Diploma Thesis

Metropolis-Hastings based synthesis of all-day round trips from origin-destination matrices and limited population segmentation

Submitted in satisfaction of the requirements for the degree of
Diplom-Ingenieur
of the TU Wien, Faculty of Civil and Environmental Engineering

---

Diplomarbeit

Synthese von Wegeketten auf Basis von OD-Matrizen und Bevölkerungssegmentierung unter Zuhilfenahme des Metropolis-Hastings Algorithmus

ausgeführt zum Zwecke der Erlangung des akademischen Grads
Diplom-Ingenieur
eingereicht an der TU Wien, Fakultät für Bau- und Umweltingenieurwesen

von

Franz-Xaver Rupprecht
Matr.Nr.: 01516775

Betreuung: Univ.-Prof. Mag.rer.soc.oec. Dr.rer.soc.oec. **Günter Emberger**
Institut für Verkehrswissenschaften
Forschungsbereich Verkehrsplanung und Verkehrstechnik
Technische Universität Wien,
Karlsplatz 13/230-01, 1040 Wien, Österreich

Wien, im Mai 2024

Abstract

The objective of this work is to generate a synthetic agent population for the traffic simulation software MATSim. This represents the initial stage in the creation of an agent-based traffic model, which is expected to offer advantages over some of the weaknesses of the four-step model. To this end, sociodemographic, socioeconomic, and transport demand data are extracted from an existing transport model and converted into a format that is suitable for use within MATSim. This is achieved by disaggregating OD matrices into coherent and distinguishable round trips and assigning them to five distinct population groups. Furthermore, the time structure resulting from each round trip's timestamps is refined. Initially, the methodology employed is contextualised within the existing literature. This is followed by the description of the probability distribution definition and the explaining of the sampling machinery used. By applying the methodology to the city of Vienna, a population of 1000 agents is generated. The process itself and its results are presented in detail. Finally, an outlook on future research work is given.

Kurzfassung

Im Rahmen dieser Arbeit wird eine synthetische Agentenbevölkerung für die Verkehrssimulationssoftware MATSim erzeugt. Dies stellt den ersten Schritt beim Erstellen eines agentenbasierten Verkehrsmodells dar, welches im Hinblick auf Schwächen des Vier-Stufen-Modellen Vorteile verspricht. Dazu werden soziodemografische, sozioökonomische und Verkehrsnachfragedaten aus einem bestehenden Verkehrsmodell entnommen und in ein Format überführt, das für die Verwendung innerhalb MATSims geeignet ist. Zu diesem Zweck werden OD-Matrizen zu zusammenhängenden und unterscheidbaren Wegeketten disaggregiert und verschiedenen Bevölkerungsgruppen zugeordnet. Des Weiteren wird die sich ergebende Zeitstruktur des Verkehrsgeschehens hinsichtlich ihrer Plausibilität verfeinert. Im Folgenden wird zunächst eine Einordnung der verwendeten Methodologie in den Kontext bestehender Arbeiten vorgenommen. Im Anschluss erfolgt eine ausführliche Erläuterung der verwendeten Sampling-Maschine sowie der vorab notwendigen Definition der Wahrscheinlichkeitsverteilung der Wegeketten. Im Rahmen der Anwendung der Methode auf die Stadt Wien wurde eine Population von 1000 Agenten erzeugt. Der Prozess und seine Ergebnisse werden detailliert dargelegt. Abschließend wird ein Ausblick auf zukünftig notwendige Forschungsarbeit gegeben.

Table of Contents

Abstract I

Kurzfassung II

Table of Contents III

Table of Figures IV

List of Tables V

List of Abbreviations VI

1 Introduction 1

2 Context 3

3 Metropolis-Hastings based synthesis of round trips 7

3.1 Overview 7

3.2 Modelling the probability distribution 8

3.3 Sampling with Metropolis-Hastings algorithm 9

4 Application to the city of Vienna 11

4.1 Scenario 11

4.2 Modelling case 1: No behavioural information 12

4.3 Modelling case 2: Adding OD reproduction 15

4.4 Modelling case 3: Adding population segmentation and home location reproduction 19

4.5 Modelling case 4: Adding staying-at-home-over-night condition 22

4.6 Modelling case 5: Adding work location reproduction and staying-at-work-during-day condition 24

5 Summary 27

6 Outlook 29

7 References 30

Table of Figures

Figure 1-1: Development of the CO_2 emissions of the four main sectors in the EU-27 between 1990 and 2022. 1

Figure 2-1: Comparison of the population synthesis approaches described by Luger (2017) and Hört & Balac (2021). 4

Figure 2-2: Comparison of three distict levels of abstraction of mobility behaviour (Ortuzar & Willumsen, 2011). 5

Figure 4-1: Mapping of all 23 traffic analysis zones within the study area. 12

Figure 4-2: Number of departures per traffic analysis zone for 10 different samples. 13

Figure 4-3: Number of departures per hour for 10 different samples. 14

Figure 4-4: Development of the shares of round trips visiting two, three resp. four locations over thousand iterations. 15

Figure 4-5: Four scatterplots that show the quality of OD reproduction. Each plot refers to a different weighting on OD reproduction. 17

Figure 4-6: Four different plots that show the development of the shares of round trips visiting two, three resp. four locations over thousand iterations. Each plot refers to a different weighting on OD reproduction. 18

Figure 4-7: Scatterplots that show the quality of the home location reproduction of population groups 06, 07, 08, 09, and 14. 21

Figure 4-8: Comparison of the time structure of the population's travel behaviour with staying-at-home-over-night condition (left) and without staying-at-home-over-night condition (right). 23

Figure 4-9: Comparison of the shares of round trips visiting two, three resp. four locations of the model with at-home-over-night condition (left) and without at-home-over-night condition (right). 23

Figure 4-10: Comparison of time structure of the model with at-work-during-day condition (right) and without at-work-during-day (left). 25

Figure 4-11: Scatterplots that show the quality of the work location reproduction of all working population groups. 26

List of Tables

Table 4-1: Considered population groups in modelling case 3 (Verkehrsverbund Ost-Region (VOR) GmbH, 2024). 19

Table 4-2: Comparison of the shares of round trips visiting two, three resp. four locations in the model and in the household travel survey Österreich Unterwegs (Aschauer, 2016; Bundesministerium für Verkehr, Innovation und Technologie, 2016). 24

List of Abbreviations

ABM	Agent-based model
FSM	Four-step model
HTS	Household travel survey
IPF	Iterative proportional fitting
MATSim	Multi-Agent Transport Simulation Toolkit
MHA	Metropolis-Hastings algorithm
OD	Origin-destination
ÖU	Österreich Unterwegs
PMT	Private motorised transportation
PT	Public transport
Resp.	Respectively
RT	Round trips
TAZ	Traffic analysis zone
VOR	Verkehrsverbund Ost-Region

1 Introduction

Recent data from the European Environmental Agency shows that the total greenhouse gas emissions of the EU-27 have declined by approximately 30% since 1990. A further breakdown of emissions by sector reveals that within the four most important sectors the transport sector is the only one to show an opposing trend (Figure 1-1) (European Environment Agency, 2019).

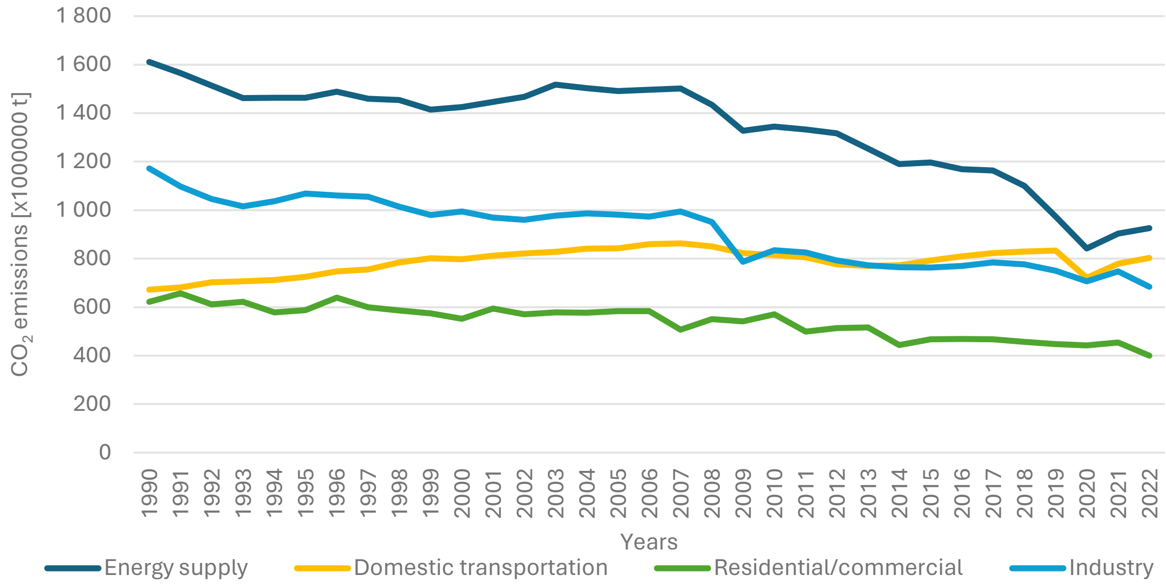

Figure 1-1: Development of the CO_2 emissions of the four main sectors in the EU-27 between 1990 and 2022.

In order to achieve the European Union's sustainability goals, this trend must be reversed. An important step towards this goal is to reduce the use of high-emission transport modes especially in passenger transport. In order to achieve such a modal shift, there is a variety of possible push and pull measures that differ in terms of effort and effectiveness (Beil & Putz, 2023). In order to be able to recommend effective and, above all, efficient measures within decision-making processes, policy sensitive transport models are required. In this context, agent-based traffic models (ABM) are becoming more and more important since they offer a number of advantages, particularly with regard to mode choice and traffic assignment (Flügel & Flötteröd, 2014; Luger, 2017).

In contrast to trip-based models, which consider individual trips to be independent of one another, ABMs are able to understand them as a coherent unit with a fixed order. This allows to capture the temporal dimension of traffic movements and to analyse it at any given time of the day. Another advantage is the ability to simulate interactions within the traffic system in a more realistic manner. While the classic four-step model (FSM) captures congestion purely in terms of time delay, disaggregated models are also capable to map its distribution across the space. Additionally, disaggregated models are less limited in terms of modelling the heterogeneity of the population. They allow for a clearer distinction between people and vehicles. Moreover, the fact that disaggregated models are typically probabilistic models enables the incorporation of uncertainty in their predictions (Flügel & Flötteröd, 2014). With regard to the evaluation of measures in the context of a modal shift, the clear differentiation between people and vehicles is a promising advantage offered by ABMs since it allows for the simulation of multi-modal trips (Monteiro et al., 2014). In the context of the planning of shared mobility services, whether motorised or non-motorised, this is essential (Ramadan & Sisiopiku, 2020).

The aim of this work is to take the first step towards such an agent-based transport model. More precisely, a population of agents for the Multi-Agent Transport Simulation Toolkit (MATSim) will be generated. MATSim is a modular software that offers, among other things, a ready-made implementation for the simulation of mode choice, route choice and departure time choice. It should be noted that modules for carrying out trip generation and destination choice are not included a priori in MATSim (Axhausen & ETH Zürich, 2016). This emphasises the necessity to perform these steps in advance of such a model, namely in the process of population synthesis.

This introductory chapter is followed by a discussion of related work in chapter 2. Chapter 3 provides a detailed explanation of the Metropolis-Hastings based synthesis of round trips, which forms the basis of this study. The application of this approach to the city of Vienna and the resulting population of agents is described in chapter 4. Chapter 5 presents a summary of the work, while chapter 6 gives an outlook for further research on this topic.

2 Context

The four-step model is the most widely used traffic model both in the past and today. It dates back to the 1960s and, as its name suggests, consists of four successive sub-models (Ortuzar & Willumsen, 2011). In the first step the number of trips that the population of each traffic analysis zone (TAZ) will undertake to satisfy their needs is calculated. The trips are summarised and therefore not individually distinguishable (Rasouli & Timmermans, 2014). In the second step, destinations are defined for each generated trip. This is typically done using a gravity model, which states that the probability of a destination is proportional to its attractiveness and inversely proportional to its distance from the respective origin (Rasouli & Timmermans, 2014). In the third step, the resulting origin-destination (OD) matrices are allocated to the various transport modes based on travel time and distance. In the final step the resulting traffic patterns are assigned to the traffic network of the study area (Ortuzar & Willumsen, 2011). When iterating through the four stages the obtained results of each iteration serve as input for the next iteration. Once the results have reached a state of mutual consistency, the calculation is considered complete.

In order to conduct mode choice and trip assignment in MATSim, the results of the first two steps of the FSM, namely trip generation and trip distribution, must be provided in disaggregated form (Axhausen & ETH Zürich, 2016). Therefore, the generated trips must be assigned to individual agents and not summarised in OD matrices. This is achieved through a population synthesis, which is carried out upstream of the ABM. A number of different approaches are available for this purpose, the most notable of which is iterative proportional fitting (IPF). Despite certain limitations, including the restricted number of assignable attributes, it has been the most frequently used method for population synthesis for over two decades. During IPF the rows and columns of an n-dimensional matrix are adjusted until the marginal sums of the matrix coincide with the target data. Each dimension corresponds to an assigned attribute, the values along that dimension are the possible attribute values, and the cells are the number of travellers exhibiting a certain attribute value combination (Ramadan & Sisiopiku, 2020).

Luger (2017) employed this methodology in the course of his population synthesis for the Viennese district of Floridsdorf. An agent population was initially synthesised based on sociodemographic information using IPF, and then socioeconomic information was added with the help of a discrete choice model. Subsequently, the agent's home locations were localised with the assistance of GIS software. However, in order to use this population for simulations in MATSim, information on the agent's mobility demand must be assigned (Luger, 2017).

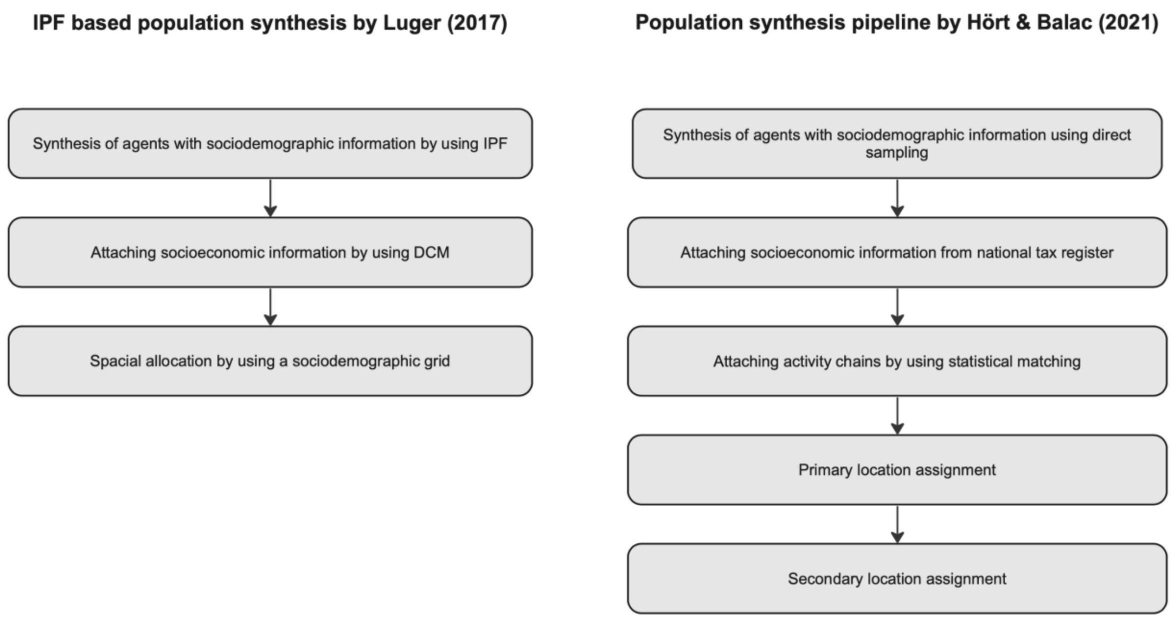

Figure 2-1: Comparison of the population synthesis approaches described by Luger (2017) and Hört & Balac (2021).

Hörl & Balac (2021) describe an extended population synthesis pipeline in respect to this need. As in Luger (2017), a representative agent population is first generated (Figure 2-1), which is then supplemented with demand data. In order to achieve this, each agent is assigned an activity chain derived from household travel surveys (HTS). The primary activities (work and education) are then allocated first, followed by the secondary activities (shopping and leisure). This process is based on commuting matrices that contain concrete OD relations and are also derived from HTSs (Hörl & Balac, 2021; Luger, 2017).

A comparison with the FSM reveals that the described population synthesis methodology is not concerned with generating transport demand. Instead, it merges population data with existing transport demand data and converts it into a format that can be used for MATSim. Furthermore, a population synthesis pipeline is characterised by the large number of consecutive steps (Figure 2-1). In contrast to the FSM, however, the number and order of these sub-models is not clearly defined and can vary depending on the scenario. This undefined and sequential character also harbours disadvantages. For instance, the sequential pipeline design lacks feedback loops that would be necessary to condition upstream computed population attributes on downstream attributes. The concatenation of many, possibly structurally different, processing steps makes it difficult to trace back the effect a pipeline parameter has onto the synthetized population.

The methodology employed in this study is analogous to the pipeline in that it is suitable for the integration of different data sets. However, the process differs significantly. The disaggregation of data from the population level to the agent level does not occur sequentially, but simultaneously, in that a joint distribution of all population attributes of interest is specified before sampling round trips from it. A clear specification that is independent of the rather technical sampling process is hence possible.

The term 'round trip' (RT) employed here refers to activity-based models, which were developed in the early 1990s in response to identified weaknesses in FSMs. This type of model abstracts individuals' mobility patterns based on activity chains, which summarise all of an individual's intended activities within a day in chronological order (Rasouli & Timmermans, 2014). Once the spatial and temporal localisation of these activities has been achieved, it is then possible to simulate the transport behaviour of individuals over time.

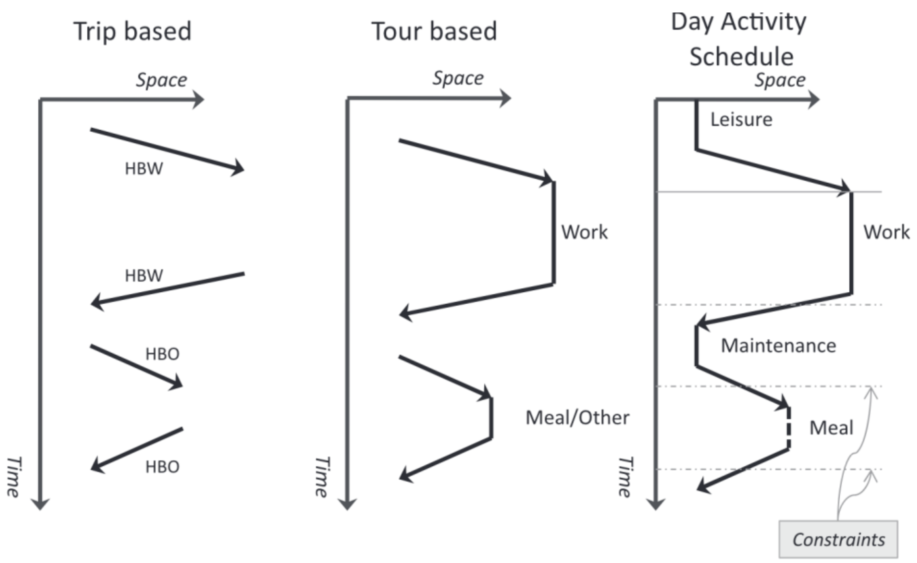

Figure 2-2: Comparison of three distict levels of abstraction of mobility behaviour (Ortuzar & Willumsen, 2011).

Figure 2-2 was taken from (Ortuzar & Willumsen, 2011) and illustrates three distinct levels of abstraction pertaining to the same mobility behaviour. The aforementioned FSM is mostly relying on a trip-based abstraction, wherein trips are regarded as independent and indistinguishable units. In contrast, tour-based models adopt a considerably more realistic perspective. They conceptualise one individual's mobility as a coherent sequence of trips that only concludes when the starting point (home location) is reached. Nevertheless, further tours by the same person during the same day are still construed as independent (Ortuzar & Willumsen, 2011). In activity-based models, day activity schedules summarise all home-based tours of one person within one day to one coherent unit. All distinct tours are linked by the activity carried out in the meantime (Ortuzar & Willumsen, 2011). The benefit of this feature can be illustrated by considering for instance a toll implementation during the morning peak hour. While in a trip-based model this will only affect morning traffic, in an activity-based model, travellers will also spread out their evening return travel times, as the activities they have to implement in between have not changed.

However, FSMs have been widely used for many decades and are therefore widely accepted, especially in practice. Very detailed and well calibrated versions of FSMs are available for many cities, providing a good representation of the prevailing travel demand. It therefore makes sense to combine the advantages of activity-based models and FSMs. An approach that allows this is used for population synthesis in the remainder of this thesis. For this purpose, an OD matrix from an existing FSM is disaggregated into distinguishable round trips. In this context a RT is defined as a sequence of departure locations and associated departure times. After departing from the last location, the agent returns to the first location in the sequence (home location). It is possible to visit one location twice within a RT, which allows agents to make several home-based tours per day. The consideration underlying this model is therefore similar to the daily activity schedule visualised in Figure 2-2. The time between two departures thus corresponds to the time spent at the location, which in the model developed here is at least partially allocated to activities. However, unlike an activity-based model, this is not mandatory. The resulting population of round-trip makers is supplemented by sociodemographic aspects and refined with regard to the time structure of its travel behaviour. At this stage, the work does not claim to be free of behavioural errors but should rather be seen as a proof of concept.

3 Metropolis-Hastings based synthesis of round trips

3.1 Overview

As previously outlined, the synthesis of all-day round trips based on a calibrated FSM offers several advantages. The code used to conduct such synthesis can be found online and was previously deployed for a charging demand analysis in Skelleftea region in northern Sweden (Flötteröd, 2024; Swedish National Road and Transport Research Institute, 2024).

The approach comprises two consecutive steps. Initially, a probability distribution is defined, which assigns a specific weight to all physically possible RTs. The level of this weight is defined based on the data that shall be reproduced. The exact procedure for this definition is illustrated later in this chapter. Once the weights have been defined, the next step is to draw random round trips from this probability distribution. In order to achieve this, an available sampling machinery, namely Metropolis-Hastings algorithm, is employed (Hastings, 1970).

The objective is to generate a population of round-trip makers whose collective traffic behaviour mirrors that observed in the study region. Therefore, traffic analysis zones, that serve as possible origin and destination locations for the agents and a temporal resolution need to be defined. This introduction considers a scenario with 23 TAZs and 24 time-bins, enabling an hourly resolution over the day. The following example shows three round trips resulting from three iteration steps of Metropolis-Hastings algorithm:

MH Iteration 1000
    state = locs[d1070,d1010], bins[13,19]
MH Iteration 2000
    state = locs[d1220,d1190,d1210,d1020], bins[1,6,12,22]
MH Iteration 3000
    state = locs[d1060,d1090,d1210], bins[4,11,12]

It can be seen that the generated RTs comprise a sequence of departure locations and departure time bins. In case of the round trip generated in iteration step 3,000 this means that the attached agent starts his travel at 4 am in zone d1060 and subsequently travels to zone d1090. He remains there until 11 am, after which he travels to zone d1210. He then remains at this location for one hour before returning to his home location (d1060).

During the course of this work not a single RT but a list of many RTs is sampled at each iteration. This will help increase the quality of the data reproduction especially for the later on carried out OD reproduction.

3.2 Modelling the probability distribution

The objective is to minimise the deviation between the sampled list of round trips and round trips in the reality in regard to a certain property. This could be, for instance, the average time people use for traveling during one day. Therefore, each generated round trip of the current sample is compared with a target value. The deviation between the sample and the target value is then calculated and summed up over all of the n round trips of the sample x, which leads to the following error function $E(x)$:

$$E(x) = \sum\nolimits_n |t - s|$$

The value t represents the target value, while s represents its reproduction achieved in the considered sample. This error function can then be used to determine the sampling weight $b(x)$:

$$b(x) \sim e^{-\mu \cdot E(x)}$$

As the equation implies the weight of the sample x is proportional to the e to the power of the negative error function. More precisely, the MHA accepts a possible sample with a probability proportional to $e^{-\mu \cdot E(x)}$. To illustrate this, if μ is set to one, the weight of x with increasing error $E(x)$ falls exponentially towards zero. However, the aim is not the strict minimisation of $E(x)$ but to favour samples with small error functions. If μ is set to a value greater than one, the exponential function falls to zero more quickly, which results in a stronger preference for small $E(x)$ values and thus a better OD reproduction. Conversely, if μ is set to zero, all possible samples are assigned the same weight, which is equivalent to $e^{-0 \cdot E(x)} = 1$. In this case, results are distributed uniformly.

3.3 Sampling with Metropolis-Hastings algorithm

The probability for Metropolis-Hastings algorithm to draw a sample x is $\pi(x)$:

$$\pi(x) = \frac{b(x)}{B}$$

The dividend $b(x)$ is the sampling weight discussed in the previous chapter. The divisor B represents the normalisation constant of the problem – in this case the sum of the sampling weights of all possible lists of round trips. It is determined by:

$$B = \sum_x b(x)$$

Due to the enormous number of possible combinations of lists of round trips, B is very large and therefore hard to calculate (Flötteröd, 2024; Hastings, 1970; Ross, 2023). The number of possible realisations of a single round trip with a maximum of J stops within L available locations and K different time bins is:

$$L^J \cdot \binom{K}{J}$$

For the scenario presented here this results in:

$$23^4 \cdot \binom{24}{4} + 23^3 \cdot \binom{24}{3} + 23^2 \cdot \binom{24}{2} = 2{,}5 \cdot 10^9$$

different possibilities for a single round trip. When sampling lists of round trips with a distinguishable order, this number increases exponentially with the size of the list. With regard to this, the MHA offers a significant advantage. It enables the sampling of the probability distribution $\pi(x)$ without the necessity of knowing its normalisation constant by constructing a Markov chain whose stationary distribution converges to $\pi(x)$. After initiating the Markov chain with an arbitrary starting value, at each iteration step the attempt is made to generate a valid sample. The Markov chain is driven by a proposal distribution q where $q(y, x)$ is the probability that the process considers the next state y given that it is currently in state x. The probability of a considered transition being accepted is determined by:

$$\alpha(x_n, x_{n+1}) = \min\left(\frac{\pi(x_{n+1}) \cdot q(x_{n+1}, x_n)}{\pi(x_n) \cdot q(x_n, x_{n+1})}, 1\right)$$

α is then compared to a randomly generated number U between 0 and 1. For

$$U \leq \alpha(x_n, x_{n+1})$$

the sample is accepted, and the algorithm continues with the next iteration step. For

$$U > \alpha(x_n, x_{n+1})$$

the sample is rejected. The algorithm dwells at the current iteration step and generates new sample candidates until one of them is finally accepted (Ross, 2023).

4 Application to the city of Vienna

This chapter presents the application of Metropolis-Hastings based synthesis of round trips to the city of Vienna. The modelling process was started with a very general model which was then step by step refined.

4.1 Scenario

To reduce the initial computational effort, a total of 23 relatively large TAZs was defined within the study region. These serve as possible origin and destination locations for the agents and were chosen to coincide with Vienna's municipal districts (Figure 4-1). The time division of a day was set to 24 hourly time-bins. Travel time and distance calculations were based on skim matrices from the existing PTV-Visum model of the Verkehrsverbund Ost-Region (VOR) (PTV Planung Transport Verkehr GmbH, 2022; Verkehrsverbund Ost-Region (VOR) GmbH, 2024). This model includes the Austrian federal states of Vienna, Lower Austria, Burgenland and parts of Carinthia and Upper Austria. However, only the section of the model covering Vienna was employed for this study. As the spatial resolution of the VOR model is considerably higher, the skim matrix values within one TAZ were averaged (Verkehrsverbund Ost-Region (VOR) GmbH, 2024).

Figure 4-1: Mapping of all 23 traffic analysis zones within the study area.

The round-trip makers generated are a pure population of car drivers. This is due to the fact that the OD matrices used later on to calibrate the traffic behaviour of the agents only contain trips made by private motorised transportation (PMT). Furthermore, it was specified that a RT can visit a maximum of 4 traffic zones out of the 23 existing. This definition is particularly appropriate given that RTs with more than four visited locations are rather rare occurrence in reality (Bundesministerium für Verkehr, Innovation und Technologie, 2016).

4.2 Modelling case 1: No behavioural information

As further data sets can be successively included in the sampling weights, it is advisable to first ensure the physical feasibility of the generated round trips. Thus, the initial modelling stage does not include any behavioural information. Subsequently, this basic model can then be refined in further stages (Flötteröd, 2024). Hence, the number of departures per zone and the number of departures per time bin in modelling case 1 is not based on any behavioural information and thus is distributed uniformly. The round trips generated must also be uniformly distributed over their number of visited locations. However, as the number of possible combinations of round trips increases with the number of visited locations, it is not sufficient to leave this property uninformed. Instead, it must be explicitly stated that longer round trips are not predominantly sampled due to their greater frequency. In order to enable the daily repeatability of a day plan, it is required that a round trip must be completed within 24 hours and that the arrival time at each location visited must be before the departure time at the same location (Flötteröd, 2024).

These three definitions are sufficient to generate uniformly distributed and behaviourally uninformed results in the first modelling stage. The mathematical realisation of this was a priori implemented in the machinery used (Swedish National Road and Transport Research Institute, 2024).

The generated results can be observed in the following three diagrams. Figure 4-2 illustrates the distribution of departure locations across the 23 TAZs for 10 different samples. As the algorithm was not provided with more precise information on the frequency of the locations, the results produced are, apart from random fluctuations, uniformly distributed.

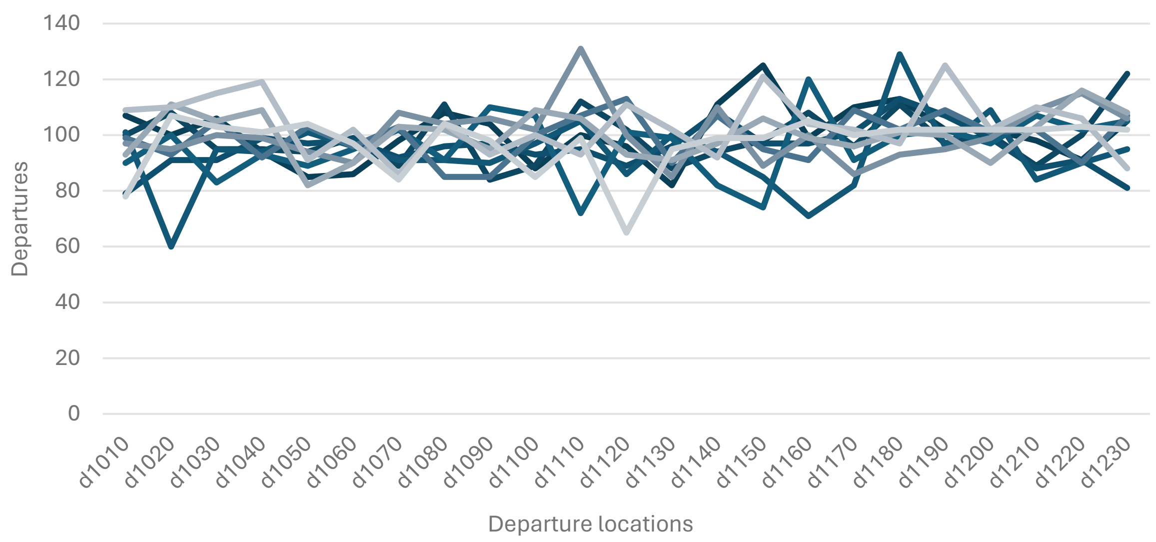

Figure 4-2: Number of departures per traffic analysis zone for 10 different samples.

Figure 4-3 shows the distribution of departure times over the course of the day. As previously stated, each of the ten plotted curves represents a sample of 1,000 RTs. As no information has yet been provided on the time structure of departures, they are also distributed uniformly.

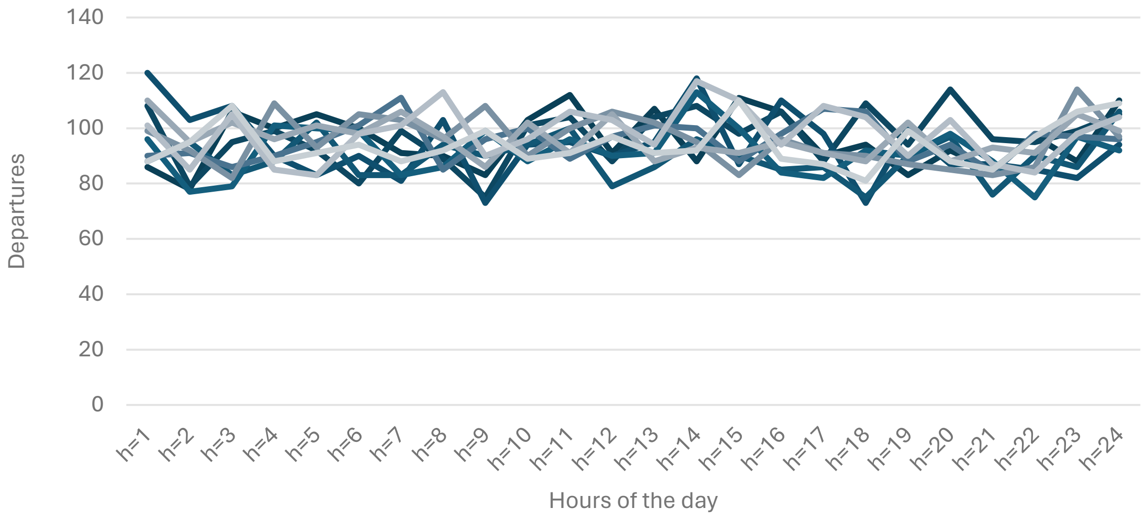

Figure 4-3: Number of departures per hour for 10 different samples.

Figure 4-4 illustrates the shares of round trips visiting two, three respectively (resp.) four locations. As indicated by the green curve, the initial value for the Markov chain was a RT with two visited locations. Consequently, the algorithm begins with a share of 100% for this category. Over the course of approximately 30,000 iterations, the distribution gradually approaches the stated uniform distribution. The vanishing trend after approximately 30,000 iterations suggests that the Markov chain underlying the algorithm has reached its stationary distribution. It can be assumed that, once stationarity is reached, the MH process it is no longer influenced by the initial state (Hastings, 1970).

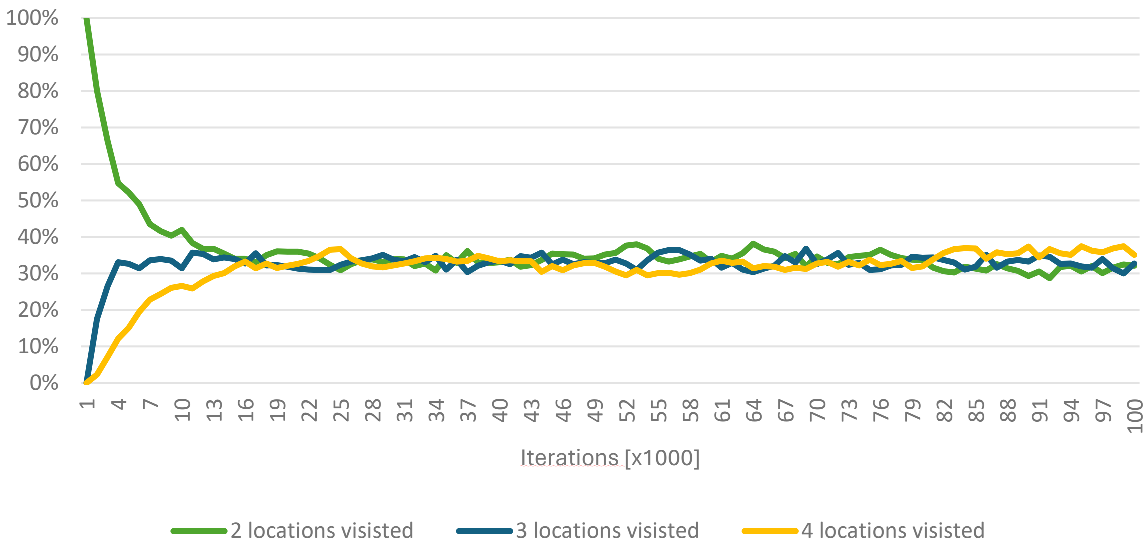

Figure 4-4: Development of the shares of round trips visiting two, three resp. four locations over thousand iterations.

The results show the desired uniform distribution with regard to all considered traffic-specific characteristics. This is far from realistic traffic behaviour, but since no behavioural meaningful information was used for generation, the results are as desired.

4.3 Modelling case 2: Adding OD reproduction

In modelling case 2, the traffic flows generated by the synthetic population were calibrated with the help of OD matrices. As previously mentioned, only OD matrices containing trips of the PMT were extracted from the VOR model, whose traffic demand estimation is based on the EVA algorithm included in PTV-Visum (PTV Planung Transport Verkehr GmbH, 2022; Verkehrsverbund Ost-Region (VOR) GmbH, 2024). Like with the skim matrices in the previous modelling case, the required section of the OD matrix was cut out. Consequently, the resulting matrices only encompass trips that start and end within the city of Vienna. Journeys that have either origin or destination outside this area were not considered. Due to the differing resolutions, it was again necessary to summarize all TAZs of the VOR model within a TAZ of the model under discussion.

The VOR model comprises 25 distinct PMT matrices, deriving from the combination of 15 trip purposes with 15 population groups (Verkehrsverbund Ost-Region (VOR) GmbH, 2024). In order to sample plausible round trips, it was necessary to summarise all trips carried out in one single matrix.

The definition of the sampling weights is such that the synthetic population's traffic flows shall approximate the OD matrix, which from now on will be called the target matrix. For this purpose, $n = 1{,}000$ round trips are generated in each iteration step and their sub-trips are entered into a sample OD matrix. Subsequently, the traffic flows contained within the sample OD matrix are compared with those of the target matrix. This leads to the following error function, $E_{od}(x)$:

$$E_{od}(x) = \sum\nolimits_{n=1000} \sum\nolimits_{rs} |t^{rs} - h \cdot s^{rs}|$$

The value t^{rs} represents the number of trips from zone r to zone s in the target matrix. It is compared with the value s^{rs} which represents the number of trips from zone r to zone s in the sample matrix. As a sample consists of 1,000 round trips and these in turn consist of an average of 2 to 3 sub-trips, the sample matrices contain around 2,000 to 3,000 trips. To ensure comparability with the target matrix, which contains significantly more trips, the scaling factor h is employed:

$$h = \frac{\sum_{rs} t^{rs}}{\sum_{rs} s^{rs}}$$

h equals the quotient of the sum of all trips in the target matrix and the sum of all trips in the sample matrix. Subsequently, the error function $E_{od}(x)$ is used to determine the weight $b(x)$ for any given list of round trips x:

$$b(x) \sim e^{-[\mu_{od} \cdot E_{od}(x)]}$$

As the deviation from the target matrix increases for a given sample, the probability of drawing this sample decreases towards zero. The equation above was evaluated for different weightings on OD reproduction μ_{od}. The corresponding results can be seen in Figure 4-5.

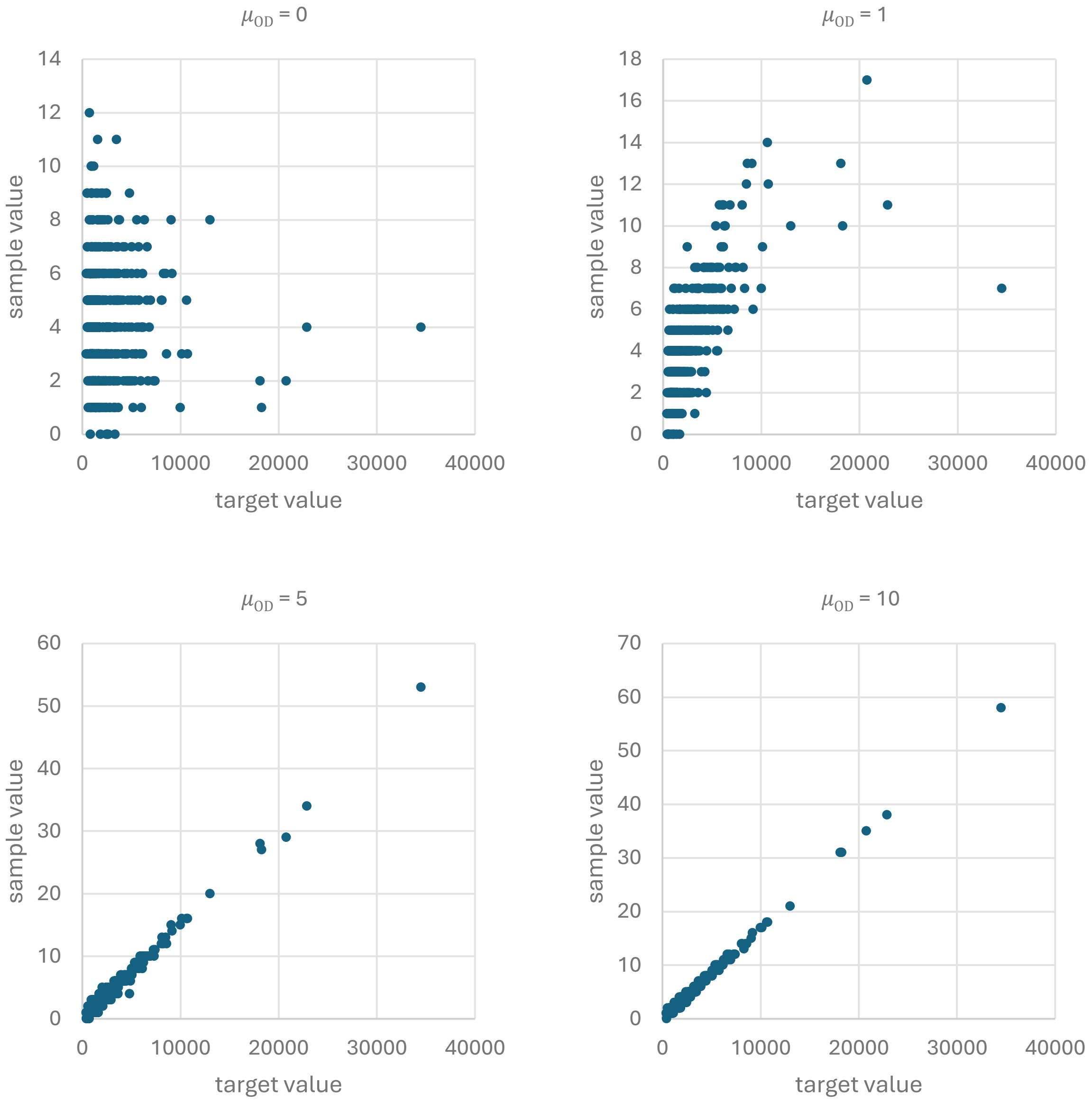

Figure 4-5: Four scatterplots that show the quality of OD reproduction. Each plot refers to a different weighting on OD reproduction.

Each marker within the four plots represents one OD relation of the scenario. The values on the x-axis indicate the number of trips assigned to the respective OD relation in the target matrix, while the values along the y-axis indicate the number of trips in the sample matrix. As the weighting on OD reproduction increases, the markers align along the diagonal of the quadrant, indicating an increasing proportionality between sample and target data. The scatterplot for $\mu_{od} = 10$ shows a near-perfect proportionality.

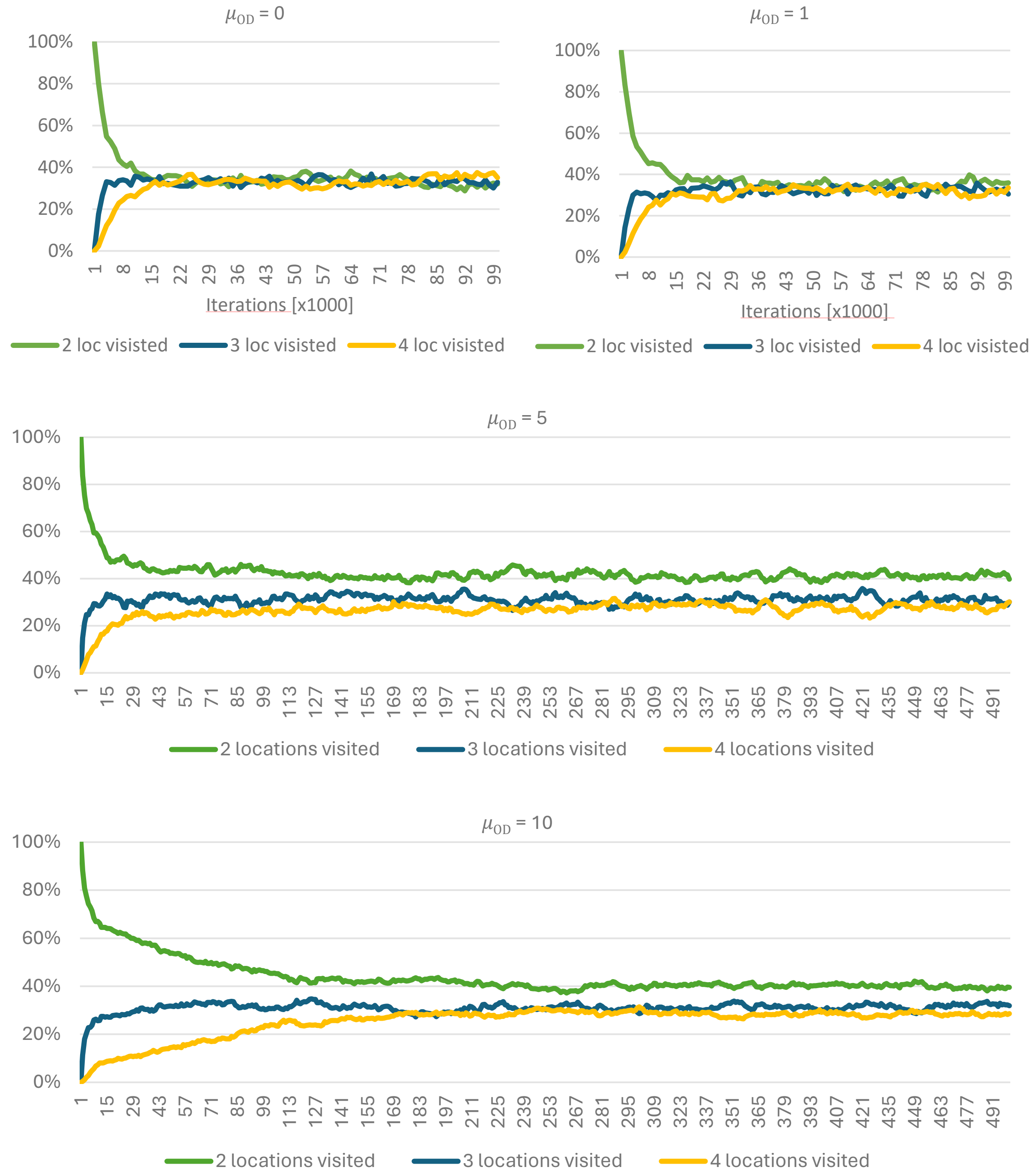

Figure 4-6: Four different plots that show the development of the shares of round trips visiting two, three resp. four locations over thousand iterations. Each plot refers to a different weighting on OD reproduction.

Figure 4-6 presents four distinct plots, which illustrate the effect of the increasing weight on OD reproduction on the shares of round trips visiting two, three, resp. four locations. When comparing the plots with each other differences become evident. While for $\mu_{od} = 0 \; and \; 1$ the stationarity of the Markov chain appears to be reached after approximately 30,000 iterations, for $\mu_{od} = 5\ and\ 10$ this state only occurs after approximately 100,000 resp. 200,000 iterations, respectively. Furthermore, there are differences with regard to the uniform distribution to be achieved. While it is achieved for $\mu_{od} = 0\ and\ 1$, it becomes apparent that for $\mu_{od} = 5\ and\ 10$ it easier for the algorithm to fulfil the OD reproduction by using a larger proportion of shorter RTs. This represents a contradiction to the stated uniform distribution but does make sense in terms of travel behaviour. As the behaviour captured in the target matrix is also characterised by a larger proportion of shorter round trips, the algorithm reproduces a property that is not stated explicitly but is certainly present within the target data.

4.4 Modelling case 3: Adding population segmentation and home location reproduction

In the course of modelling step 3, the agent population was divided into five distinct population groups and the distribution of the home locations of each group was specified for each TAZ using the structural data of the VOR model (Verkehrsverbund Ost-Region (VOR) GmbH, 2024). Since the population generated here is exclusively comprised of car drivers, only population groups with access to a vehicle were considered. They are listed in Table 4-1.

pop group	lower limit	upper limit	occupation	car access
pop group 06	19 years	35 years	working	yes
pop group 07	19 years	35 years	other	yes
pop group 08	35 years	65 years	working	yes
pop group 09	35 years	65 years	other	yes
pop group 14	65 years	no upper limit	not specified	yes

Table 4-1: Considered population groups in modelling case 3 (Verkehrsverbund Ost-Region (VOR) GmbH, 2024).

From now on each sampled list of 1,000 RTs consists of five population groups, with the size of each group being proportional to its size in reality. The number of home locations per zone of each population group in the sample was then compared to the number of home locations per zone of each population group in the target data. The difference between these values was summarised over the 23 zones and the 5 considered groups to obtain the error function $E_{home}(x)$:

$$E_{home}(x) = \sum_n \sum_r |t^r - h \cdot s^r|$$

Inside this equation t^r represents the number of people of population group n that have their home location at zone r in the target data. Respectively, s^r represents the number of agents inside the synthetic population group n that have their home location at zone r. The scaling factor h is defined as:

$$h = \frac{\sum_r t^r}{\sum_r s^r}$$

It becomes clear, that the mathematical formulation of $b(x)$ is identical to that of the preceding modelling step. The sole distinction is that the sample and the target are now compared in regard to the home locations of RTs. The resulting sampling weight is as follows:

$$b(x) \sim e^{-[\,\mu_{od} \cdot E_{od}(x) + \mu_{home} \cdot E_{home}(x)\,]}$$

The effect of this updated sampling weight is illustrated in the following figures. The results were obtained with $\mu_{home} = 10$.

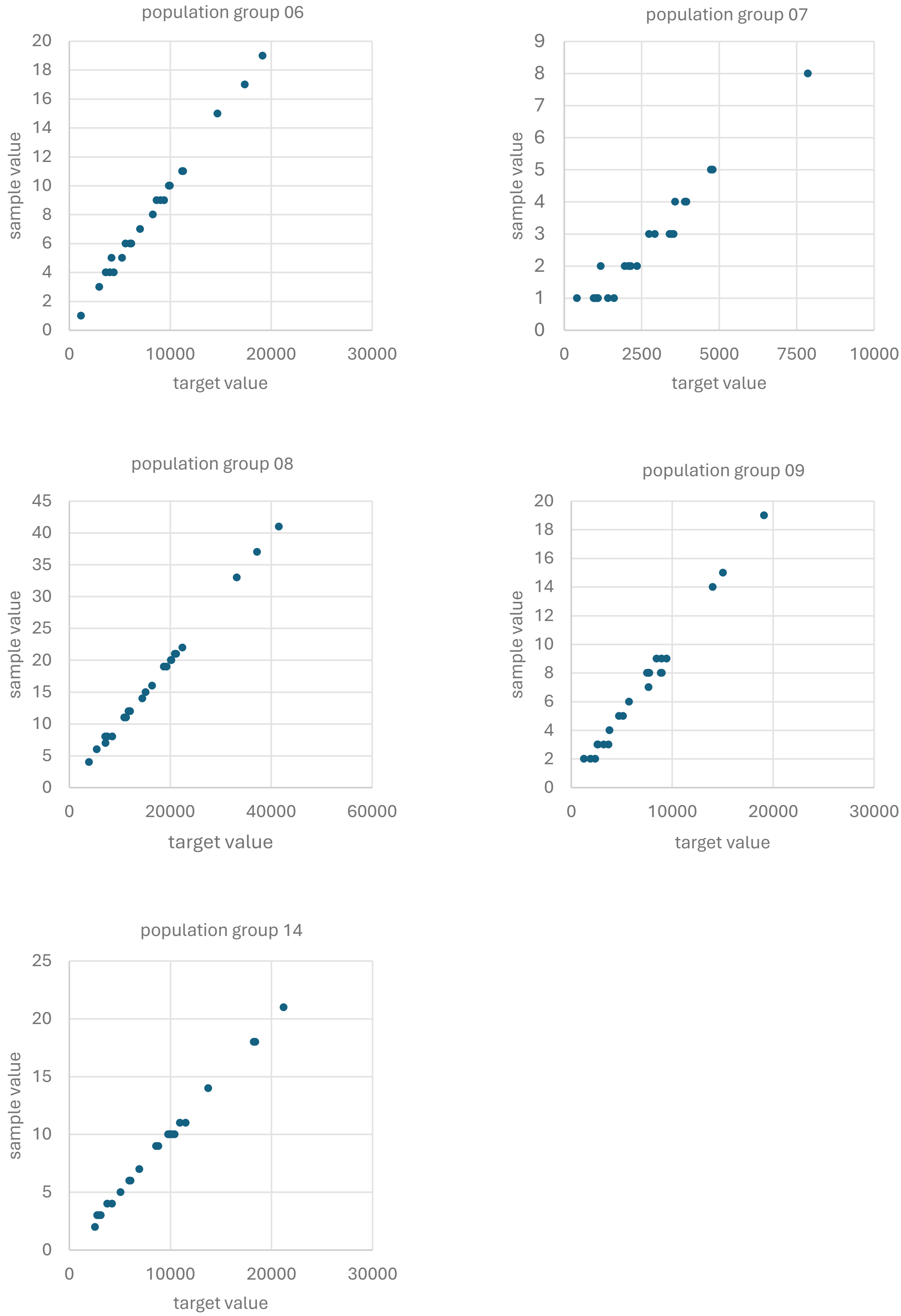

Figure 4-7: Scatterplots that show the quality of the home location reproduction of population groups 06, 07, 08, 09, and 14.

Figure 4-7 presents five scatterplots each for one population groups. The markers within each plot represent the 23 traffic zones. The values along the x-axis indicate the number of home locations in the target data, while the values along the y-axis show the number of home locations in the generated sample. The alignment of the markers along the diagonal of each plot indicates that the MHA was able to reproduce the specified home location distribution, since sample and target appear to be proportional. The quality of the OD reproduction and of the uniform distribution of the shares of round trips visiting two, three resp. four locations is not affected by the home location reproduction.

4.5 Modelling case 4: Adding staying-at-home-over-night condition

In modelling case 4, the uniformly distributed time structure of traffic behaviour (Figure 4-3) was modified to a more behavioural meaningful structure. Consequently, an additional term was added to the calculation of the sampling weights to address the disproportionately high traffic volume during nighttime hours. The assumption was made that every round-trip maker spends at least 10 consecutive hours at their home location within the time interval between 7 pm and 7 am. The mathematical realisation was identical to that employed for the previous sampling weights. With the only difference being that sample and target are now compared in regard to the number of round trips that stays at their home location overnight. From this, the error function $E_{night}(x)$ and subsequently the updated $b(x)$ was obtained:

$$b(x) \sim e^{-[\mu_{od} \cdot E_{od}(x) + \mu_{home} \cdot E_{home}(x) + \mu_{night} \cdot E_{night}(x)]}$$

The effect of this additional term on the time structure of the traffic can be seen in the figure 4-8. The results were obtained with $\mu_{night} = 10$.

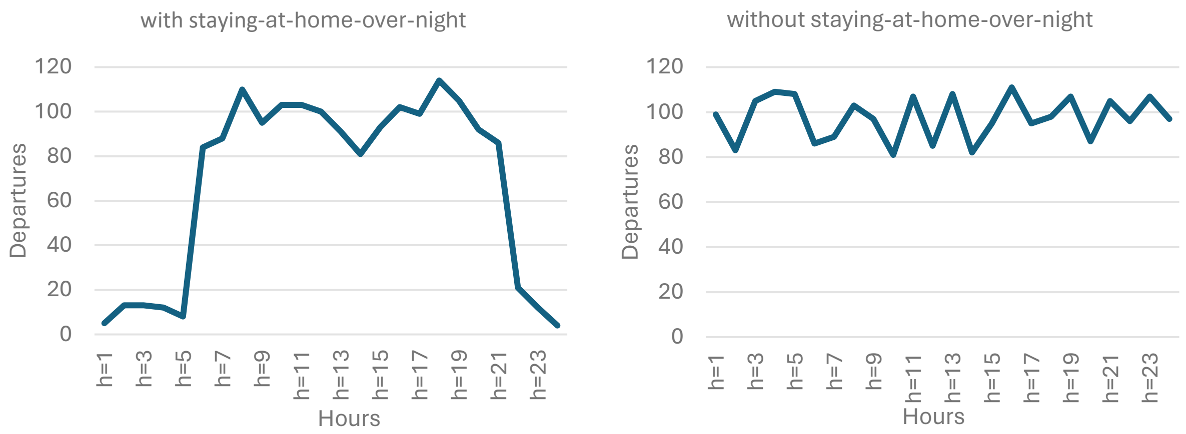

Figure 4-8: Comparison of the time structure of the population's travel behaviour with staying-at-home-over-night condition (left) and without staying-at-home-over-night condition (right).

While daytime traffic remains at a similar level, night-time traffic decreases significantly. The lack of increase in daytime traffic is due to the fact that while the number of 1,000 round trips remains unchanged, the number of sub-trips falls from approximately 2,200 in the previous model to approximately 1,600 here. As illustrated in Figure 4-9, this phenomenon can be attributed to a decline in the proportion of round trips with three and four locations visited.

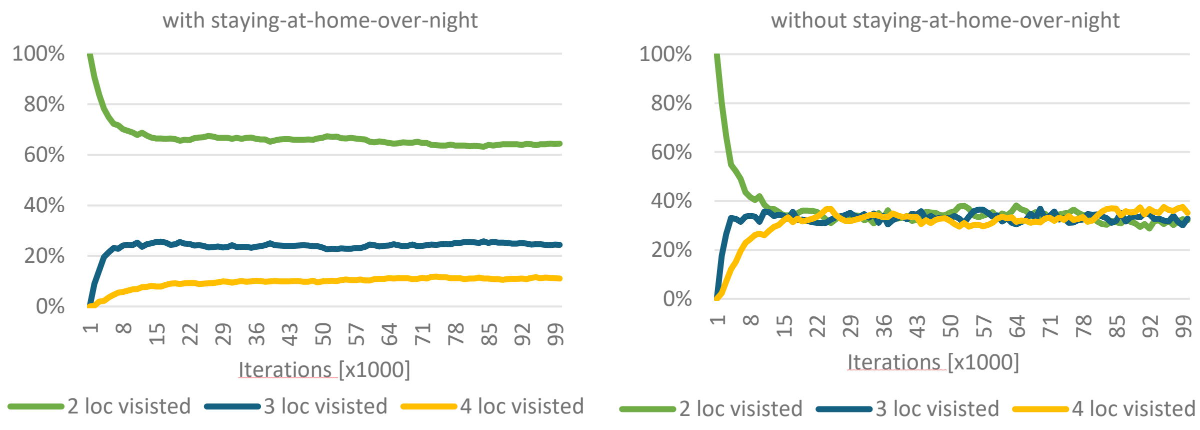

Figure 4-9: Comparison of the shares of round trips visiting two, three resp. four locations of the model with at-home-over-night condition (left) and without at-home-over-night condition (right).

The definition that 10 out of the 24 hours may not be utilized for traveling has resulted in a stronger reduction of potential realizations of longer RTs than of shorter ones. The resulting distribution of round trips containing two, three, resp. four stops again contradicts the initially adopted uniform distribution because more and more behavioural information is used. Table 4-2 illustrates a comparison of the distribution within the model with the distribution derived from the household travel survey Österreich Unterwegs (ÖU). Given that the target data is based on the traffic behaviour described by ÖU, the approximation of the distributions represents a plausible development. Furthermore, the quality of the OD reproduction remained unchanged.

number of visited locations	2	3	4
share in the model	67%	24%	11%
share in the ÖU travel survey	72%	16%	8%

Table 4-2: Comparison of the shares of round trips visiting two, three resp. four locations in the model and in the household travel survey Österreich Unterwegs (Aschauer, 2016; Bundesministerium für Verkehr, Innovation und Technologie, 2016).

4.6 Modelling case 5: Adding work location reproduction and staying-at-work-during-day condition

Modelling case 5 is the final stage of this work. In order to refine the choice of work locations for agents of working population groups, the number of work locations within each of the 23 TAZ was specified. The data required for this was again taken from the structural data of the VOR model (Verkehrsverbund Ost-Region (VOR) GmbH, 2024). Consequently, it was postulated that the greater the number of work locations within a zone, the greater the probability that agents of the working population groups will spend their longest out-of-home activity (working) there. The mathematical implementation of this is identical to that employed in the previous modelling cases. The only difference is that now sample and target are compared in regard to the reproduction of the work location distribution.

Beyond that, the temporal structure of travel behaviour was further refined during this modelling step. Consequently, it was postulated that all agents of a working population group spend a minimum of nine consecutive hours within the time interval between 7 am and 7 pm at their longest out-of-home location (work location). This results in the updated sampling weight:

$$b(x) \sim e^{-[\,\mu_{od} \cdot E_{od}(x) + \mu_{home} \cdot E_{home}(x) + \mu_{night} \cdot E_{night}(x) + \mu_{work} \cdot E_{work}(x)\,]}$$

The enhanced temporal structure resulting from this additional sample condition is illustrated in Figure 4-10. The results were obtained with a weight of $\mu_{work} = 10$.

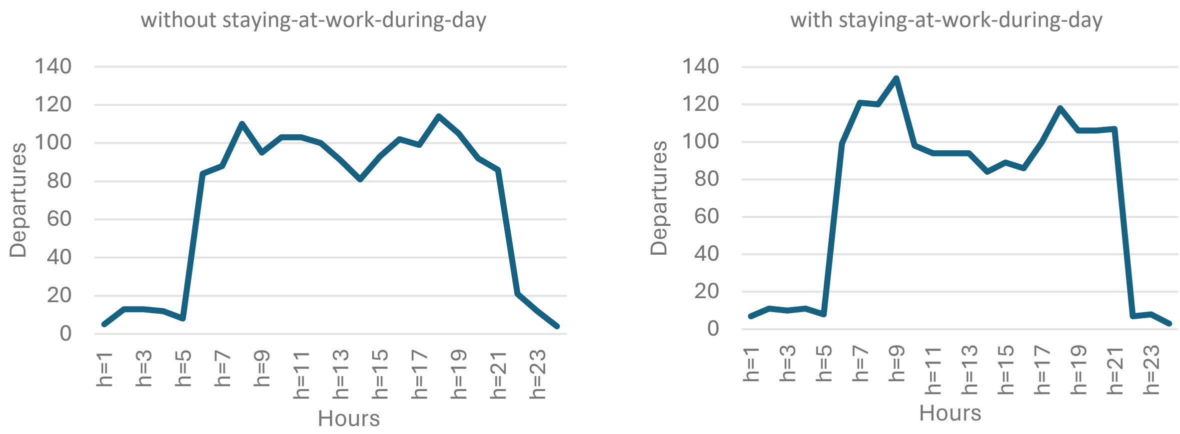

Figure 4-10: Comparison of time structure of the model with at-work-during-day condition (right) and without at-work-during-day (left).

The daily traffic pattern now exhibits a more distinct morning and evening peak, accompanied by a reduced traffic volume in the interim period. This is attributable to the fact that agents of the working population groups are highly likely to be at a work location during this time, and therefore do not utilise the time to move around. As expected, night-time traffic remains unchanged.

Figure 4-11 presents scatterplots illustrating the quality of work location reproduction for the two working population groups 06 and 08. Each marker represents one of the 23 TAZs within the scenario. The values along the x-axis indicate the number of work location of each zone in the target data, while the values along the y-axis indicate the number of agents of the respective population group that have their working location there. The alignment of the markers along the diagonals indicates again that the required proportionality between sample and target has been achieved. The additional error function does not affect the reproduction quality of the previously employed sampling weights.

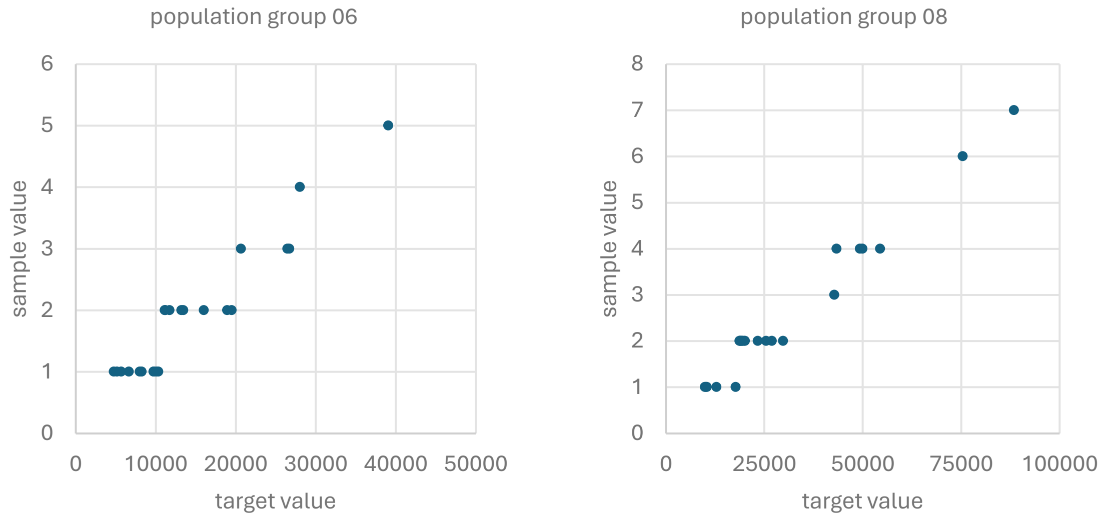

Figure 4-11: Scatterplots that show the quality of the work location reproduction of all working population groups.

5 Summary

The results demonstrate that the Metropolis-Hastings based synthesis of round trips is an effective method for generating a population of agents for simulation in MATSim based on OD matrices. The runtime for generating a list of 1,000 round trips on a 2.7 GHz Intel Core i5 processor is a few minutes, which seems promising towards further refinements.

During the synthesis process, each of the 1,000 generated agents received an exact day-plan in the form of a RT, which was allocated both temporally and spatially. This approach enables the trips of the agents to be distinguished, facilitating their tracking during a later MATSim simulation. The properties of each round-trip maker can be accessed at any time during the simulation, enabling the precise modelling of mode and route choice decisions. This represents a significant enhancement in quality when compared to the underlying FSM.

In addition, the agent population of the final modelling case is divided into five segments. Their size and home location distribution is the same as in the target data, which alone does not represents an extension of the VOR model. However, the 15 population groups contained in the VOR model (5 relevant for PMT) reach the limits of what is computationally feasible for a model of this size which makes finer population segmentation difficult to achieve within PTV-Visums (PTV Planung Transport Verkehr GmbH, 2022; Verkehrsverbund Ost-Region (VOR) GmbH, 2024). Therefore, the advantage of the Metropolis-Hastings based synthesis of round trips is evident. Building on the successful reproduction achieved here, the population can now be segmented further by, for example, linking sociodemographic and socioeconomic characteristics of agents to the distribution of these aspects prevailing at their home location. This enables the population to be constructed in a way that reflects the heterogeneity of the real population in a better way.

By linking more detailed departure time decisions during the simulation to sociodemographic and socioeconomic characteristics of the agents, the time structure of traffic events will also benefit from the increasing heterogeneity of the population. This will enable precise dynamic analyses of traffic flows. For instance, this seems very promising for capacity measurement in public transport (PT) outside of peak hours, which is difficult to achieve within PTV-Visum. It is noteworthy that very few modelling assumptions were sufficient to generate a realistic time structure in the course of this work.

In general, it can be said that the results respond plausibly to the modelling assumptions made. Furthermore, the quality of the data reproduction is adjustable and more than satisfactory, even when reproducing several data sets simultaneously. Throughout this work, deviations from the stated properties could always be explained conclusively and at no time represented a contradiction to behavioural facts.

The advantages of the Metropolis-Hastings based synthesis of round trips over the population synthesis pipeline described in chapter 2 were confirmed in the course of this work. As with Hörl & Balac (2021), additional demands can be added incrementally but without the necessity of additional (sub-) models. The primary advantages of summarising all requirements in one probabilistic model and solving them simultaneously are twofold: firstly, it improves the traceability of assumptions and secondly, the sampling process is clearly decoupled from model specification.

6 Outlook

In order to be able to carry out reliable analysis with the population produced, a more refined synthesis is still required. In particular, a higher spatial resolution is essential. Therefore, within the VOR model, data is available with a division of the study area into 250 resp. 1,000 TAZs (Verkehrsverbund Ost-Region (VOR) GmbH, 2024).

Furthermore, it would be beneficial to incorporate information that goes beyond what is already included in the VOR model. With regard to mode choice decisions, socioeconomic aspects of the agents are particularly valuable. Depending on the scenario, it will also make sense to carry out the synthesis for PT, cycling, and walking OD matrices in order to achieve an overall representation of the population.

It should be noted that, in contrast to Hörl & Balac (2021), there is no composition of individual agents into households. An expansion of the population in this direction should be considered in any case.

The most significant limitation of this approach and the discussed population synthesis pipeline is undoubtedly the necessity for travel demand data as an input. In both cases, demand data is not generated independently, as it is the case within most FSMs. A further development of the Metropolis-Hasting based synthesis of round trips to a stand-alone model is a goal worth pursuing. However, further research is required to achieve this objective.

7 References

Aschauer, F. (2016). Travel and activities in different data sources derived from travel surveys, time use surveys and a combined approach.

Axhausen, K. W. & ETH Zürich. (2016). The Multi-Agent Transport Simulation MATSim (ETH Zürich, A. Horni, K. Nagel, & TU Berlin, Eds.). Ubiquity Press. https://doi.org/10.5334/baw

Beil, D., & Putz, L.-M. (2023). Modal shift measures to increase the use of eco-friendly transport modes: A literature review. Transportation Research Procedia, 72, 4279–4286. https://doi.org/10.1016/j.trpro.2023.11.343

Bundesministerium für Verkehr, Innovation und Technologie. (2016). Österreich unterwegs 2013/2014.

European Environment Agency. (2019, March 22). CO2 emissions from cars: Facts and figures (infographics). Topics | European Parliament. https://www.europarl.europa.eu/topics/en/article/20190313STO31218/co2-emissions-from-cars-facts-and-figures-infographics

Flötteröd, G. (2024). Charging demand analysis in the face of overwhelming uncertainty.

Flügel, S., & Flötteröd, G. (2014). EVALUATION of methods for calculating traffic assignment and travel times in congested urban areas with strategic transport models. TOI.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Hörl, S., & Balac, M. (2021). Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. Transportation Research Part C: Emerging Technologies, 130, 103291. https://doi.org/10.1016/j.trc.2021.103291

Luger, B. (2017). Generation of a Synthetic Population for MATSim Models Using Multidimensional Iterative Proportional Fitting and Discrete Choice Models.

Monteiro, N., Rossetti, R., Campos, P., & Kokkinogenis, Z. (2014). A Framework for a Multimodal Transportation Network: An Agent-based Model Approach. Transportation Research Procedia, 4, 213–227. https://doi.org/10.1016/j.trpro.2014.11.017

Ortuzar, J. de D., & Willumsen, L. G. (2011). Modelling Transport.

PTV Planung Transport Verkehr GmbH. (2022). PTV Visum 2022—Manual.

Ramadan, O., & Sisiopiku, V. (2020). A Critical Review on Population Synthesis for Activity- and Agent-Based Transportation Models. In S. De Luca, R. Di Pace, & B. Djordjevic (Eds.), Transportation Systems Analysis and Assessment. IntechOpen. https://doi.org/10.5772/intechopen.86307

Rasouli, S., & Timmermans, H. (2014). Activity-based models of travel demand: Promises, progress and prospects. International Journal of Urban Sciences, 18(1), 31–60. https://doi.org/10.1080/12265934.2013.835118

Ross, S. M. (2023). Simulation (Sixth edition). Academic Press, an imprint of Elsevier.

Swedish National Road and Transport Research Institute. (2024). Matsim-projects/od2roundtrips at master · vtisweden/matsim-projects · GitHub. https://github.com/vtisweden/matsim-projects/tree/master/od2roundtrips

Verkehrsverbund Ost-Region (VOR) GmbH. (2024). Verkehrsverbund Ost-Region. Verkehrsverbund Ost-Region. https://www.vor.at/en/