A Kernel-Based Approach for Survey-Informed Generation of Synthetic Travel Plans

Michael Sederlin*[1,2] and Gunnar Flötteröd[1,2]

[1]Linköping University, Department of Communications and Transport Systems, Bredgatan 33, Norrköping, Sweden
[2]The Swedish National Road and Transport Research Institute, Malvinas Väg 6, Stockholm, Sweden

Short summary

Modeling dynamic urban traffic accurately requires travel demand data at high temporal and spatial resolution. Congestion results from overlapping and time-dependent trips that cannot be represented adequately by static demand formats such as temporally aggregated OD matrices. At the same time, many national forecasting tools estimate demand using aggregated data and simplified behavioral representations. We propose a kernel-based framework for incorporating individual-level travel survey information into the simulation of agent travel plans. Kernel similarity measures combine socio-demographic characteristics, trip attributes, and activity–location information, making it possible to compare heterogeneous objects such as travelers, trips, and zones across different feature spaces. Such an approach is interpretable and modular, making it attractive for travel-demand synthesis, where comparisons across diverse domains are required.

We demonstrate the approach in a small case study for Stockholm, using the national travel survey and a synthetic population applied in real-world transport policy analysis. The results indicate that key marginal characteristics of the survey data are reproduced in the simulated trips. Departure-time distributions, which are crucial for urban congestion analysis, were represented well despite not being explicitly encoded in the model. The activity–location component further combines survey-based activity participation with external land-use information to guide where activities are undertaken.

Keywords: Activity-based Modelling; Kernel methods; Travel demand synthesis; Metropolis-Hastings; Travel Survey

1 Introduction

Modeling dynamic urban traffic accurately requires travel demand data at high temporal and spatial resolution. Congestion results from overlapping and time-dependent trips that cannot be represented adequately by static demand formats such as temporally aggregated OD matrices. At the same time, national forecasting tools such as Sampers Kristoffersson et al. (2018) estimate demand using aggregated data and simplified behavioral representations. Using such forecasts for congestion analysis, capacity assessment, or the evaluation of time-dependent mobility policies therefore requires a disaggregation step. That disaggregation must preserve observed travel behaviour and physical constraints, remain consistent with the assumptions of the underlying forecast model, and accommodate changes in both transport supply and land use. A natural way to do this is to model demand at the level of individual travelers and their daily activity patterns, as in agent-based models such as MATSim Axhausen & ETH Zürich (2016), which represent individual decision-making and travel behaviour directly. In such models, travel demand is represented as a

*Corresponding author

schedule over the modeling horizon, where movement arises from the utility associated with visiting different locations.

Existing work typically either (i) reconstructs incomplete activity sequences for partially observed populations, or (ii) synthesizes synthetic populations and travel plans jointly. Cui et al. (2021) impute missing activities in long-term GPS traces using a probabilistic multi-level framework that models visit probabilities based on observed sequences, durations, and frequencies. Their approach requires individual-level trace data and does not incorporate traveler attributes, which limits its applicability to synthetic populations. Joubert & De Waal (2020) use Bayesian Networks to generate trip–activity–time sequences conditional on socio-demographic characteristics, but they do not represent the spatial dimension and therefore cannot incorporate network locations or aggregated flow information. Tozluoğlu et al. (2024) combine mobility surveys with mobile-phone location data to synthesize travel patterns for phone users using a generative model. Similarly, Zilske & Nagel (2015) construct full-day plans for mobile-phone users by enriching call records with traffic counts within a traffic-simulation calibration framework. More generally, Hörl & Balac (2021) present an open and reproducible pipeline for jointly synthesizing population and daily travel schedules, demonstrated for Paris, but based on a sequence of deterministic and randomized steps rather than on a unified likelihood-based formulation.

We propose a kernel-based framework for incorporating individual-level survey information into the simulation of agent travel plans. Kernel similarity measures combine socio-demographic characteristics, trip attributes, and activity–location information, making it possible to compare heterogeneous objects such as travelers, trips, and zones across different feature spaces. Because kernel functions remain valid under addition and multiplication, corresponding to OR and AND logic Duvenaud (2014), the framework can represent several behavioural dimensions jointly while remaining interpretable. This makes such an approach attractive for travel-demand synthesis, where comparisons across domains are required. The framework builds on the Metropolis-Hastings (MH) travel plan simulator introduced in Flötteröd (2025), with inter-zonal OD information incorporated in Rupprecht (2024).

2 Methodology

We use the MH sampler described in Flötteröd (2025) and keep its proposal mechanism unchanged. We require that all trips are roundtrips, starting and ending at the same location. A state corresponds to a set of roundtrips Y^{sim} for all n agents. At each iteration a proposal Y^{sim}_{*} is generated by modifying a small subset of roundtrips. Since the MH acceptance probability only depends on the ratio of target densities, normalizing constants cancel. We therefore specify the target density up to proportionality as

$$\pi(Y^{\text{sim}} \mid X^{\text{sim}}, X^{\text{obs}}, Y^{\text{obs}}) \sim W(Y^{\text{sim}}; X^{\text{sim}}, X^{\text{obs}}, Y^{\text{obs}}) P(Y^{\text{sim}} \mid X^{\text{sim}}), \tag{1}$$

where "$\sim$" means "equal up to normalization". Here $P(Y^{\text{sim}} \mid X^{\text{sim}})$ constitutes a behavioural prior that weights the density in the absence of survey information. The weights $W(Y^{\text{sim}}; X^{\text{sim}}, X^{\text{obs}}, Y^{\text{obs}})$ are survey-based compatibility weights used in the MH-sampling scheme. We define the weights using a 'product-of-agents' formulation

$$W(Y^{\text{sim}}; X^{\text{sim}}, X^{\text{obs}}, Y^{\text{obs}}) = \prod_{i=1}^{n} \sum_{j=1}^{m} \kappa((X^{\text{sim}}_i, Y^{\text{sim}}_i), (X^{\text{obs}}_j, Y^{\text{obs}}_j)), \tag{2}$$

where $\kappa((X^{\text{sim}}_i, Y^{\text{sim}}_i), (X^{\text{obs}}_j, Y^{\text{obs}}_j))$ is a soft-matching kernel function measuring how well agent i aligns with respondent j. For each agent i, the inner sum goes over all respondents, meaning that every agent is matched to every respondent.

The product over i factorizes the weight W over the full set of roundtrips into per-agent contributions.

Construction of the kernel function

Let X_i^{sim} and X_j^{obs} be the socio-demographic attributes of synthetic agent i and survey respondent j, respectively. Let Y_i^{sim} and Y_j^{obs} be their simulated and reported travel plans, respectively. We evaluate the similarity between agent i and survey record j using the composite kernel

$$\kappa\left((X_i^{\text{sim}}, Y_i^{\text{sim}}), (X_j^{\text{obs}}, Y_j^{\text{obs}})\right) = \kappa_{\text{trav}}(X_i^{\text{sim}}, X_j^{\text{obs}})\ \kappa_{\text{sched}}(Y_i^{\text{sim}}, Y_j^{\text{obs}})\ \kappa_{\text{loc}}(Y_i^{\text{sim}}, Y_j^{\text{obs}}), \quad (3)$$

where $\kappa_{\text{trav}}(X_i^{\text{sim}}, X_j^{\text{obs}})$ is a *traveler kernel* comparing socio-demographic attributes (X_i^{sim} vs. X_j^{obs}), $\kappa_{\text{sched}}(Y_i^{\text{sim}}, Y_j^{\text{obs}})$ is a *schedule kernel* comparing travel-plan summary characteristics (Y_i^{sim} vs. Y_j^{obs}), and $\kappa_{\text{loc}}(Y_i^{\text{sim}}, Y_j^{\text{obs}})$ is an *activity-location kernel* comparing reported activities in Y_j^{obs} with the locations visited in Y_i^{sim}.

Traveler Kernel Similarity between travelers is computed as a product of kernels measuring deviations in income and age as well as a fixed penalty for individuals of different genders as

$$\kappa_{\text{trav}}(X_i^{\text{sim}}, X_j^{\text{obs}}) = \gamma\left(\Delta\text{inc}_{ij}, \ell_{\text{inc}}\right) \gamma\left(\Delta\text{age}_{ij}, \ell_{\text{age}}\right) \eta^{\mathbb{1}\left\{\text{gender}(X_i^{\text{sim}}) \neq \text{gender}(X_j^{\text{obs}})\right\}}, \quad (4)$$

where $\mathbb{1}\{\cdot\}$ is the 0/1 indicator function. The length scales $\ell_{\text{inc}}, \ell_{\text{age}}$ and mismatch factor η each control the strictness of the matching for their respective dimensions. We use the shorthand

$$\gamma(x, \ell) = \exp\left(-\frac{x^2}{\ell^2}\right), \qquad \ell > 0, \quad (5)$$

i.e. a squared-exponential transformation of the discrepancy x with scale parameter ℓ.

Location-Activity Kernel We model the system as an activity-expanded network, and do not rely on direct correspondence between simulated and observed locations. Instead, the location-activity kernel measures whether simulated round trip i visits locations that support the activities reported by respondent j.

Let L_i denote the set of locations visited in the round trip i. Each location is associated with one activity type, a^l and the set of feasible activities is then

$$A_i^{\text{sim}} = \{a^l : l \in L_i\}. \quad (6)$$

We let A_j^{obs} denote the set of activities reported by respondent j.

A basic measure of the compatibility between the two activity sets is the Jaccard similarity

$$J(A_i^{\text{sim}}, A_j^{\text{obs}}) = \frac{|A_i^{\text{sim}} \cap A_j^{\text{obs}}|}{|A_i^{\text{sim}} \cup A_j^{\text{obs}}|}. \quad (7)$$

This term captures whether visited locations support the observed activities, but treats all activity-location pairs equally, failing to distinguish between locations offering more opportunities (or a higher quality) for a given activity.

To account for this we define the opportunity-weighted score

$$S(L_i, A_j^{\text{obs}}) = \sum_{l \in L_i} \alpha_l \mathbb{1}\left\{a^l \in A_j^{\text{obs}}\right\} \quad (8)$$

where α_l denotes the attractiveness for a given location in its associated activity a^l. This attractiveness score can be constructed from, for example, land-use information or observed location-specific visit patterns.

The location-activity kernel combines these two terms as

$$\kappa_{\text{loc}}(Y_i^{\text{sim}}, Y_j^{\text{obs}}) = \exp\left(-\frac{1 - J(A_i^{\text{sim}}, A_j^{\text{obs}})}{\ell^J}\right) \exp\left(\frac{S(L_i, A_j^{\text{obs}})}{\ell^S}\right), \quad (9)$$

where $\ell^J > 0$ and $\ell^s > 0$ control the sharpness of the matching for the two similarity scores. Similar extensions can incorporate factors such as the time spending, ordering of activities, or other structural characteristics of activity–location pairs, as long as they can be expressed through pair-wise scores that are cheap to compute.

Schedule Kernel The schedule kernel represents the similarity between a given respondent's schedule and a simulated trip chain. This is done in an aggregated manner taking into account the total travel time *ttt* and the number of locations *nStops* visited in a day, and the time windows when certain activities are performed, as follows:

$$\kappa_{\text{sched}}(Y_i^{\text{sim}}, Y_j^{\text{obs}}) = \gamma\left(\Delta ttt_{ij}, \ell_{ttt}\right)\gamma\left(\Delta n\text{Stops}_{ij}, \ell_{n\text{Stops}}\right)\prod_{a\in\mathcal{A}}\gamma\left(\left|\vartheta_{ia}^{\text{sim}}\cup\vartheta_{ja}^{\text{obs}}\right| - \left|\vartheta_{ia}^{\text{sim}}\cap\vartheta_{ja}^{\text{obs}}\right|, \ell_{\vartheta}\right) \tag{10}$$

where Δttt_{ij} and $\Delta n\text{Stops}_{ij}$ are the differences between total travel times and number of stops between agent i and respondent j. The final component combines the time during which a reported activity a is performed and when simulated trip visits locations where activity a is possible, where $\vartheta_{ia}^{\text{sim}} \subseteq [0, 24\,\text{h})$ and $\vartheta_{ja}^{\text{obs}}$ represents all times during which simulated round trip i and survey respondent j performs activity $a \in \mathcal{A}$ respectively. Furthermore, $|\ldots\cup\ldots|$ and $|\ldots\cap\ldots|$ represent the time length of the union and overlap of two activity time allocations respectively.

3 Results

Experiment Design

The empirical evaluation presented here presents a small-scale proof of concept for the Stockholm area, with large-scale demonstrations forthcoming. The experiments combine information from the Swedish national travel survey (Resvaneundersökningen (RVU)) and a given synthetic population derived from the Swedish national forecast model Sampers. Only car-based roundtrips with trip legs remaining within the study area are considered from the survey. Furthermore, detailed trip purpose information is aggregated into the five categories HOME, WORK, LEISURE, EDUCATION, SERVICES to ensure coverage and a manageable network size.

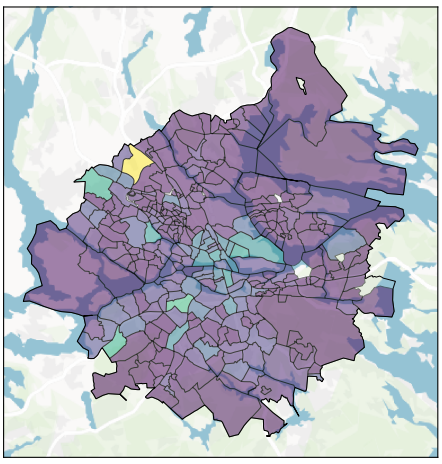
Figure 1a shows the survey zones of the considered case study area. An example of a round-trip from the survey data is shown in Figure 1b. The simulation is performed on an activity-expanded network in which each node represents a location–activity pair. This representation makes it possible to evaluate not only whether a simulated roundtrip has a plausible overall schedule, but also whether the visited locations support the activities reported in the survey.

The proposal mechanism is unchanged from the underlying simulator, and the prior includes strict enforcement of home-location consistency, ensuring that the synthetic agents retain their already assigned home locations throughout the simulation, and a schedule period of 24 hours. The MH sampler is initialized from home-based roundtrips and run for 5 000 000 iterations to ensure sampling from a stationary distribution and to allow sufficient between-sample distance to mitigate the autocorrelation present in Monte-Carlo methods. All parameters are given in Table 1.

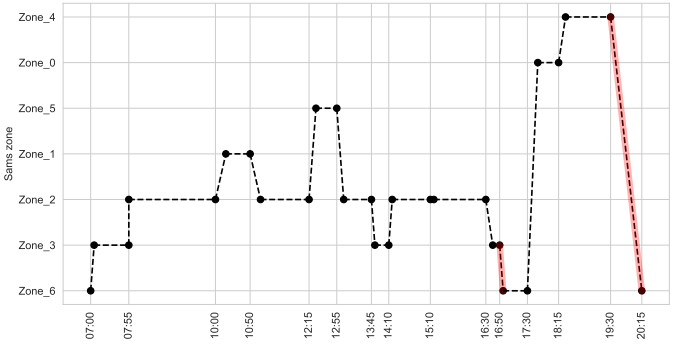
The empirical evaluation focuses on the questions.

1. To what extent do the synthesized plans reproduce the main travel-pattern characteristics observed in the survey?
2. Are the resulting activity-location patterns spatially plausible given the underlying distribution of activity opportunities?

To answer these questions, we compare synthesized and observed travel patterns using aggregate indicators that are directly related to the kernel design. The main outputs are the distributions of total daily travel time and number of stops, the temporal structure of activities over the day, and the degree to which simulated visited locations support the activity types observed in the survey. As well as the inclusion of available land-use data in the spatial distribution of these activities. In the present study, the opportunity scores α_l were constructed from building floor area, where the total area of buildings with classifications associated with a given activity type was used as the corresponding location attractiveness measure.



(a) Case study area in Stockholm showing survey zones and derived WORK attractiveness



(b) Diagram of RVU trip structure

Table 1: Simulation and kernel parameter values used in the experiments

Settings			
Component	Description	Value	
Simulation	Horizon	24 h	
Simulation	Survey respondents	200	
Simulation	Synthetic agents	≈ 400	
Simulation	MH iterations	5,000,000	
Parameters			
Component	Description	Parameter	Value
Traveler kernel	Age scale	ℓ_{age}	5 years
Traveler kernel	Income scale	ℓ_{inc}	20,000 SEK
Traveler kernel	Gender mismatch factor	η	0.2
Schedule kernel	Number-of-stops scale	$\ell_{n\text{Stops}}$	0.3 stops
Schedule kernel	Total-travel-time scale	ℓ_{ttt}	20 mins
Schedule kernel	Schedule-overlap scale	ℓ_{ϑ}	20 mins
Location kernel	Jaccard scale	ℓ^J	0.1
Location kernel	Opportunity-weighted scale	ℓ^S	0.2

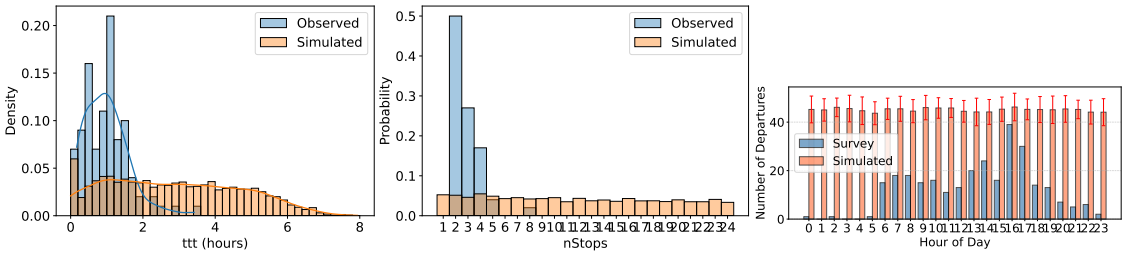
Results

Figure 2 shows the uninformed prior baseline when mobility behaviour is sampled without considering the survey information through the kernel weights. All behaviour in these plots come from the proposal mechanism and weights ensuring uniformity in roundtrip lengths and a 24 hour periodicity. Figure 2c shows no structure in the departure-time distribution. In the kernel-weighted results shown in Figures 3, marginal distributions of the travel time and number of stops follow the survey distribution.

illustrates how the marginal distributions of total travel time and number of stops are reproduced even though all weights are on an individual respondent-agent-level. In particular, Figures 2c and 3c show the distribution of departure times over the day in the survey sample and in the simulated trips. Figure 4 provides a more direct view of the observed and simulated number of hourly departures.

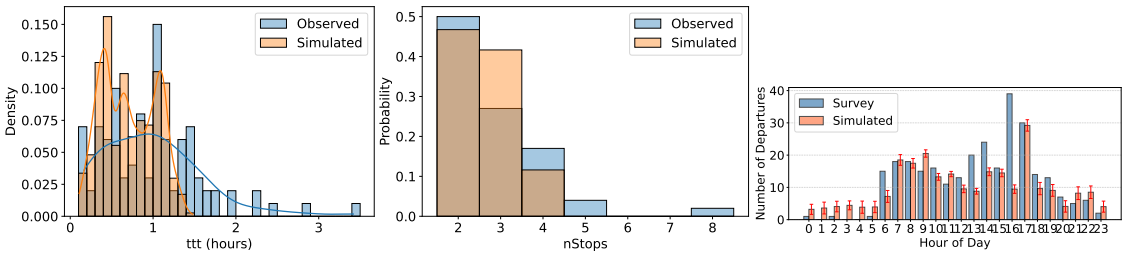
Given the small sample size and large number of zones, noise and a large share of unvisited zones make direct comparison between the zone attractiveness and visit counts difficult to interpret. Instead, the effect of the land-use component in kernel (9) is assessed by comparing the correlation

between opportunity scores in the unweighted and kernel-weighted scenarios. There is a stronger correlation between the opportunities in a zone and simulated visits when the location–activity kernel is used. In the prior configuration, the strongest correlation is seen in the HOME activity, likely as a result of requiring all travel plans to start and end in a HOME node. By randomness, the used survey sample did not contain EDUCATION activities, and as a consequence, no trips to EDUCATION nodes are generated in the simulated trip plans. The final column confirms that the positive correlation between zone attractiveness is not entirely generated by a single very strong zone attracting all trips.



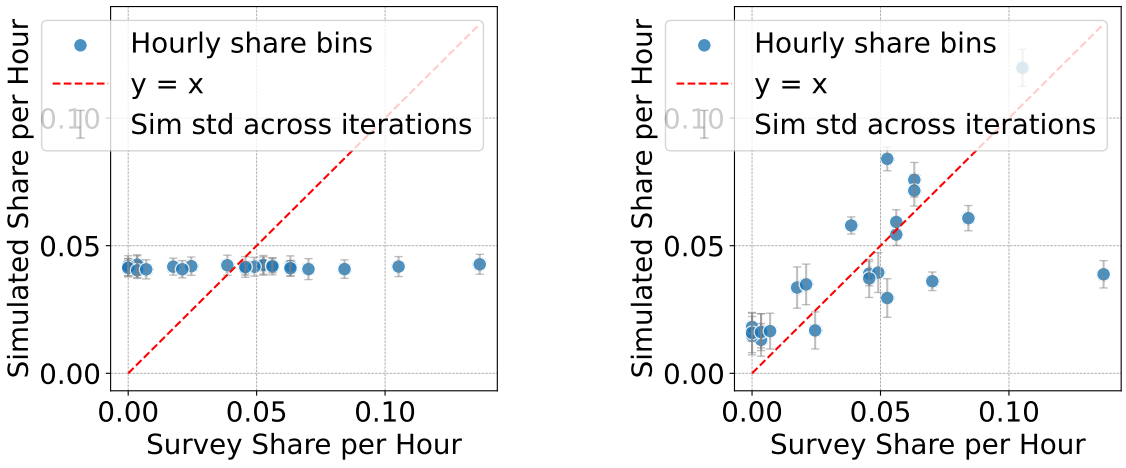
(a) Total travel time [hours] (b) Number of visited locations (c) Departure time distribution

Figure 2: Simulated behaviour without kernel smoothing weights



(a) Total travel time [hours] (b) Number of visited locations (c) Departure time distribution

Figure 3: Simulated behaviour with kernel smoothing weights



(a) Unweighted specification (b) Kernel-weighted specification

Figure 4: Correspondence between survey and simulated hourly departure shares

Table 2: Spearman rank correlation between activity-specific opportunity score and simulated mean visits, comparing the prior and the kernel-weighted specification.

Activity	Prior ρ	Kernel ρ	Kernel ρ dropping top zone
EDUCATION	-0.048	–	–
HOME	0.278	0.446	0.439
LEISURE	0.027	0.208	0.170
SERVICES	-0.040	0.369	0.356
WORK	0.077	0.336	0.323
Mean	0.059	0.340	0.322

4 Conclusions

This paper presented a kernel-based approach for survey-informed synthetic travel pattern generation within an existing Metropolis–Hastings framework for simulating travel plans. The main contribution is a composite similarity formulation that combines traveler attributes, schedule characteristics, and activity–location information in a unified framework. In a small-scale study of Stockholm, the kernel-weighted specification reproduced key marginal travel characteristics better than the uninformed prior and induced a clearer positive association between activity-specific opportunity scores and simulated visits. In particular, key marginal properties of the survey were reproduced through individual-level associations. Departure-time distributions, crucial for urban congestion analysis, were represented well despite not being explicitly encoded in the model. In addition, the location–activity kernel allows the model to combine survey-based activity participation with external land-use data to guide where activities are undertaken. A broader analysis will consider the spatial behavior of the model in greater detail, its performance in a full regional application, and the effects of structured parameter calibration.

Acknowledgements

Funding: This research was funded by the Swedish Transport Administration (Trafikverket), grant: TRV 2023/33572.

References

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

Cui, Y., He, Q., & Bian, L. (2021, November). Generating a synthetic probabilistic daily activity-location schedule using large-scale, long-term and low-frequency smartphone GPS data with limited activity information. *Transportation Research Part C: Emerging Technologies*, *132*, 103408. doi: 10.1016/j.trc.2021.103408

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes* (Unpublished doctoral dissertation). University of Cambridge.

Flötteröd, G. (2025). An operational alternative to origin/destination matrices. In *hEART 2025*. Aalto University, Helsinki, Finland.

Hörl, S., & Balac, M. (2021, September). 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. doi: 10.1016/j.trc.2021.103291

Joubert, J. W., & De Waal, A. (2020, November). Activity-based travel demand generation using Bayesian networks. *Transportation Research Part C: Emerging Technologies*, *120*, 102804. doi: 10.1016/j.trc.2020.102804

Kristoffersson, I., Berglund, S., Samuelsson, S., Almström, P., & Algers, S. (2018). *Sampers4 - Skattning av regionala efterfrågemodeller* (Tech. Rep.).

Rupprecht, F. (2024). *Metropolis-Hastings based synthesis of all-day round trips from origin-destination matrices and limited population segmentation* (Unpublished master's thesis). Technische Universität Wien.

Tozluoğlu, Ç., Liao, Y., & Sprei, F. (2024, October). *Mobile Phone Application Data for Activity Plan Generation* (No. arXiv:2410.22386). arXiv.

Zilske, M., & Nagel, K. (2015). A Simulation-based Approach for Constructing All-day Travel Chains from Mobile Phone Data. *Procedia Computer Science*, *52*, 468–475. doi: 10.1016/j.procs.2015.05.017