ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2024) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

26th EURO Working Group on Transportation Meeting, EWGT 2024, 4th-6th September 2024, Lund, Sweden

Advancing the MATSim Open Berlin Scenario: Improvements in transport scenario generation and calibration methods

Christian Rakow[a,*], Moritz Kreuschner[a], Kai Nagel[a]

[a] *Technische Universität Berlin, Chair of Transport Systems Planning and Transport Telematics, Straße des 17. Juni 135, 10623 Berlin, Germany*

Abstract

This study enhances the MATSim Open Berlin Scenario, a multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data. Shortcomings of the existing model, such as outdated data sources or limited granularity of activity patterns, are addressed. Another major improvement is the implementation of an automated calibration process, which not only reduces the time required for manual calibration but also enhances the overall quality of the model. Additionally, recent advancements, including the integration of commercial traffic and web-based visualization tools, have been incorporated into the model. The scenario is evaluated using a range of metrics and compared against survey data, travel times from online routing services, and data from stationary counting stations. The evaluation demonstrates that the model effectively captures personal traffic, surpassing the previous MATSim Open Berlin Scenario in many aspects. However, achieving the same level of representation for commercial traffic and transit traffic from outside the study area is challenging due to the lack of detailed data.

Keywords: agent-based simulation; transport modelling; MATSim

---

1. Introduction

The Multi-Agent Transport Simulation (MATSim) offers an agent-based platform for modeling and analyzing complex urban mobility systems. With MATSim, it is possible to model and evaluate the impacts of changes in infrastructure, pricing policies, and other interventions on travel behavior and the overall transportation system. This information can then be used to make informed decisions about future investments and policies. Moreover, the open-source nature of MATSim allows for a high degree of customization and flexibility, making it suitable for a wide range of applications and research questions [11].

---

* Presenting author
E-mail address: rakow@vsp.tu-berlin.de

MATSim originated from collaborative efforts between TU Berlin and ETH Zürich. Consequently, numerous agent-based transport models of Berlin have been established over the years. In 2006, an initial Berlin model was initiated, utilizing a road network originally devised by the planning department of the city of Berlin. This early model only represented car drivers, with a sample size of 10%, resulting in about 205,000 simulated agents [4]. In 2010, a new model was created using survey data to accurately replicate travel behaviors in the Berlin area, incorporating the entire public transit network [23]. However, that model was built on proprietary data and thus not publicly accessible. Subsequently, the MATSim Open Berlin Scenario was introduced in 2017 [29] and later modified and improved [28]. That model was openly available, leading to its widespread use in research and academic studies [14]. As to this date, it has been referenced in more than 100 published papers. Nevertheless, the existing Open Berlin scenario has its limitations, some of which are addressed in this paper.

While conventional trip-based four-step travel demand models are still widely used in practice, activity-based models have gained popularity in recent years. Activity-based models are based on the concept that travel is a derived demand, driven by the need to participate in activities at different locations. In the literature, there are various methodologies and approaches for generating activity-based demand. One of the early advocates of activity-based demand modeling is *TASHA* [21], which has been applied in different regions, including the Greater Toronto Area. Other notable models include *FEATHERS* [13] and *mobiTopp* [18]. The scenario generation pipeline *eqasim* [10], is specifically designed for use with MATSim and has been successfully applied in regions in France. It can also be extended for use in other countries [17]. Many recent developments and improved methodologies in activity-based demand modeling have not yet been reflected in the OpenBerlin scenario. The contributions of this paper are twofold: First, we update the data sources and methodologies used in the OpenBerlin scenario to incorporate the latest developments in the field. Second, we address one of the major challenges in transport scenario creation, which is the calibration of the model. In this paper, we present an automated calibration approach that reduces the manual effort required for calibration.

2. Scenario generation and calibration

We have made several improvements to the scenario generation process in order to enhance the quality of the resulting model. The updated scenario generation process is illustrated in Figure 1, and it is explained in detail in the subsequent sections.

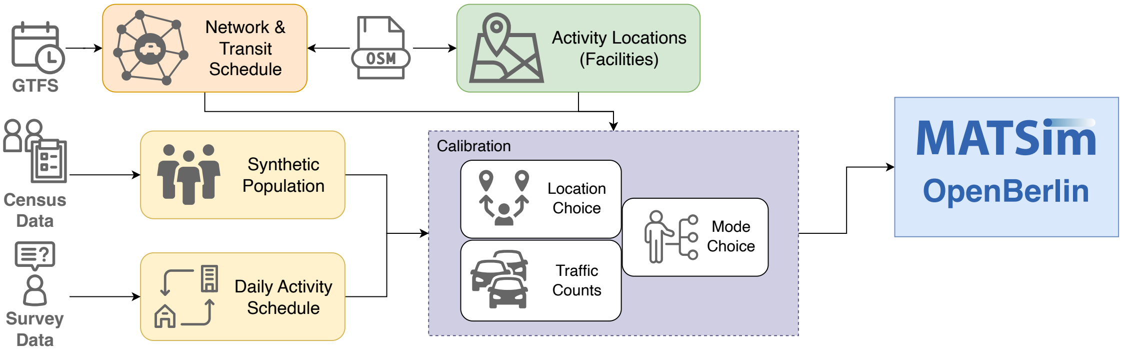

Fig. 1: Overview of the updated scenario generation process.

2.1. *Road Network & Transit Schedule*

Information about the road network is obtained from OpenStreetMap (OSM), which has proven to be a reliable and freely available source for road infrastructure. The updated network generation process involves first creating a microscopic representation of the road network for *SUMO* [15]. Flow capacities on intersections are determined using microscopic simulation, as demonstrated in [24]. Additionally, this coupling greatly facilitates further studies that require the use of microscopic traffic simulation, such as studies related to connected and automated driving [24, 27].

In a mesocopic traffic simulation, traffic lights are typically not modeled explicitly. Instead, the free speed of each link must account for the delays caused by traffic lights and other traffic signs even in uncongested conditions. The free speed is determined by combining microscopic simulation and routing data obtained from online APIs [26].

Information about public transit network is imported from the General Transit Feed Specification (GTFS) feed provided by [1]. The day “2023-06-06” is used as a reference day for the transit schedule.

2.2. Synthetic Population

Population data is obtained from two different data sources for Berlin and Brandenburg. The *Kommunalatlas* for 2020 [2] provides detailed population data for Berlin with a resolution of 531 zones. For Brandenburg, German wide census data [3], with a more coarse resolution is used. This is a huge improvement over the old model version in terms of geospatial resolution in Berlin. These datasets provide the number persons per age group, gender and employment status living in each zone. Each person is assigned a random home location within their respective zone and a building classified for residential use (cf. subsection 2.3).

The next step involves assigning daily plains i.e. daily activity patterns to all agents. The original version of the model used the activity scheduling model Comprehensive Econometric Micro-simulator for Daily Activity-travel Patterns (CEMDAP) to generate daily plans for each agent [30]. However, two major drawbacks became apparent over the years: (i) CEMDAP is no longer actively maintained and is therefore based on outdated survey data and (ii) there is no model specifically for Berlin or Germany. Instead, it uses the same behavioral parameters as a metropolitan model for Los Angeles. Also, the CEMDAP model contained only five distinct types of activities.

To address these issues, daily activity patterns are now derived from data collected through a travel diary survey [System repräsentativer Verkehrsbefragungen (SrV); 12]. In order to combine the survey data with the synthetic population (census data), we employ statistical matching [7]. This procedure selects a synthetic person and randomly matches them with a person from the survey data based on similar attributes. The selected person is assigned all attributes that are not present in the census data, as well as the daily plan of the survey person. Persons are matched by age groups (0-6, 6-10, 10-18, 65+), gender and employment status. Persons between 18 and 65 are considered similar if the age difference is not more than 6 years. The daily plan consists of a sequence of activities, each with a start time, end time, and activity type. Additionally, the travelled distance to the activity location is preserved and used in the location choice process.

At this stage, the daily plan does not contain any information about the location of the activities, except for the home location. This information is added in the next step by assigning specific activity locations to each activity in the plan.

2.3. Location choice

The activity matching process in the updated scenario generation process involves extracting detailed activity locations from OSM. The previous method relied on Coordination of Information on the Environment (CORINE) land cover data [5], which lacks precise building coordinates and facility details.

The new approach considers relevant OSM entities such as buildings and Point of Interests (POIs) and assigns activity types based on specific tags associated with these entities. These extracted activity locations are then used in the location choice process to determine suitable activity locations that align with the intended purpose of a trip. Figure 2 illustrates the extracted facilities for shopping, dining, and higher education.

The available activity types align with those in the survey data. The distribution of trip purposes throughout the day, as depicted in Figure 3, follows expected patterns. Specifically, the morning hours are characterized by trips for work or education, while the afternoon and evening hours are primarily associated with home or leisure trips.

Exact activity locations are not known from the survey data. Often, many locations fulfill the criteria of having the correct type and being at the right distance as reported in the survey. To calibrate the choice of locations, the approach from [28] is kept: each agent is given five plausible plans, where different locations have been selected. Plan selection is then calibrated against counting data, such that the error is minimized [9, 22].

Fig. 2: Activity locations extracted from OSM data. Left: Shopping stores. Middle: Dining facilities. Right: Higher education facilities.

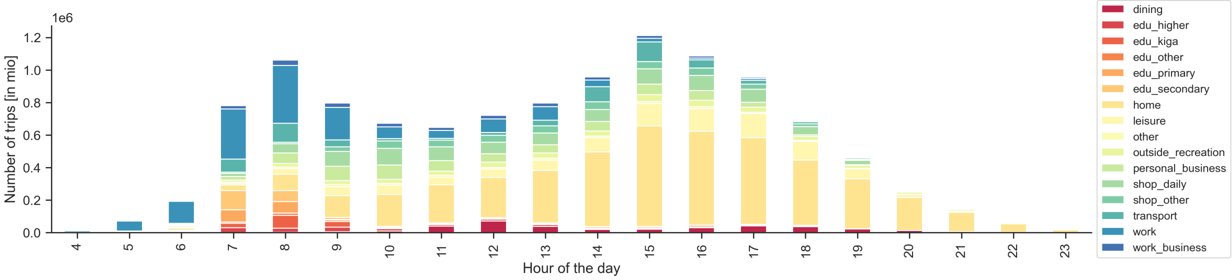

Fig. 3: Trip purposes by hour throughout one day. Hours 0-3 with less than 1000 reported trips have been excluded.

2.4. Mode Calibration

A major advantage of the updated scenario generation process is the automated calibration of the mode-specific Alternative Specific Constants (ASCs). This is also a major improvement over existing scenario pipelines, such as *eqasim* [10], where manual calibration is still necessary to fine-tune the obtained mode shares.

MATSim utilizes a co-evolutionary algorithm, where agents maximize their utility over multiple iterations. In each iteration, a plan is executed and scored. Some agents select a known plan from their memory based on its utility, while others try out new plans by changing the mode of transport. The utility of a plan is computed by summing up all activity utilities plus all travel (dis)utilities [11, p. 26]. The travel disutility of using a certain mode is given by its attributes, such as travel time, costs, distance, etc. plus an ASC. The ASC is a constant that is added to the utility of a mode to account for additional factors that are not explicitly modeled. Since utilities are relative to each other, one ASC needs to be fixed (usually at zero), otherwise there would not be a unique solution.

We calibrate the mode-specific ASCs using the following approach: After the simulation finished, the mode shares of all relevant trips are collected. Let z_i be the observed mode share for mode i and m_i the simulated mode share. The ASCs are then adjusted by Δ_i using the following formula:

$$\widehat{asc_i} = asc_i + \underbrace{\ln z_i - \ln m_i - (\ln z_0 - \ln m_0)}_{\Delta_i} \tag{1}$$

where z_0 and m_0 refer to the mode with fixed ASC.

The calibration process is iterated until the mode shares converge with the observed values, i.e., the absolute difference in percentage points falls below a certain threshold. For the derivation and mathematical background of the formula refer to the supplementary material [25].

Table 1: Mode parameter and comparison of simulated and reported mode statistics. Trip speed and distance refers to the average across all trips of a certain mode. Reported data is based on [SrV; 12].

	Config Parameter			Share [%]		Speed [km/h]	
Mode of transport	ASC	[€/km]	[€/day]	sim.	ref.	sim.	ref.
Car	−0.53	0.149	5	19.93	20.07	23.52	22.25
Ride	−1.24	0.149	-	5.94	5.96	25.62	22.78
Public transport	0.40	-	3	26.97	26.51	14.78	14.10
Bicycle	−1.35	-	-	17.81	17.79	11.31	10.30
Walk	0	-	-	29.35	29.68	4.44	4.43

2.5. Further improvements

The cost parameters have been updated to 2022. The cost of public transport is now set at 3 EUR per day, which corresponds to the average cost of a yearly ticket (761 EUR) divided by 250 working days. The updated parameters and resulting (= calibrated) ASCs are shown in Table 1. The ride mode (i.e. passenger on a car that is not the driver) incurs the same distance costs as driving a car, plus an additional time costs, to account for the opportunity costs of the driver. For a detailed explanation of how these costs are calculated, refer to [19].

Trips induced by commercial traffic are not represented in the survey data, however they constitute for a significant portion of car traffic. To model this part more accurately recent approaches for long-haul freight [16] and small-scale commercial traffic have been incorporated [8]. The generation of commercial traffic is based on structural data, such as population and employment figures, to create realistic routes. The model is divided into two main segments: commercial person traffic and freight traffic. Each segment serves a different transportation purpose and utilizes appropriate vehicles for their respective operations.

By utilizing *SimWrapper* [6], a web-based application, the model output can be visualized, inspected, and shared more effectively. SimWrapper has been seamlessly integrated into the model workflow, generating user-friendly dashboards with each model run.

3. Results and discussion

The automated calibration procedure described in subsection 2.4 performs a whole simulation run and iteratively adjusts the ASCs to match the simulated and reported mode shares. The calibration process converges to a stable solution after a few runs as demonstrated in Figure 4.

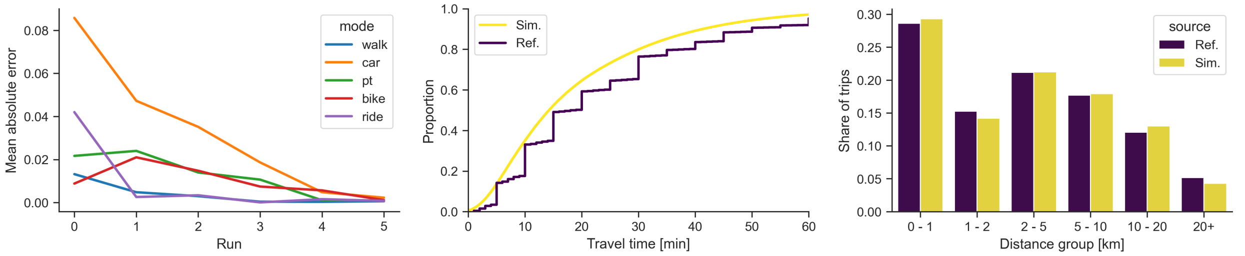

Fig. 4: Mean absolute error between simulated and reported mode share (left); Cumulative distribution of trip durations (middle); Trip distance distribution (right). All figures refer to persons with a home location in Berlin.

The scenario has also been evaluated in terms of various other metrics, based on the reference data [12], which are shown in Table 1.

The modal split is closely aligned with the survey data. The mean absolute error for all modes is below 0.5 percentage points, indicating a substantial improvement compared to the previous model. In the previous model, manual calibration was necessary, and the mean absolute error was reported to be as high as 1.4 percentage points (cf. [28, p. 875; Table 3]). Also, the trip travel time and distance distributions (shown in Figure 4) are in line with the reference data. Especially the trip distances are now much closer to the reference data compared to the previous model based on CEMDAP. A discrepancy in simulated speeds is observed for the car and ride modes. For the survey the whole trip duration needed to be reported, which includes access and egress to the car and in particular parking time. As parking search is not included in the model, the car speeds are generally too high, however the negative ASCs of these modes will account for these effects.

While the above comparison is based on survey data, the car travel times have also been evaluated against online routing services. For this purpose, 1000 routes have been randomly sampled and the expected travel time and distance queried from the routing service of Google Maps, HERE, TomTom and Mapbox. This has been done for different times of the day. Aggregating the results of these services gives a minimum and maximum expected travel speed, which can be compared against the simulated car speeds. The results are shown in Figure 5 (left).

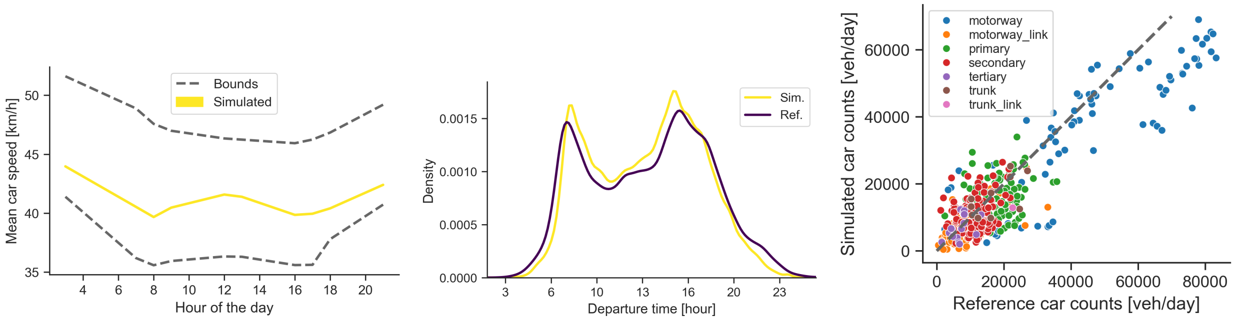

Fig. 5: Mean car speed and reference speeds obtained from online APIs (left); Distribution of departure times throughout the day (middle); Measured vs. simulated daily traffic volumes, categorized by OSM road type (right).

The plot illustrates that the car travel speed is well within the plausible range. Additionally, the model accurately captures the drop in speed during peak hours. However, it is worth noting that the speed increases again in the middle of the day, which is not observed in the reference data. The distribution of departure times throughout the day is depicted in the middle plot of Figure 5. Clear morning and evening peaks can be observed in both the model and the reference data.

To compare traffic counts, we utilized data from the local traffic management authority. The traffic counts for an entire year 2018 were aggregated into daily traffic volumes, considering only weekdays. Analyzing the traffic counts shown in the right plot of Figure 5, it is evident that the daily simulated traffic volumes are generally too low, particularly on motorways. There could be several reasons for this discrepancy. Firstly, the survey data focuses only on individual traffic. Although commercial traffic is included, it is based on much older and less detailed data. For instance, very little is known about the types of vehicle in use, their load, empty runs, etc. Without this information it is difficult to estimate the total vehicle kilometers driven. Other types of traffic that are missing include transit coming from outside of Berlin/Brandenburg and any tourist activities.

To summarize, the evaluation demonstrates that the model effectively captures individual traffic, which accounts for the majority of the traffic in Berlin. Several key metrics are represented more accurately than in the previous model version. The availability of detailed survey data allows for a comprehensive understanding and modeling of individual traffic. However, the model currently lacks the same level of representation for commercial traffic, which is the same limitation as in the previous version of the Open Berlin Scenario.

4. Conclusion and Outlook

Although there have been significant advancements in agent-based transport models, there are still areas where further improvements can be made. Most notably, even the updated model relies on data that precedes the COVID-19 pandemic. The pandemic had a significant impact on mobility behavior and the model should be updated to reflect these changes as soon as more recent survey data becomes available. Due to more streamlined scenario generation and automatic calibration, this should be easier to achieve than in the past. Other areas for improvement, as highlighted by [20], involve weekly activity modeling and capturing intra-household interactions. The possibility to model these aspects is currently limited by the available survey data. Similarly, the representation of commercial traffic is hindered by the lack of detailed data.

Lastly, calibration and validation of location choice remains a significant challenge, as essential data on facility visitation is often unavailable and proprietary. Further improvements may require exploring alternative data sources, even if not openly accessible.

Acknowledgements

Code and data to run the scenario and based on which own functionality can be added, is available online under `https://github.com/matsim-vsp/matsim-berlin`. The version documented and analyzed in this paper is the 10% scenario with release number 6.3 (berlin-v6.3-10pct). All required input files and the ready scenario are accessible via `https://svn.vsp.tu-berlin.de/repos/public-svn/matsim/scenarios/countries/de/berlin/berlin-v6.3/input/`. Visualizations with Simwrapper are available via `https://vsp.berlin/simwrapper/public/de/berlin/berlin-v6.3/output/berlin-v6.3-10pct`. The authors wish to thank the Team of VSP who keeps contributing to the MATSim Open Berlin Scenario and providing valuable feedback throughout the development process.

During the preparation of this work the authors used ChatGPT in order to improve the readability and wording of the paper. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

This work was funded by the German Research Foundation (DFG) (project numbers: 323900421, 398051144) and by the Federal Ministry of Education and Research (BMBF) (project number: 03SF0674A).

References

[1] GTFS.DE - GTFS für Deutschland, . URL: `https://gtfs.de/`.

[2] Kommunalatlas Berlin 2020, . URL: `https://instantatlas.statistik-berlin-brandenburg.de/instantatlas/interaktivekarten/kommunalatlas/atlas.html`.

[3] Regionaldatenbank Deutschland, . URL: `https://www.regionalstatistik.de/genesis/online`.

[4] Beuck, U., Nagel, K., Rieser, M., Strippgen, D., Balmer, M., 2007. Preliminary results of a multi-agent traffic simulation for Berlin. Advances in Complex Systems (ACS) 10, 289–307. doi:`10.1142/S0219525907001367`.

[5] Büttner, G., Kosztra, B., Maucha, G., Pataki, R., Kleeschulte, S., Hazeu, GW., Vittek, M., Schroder, C., Littkopf, A., 2021. Copernicus land monitoring Service—CORINE land cover. User Manual; Copernicus Publications: Göttingen, Germany .

[6] Charlton, W., Sana, B., 2023. SimWrapper, an open source web-based platform for interactive visualization of microsimulation outputs and transport data. Procedia Computer Science 220, 724–729. doi:`10.1016/j.procs.2023.03.095`.

[7] D'Orazio, M., Zio, M.D., Scanu, M., 2006. Statistical Matching: Theory and Practice. John Wiley & Sons.

[8] Ewert, R., Nagel, K., 2024. Agentenbasierte modellierung des kleinräumigen wirtschaftsverkehrs, in: Proceedings of Heureka 2024. Forschungsgesellschaft für Straßen-und Verkehrswesen. URL: `https://verlag.fgsv-datenbanken.de/tagungsbaende?kat=HEUREKA&p=3&tagungsband=2530&_titel=Agentenbasierte+Modellierung+des+kleinrÃďumigen+Wirtschaftsverkehrs`.

[9] Flötteröd, G., 2010. Cadyts–Calibration of dynamic traffic simulations–Version 1.1. 0 manual. Transport and Mobility Laboratory, École Polytechnique Fédérale de Lausanne .

[10] Hörl, S., Balac, M., 2021. Introducing the eqasim pipeline: From raw data to agent-based transport simulation. Procedia Computer Science 184, 712–719. doi:`10.1016/j.procs.2021.03.089`.

[11] Horni, A., Nagel, K., Axhausen, K.W., 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press. doi:`10.5334/baw`.

[12] Hubrich, S., Ließke, F., Wittwer, R., Wittig, S., Gerike, R., 2019. Methodenbericht zum Forschungsprojekt. „Mobilität in Städten – SrV 2018". URL: `https://orlis.difu.de/handle/difu/259350`.

[13] Janssens, D., Wets, G., Timmermans, H., Arentze, T.A., 2007. Modelling short-term dynamics in activity-travel patterns: The feathers model URL: `https://documentserver.uhasselt.be//handle/1942/10567`.

[14] Kaddoura, I., Bischoff, J., Nagel, K., 2020. Towards welfare optimal operation of innovative mobility concepts: External cost pricing in a world of shared autonomous vehicles. Transportation Research Part A: Policy and Practice 136, 48–63. doi:`10.1016/j.tra.2020.03.032`.

[15] Lopez, P.A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P., Wiessner, E., 2018. Microscopic Traffic Simulation using SUMO, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), pp. 2575–2582. doi:`10.1109/ITSC.2018.8569938`.

[16] Lu, C., Martins-Turner, K., Nagel, K., 2022. Creating an agent-based long-haul freight transport model for Germany. Procedia Computer Science 201, 614–620. doi:10.1016/j.procs.2022.03.080.

[17] Maia Pereira, A., Dingil, A., Přibyl, O., Myška, V., Vorel, J., Kříž, M., 2022. An Advanced Travel Demand Synthesis Process for Creating a MATSim Activity Model: The Case of Ústí nad Labem. Applied Sciences 12, 10032. doi:10.3390/app121910032.

[18] Mallig, N., Vortisch, P., 2017. Modeling travel demand over a period of one week: The mobiTopp model. doi:10.48550/arXiv.1707.05050, arXiv:1707.05050.

[19] Martins-Turner, K., Leich, G., Rakow, C., Schlenther, T., Rybczak, G., Kreuschner, M., Nagel, K., 2024. Calculation for Car Costs in MATSim Open Berlin Scenario. Technical Report. TU Berlin, Transport Systems Planning and Transport Telematics. Berlin. URL: https://doi.org/10.14279/depositonce-20678.

[20] Miller, E., 2023. The current state of activity-based travel demand modelling and some possible next steps. Transport Reviews 43, 565–570. doi:10.1080/01441647.2023.2198458.

[21] Miller, E.J., Roorda, M.J., 2003. Prototype Model of Household Activity-Travel Scheduling. Transportation Research Record 1831, 114–121. doi:10.3141/1831-13.

[22] Nagel, K., Zilske, M., Flötteröd, G., 2016. CaDyTS: Calibration of Dynamic Traffic Simulations, in: Horni, A., Nagel, K., Axhausen, K.W. (Eds.), The Multi-Agent Transport Simulation MATSim. Ubiquity, London. chapter 32. doi:10.5334/baw.

[23] Neumann, A., Balmer, M., 2011. Micro meets macro: A combined approach for a large-scale, agent-based, multi-modal and dynamic transport model for Berlin. VSP Working Paper 11-14. TU Berlin, Transport Systems Planning and Transport Telematics. URL http://vsp.berlin/publications.

[24] Rakow, C., Kaddoura, I., Nippold, R., Wagner, P., 2021. Investigation of the system-wide effects of intelligent infrastructure concepts with microscopic and mesoscopic traffic simulation, in: 27th ITS World Congress, Hamburg, Germany, Hamburg, Germany. URL: https://elib.dlr.de/144810/.

[25] Rakow, C., Nagel, K., 2024a. Automated Calibration of Alternative Specific Constants in Agent Based Transport Models. Technical Report. TU Berlin, Transport Systems Planning and Transport Telematics. Berlin. URL: https://doi.org/10.14279/depositonce-20593.

[26] Rakow, C., Nagel, K., 2024b. Road network free flow speed estimation using microscopic simulation and point-to-point travel times. Procedia Computer Science 238, 698–705. doi:10.1016/j.procs.2024.06.080.

[27] Schrab, K., Protzmann, R., Radusch, I., 2023. A Large-Scale Traffic Scenario of Berlin for Evaluating Smart Mobility Applications, in: Nathanail, E.G., Gavanas, N., Adamos, G. (Eds.), Smart Energy for Smart Transport, Springer Nature Switzerland, Cham. pp. 276–287. doi:10.1007/978-3-031-23721-8_24.

[28] Ziemke, D., Kaddoura, I., Nagel, K., 2019. The MATSim Open Berlin Scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data. Procedia Computer Science 151, 870–877. doi:10.1016/j.procs.2019.04.120.

[29] Ziemke, D., Nagel, K., 2017. Development of a fully synthetic and open scenario for agent-based transport simulations – The MATSim Open Berlin Scenario. URL: https://www.semanticscholar.org/paper/Development-of-a-fully-synthetic-and-open-scenario-Ziemke-Nagel/84da13211919457ba2579aa18ee86d1d692cb23f.

[30] Ziemke, D., Nagel, K., Bhat, C., 2015. Integrating CEMDAP and MATSim to increase the transferability of transport demand models. Transportation research record 2493, 117–125.