characteristics of the car ownership model. The estimation of the model parameters is done by deploying BIOGEME [4] on household survey data of EU-SILC [21].

4.2.1. Assign Income to Persons

In order to depict the real-world population best, it is important to reassemble the population not only with respect to socioeconomic characteristics (like “sex”, “age” and “employment status”), but also with respect to income information. As stated by Taylor & Morris [22], the income of the persons is important when it comes to their mobility pattern. Wealthy persons are more likely to choose their own car for making trips to work, shopping or private appointments. In contrast persons which may not be able to even effort a car on their own in first place and may thus more often choose public transportation systems for satisfying their mobility demand. To allocate income levels to the generated persons, we chose to estimate the impact of the parameters given in Table 4.1 on five different income levels. The estimation was done by defining the model with these parameters using input data of the EU-SILC household survey and estimating the parameters with BIOGEME [4]. The summary of the income model estimation is given in Table 4.2, whereas the actual estimated parameters for the model are given in Table 4.3.

Parameter	Description
ASC_INC1	Constant for Income Level 0-750 €/ Month
ASC_INC2	Constant for Income Level 750-1850 €/ Month
ASC_INC3	Constant for Income Level 1850-3000 €/ Month
ASC_INC4	Constant for Income Level 3000-5000 €/ Month
ASC_INC5	Constant for Income Level 5000-9999+ €/ Month
AGE_INC_X	Influence of Age on alt X
$JOB1\_INC_X$	Dummy Job Status “Employed” on alt X
$JOB2\_INC_X$	Dummy Job Status “Self-Employed” on alt X
$SCHOOL1\_INC_X$	Dummy School Qualification “Compulsory” on alt X
$SCHOOL2\_INC_X$	Dummy School Qualification “Apprenticeship” on alt X
$SCHOOL3\_INC_X$	Dummy School Qualification “Middle/High School” on alt X
$SCHOOL4\_INC_X$	Dummy School Qualification “University” on alt X
SEX_INC_X	Dummy Gender (0 = male, 1 = female) on alt X

Table 4.1.: Description of estimated parameters for the income model.

Model	:	Logit
Number of estimated parameters	:	3
Number of observations	:	5909
Number of individuals	:	5909
Null log likelihood	:	-4095.807
Cte log likelihood	:	-2964.072
Init log likelihood	:	-4095.807
Final log likelihood	:	-2776.853
Likelihood ratio test	:	2637.908
Rho-square	:	0.322
Adjusted rho-square	:	0.321
Final gradient norm	:	+1.678e-003
Diagnostic	:	Convergence reached...
Iterations	:	5
Runtime	:	00:00
Variance-covariance	:	from analytical hessian

Table 4.2.: Summary of the parameter estimation for the income levels of citizen of Austria.

The estimated model shows how certain characteristics influence the decision situation. The parameters for “sex” (i.e. the estimated parameters “SEX_INC1”, “SEX_INC2”, “SEX_INC3”, “SEX_INC4”, “SEX_INC5”) show a negative utility for each income level, with the exception of the first one, where there is even given a positive utility. That means female persons are more likely to "choose" a rather low income level than male persons. The parameters for "school" (i.e. the estimated parameters "SCHOOL_INCX") show that people with higher levels of education are more likely to "choose" higher income levels. Furthermore parameters for "age" (i.e. the estimated parameters "AGE_INCX") indicate, the higher the age of persons is, the higher is their income.

4.2.2. Assign Car Ownership to Households

As there was no information on car ownership given in the cross-classification tables of Statics Austria, we deduced parameters for estimating the car ownership of households. Car ownership has of course a huge impact on the mobility pattern of persons living in a household. Whether or not a car is available has a strong link to choosing car or car passenger as mode of transport for performing trips to activities, such as work, shopping and private errands. The EU-SILC household survey contains information, whether or not a household has at least one car available [21]. Therefore we estimated parameters, given in Table 4.4, which explain whether or not a household may have a car available or not. The summary of the car ownership model estimation is given in Table 4.5, whereas the actual estimated parameters for the model are shown in Table 4.6.

The estimated parameters for "KIDS_CAR" indicate, the more kids are living in the household, the more likely it is, that there is at least one car given in the household. An even stronger link is given between the number of employed persons and car ownership. The more employed persons are living in a household, the more likely there is at least one car available. This link is approximately three times stronger than it is the case with the number of kids.

4.3. Allocate Households to ACDs

Once the persons are allocated to families and subsequently to households, each household has a so-called sociodemographic profile. This profile holds counts on certain characteristics(-combinations), such as how many infants and employed persons are living there. Liable to the information given in the sociodemographic information grid layers, this profile needs to be customized. We make profiles for each household with respect to the sociodemographic combinations of "age-sex" and "employment status-sex" in order to fit the sociodemographics given in the grid layer. We randomly pick a grid element and check if the household fits the characteristics given in the grid and place the household on an ACD within the grid, whereas we first check if the inhabitants count on the ACD is also fitting. In case the inhabitants count does not fit, another ACD within the grid is randomly chosen. If there can not be found an appropriate ACD in the grid, another grid element is chosen. If there can not be found any grid to place the household in 100 tries, the household is neglected.

			Robust		
Parameter		Coeff.	Asympt.		
number	Description	estimate	std. error	*t*-stat	*p*-value
1	AGE_INC1	-0.0545	0.00224	-24.36	0.00
2	AGE_INC3	0.0283	0.00253	11.19	0.00
3	AGE_INC4	0.0545	0.00379	14.39	0.00
4	AGE_INC5	0.0773	0.00762	10.14	0.00
5	ASC_INC1	4.39	0.137	32.09	0.00
6	ASC_INC3	-0.408	0.205	-1.99	0.05
7	ASC_INC4	-1.88	0.298	-6.29	0.00
8	ASC_INC5	-5.03	0.597	-8.43	0.00
9	JOB1_INC1	-2.04	0.0930	-21.94	0.00
10	JOB1_INC3	1.47	0.0871	16.94	0.00
11	JOB1_INC4	2.70	0.139	19.46	0.00
12	JOB1_INC5	3.45	0.247	13.99	0.00
13	JOB2_INC1	-1.61	0.152	-10.65	0.00
14	JOB2_INC3	0.474	0.145	3.27	0.00
15	JOB2_INC4	1.37	0.184	7.43	0.00
16	JOB2_INC5	2.48	0.273	9.08	0.00
17	SCHOOL1_INC1	-1.43	0.181	-7.90	0.00
18	SCHOOL1_INC3	-2.38	0.244	-9.77	0.00
19	SCHOOL1_INC4	-4.56	0.308	-14.79	0.00
20	SCHOOL1_INC5	-4.23	0.584	-7.25	0.00
21	SCHOOL2_INC1	-2.32	0.178	-13.04	0.00
22	SCHOOL2_INC3	-1.45	0.230	-6.32	0.00
23	SCHOOL2_INC4	-3.03	0.248	-12.19	0.00
24	SCHOOL2_INC5	-3.07	0.333	-9.23	0.00
25	SCHOOL3_INC1	-1.71	0.174	-9.83	0.00
26	SCHOOL3_INC3	-1.18	0.230	-5.14	0.00
27	SCHOOL3_INC4	-2.17	0.245	-8.88	0.00
28	SCHOOL3_INC5	-1.59	0.320	-4.96	0.00
29	SCHOOL4_INC1	-1.47	0.205	-7.20	0.00
30	SCHOOL4_INC3	-1.01	0.249	-4.04	0.00
31	SCHOOL4_INC4	-1.20	0.257	-4.66	0.00
32	SCHOOL4_INC5	0.283	0.327	0.87	0.39
33	SEX_INC1	0.642	0.0800	8.03	0.00
34	SEX_INC3	-1.13	0.0642	-17.54	0.00
35	SEX_INC4	-1.72	0.0867	-19.80	0.00
36	SEX_INC5	-2.41	0.142	-16.96	0.00

Table 4.3.: Parameters estimated using BIOGEME in order to deduce income for citizen of Austria.

Parameter	Description
ASC_CAR0	Constant for Zero Cars in HH
ASC_CAR1	Constant for One or more Cars in HH
$KIDS\_CAR_X$	Influence of Number of Kids in HH on alt X
$WORKERS\_CAR_X$	Influence of Number of employed Persons in HH on alt X

Table 4.4.: Description of estimated parameters for the car ownership model.

Model	:	Logit
Number of estimated parameters	:	44
Number of observations	:	12982
Number of individuals	:	12982
Null log likelihood	:	-20875.875
Cte log likelihood	:	-18671.999
Init log likelihood	:	-20875.875
Final log likelihood	:	-12820.462
Likelihood ratio test	:	16110.825
Rho-square	:	0.386
Adjusted rho-square	:	0.384
Final gradient norm	:	+1.304e-001
Diagnostic	:	Convergence reached...
Iterations	:	19
Runtime	:	00:13
Variance-covariance	:	from analytical hessian

Table 4.5.: Summary of the parameter estimation for the car ownership of citizen in Austria.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	t-stat	p-value
1	ASC_CAR1	0.426	0.0447	9.53	0.00
2	KIDS_CAR1	0.353	0.120	2.94	0.00
3	WORKERS_CAR1	1.03	0.0698	14.76	0.00

Table 4.6.: Parameters estimated using BIOGEME in order to deduce car ownership for citizen of *Floridsdorf* with respect to number of workers and kids in the household. The more kids and workers are living in the same household, the more the utility for car ownership increases.

5. Synthetic Population of the 21^{st} District of Vienna

In this chapter we present the population of the 21^{st} district of Vienna, *Floridsdorf*, which we created by deploying the design described in the chapters before. The population consists of approximately 148,000 persons, 40,000 families and 70,000 households. Person entities do have a characteristics set of nine different demographic attributes, families differ in eight attributes and finally households differ in seven attributes. The characteristics count is important when it comes to synthetic population generation. We will highlight the impact of additional characteristics on the computational runtime, as well as on the need for input data. The resulting population of the 21^{st} district is compared to the information present in the input data and the differences of the synthetic population to the real-world population are highlighted. Furthermore the characteristics, which we deduced by deploying discrete choice models, on income per capita and car ownership per household, are presented.

Firstly we introduce the processed number of margin constraints, as well as the total number of templates for persons, families and households, besides the runtime for 100 iterations when extrapolating the templates. Secondly we compare the synthetic population to the real-world population by comparing the input margin constraints to the final population. Subsequently we demonstrate the outcomes of the synthetic population generation in a spatial context. There is highlighted how the synthetic population reassembles the information given in the sociodemographic information grid layers. Finally we show how the estimated parameters of the income and car ownership model have been applied to the population and will present the outcomes of the procedure.

5.1. Processed Data

This section presents information on the data processed in the extrapolation procedure, for generating the baseline population. Figure 5.1 contains information on the runtime for extrapolating, number of templates and number of margin constraints (full-conditional status). This figure shows, that each of the values significantly rises, when the number of sociodemographics we want to generate rises, by comparing the different values of the domains of persons, families and households, where nine, eight and seven characteristics were generated. The number of margin constraints we have to process is highly dependent on the characteristics we want to add to our population as well as the number of attributes within the characteristics. As a result of introducing further characteristics, also the number of possible templates, i.e. sets of possible combinations of sociodemographic attributes, rises significantly. This is due to the pool of sets of combinations, which multiplies by the number of attributes of a further introduced characteristics.

5.2. Real-World vs. Synthetic Population

We present population specific counts of persons, families and households and compare those to the input counts. By considering the total counts of persons, families and households, the synthetic population meets the real-world population almost perfectly. Figure 5.2 illustrates that there are hardly any deviations in any of these domains.

Nevertheless the more interesting part is how the generated synthetic population meets the input margin constraints of the real-world population. The scatter plots in Figure 5.3 indicate how the synthetic population reassembles the margin constraints of the real-world population for persons, families and households respectively. There is a good fit of the margin counts deduced from the real-world population to the margin counts deduced from the synthetic population. On average there is a deviation of approximately 0.84 persons, 0.46 families and 0.095 households per margin constraint. If we had perfectly reassembled the population, there would be no deviation at

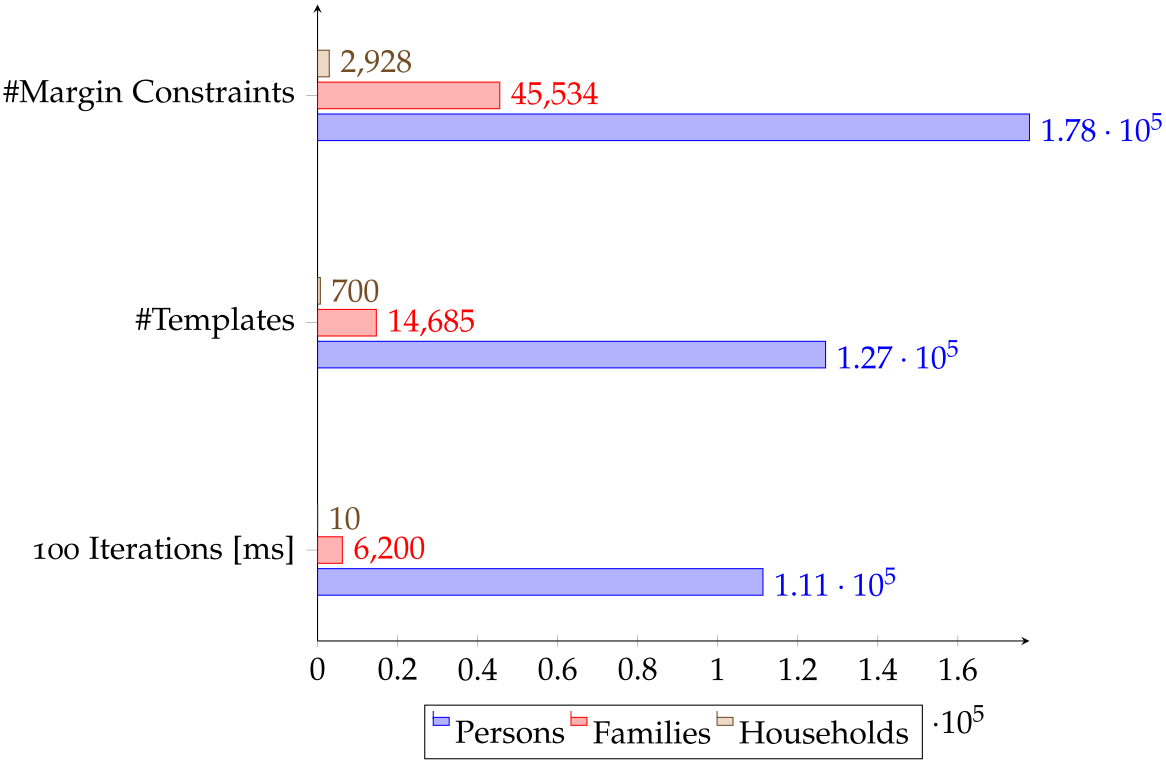


Figure 5.1.: Comparison of the number of constraints, number of templates and average runtime for 100 iterations for the domains of persons, families and households.

all. Moreover all the data points would be placed on the diagonal. Nevertheless such deviations are hard to cope with, because there might always be small deviations due to the applied sampling strategy, which is often probability-based.

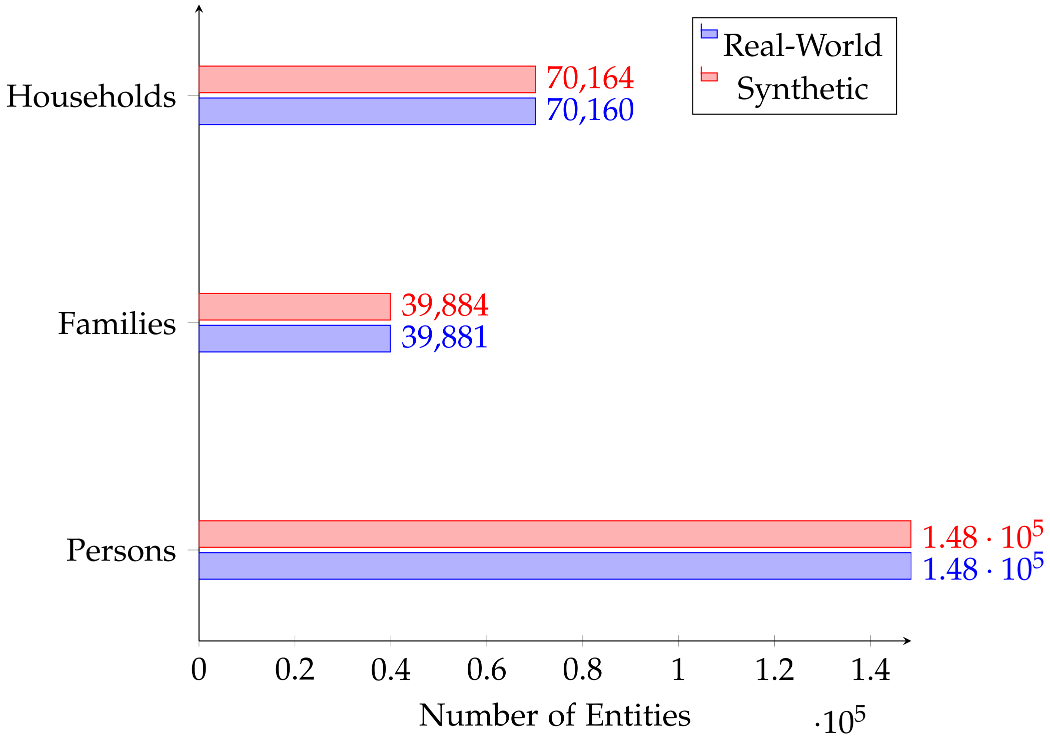


Figure 5.2.: Comparison of the real-world population to the synthetic generated population. The synthetic population almost perfectly meets the real-world population with respect to total created counts of person, family and household entities.

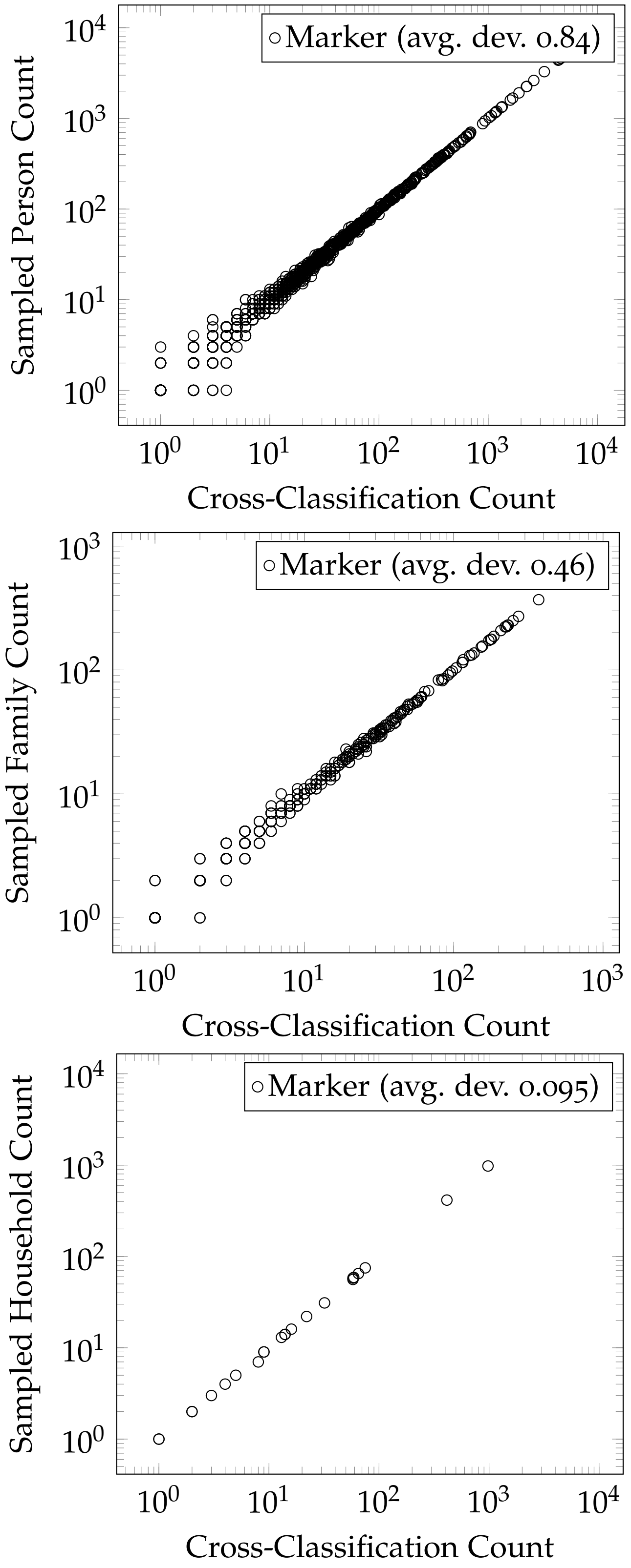


Figure 5.3.: Comparison of sampled counts to input margin counts. There is a higher average deviation for domains with more characteristics than for those with less. The average deviation corresponds to 0.84 persons, 0.46 families and 0.095 households per margin constraint.

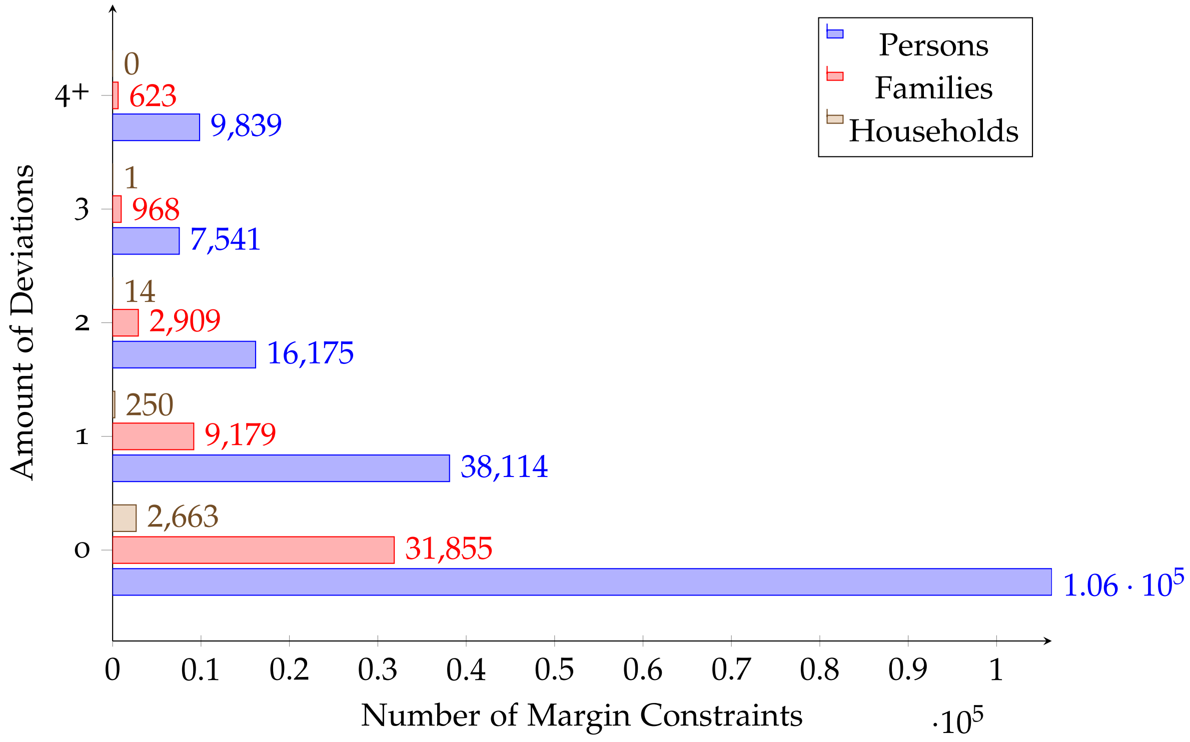


Figure 5.4.: Absolute count of margin constraints with deviation to the final population.

For demonstration purpose Figure 5.4 highlights the overall fitting of the resulting generated entities (persons, families and households) to the multiple input constraints. Almost 105,000 input constraints with respect to the person domain are perfectly met, i.e. have zero deviation. This corresponds to a perfectly fit for approximately 60% of the total given real-world constraints as shown in Figure 5.5. By considering Figure 5.4 and Figure 5.5, there is a worse fit for domains which have a higher count of characteristics, like the person domain, in comparison to those, which have less characteristics, such as the family and the household domain. By performing the population generation process again and again, the deviations of the resulting population to the input data hardly change. This might be due to the implemented sampling strategy for creating actual entities, based upon a probability based procedure.

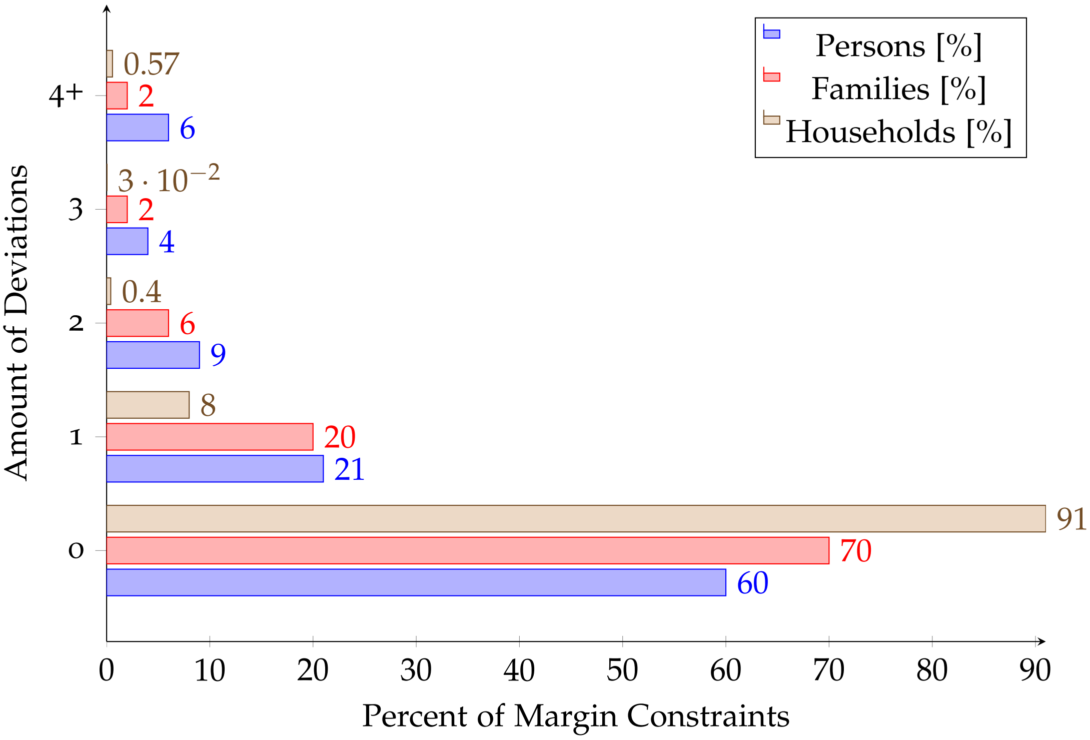
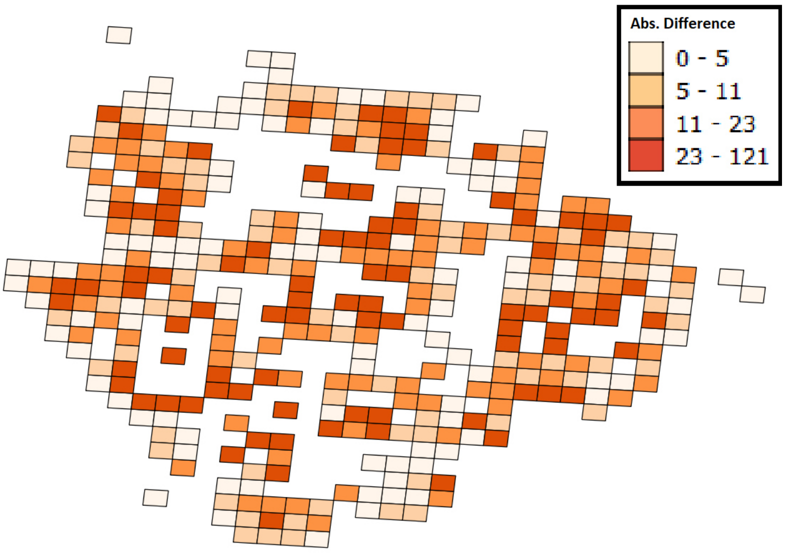


Figure 5.5.: Relative value of margin constraints with deviation to the final population.

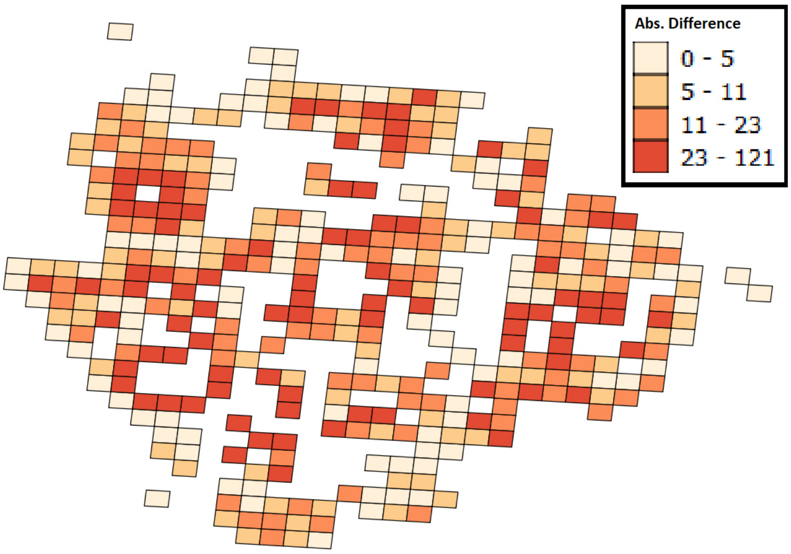
5.3. Spatial Allocation

As presented in the preceding section the sampled entities for all of the three domains almost perfectly meet the actual totals given in the input data. Nevertheless when it comes to spatial allocation, we can not reassemble the population that well. The complete set of entities of persons, families and households (148,493, 39,881, 70,160) could not be allocated a spatial reference. We allocated 128,808, 35,957, 66,288 entities. These are approximately 87%, 90% and 95% of the overall given entities.

We show how robust the counts of the grid layers are met, even when re-sampling and re-allocating the population once again. Figure 5.6 compares the population deviations with respect to the counts given in the sociodemographic information grid layers, "age by sex" and "employment status by sex" in two runs of population synthesis. Each grid element is colored according to the difference between the counts given in spatial input grid and the finally allocated synthetic population. The higher the difference is, the worse the real-world population is reassambled in this spatial entity. By re-sampling the population, Figure 5.6a and Figure 5.6b indicate similar patterns regarding the distribution of the deviation counts. This implies the population is quite similar allocated to the grids in two completely different turns. Nevertheless when considering characteristics, which are not given in a sociodemographic information grid, such as information on school leaving qualification, the patterns differ more likely as demonstrated in Figure 5.7. There is shown the distribution of persons with an university degree. In the first allocation procedure illustrated in Figure 5.7a, the persons are allocated more likely in the south-west. In the second allocation procedure illustrated in Figure 5.7b those people are more likely to reside at the north, whereas the overall distribution seems to be more random.

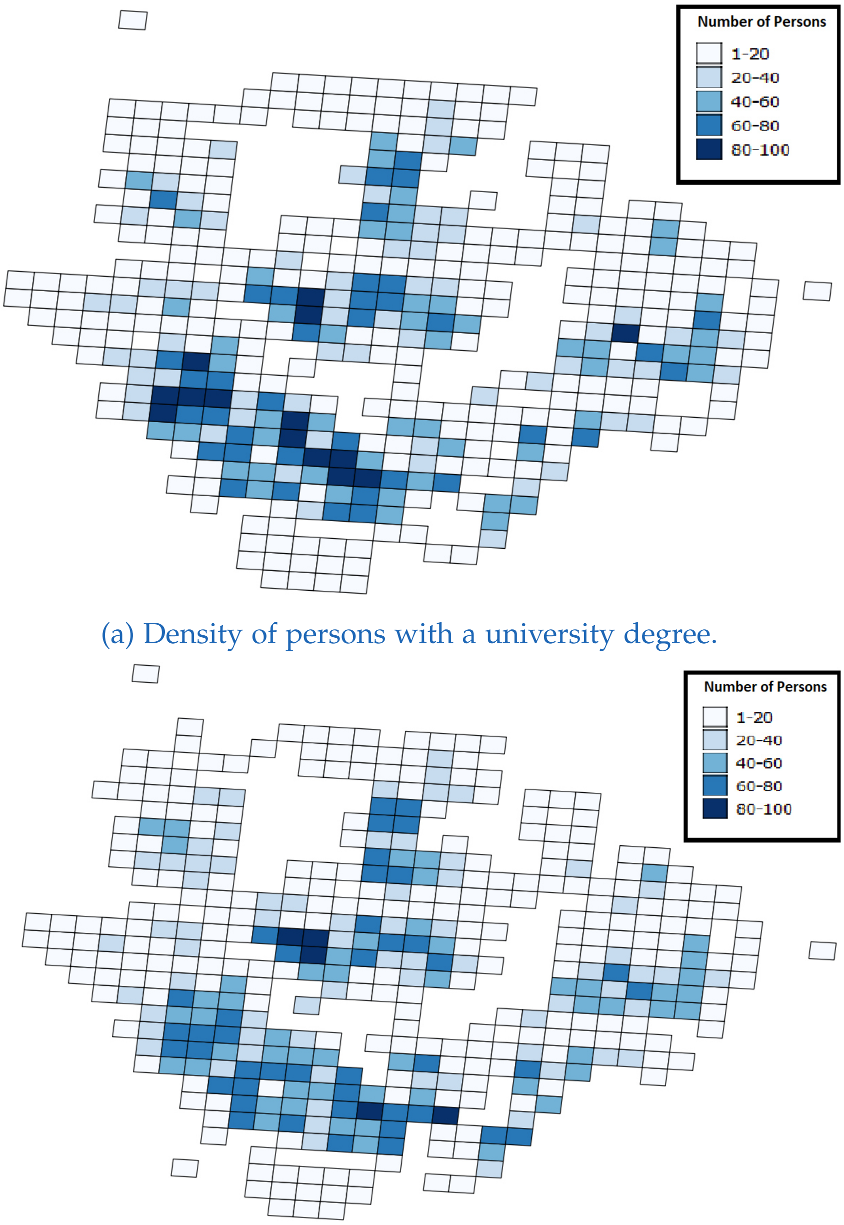


(a) Difference of counts of sociodemographic data provided by Statistics Austria for "age and sex" and "sex and employment status" to the synthetic population.



(b) The pattern indicates a quite similar allocation of the population, probably due to the information given in the grids, which we process in the allocation procedure.

Figure 5.6.: Comparison of differences of grid counts to different generation turns.



(a) Density of persons with a university degree.

(b) There is a rather arbitrary allocation of the persons with a university degree.

Figure 5.7.: Comparison of density of persons with university degree.

5.4. Runtime Evaluation

For generating the population of Vienna *Floridsdorf*, we processed 2,928, 45,534, 177,910 margin constraints on 700, 14,685, 126,939 templates for households, families and persons for at least 50,000 iterations. The generation process took approximately 5 seconds, 3,100 seconds and 15 hours for extrapolating all the templates. Due to the concurrent allocation of extrapolation factors to templates, the design is based upon a single-threaded architecture. The gain with respect to the runtime, did not seem to justify the additional implementation effort for a multithreaded architecture.

We implemented and tested our design on a system with 32GB Ram, 512GB SSD and a i7-6820HQ CPU @ 2.70GHz CPU.

A great effort was gaining input data, as these had to be fetched from Statistics Austria by hand. We made approximately 15, 115, 178 queries to the Statcube portal for recovering the input data for our household, family and person synthetic population generation process, whereas these queries comprise a full-conditional input data set.

5.5. Outcomes of the Discrete Choice Models

In this section we highlight the impacts of the estimated parameters of the car ownership and income model on the synthetic population. We defined multinominal logit models and estimated those with BIOGEME [4]. Figure 5.8 shows the overall car ownership level within households in the 21^{st} district of Vienna, *Floridsdorf*. According to our estimation, approximately 51,000 (i.e. 72%) of our generated households do have at least one car available, which the residents may use. Figure 5.9 presents the impact of the estimated parameters of the income model. The average income levels are plotted with respect to the age of the persons. We can see, there are on average lower incomes for female and young persons. In addition the average incomes significantly decrease at the age of approximately 60 years.

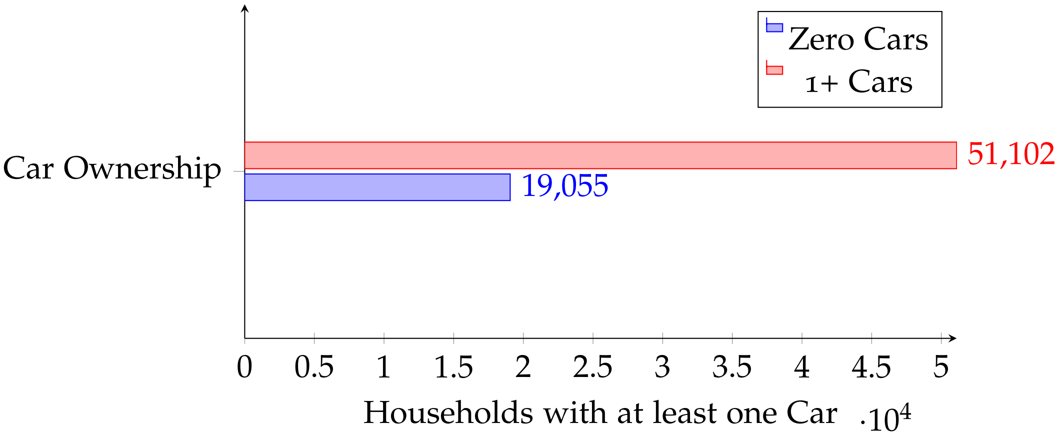


Figure 5.8.: Outcomes of the car ownership model.

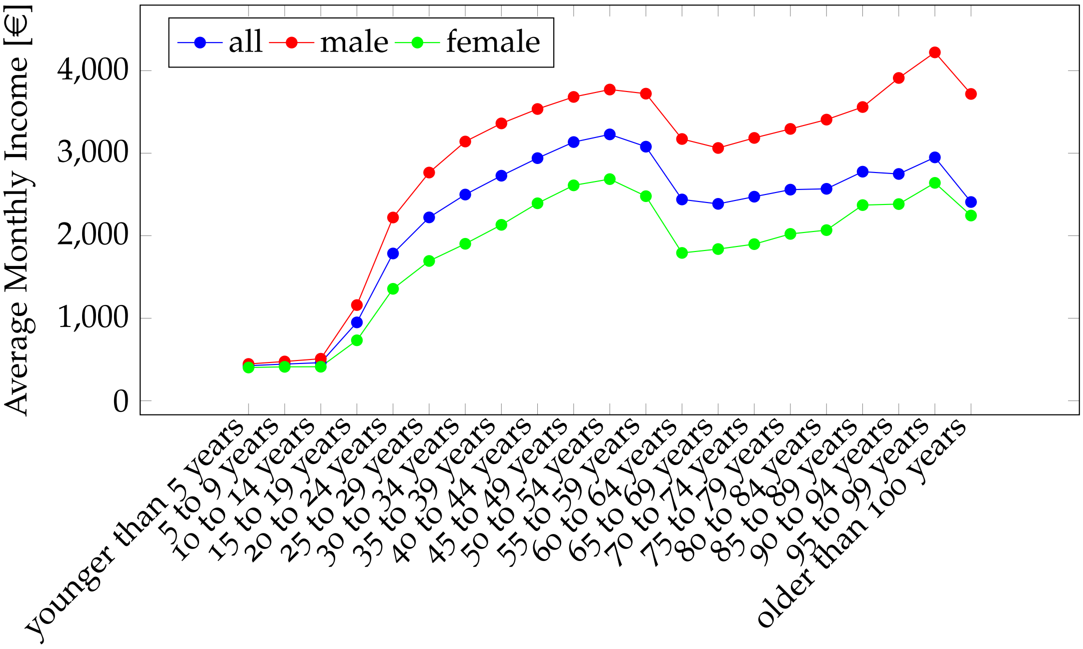


Figure 5.9.: Average monthly income of Austrian citizen before taxes by age after assigning the income levels with respect to the prior estimated parameters.

6. Discussion and Conclusion

We created a synthetic population of the 21^{st} district of Vienna by deploying multidimensional iterative proportional fitting (on the domains of households, families and persons), discrete choice models for deducing further sociodemographic attributes and GIS-based methods for allocating the population a spatial context. In this chapter we highlight the limitations and strengths of the design and will evaluate those.

The results presented in Chapter 5 indicate that it is possible to deduce a fine-grained population on a single entity level from aggregated margin constraints by deploying our introduced approach. The approach to exploit the mutual dependencies in the input data within sets of characteristics, allowed us creating a near real-world population of *Floridsdorf* with respect to the given input margin constraints given in cross-classification tables. Although the results indicate a good fit to reality for a population with nine different characteristics for persons, eight characteristics for families and seven characteristics for households, it is evident to take in consideration that this demands a great effort of providing sound input data. As shown in Section 5.4, the best results can be gained by providing a full-conditional input set of margin constraints. However when adding another sociodemographic characteristic, we need more and more input constraints (i.e. more and more cross-classification tables) to keep the input data in a full-conditional state. For our design to work out properly, it is important to provide information on infeasible combinations of sociodemographic attributes. Thus there is a great need for margin constraints with information on combinations on sociodemographics which occur zero times in the real-world population. We process this information, s.t. we can shrink the pool of possible sets of sociodemographics, i.e. templates, in the creation process. Recovering this kind of information from Statistics Austria is tedious, as the free user is limited with respect to the size and also with respect to the level of detail of queries the user can perform. To overcome this, Statistics Austria offers a premium account which enables for more convenient querying marginal information in order to more proficient generate input data. The premium account also enables querying more fine-grained information. We need to pay attention of keeping the possible template count as small as possible as the runtime is highly related to the count of processed templates as highlighted in Section 5.4.

As long as there are given sound input sources for statistical data, there can be gained an advantage by deploying this introduced design for population synthesis. We not only assign persons to spatial areas on their own, but rather assign persons to families and subsequently families to households and eventually households to spatial areas. This procedure brings the synthetic population one step closer to reality in contrast to implementations of Farooq et al. [7] and Adiga et al. [1], as we can address relationships within households with respect to the mobility behavior. The design is open to further data sources such as additional grid layers with sociodemographic information on a spatial level, which can improve the quality of the resulting population regarding the spatial allocation by adding further and further layers. Nevertheless the layers should be compatible to each other. This is not the case in our implementation as we are processing grid layers and ACDs with information of different years leading to problems by allocating the households to spatial entities. The cross-classification tables hold information of the population of the year 2013, the ACDs are from the year 2015 and the sociodemographic grid information are from the year 2014 (“age by sex”) and 2015 (“sex by employment status”). This certainly causes problems when allocating the generated population to a spatial level as, at some point, the counts given in the two grid layers may be contradictory. However the resulting population meets the grid counts quite good, though there is certainly a problem with the grid data sources, as these are from two different years. The resulting population reassembles the counts (given in the grid layers) quite good, even by performing the population generation process over and over again. The demographics given in the sociodemographic information grids, as shown in Figure 5.6, are also present in the final allocated population, though the grid counts are not perfectly met. Taking a look at characteristics not given in the grids, it is evident, that such information are nearly randomly allocated, as the link of the sociodemographic data given in the grids to other characteristics (e.g. school-leaving qualification) is strong, as shown in Figure 5.7. When it comes to extrapolating the sets of templates, presented in Section 3.4, the sum of deviations is rapidly decreasing in the first 1000 iterations. The sum seems to converge to zero, which indicates the overall compliance of the input data. For domains with more sociodemographics there is a much higher sum of deviation in the whole extrapolation procedure, in contrast to domains with less sociodemographics. This might be due to the significantly higher demand for input data (i.e. margin constraints). Therefore, as domains depict more characteristics, they also expect more input data and as a result evoke higher iteration count for extrapolating and thus more time for the whole procedure.

Liable to the full-conditional status of the input data, it is necessary to perform a huge effort on queries of multiple sets of characteristics in order to gain a full-conditional input data set, which is needed for depicting the real-world population best. The exact count of queries we need to perform by hand is given in Formulas 6.1 and 6.2. For each characteristics CHA_X the number of attributes given in this characterisitcs has a large impact on the necessary amount of queries, which need to be done. If there are given seven attributes in the characteristics, then the corresponding characteristics is given as $CHA_X(7)$. By using this information and the formulas we can calculate the necessary amount of queries for providing the design with full-conditional demographics, in the case of lacking a premium user account for Statistics Austria portal, Statcube. A free user is limited to query 25,000 cell values at once, thus managing full-conditional input data may become a burden, when the sociodemographics count rises [18].

$$\lceil queries \rceil = \frac{CHA_1(22) * CHA_2(21) * CHA_3(8) * CHA_4(2) * CHA_5(7)}{25,000} \tag{6.1}$$

$$\lceil queries \rceil = \frac{73,920}{25,000} = 3 \tag{6.2}$$

Our presented design is applicable for population synthesis by processing single characteristic margin sums but also composite margin sums ("sex by age by employment status"), such as given in cross-classification tables. It is possible to add further margin information besides those of Statistics Austria. But only by providing data in cross-classification tables, thus offering the possibility to deploy this mechanism in other fields of research rather than transportation planning. This design may be auspicious in the field of socioeconomic studies where there is also a great need for deducing single individuals from margin sums. As this design allows for extending or replacing Statistics Austria data, the resulting population can get closer to reality by each cross-classification table (i.e. set of margin constraints) we add. In order to allow for adding further characteristics to this population, we deployed discrete choice models. The results show that there are created incomes for each person which seem to be plausible, when considering average incomes [6]. Nevertheless we have to add, that there is a need for further research to examine, if these assumptions are also fitting the real-world when considering single persons rather than average sums over the population. Due to the limited counts of observations in the EU-SILC household survey, we based our models on the observations of the total citizenship of Austria. This decision neglects the differences related to car ownership and income in a regional context. The outcome of this simplification can be observed in the results of the car ownership model, where there are approximately 72.8% of households in Vienna *Floridsdorf* with at least one car, though the official data of Statistics Austria indicate otherwise [24]. However for the total of Austrian households this estimation may seem reasonable. Within the EU-SILC household survey there was questioned whether the family owned at least one car and not the actual amount of cars in the household. As a result the given car ownership model is not conclusive, as we only estimate whether there is a car or not. It would be certainly more interesting if it is possible to estimate the actual amount of cars in the household as it was done by Farooq et al. [7].

Another limitation is given, as the estimation of incomes does not consider part-time jobs by now. Therefore the synthetic population does not depict reality for these kind of people with respect to their income. As shown in Figure 5.9 there are already incomes at younger ages. This certainly may not be true as infants in Austria are certainly not permitted to work. Nevertheless we also allocated those an income, as we wanted to allocate the total of the persons in our population an income. These may therefore be interpreted as alimony payments and child allowances.

Our proposed design for population synthesis comprises different techniques for population synthesis introduced in Chapter 2. We combined the introduced approaches, such as IPF, discrete choice models and GIS techniques. In our case we used these techniques to generate a synthetic population of the 21^{st} district of Vienna, *Floridsdorf* and additionally added information on household compositions. None of the introduced approaches have dealt with this combination of deploying various techniques at the same time and depicting the household compositions before. This design can be used for creating populations for each Austrian city, district and municipality. Moreover our approach allows to generate a synthetic baseline population based on free statistical data of Statistics Austria, by processing cross-classification tables on socioeconomics and demographics. For refining the population, we also used survey data of EU-SILC which is not openly accessible. Nevertheless these data may be requested for free from Statistics Austria for research purposes. Finally for allocating the population to spatial entities, it is necessary to process grid layers with information on sociodemographics. These are provided by Statistics Austria against payment of a fee. With respect to the data sources it is therefore possible to generate a population for free by deploying our design, though for refining the population there may be costs.

7. Outlook

The population which can be gained by deploying the design introduced in this master's thesis is static. That means, there are no mobility profiles assigned to the single citizens and the population simply depicts the citizenship of 2013 of the 21st district of Vienna, *Floridsdorf*. For depicting the real-world population better, it would be promising to add day-plans and activities to the population.

Day-plans are a set of activities a person carries out a day. An activity in turn is a location, where a person may execute a certain activity type, such as "shopping". A common day-plan is the sequence of activities "home-work-home". That means at a certain time the person is located at his "home" activity location, some time later the person is located at his "work" activity location and finally the person returns to his "home" location. In the scope of this thesis we already allocated activities of type "home" to all citizen. To bring the population closer to reality, it would be promising to additionally assign the primary and secondary activities to the persons. Primary activities like "work" and "school" are those, which typically are carried out at the same location every weekday. In addition to the primary activities, there are also secondary activities, such as "shopping", "leisure", etc. These kind of activities, i.e. the locations of these activities, may be arbitrarily (probability-based) chosen from day to day (i.e. in each simulation run).

Assigning such day-plans would make the population more compliant with MATSim models, as it would be possible, on the one hand to process the sociodemographic patterns of the persons within the transport simulation and on the other hand also their mobility patterns. Examining the outcomes of the transport simulation, based upon the combination of these two information types, might enable for more conclusive results with respect to traffic flow analysis and site-evaluation. For extending the population with such plans, it would be necessary to process additional information pools. Information pools related to primary activities may hold information on companies and schools, such as employee/pupil count, geographical position and field of business/type of school. These kind of information would make it possible to assign persons to geographical working/school places with respect to their field of business/school type. Information pools related to secondary activities, may hold information on the geographic position, activity type and a value indicating the attractiveness of this location to the corresponding activity type. These kind of information might be provided by business directories, regional education authorities and yellow pages.

For putting these recommendations for MATSim models into practice, it is necessary to continue researching in the field of generating synthetic populations on an agent-level.

Appendix

Appendix A.

Attributes of Persons

German (Original) **Wohnort (Ebene +1)**	English (Translated) **Place of residence (level +1)**
Wien-Floridsdorf	Wien-Floridsdorf

Table A.1.: “Place of residence” with 1 attribute.

German (Original) **Geschlecht**	English (Translated) **Gender**
Männlich	Male
Weiblich	Female

Table A.2.: “Gender” with 2 attributes.

German (Original) **Anzahl der Kinder in der Familie (Ebene +1)**	English (Translated) **Number of children in the family (level +1)**
1 Kind	1 child
2 Kinder	2 children
3 Kinder	3 children
4 Kinder	4 children
5 Kinder	5 children
6 und mehr Kinder	6 or more children
Keine Kernfamilie	No nuclear family
Keine Kinder	No children

Table A.3.: “Number of children” with 8 attributes.

German (Original) **Alter in Jahren (Ebene +2)**	English (Translated) **Age in years (level +2)**
Unter 5 Jahre	Younger than 5 years
5 bis 9 Jahre	5 to 9 years
10 bis 14 Jahre	10 to 14 years
15 bis 19 Jahre	15 to 19 years
20 bis 24 Jahre	20 to 24 years
25 bis 29 Jahre	25 to 29 years
30 bis 34 Jahre	30 to 34 years
35 bis 39 Jahre	35 to 39 years
40 bis 44 Jahre	40 to 44 years
45 bis 49 Jahre	45 to 49 years
50 bis 54 Jahre	50 to 54 years
55 bis 59 Jahre	55 to 59 years
60 bis 64 Jahre	60 to 64 years
65 bis 69 Jahre	65 to 69 years
70 bis 74 Jahre	70 to 74 years
75 bis 79 Jahre	75 to 79 years
80 bis 84 Jahre	80 to 84 years
85 bis 89 Jahre	85 to 89 years
90 bis 94 Jahre	90 to 94 years
95 bis 99 Jahre	95 to 99 years
100 Jahre und älter	100 years or older

Table A.4.: “Age in Years” with 21 attributes.

German (Original) **Typ der Familie (Ebene +1)**
Ehepaar mit mind. 1 Kind unter 25 Jahren
Lebensgemeinschaft mit mind. 1 Kind unter 25 Jahren
Ein-Eltern-Familie (nur Vater) mit mind. 1 Kind unter 25 Jahren
Ein-Eltern-Familie (nur Mutter) mit mind. 1 Kind unter 25 Jahren
Keine Kernfamilie
Ehepaar ohne Kinder
Lebensgemeinschaft ohne Kinder
Ehepaar mit jüngstem Kind ab 25 Jahren
Lebensgemeinschaft mit jüngstem Kind ab 25 Jahren
Ein-Eltern-Familie (nur Vater) mit jüngstem Kind ab 25 Jahren
Ein-Eltern-Familie (nur Mutter) mit jüngstem Kind ab 25 Jahren

English (Translated) **Type of family (level +1)**
Married couple with at least 1 child under the age of 25
Cohabitation with at least 1 child under the age of 25
Single-parent family (only father) with at least 1 child under the age of 25
Single-parent family (only mother) with at least 1 child under the age of 25
No nuclear family
Married couple without children
Cohabitation without children
Married couple, whose youngest child is older than 25
Cohabitation, whose youngest child is older than 25
Single-parent familiy (only father), whose youngest child is older than 25
Single-parent familiy (only mother), whose youngest child is older than 25

Table A.5.: "Type of family" with 11 attributes.

German (Original) **Größe des Privathaushalts (Ebene +1)**	English (Translated) **Private household size (level +1)**
1 Person	1 person
2 Personen	2 persons
3 Personen	3 persons
4 Personen	4 persons
5 Personen	5 persons
6 Personen	6 persons
7 Personen	7 persons
8 und mehr Personen	8 or more persons
Nichtprivathaushalte	Non-private households

Table A.6.: "Private household size" with 9 attributes.

German (Original) **ÖNACE 2008 Arbeitsstätte (Ebene +4)**	English (Translated) **Field of business (level +4)**
Entfällt	Does not apply
Erziehung und Unterricht	Education
Handel	Trade
Freiberufliche/techn. Dienstleistungen	Freelance/technical services
Gesundheits- und Sozialwesen	Health and social services
Verkehr	Transport
Sonst. wirtschaftl. Dienstleistungen	Other economical service
Öffentliche Verwaltung	Public administration
Bau	Construction
Beherbergung und Gastronomie	Accommodation and gastronomy
Sonst. Dienstleistungen	Other services
Wasserversorgung/Abfallentsorgung	Water supply and waste management
Grundstücks- und Wohnungswesen	Real estate and housing
Herstellung von Waren	Manufacturing
Information und Kommunikation	Information and communication
Finanz- und Versicherungsleistungen	Finance and insurance services
Kunst, Unterhaltung und Erholung	Arts, entertainment and recreation
Unbekannte Wirtschaftstätigkeit	Unknown economic activity
Energieversorgung	Energy supply
Private Haushalte	Private households
Land- und Forstwirtschaft	Agriculture and forestry
Bergbau	Mining
Exterritoriale Organisationen	Extra-territorial organizations

Table A.7.: “Field of business” with 23 attributes.

German (Original) **Höchste abgeschlossene Ausbildung**	English (Translated) **Highest educational attainment**
Akademie	Academy
Allgemein bildende höhere Schule	Grammar school
Berufsbildende höhere Schule	Higher vocational school
Berufsbildende mittlere Schule	Vocational school
Entfällt	Does not apply
Hochschule	University
Kolleg	Course of lectures
Lehrabschluss	Apprenticeship certificate
Pflichtschule	Compulsory school

Table A.8.: "Highest educational attainment" with 8 attributes.

German (Original) **Stellung im Beruf (Ebene +2)**	English (Translated) **Occupational status (level +2)**
Nicht-Erwerbspersonen	Inactive persons
Arbeiterinnen, Arbeiter (inklusive Lehrlinge)	Workers (inclusive apprentices)
Angestellte (inklusive Lehrlinge)	Employees (inclusive apprentices)
Sonstige unselbständig Erwerbstätige	Other paid employment
Selbständige	Self-employed
Arbeitslos	Unemployed
Arbeitgeber	Employers
Mithelfende Familienangehörige	Family workers

Table A.9.: "Occupational status" with 8 attributes.

Appendix B.

Attributes of Families

German (Original)	English (Translated)
Wohnort (Ebene +1)	**Place of residence (level +1)**
Wien-Floridsdorf	Wien-Floridsdorf

Table B.1.: "Place of residence" with 1 attribute.

German (Original)	English (Translated)
Alter des Mannes (Ebene +2)	**Age of the man (level +2)**
15 bis 19 Jahre	15 to 19 years
20 bis 24 Jahre	20 to 24 years
25 bis 29 Jahre	25 to 29 years
30 bis 34 Jahre	30 to 34 years
35 bis 39 Jahre	35 to 39 years
40 bis 44 Jahre	40 to 44 years
45 bis 49 Jahre	45 to 49 years
50 bis 54 Jahre	50 to 54 years
55 bis 59 Jahre	55 to 59 years
60 bis 64 Jahre	60 to 64 years
65 bis 69 Jahre	65 to 69 years
70 bis 74 Jahre	70 to 74 years
75 bis 79 Jahre	75 to 79 years
80 bis 84 Jahre	80 to 84 years
85 bis 89 Jahre	85 to 89 years
90 bis 94 Jahre	90 to 94 years
95 bis 99 Jahre	95 to 99 years
Trifft nicht zu	Does not apply

Table B.2.: “Age of the man” with 18 attributes.

German (Original) **Alter der Frau (Ebene +2)**	English (Translated) **Age of the woman (level +2)**
15 bis 19 Jahre	15 to 19 years
20 bis 24 Jahre	20 to 24 years
25 bis 29 Jahre	25 to 29 years
30 bis 34 Jahre	30 to 34 years
35 bis 39 Jahre	35 to 39 years
40 bis 44 Jahre	40 to 44 years
45 bis 49 Jahre	45 to 49 years
50 bis 54 Jahre	50 to 54 years
55 bis 59 Jahre	55 to 59 years
60 bis 64 Jahre	60 to 64 years
65 bis 69 Jahre	65 to 69 years
70 bis 74 Jahre	70 to 74 years
75 bis 79 Jahre	75 to 79 years
80 bis 84 Jahre	80 to 84 years
85 bis 89 Jahre	85 to 89 years
90 bis 94 Jahre	90 to 94 years
95 bis 99 Jahre	95 to 99 years
100 Jahre und älter	Older than 100 years
Trifft nicht zu	Does not apply

Table B.3.: “Age of the woman” with 19 attributes.

German (Original) **Typ der Kernfamilie (Ebene +2)**	English (Translated) **Type of the nuclear family (level +2)**
Ehepaar ohne Kinder	Married couple without children
Ehepaar mit mind. 1 Kind	Married couple with at least 1 child
Lebensgemeinschaft ohne Kinder	Cohabitation without children
Lebensgemeinschaft mit mind. 1 Kind	Cohabitation with at least 1 child
Ein-Eltern-Familie (nur Vater)	Single-parent family (only father)
Ein-Eltern-Familie (nur Mutter)	Single-parent family (only mother)

Table B.4.: “Type of the nuclear family” with 6 attributes.

German (Original) **Höchste abgeschlossene Ausbildung des Mannes**	English (Translated) **Highest educational attainment of the man**
Akademie	Academy
Allgemein bildende höhere Schule	Grammar school
Berufsbildende höhere Schule	Higher vocational school
Berufsbildende mittlere Schule	Vocational school
Entfällt	Does not apply
Hochschule	University
Kolleg	Course of lectures
Lehrabschluss	Apprenticeship certificate
Pflichtschule	Compulsory school
Trifft nicht zu	Does not apply

Table B.5.: "Highest educational attainment of the man" with 10 attributes.

German (Original) **Höchste abgeschlossene Ausbildung der Frau**	English (Translated) **Highest educational attainment of the woman**
Akademie	Academy
Allgemein bildende höhere Schule	Grammar school
Berufsbildende höhere Schule	Higher vocational school
Berufsbildende mittlere Schule	Vocational school
Entfällt	Does not apply
Hochschule	University
Kolleg	Course of lectures
Lehrabschluss	Apprenticeship certificate
Pflichtschule	Compulsory school
Trifft nicht zu	Does not apply

Table B.6.: "Highest educational attainment of the woman" with 10 attributes.

German (Original) **Anzahl der Kinder in der Familie (Ebene +1)**	English (Translated) **Number of children in the family (level +1)**
1 Kind	1 child
2 Kinder	2 children
3 Kinder	3 children
4 Kinder	4 children
5 Kinder	5 children
6 und mehr Kinder	6 or more children
Keine Kinder	No children

Table B.7.: "Number of children in the family" with 7 attributes.

German (Original) **Anzahl der Kinder unter 25 Jahren in der Familie (Ebene +1)**	English (Translated) **Number of children under the age of 25 in the family (level +1)**
1 Kind unter 25 Jahren	1 child under the age of 25
2 Kinder unter 25 Jahren	2 children under the age of 25
3 Kinder unter 25 Jahren	3 children under the age of 25
4 Kinder unter 25 Jahren	4 children under the age of 25
5 Kinder unter 25 Jahren	5 children under the age of 25
6 Kinder und mehr unter 25 Jahren	6 or more children under the age of 25
Keine Kinder unter 25 Jahren	No children under the age of 25

Table B.8.: "Number of children under the age of 25 in the family" with 7 attributes.

Appendix C.

Attributes of Households

German (Original) **Wohnort (Ebene +1)**	English (Translated) **Place of residence (level +1)**
Wien-Floridsdorf	Wien-Floridsdorf

Table C.1.: "Place of residence" with 1 attribute.

German (Original) **Anzahl der Personen ab 65 Jahren im Privathaushalt (Ebene +1)**	English (Translated) **Number of persons older than 65 years in the private household (level +1)**
1 Person ab 65 Jahren	1 person older than 65 years
2 Personen ab 65 Jahren	2 personen older than 65 years
3 und mehr Personen ab 65 Jahren	3 und mehr personen older than 65 years
Anstaltshaushalte	Institutional households
Keine Person ab 65 Jahren	No person older than 65 years

Table C.2.: "Number of persons older than 65 years in the private household" with 5 attributes.

German (Original) **Typ des Einfamilienhaushalts**	English (Translated) **Type of the single-family household**
Einfamilienhaush. o. weitere Pers.	Single-family HH without add. pers.
Einfamilienhaush. m. weiteren Pers.	Single-family HH with add. pers.
Kein Einfamilienhaushalt	No single-family household

Table C.3.: “Type of the single-family household” with 3 attributes.

German (Original) **Geschlecht**	English (Translated) **Gender**
Männlich	Male
Weiblich	Female

Table C.4.: “Gender” with 2 attributes.

German (Original) **Größe des Privathaushalts (Ebene +1)**	English (Translated) **Private household size (level +1)**
1 Person	1 person
2 Personen	2 persons
3 Personen	3 persons
4 Personen	4 persons
5 Personen	5 persons
6 Personen	6 persons
7 Personen	7 persons
8 und mehr Personen	8 or more persons
Anstaltshaushalte	Institutional households

Table C.5.: “Private household size” with 9 attributes.

German (Original) **Anzahl der Personen unter 18 Jahren im Privathaushalt (Ebene +1)**	English (Translated) **Number of persons under the age of 18 in the private household (level +1)**
Keine Person unter 18 Jahren	No person under the age of 18
1 Person unter 18 Jahren	1 person under the age of 18
2 Personen unter 18 Jahren	2 persons under the age of 18
3 Personen unter 18 Jahren	3 persons under the age of 18
4 Personen unter 18 Jahren	4 persons under the age of 18
5 und mehr Personen unter 18 Jahren	5 or more persons under the age of 18
Anstaltshaushalte	Institutional households

Table C.6.: "Number of persons under the age of 18 in the private household" with 7 attributes.

German (Original) **Typ des Haushalts (Ebene +1)**
Ehepaar ohne Kinder
Ehepaar mit mind. 1 Kind unter 25 Jahren
Ehepaar mit jüngstem Kind ab 25 Jahren
Lebensgemeinschaft ohne Kinder
Lebensgemeinschaft mit mind. 1 Kind unter 25 Jahren
Lebensgemeinschaft mit jüngstem Kind ab 25 Jahren
Ein-Eltern-Familie (nur Vater) mit mind. 1 Kind unter 25 Jahren
Ein-Eltern-Familie (nur Vater) mit jüngstem Kind ab 25 Jahren
Ein-Eltern-Familie (nur Mutter) mit mind. 1 Kind unter 25 Jahren
Ein-Eltern-Familie (nur Mutter) mit jüngstem Kind ab 25 Jahren
Zwei- oder Mehrfamilienhaushalte
Einpersonenhaushalte
Nichtfamilien-Mehrpersonenhaushalte
Internat, Schüler, Studentenheim sowie Heim für Berufstätige in Ausbildung
Heil- Pflegeanstalt sowie Pensionisten- bzw. Altersheim
Einrichtung für Behinderte
Einrichtung für Kinder und Jugendliche
Kloster oder ähnliche Anstalt
Kaserne
Justizvollzugsanstalt
Einrichtung für Flüchtlinge
Sonstige Anstalt
Gemeinschaftsunterkunft
Einrichtung für sozial Bedürftige bzw. Wohnungslose

Table C.7.: “Type of the household” with 24 attributes, continued in next table.

English (Translated) **Type of the household (level +1)**
Married couple without children
Married couple with at least 1 child under the age of 25
Married couple whose youngest child is older than 25 years
Cohabitation without children
Cohabitation with at least 1 child under the age of 25
Cohabitation whose youngest child is older than 25 years
Single-parent family (only father) with at least 1 child under the age of 25
Single-parent family (only father) whose youngest child is older than 25 years
Single-parent family (only mother) with at least 1 child under the age of 25
Single-parent family (only mother) whose youngest child is older than 25 years
Two- or multi-family households
Single-person households
Non-family- multi-person households
Boarding school, student home or dormitory as well as homes for apprentices
Nursing home as well as retirement home
Facility for the disabled
Facility for children and young people
Monastery or similar institutions
Barrack
Prison
Facility for refugees
Other institutions
Shared accommodation
Facility for socially needy or homeless people

Table C.8.: "Type of the household" with 24 attributes (english).

Bibliography

[1] Abhijin Adiga et al. "Generating a synthetic population of the United States." In: (2015) (cit. on pp. 2, 5, 7, 12, 63).

[2] Paul Anderson et al. "Associations Generation in Synthetic Population for Transportation Applications." In: *Transportation Research Record: Journal of the Transportation Research Board* 2429 (2014), pp. 38–50. ISSN: 0361-1981. DOI: 10.3141/2429-05 (cit. on pp. 2, 8, 9, 12, 13).

[3] Moshe E Ben-Akiva and Steven R Lerman. *Discrete Choice Analysis*. MIT Press, 1985, p. 390 (cit. on p. 33).

[4] Michel Bierlaire. "BIOGEME: A free package for the estimation of discrete choice models." In: *Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland*. 2003 (cit. on pp. 4, 33, 44, 60).

[5] W Edwards Deming and Frederick F Stephan. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." In: *The Annals of Mathematical Statistics* 11.4 (1940), pp. 427–444. ISSN: 00034851. URL: http://www.jstor.org/stable/2235722 (cit. on p. 6).

[6] DerStandard.at. *Alter, Bildung, Geschlecht: Wer wie viel verdient*. 2014. URL: http://derstandard.at/2000041068013/Wie-viel-Menschen-wie-Sie-verdienen (cit. on p. 65).

[7] Bilal Farooq, Michel Bierlaire, and Gunnar Flötteröd. "Simulation based Population Synthesis." In: *Seventh workshop on discrete choice models, EPFL* (2011), pp. 1–37 (cit. on pp. 2, 3, 5, 8–12, 63, 65).

[8] Pierre Guerin and Massimiliano Marcellino. "Markov-Switching MIDAS Models." In: *Journal of Business & Economic Statistics* 31.1 (2012), pp. 45–56. ISSN: 0735-0015. DOI: 10.1080/07350015.2012.727721 (cit. on p. 32).

[9] Harold W. Kuhn. "The Hungarian Method for the assignment problem." In: *Naval Research Logistics Quarterly* 2 (1955), pp. 83–97 (cit. on p. 9).

[10] Andreas Horni, Kai Nagel, and Kay W. Axhausen. *The Multi-Agent Transport Simulation Title of Book : The Multi-Agent Transport Simulation MATSim Subtitle positioned below*. 2016, p. 618. ISBN: 978-1-909188-76-1. DOI: http://dx.doi.org/10.5334/baw (cit. on p. 14).

[11] Frank S Koppelman and Chandra Bhat. "A Self Instructing Course in Mode Choice Modeling : Multinomial and Nested Logit Models by with technical support from Table of Contents." In: *Elements* 28.3 (2006), pp. 501–12. ISSN: 15494918. DOI: 10.1002/stem.294 (cit. on p. 9).

[12] Gunther Maier and Peter Weiss. *Modelle diskreter Entscheidungen*. Springer Science & Business Media, 2013 (cit. on p. 33).

[13] R Mankiewicz. *The story of mathematics*. The story of mathematics. Princeton University Press, 2000. URL: https://books.google.at/books?id=JXxrpwAACAAJ (cit. on p. 34).

[14] Rolf Moeckel. "Creating a synthetic Population." In: *Proceedings of the 8th International Conference on Computers in Urban Planning and Urban Management (CUPUM)* May (2003), pp. 1–18. URL: http://www.spiekermann-wegener.com/pub/pdf/CUPUM%7B%5C_%7D2003%7B%5C_%7DSynpop.pdf (cit. on pp. 2, 3, 5, 6, 11–13).

[15] Kirill Müller and Axhausen Kay. "Hierarchical IPF: Generating a synthetic population for Switzerland." In: *51st Congress of the European Regional Science Association June (2011)* (2011) (cit. on pp. 2, 5, 8).

[16] Thomas Otterstätter. "Methoden zur Erfassung von Verkehrsströmen und Fahrzeiten mit stationären fahrzeugwiedererkennenden Detektoren." In: 93.August (2013) (cit. on p. 29).

[17] Klaus Spiekermann and Michael Wegener. "Freedom from the Tyranny of Zones : Towards New GIS-based Spatial Models." In: *Spatial Models and GIS: New Potential and New Models* (1999), pp. 45–61 (cit. on pp. 11, 35).

[18] Statistik Austria. *Abgestimmte Erwerbsstatistik 2013*. 2013. URL: http://statcube.at (cit. on pp. 15, 17, 22, 33, 37, 64).

[19] Statistik Austria. *Abgestimmte Erwerbsstatistik 2013 - Familien*. 2013. URL: http://statcube.at (cit. on pp. 15, 22, 37).

[20] Statistik Austria. *Abgestimmte Erwerbsstatistik 2013 - Haushalte*. 2013. URL: http://statcube.at (cit. on pp. 15, 22, 37).

[21] Statistik Austria. *EU-SILC*. 2014. URL: http://www.statistik.at (cit. on pp. 3, 15, 17, 44, 46).

[22] Brian D Taylor and Eric A Morris. "Public transportation objectives and rider demographics: are transit's priorities poor public policy?" In: *Transportation* 42.2 (2015), pp. 347–367. ISSN: 1572-9435. DOI: 10.1007/s11116-014-9547-0 (cit. on p. 44).

[23] Kenneth E Train. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511805271 (cit. on pp. 31–33).

[24] VCÖ. *VCÖ: Mehr als 850.000 Haushalte in Österreich sind ohne eigenes Auto mobil*. 2017. URL: https://www.vcoe.at/news/details/vcoe-mehr-als-850-000-haushalte-in-oesterreich-sind-ohne-eigenes-auto-mobil (cit. on p. 65).

[25] Kimmo Vehkalahti. "The Concise Encyclopedia of Statistics by Yadolah Dodge." In: *International Statistical Review* 76.3 (2008), pp. 460–461. URL: http://econpapers.repec.org/RePEc:bla:istatr:v:76:y:2008:i:3:p:460-461 (cit. on p. 34).

[26] Yi Zhu and Joseph Ferreira. "Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation." In: *Transportation Research Record: Journal of the Transportation Research Board* 2429 (2014), pp. 168–177. ISSN: 0361-1981. DOI: 10.3141/2429-18 (cit. on pp. 2, 5, 7).