Metropolis-Hasting based synthesis of all-day round-trips from OD-matrices and limited population segmentation

Franz-Xaver Rupprecht*[1], Gunnar Flötteröd[2], Yusak O. Susilo[1]

[1] Department of Landscape, Spatial and Infrastructure Sciences, Institute for Transport Studies (IVe), University of Natural Resources and Life Sciences, Vienna, Austria
[2] Department of Communication and Transport Systems, Linköping University, Sweden & The Swedish National Road and Transport Research Institute

SHORT SUMMARY

This work considers the synthesis of coherent and distinguishable round trips from existing travel demand data (OD-matrices) in order to enable person-level mobility simulations with agent-based simulations such as MATSim (Horni et al., 2016). Five distinct population groups are considered and the spatial distribution of their home locations over the study region is reconstructed. Also, the spatial distribution of work locations is reproduced. The round trips are equipped with a plausible temporal structure by reproducing activity durations and departure times derived from a household travel survey. The Metropolis-Hastings (Hastings, 1970) algorithm is used to sample from a probability distribution defined based on input data drawn from an upstream four-step travel demand model. The method is applied to the city of Vienna, resulting in the generation of a population of 50,000 representative round trip makers.

Keywords: agent-based modelling; population synthesis; travel demand; Metropolis-Hastings algorithm; OD-matrices

1. INTRODUCTION

Agent-based models are widely used to simulate the behaviour of individuals, their interactions with each other and the surrounding transport system (Department for Transport, 2024). With regard to policy evaluation, agent-based models outperform traditional four-step models (Tozluoglu, 2024). Due to their disaggregated nature, they are sensitive for a much wider range of policy measures (Rasouli & Timmermans, 2014). To perform reliable simulations with agent-based models, an accurate representation of the traveller population is required. This comprises a correct representation of socioeconomic and sociodemographic aspects of the population, as well as those aspects of individual-level travel behaviour that are endogenous to the agent-based model (Müller & Axhausen, 2010; Pougala, 2024).

Population synthesis

The most frequently used approach for synthesising a population of agents is iterative proportional fitting (IPF). It is used to disaggregate data from the population level to the individual-level by adjusting the rows and columns of a multi-dimensional matrix until the marginal sums match the aggregate data. The matrix's dimension reflects the number of reproduced attributes, along each dimension the possible values of each attribute occur, and the number of cells is equal to the number of distinct individual groups with a certain attribute-value combination (Ramadan & Sisiopiku, 2020). Information on the correlation between different attributes at the population level needs to be derived from microdata and then scaled up to the population level (Choupani & Mamdoohi, 2016). Despite its widespread use, there are three major limitations of IPF: The integer conversion problem, the zero cell problem, and the disadvantage of either synthesising a

population at the individual or household-level, but not both simultaneously (Choupani & Mamdoohi, 2016).
Simulation-based population synthesis uses Markov chain Monte Carlo (MCMC) methods to sample a synthetic population from a previously defined probability distribution. Farooq et al. (2013) propose this approach to overcome some of the drawbacks of IPF. One of their main advantages is that they truly synthesize population rather than just cloning it from microdata (Farooq et al., 2013; Ramadan & Sisiopiku, 2020). The approach used within this work also utilises a MCMC method and therefore falls in the wider field of simulation-based population synthesis.

Travel demand synthesis

For travel demand synthesis, the two predominant approaches are trip-based and activity-based models (Bastarianto et al., 2023). Trip-based models are closely related to the four-step models and capture travel in terms of numbers of independent trips. In contrast, activity-based models do not focus on individual trips but on the activities conducted in between and derive travel from the need to participate in these activities. This leads to more sensitive models that are capable of evaluating transport policies (Tozluoglu, 2024).
ALBATROSS is a prominent example of an activity-based model. The built-in scheduling engine includes a decision tree with 27 steps and models decision-making processes in a hierarchical manner including location choice, departure time choice, activity duration and mode choice (Rasouli & Timmermans, 2014).
The FEATHERS framework is based on the scheduling engine of ALBATROSS. It is a dynamic activity-based model that incorporates rescheduling, learning effects and traffic routing based on a previously existing trip-based model (Bellemans et al., 2010).

Joint synthesis of population and travel demand

Both population synthesis and travel demand generation are necessary to enable agent-based simulations at the network assignment level. It hence is useful to combine them in one model. This tends to be realized in the form of “demand generation pipelines” that concatenate several population construction steps and travel behavioural models (Hörl & Balac, 2021).
Hörl & Balac (2021) describe one such pipeline. They use direct sampling to synthesise the population from sociodemographic data, then attach supplementary socioeconomic information to the agents. Statistical matching is used to assign activity chains to each individual. Subsequently, primary and secondary activity locations are sampled from different data sources (Hörl & Balac, 2021).
Mallig et al. (2013) present MobiTopp, a further example of combining population and travel demand synthesis. It uses a modular approach comprising many different sub-models, some of which rely on utility maximisation principles (Mallig et al., 2013).
Although the described pipelines produce results of convincing quality, there are drawbacks. Due to their complexity, the effect of assumptions made is hard to trace. The concatenation of many sub-models promotes error propagation (Farooq et al., 2013). Their sequential character does not allow to adjust assumptions of previously performed modelling steps, rendering the model sensitive to the order in which modelling steps are pipelined (Pougala, 2024). Since population synthesis is a highly underspecified problem, the ability of pipelines to reproduce a given dataset lacks (model-based) indication of why the synthesized population is better than any other of the same data reproduction quality. The order, number, and type of algorithms used can vary widely between different approaches, making it difficult to compare and validate results (Hörl & Balac, 2021).
In order to address the problems identified above, this work propose a method which overcomes the sequential nature of population synthesis pipelines by using the Metropolis-Hastings (Hastings, 1970) algorithm to impute all relevant travel behavioural aspects of a given synthetic

population simultaneously. It decouples model specification (in the form of target weights attached to travel behavioural aspects) and demand synthesis (sampling according to these weights). The approach rests on interpretable parameters, is easy to overview and replicate, does not assume any particular ordering of sub-models and hence reduces possible error propagation. The case study presented here focuses on the individual-level travel demand synthesis for a given synthetic population, even though the approach has clear potential for performing both simultaneously (Rupprecht, 2024).

2. METHODOLOGY

The individual-level travel demand of a given synthetic population is represented in the form of a list of round-trips. Each round trip reflects the daily travel schedule of an agent, comprising a sequence of activity locations and the associated departure times. This round-trip list constitutes the state of the Metropolis-Hastings algorithm. At each iteration, the algorithm samples a variation of this list and accepts or rejects this proposal based on the model-informed weights specified below.

A Bayesian approach is adopted to define the probability distribution of round trips. A behaviourally uninformed prior in the form of a maximum entropy distribution over all possible round trips with a given expected number of travel episodes is chosen. Given aggregate data sources, measurement equations are then defined which yield a likelihood function that, when multiplied into the prior distribution, gives the desired posterior distribution. The following error functions are used in the measurement model:

$$E(x) = |t - s(x)| \tag{1}$$

where t is a target value, such as the number of trips in one origin-destination (OD) relation or the number of agents exhibiting a particular travel time pattern, and $s(x)$ extracts the correspondingly simulated value from the round trip list x. From each such error function, unnormalized likelihood term

$$b(x) = e^{-\mu \cdot E(x)} \tag{2}$$

is derived. It is not larger than 1, and a higher deviation between the sample and the target will lead to a lower likelihood of that sample. When reproducing several data sets simultaneously the different error functions can be weighed against each other using the scaling factor μ. The most sophisticated model presented here comprises 11 error functions that are outlined in the following. Flötteröd (2025) provides a detailed specification of the sampling machinery used here.

3. RESULTS AND DISCUSSION

The method was applied to the city of Vienna. 250 Travel Analysis Zones (TAZ) of an existing PTV Visum model (PTV Planung Transport Verkehr GmbH, 2022) were considered. A temporal resolution of 24 time bins was chosen and the expected number of locations visited during one round trip was set to 2.8 (according to Österreich Unterwegs; Bundesministerium für Verkehr, Innovation und Technologie, 2016) in the uniformed prior distribution. Figure 1 (left) shows the proportions of round trips of different lengths when only using the prior distribution sampling. The mean trip length is 2.9, which apart from random noise coincides with the expected value of the maximum entropy prior distribution. Figure 1 (centre and right) shows that round trips are uniformly distributed in space and time. Up to this point, no real data beyond the mean trip length is reproduced. By doing so, the model is gradually pushed in a more behaviourally meaningful direction.

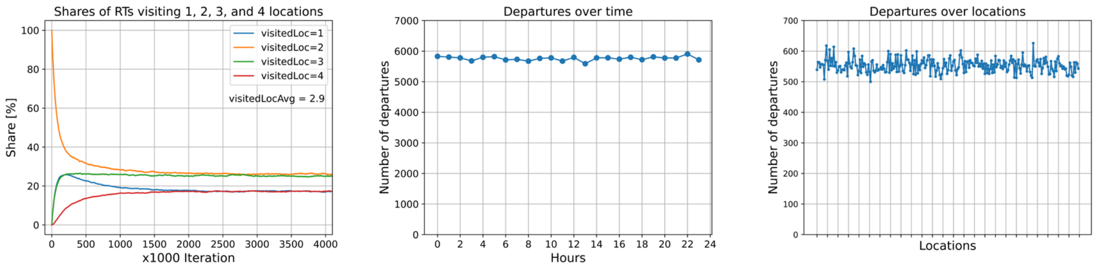

Figure 1: Shares of RT visiting 1,2,3, and 4 locations in the *uninformed model* (left); Departures over time in the *uninformed model* (centre); Departures over locations in the *uninformed model* (right)

A likelihood term representing a 24h OD-matrix is added. The error function compares each OD flow of the target OD-matrix with the corresponding OD flow of the sample OD-matrix (obtained by aggregating all time-dependent round trips into one 24h-OD-matrix). The scatter plot in Figure 2 (left) indicates that the algorithm achieves proportionality between sample and target, and shows a good quality of reproduction. Figure 2 (right) reveals the effect of OD reproduction on the distribution of round trip lengths. The algorithm generates a smaller proportion of round trips with a higher number of visited locations, lowering the mean trip length to 2.6. This phenomenon may be attributed to the fact that the combinatorial problem (sampling a list of round trips that reconstructs the target OD-matrix) becomes easier to solve for the algorithm by using shorter round trips. When removing round trips that visit only one location (traveller stays at home location) from the sample, the proportions match with the target data.

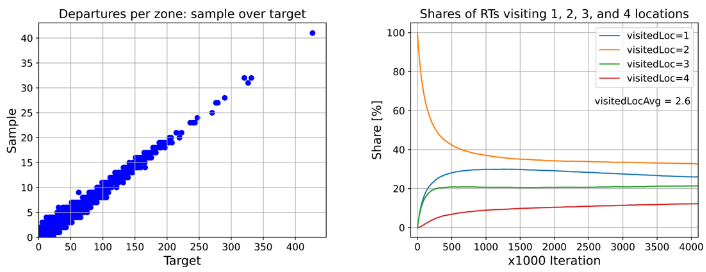

Figure 2: Quality of OD reproduction in the *OD reproduction model* (left); Shares of RT visiting 1,2,3, and 4 locations in the *OD reproduction model* (right)

In a subsequent modelling stage, five different population groups are defined, and the number of people of each group living in each TAZ is specified. Separately, for each population group an error function compares the number of home locations per zone in the sample with the target data to ensure correct reproduction of the spatial distribution of the home locations. Additionally, the number of work locations per TAZ is specified. It is stated that all members of the two working population groups spend their longest out of home activity at such a location. For each working population group another error function ensures correct reproduction of the spatial distribution of work locations. Figure 3 (left) shows good reproduction quality of the spatial distribution of home locations of population group 6 whereas figure 3 (right) shows the same for work locations of population group 8. All other population groups reach the same level of reproduction quality.

Figure 3: Quality of home location reproduction of population group 6 in the *OD and location reproduction model* (left); Quality of work location reproduction of population group 8 in the *OD and location reproduction model* (right)

In the final stage plausible time structure is added to the model. Therefore, a two-dimensional matrix holding the number of persons allocated to all possible combinations of the duration spent at their home location (number of consecutive hours) and the hourly time bin of the departure from the home location is derived from Österreich Unterwegs (Bundesministerium für Verkehr, Innovation und Technologie, 2016). By adding a further error function correct reproduction of this data is ensured. Another matrix of the same kind was derived for work durations and departures from work locations of working population groups. By again adding an error function this data is being reproduced. Figure 4 shows near perfect reproduction quality of both duration-departure matrices. Figure 5 shows the effect of adding the reproduction of the home respectively work duration-departure matrix on the time structure of the model.

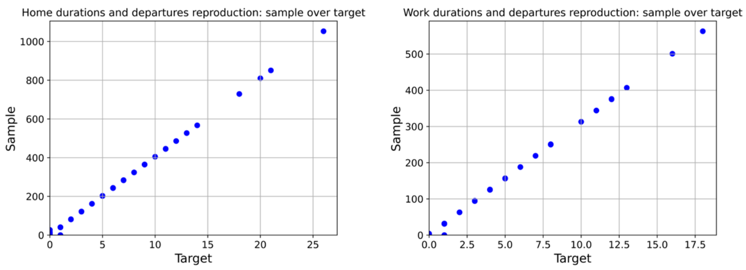

Figure 4: Quality of the home duration-departure matrix reproduction in the *OD, location and time structure reproduction model* (left); Quality of the work duration-departure matrix reproduction in the *OD, location and time structure reproduction model* (right)

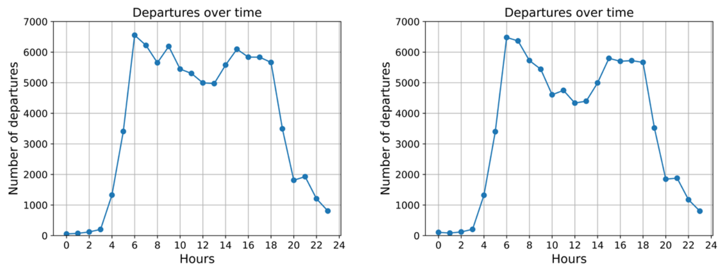

Figure 5: Departures per hour in the *OD, location and time structure reproduction model* with only the duration-departure matrix for home activities being reproduced (left); Departures per hour in the *OD, location and time structure reproduction model* with the duration-departure matrix for home and work activities being reproduced (right)

Figure 6 shows how the logarithmic sampling weights over iterations evolve with increasing complexity of the model. In the *OD reproduction model* the curve indicates stationarity of the algorithm after about 1 million iterations (Figure 6 left). In the *OD and location reproduction* *model* this takes about 3.5 million iterations (Figure 6 centre). The most complex model here does not yet indicate stationarity (Figure 6 right). Previous experiments have shown that doubling the number of iterations is enough to observe convergence.

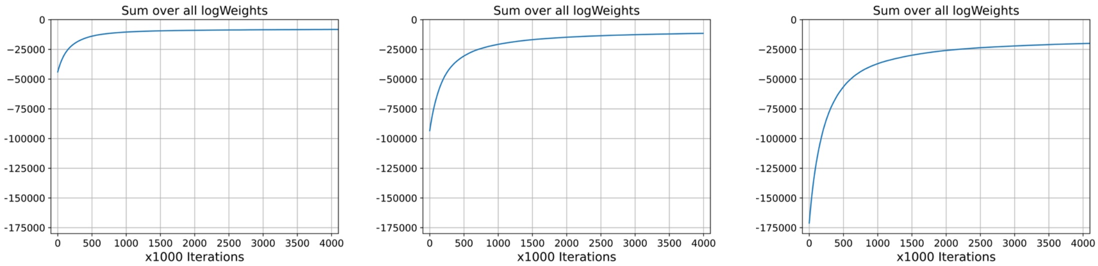

Figure 6: Logarithmic plot of the sampling weight for OD reproduction over iterations in the *OD reproduction model* (left); Logarithmic plot of the sum over all sampling weight over iterations in the *OD and location reproduction model* (centre); Logarithmic plot of the sum over all sampling weight over iterations in the *OD, location and time structure reproduction model* (right)

4. CONCLUSION

The proposed method is well suited for the unbiased disaggregation of aggregate travel demand data for agent-based simulation. The presented results correspond plausibly to the underlying modelling assumptions. The data reproduction quality is satisfactory, even when different datasets (targets) are reproduced at the same time.
The main limitation of the presented approach is its runtime requirement. The results presented here involve 4 million iterations of the algorithm. The experiment ran about 3.5 days on a 3.0 GHz AMD EPYC 7302 CPU single-threaded. Ongoing work hence focuses on speeding up the algorithm. The sequential nature of the Metropolis-Hastings algorithm hinders the straightforward deployment of multithreading techniques however great run time improvements are possible. Calderhead (2014) alleviate the computational bottleneck of MCMC methods by generating multiple proposals in parallel and then sample from a finite-state Markov chain constructed upon them to ensure matching with the target distribution. By incorporating all proposals (not just acceptable ones) in the estimation process during the sampling the statistical efficiency increases (faster movement within the state space; Calderhead, 2014). Since the number of possible rejections grows with the number of proposal processes run in parallel, it is especially useful to combine a paralleled proposal generation with an improved estimation process (Austad, 2007). The implementation of these techniques achieves run time reductions up to two orders of magnitude (Calderhead, 2014), which seems very promising for the further development of Metropolis-Hasting based synthesis of round trips as an alternative to less transparent pipeline approaches.

REFERENCES

Austad, H. M. (2007). *Parallel Multiple Proposal MCMC Algorithms*.

Bastarianto, F. F., Hancock, T. O., Choudhury, C. F., & Manley, E. (2023). Agent-based models in urban transportation: Review, challenges, and opportunities. *European Transport Research Review*, *15*(1), 19. https://doi.org/10.1186/s12544-023-00590-5

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., & Timmermans, H. (2010). Implementation Framework and Development Trajectory of FEATHERS Activity-Based Simulation Platform. *Transportation Research Record: Journal of the Transportation Research Board*, *2175*(1), 111–119. https://doi.org/10.3141/2175-13

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

Calderhead, B. (2014). A general construction for parallelizing Metropolis−Hastings algorithms. *Proceedings of the National Academy of Sciences*, *111*(49), 17408–17413. https://doi.org/10.1073/pnas.1408184111

Choupani, A.-A., & Mamdoohi, A. R. (2016). Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia*, *17*, 223–233. https://doi.org/10.1016/j.trpro.2016.11.078

Department for Transport. (2024). *Agent-based methods and activity-based demand modelling*.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. https://doi.org/10.1016/j.trb.2013.09.012

Flötteröd, G. (2025). *An operational alternative to origin/destination matrices*. In (submitted) 13th Symposium of the European Association for Research in Transportation, Munich, Germany, 2025.

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

Horni, A., Nagel, K., & Axhausen, K. W. (2016). *The Multi-Agent Transport Simulation MATSim—Book*. Ubiquity Press. https://doi.org/10.5334/baw

Mallig, N., Kagerbauer, M., & Vortisch, P. (2013). mobiTopp – A Modular Agent-based Travel Demand Modelling Framework. *Procedia Computer Science*, *19*, 854–859. https://doi.org/10.1016/j.procs.2013.06.114

Müller, K., & Axhausen, K. W. (2010). *Population synthesis for microsimulation: State of the art*.

Pougala, J. (2024). *OASIS: An integrated optimisation framework for activity scheduling*.

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

Rupprecht, F.-X. (2024). *Metropolis-Hastings based synthesis of all-day round trips from origin-destination matrices and limited population segmentation.* [Application/pdf]. 31 pages. https://doi.org/10.34726/HSS.2024.121270

Tozluoglu, Ç. (2024). *Agent-based Models for Evaluating Sustainable Transportation Systems*.