Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 47 (2020) 99–106

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18-20 September 2019, Barcelona, Spain

Development of an Agent-Based Transport Model for the City of Hanover Using Empirical Mobility Data and Data Fusion

Lasse Bienzeisler[a,*], Torben Lelke[a], Oskar Wage[b], Falco Thiel[a], Bernhard Friedrich[a]

[a]*Institute of Transportation and Urban Engineering, TU Braunschweig, Hermann-Blenk-Straße 42, 38108 Braunschweig, Germany*
[b]*Institute of Cartography and Geoinformatics, Leibniz University Hannover, Appelstraße 9a, 30167 Hannover, Germany*

Abstract

The model presented in this work is based on the agent-based simulation framework MATSim. We describe a new approach for collecting and processing the underlying data required for the development of a MATSim scenario. In cooperation with the administration of Hanover, available data was centrally collected, analysed and clustered. Using data fusion, the dataset was combined and enriched with additional information from open data sources in order to improve the model's level of detail. In combination with the German mobility survey *Mobilität in Deutschland 2017*, the developed model is specifically adapted for the regional characteristics of Hanover, but yet transferable to other scenarios.

Keywords: MATSim; Scenario Generation; Geo-spatial data

1. Introduction

Microscopic traffic models are progressively applied to expand traditionally favored macroscopic modelling approaches (Erol et al., 2000). Although the relationship between traffic models and the spatial input data is evident, it is often neglected. Instead, transportation and geographic research developed separately (Loidl et al., 2016). By collaborating, we want to close this gap and combine knowledge from both disciplines to further improve the accuracy of traffic models. Novel agent-based models utilize singular mobility decisions such as trip purpose, destination, mode and time of day, to reproduce the traffic demand. The simulation of each agent enables the consideration of complex linkages across several trips, which ultimately leads to a refined understanding of traffic flow (Vovsha et al., 2002). A key challenge of agent-based approaches is the detailed mapping of the population into an initial traffic demand input. We used the agent-based transport simulation framework MATSim (Horni et al., 2016), reproducing typical mobility patterns and approximating the traffic flow in the investigated area.

* Corresponding author.
E-mail address: l.bienzeisler@tu-braunschweig.de

10.1016/j.trpro.2020.03.073

2. Research contribution

Geo-spatial data is bound to a specific scale. Therefore, it only can be aggregated in lesser detail than originally captured. The level of aggregation defines the resolution of any operation based on the data. It is thus challenging to find optimally scaled input data within a given model context (Loidl et al., 2016). To improve the accuracy and geo-spatial resolution of agent-based transport models, we developed a methodology to determine the distribution of activity locations by fusing aggregated data sources on the level of buildings. A critical part in this process is the assignment of workplaces (Rieser et al., 2016). While municipalities typically have rich demographic data, little information on the distribution of inner city workplaces is collected, albeit commuting traffic still being a major issue for most if not every modern city. Despite our cooperation with the City of Hanover within the research project, we were not able to obtain this data.

Various methods have been developed to create the initial demand for an agent-based traffic simulation framework such as MATSim. Existing approaches are often highly specific and were adapted to a precise scenario depending on the available input data. These are usually based either on different regional mobility surveys (Kickhöfer et al., 2016), artificial approaches (Hörl, 2017), big data sources (Anda et al., 2018), or synthetic populations generated by an external system of random-utility-based-models representing the decision making behavior of individuals (Ziemke et al., 2015). This method was used to generate a synthetic population of Berlin (Ziemke et al., 2019) and the German Ruhr-Region (Ziemke et al., 2018). They created activity patterns based on a framework of a model specification (CEMDAP - Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns (Bhat et al., 2008)) from the metropolitan regions of Los Angeles, USA, and applied these to their scenarios. Simulation studies demonstrated that the approach is feasible for German scenarios as well. Nonetheless, it is desirable to include further data collected and provided by the cities directly into the model generation, as it is complex, in-transparent, and time-consuming to adapt external activity generators to regional mobility data. Therefore, the aim of this paper is to introduce spatial data as possible in a high resolution into the MATSim model generation of the city of Hanover.

3. Methodology

A common challenge in building a transport model is the access to sufficient data with respect to completeness, granularity, and reliability (Parsuvanathan, 2015). Having no all-in-one data set, a combination of various data sources is indispensable. We build up a data collection stored and managed in a relational database system with spatial capabilities enabling a joint and up-to-date data pool for our research group. Using this data sets we applied a data fusion algorithm estimating adequate base parameters for model generation and applied the results to the MATSim model generation.

3.1. Data Basis

The input road network was created using OpenStreetMap (2019) (OSM) data. To implement schedule-based public transport, open General Transit Feed Specification (GTFS) data provided by Connect Fahrplanauskunft GmbH (2019) was converted into a MATSim data. Locations for leisure activities, schools, and universities were also extracted from OSM based on common tags on this topic from OpenStreetMap Wiki contributors (2018). To identify buildings with a commercial use, the key-values in *amenity*, *building*, *office* and *shop* tags were used as evidence. We mapped all locations on building geometries to derived the location size using the building height and floor area. In combination with a granular population statistics simulation agents' home locations were generated from this data. The city of Hanover provided traffic counts and listing of retail locations with corresponding main branch as well as retail area.

Working places estimation

We assumed a total of 421.600 workplaces[1] in Hanover. To approximate a realistic distribution of this given target sum, we developed an alternative estimation process from available data. Building geometries from the city of Hanover were iteratively enriched (see figure 1) with further data to optimize the estimation. In a first step accounting

[1] https://www.hannover.de/Wirtschaft-Wissenschaft/Arbeit/Arbeitsmarkt/Arbeitsmarktdaten/Arbeitsmarkt

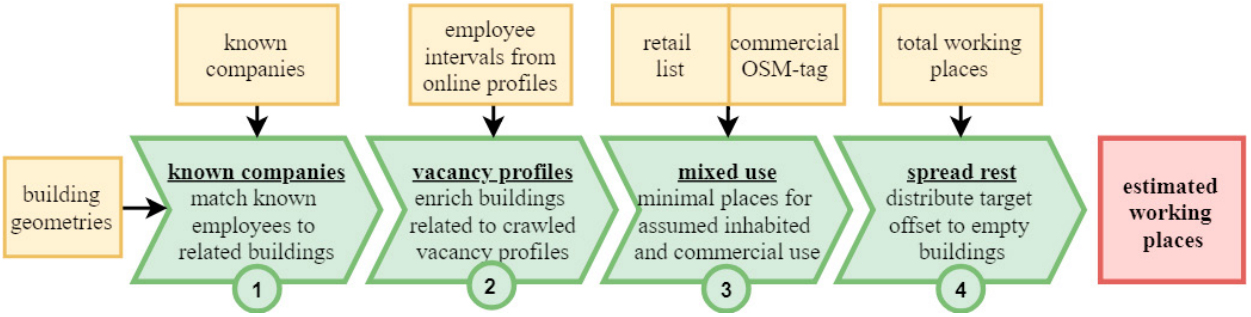

Fig. 1: Four main steps (green) to estimate working place distribution on building level by introducing multiple data sources (yellow).

for approximately 100.000 working places, public information on the number of employees for some of the largest companies[2] in Hanover were assigned to the respective buildings. The total amount of each company was split relatively to the buildings' volume. Implausible allocations were avoided by assuming a minimum of 25 m^3 per working place. This value is deduced from a minimal office space of about 10 m^2 per person (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, 2018) times 2.5 m floor height. In a second step, additional data on companies' employees was mined severel times from the federal online job portal[3]. Listed vacancies were linked to a company profile which included inter alia the company name, number of employees in rough intervals, branch and address. The company entries were matched to the buildings by geocode addresses to coordinates. From this, the sum of resulting values per building was assigned constrained by a minimum of 25 m^3 per workspace. Capturing buildings with dual use (housing and commerce), ones with home activities but without assigned working places were intersected with extracted OSM objects and retail locations from the city administration. The commercial use was assumed to be small and thus set as three places per detected building. Lastly, the remaining difference of working places to the target number was distributed among so far untouched and assumed to be uninhabited buildings relatively to their volume. Checking for plausibility, we compared our estimates to the German guideline on induced traffic volume (FGSV, 2006) which also includes information on workers per hectare. As a result buildings in areas of lesser used land development plan types (e.g. *sports)* were weighted down by the factor of ten.

Population

The population was assumed to match the yearly statistical report[4] published by the city of Hanover. Using this data, we extrapolated the exact number of agents for all 51 city districts. We divided the total number of agents into population groups. This was carried out by using the obtained data from the administration as well as other data sources, such as city-specific data from the German nationwide census of 2011[5]. We determined the amount of children and pensioners within the districts directly from the administration's age statistics. Calculating the number of full- and part-time workers posed a challenge, as the data from the statistical reports only included the number of employees contributing to social insurance in each district. We upscale the data by calculating the average proportion of these employees, the share of people working half-day and and the rate of homemakers, using data from the census. The statistical report of the districts also included the unemployment rate which we used to calculate the amount of unemployed persons for the model by applying this rate to all persons between the ages of 18 and 65. Remaining persons were identified as students. To determine the required number of commuters within a range of 100 km we

[2] https://www.hanover.ihk.de/fileadmin/data/Dokumente/Themen/Konjunktur_Statistik/Groesste_Unternehmen_IHK-Hannover_2017_01.pdf
[3] https://jobboerse.arbeitsagentur.de
[4] https://www.hannover.de/Service/Presse-Medien/Landeshauptstadt-Hannover/Meldungsarchiv-für-das-Jahr-2017/Statistische-Profile-der-Stadtteile-und-Stadtbezirke-2017
[5] https://ergebnisse.zensus2011.de/

used the German commuter statistic[6]. Since the total number of students in Hanover is known[7], the difference between this number and the extrapolated number of students was used to calculate the number of commuting students. After generating the population, the agents had to be assigned with personal attributes, such as age, income and whether they are mobile at all and have access to a car. We distributed these attributes with regard to the latest German mobility study *Mobilität in Deutschland* (MiD) (Bundesministerium für Verkehr und digitale Infrastruktur, 2018). In a last step we equipped the agents with initial daily plan pattern for the simulated day. For the model of Hanover, we used activity sequences taken from the survey *Deutsches Mobilitätspanel* (MOP) (Bundesministerium für Verkehr und digitale Infrastruktur, 2016).

3.2. Model generation

Generation of a synthetic population and assignment of initial daily activity-travel pattern

For the generation of the initial MATSim demand, we used mobility surveys as the basis of agents' travel patterns. The applied process is described in Algorithm 1. We generated the synthetic population automatically, incorporating our previous findings and assumptions. In a first step, we initialized the population for each district of the city of Hanover according to the population distribution. Specific home coordinate were assigned to each agent serving as the initial and last activity of the day. Commuting agents were set up with a approximated coordinate of their home region. Every agent's sequence of activities was based on the MOP. Due to the frequency of certain activity patterns, a sequence was selected for the respective agent. Since the focus of the modelling is on the city of Hanover, the daily schedule of commuters and commuting students was approximated by the simplified pattern *Home-Work/Education-Home*. After creating and locating the agents in the model, we collectively assigned attributes according to the MiD. Each agent was set up with an age and gender. These attributes were then used to calculate an income group. The income was utilized to determine, whether the agent has access to a car on the simulated day. The last attribute is the mobility of a agent during the simulated day and is dependent on the population group and the car availability. In order to reduce the computing time of the MATSim simulation, the population was reduced to a 10 % example. Corresponding amounts of inhabitants and working places were randomly selected for each city district independently to preserve their respective density and to prevent single areas from disproportionate thinning.

Location and time choice for the initial daily activity-travel pattern

By using the MiD trip distance distribution and starting with the first home coordinate we extrapolated agents' activities. The type of the next activity and the population group of each agent specified the search area for the location of the next activity. Due to the model size, longer trip distances were excluded and approximately represented by the commuting agents. Suitable activity coordinates must meet requirements for the different activity types. Shopping activities for example were selected based on the retail area of the shop. The more retail area a shopping category has, the more likely it is that this coordinate was selected. In some cases it may occur that there is no coordinate in the search radius satisfying the selected requirements. The search radius was altered until a coordinate with the desired attributes was found. In a last step, we defined activity times for each agent's schedule. Since the duration of each activity depends on the total number of activities, an average activity duration was determined according to the MOP. The average duration was then used to calculate a daily time contingent for every agent. Using empirical data of the average duration of specific activities, a daily duration for an agent's main activities was calculated according to the German time use survey (Statistisches Bundesamt, 2013). For plans including several main activities, the daily duration was split among these. The agent's remaining free time was divided among the remaining activities of the plan. For all agents with the type *work* or *education*, the start time was normally distributed around 8 o'clock with a standard deviation of one hour. The remaining agents started their daily schedule randomly between 8 and 16 o'clock. In addition, the typical daily travel time for each agent was included in the calculation.

[6] https://statistik.arbeitsagentur.de/Navigation/Statistik/-Statistische-Analysen/Interaktive-Visualisierung/Pendleratlas/Pendleratlas-Nav.html

[7] https://www.hannover.de/Leben-in-der-Region-Hannover/Politik/Wahlen-Statistik/Statistikstellen-von-Stadt-und-Region/Statistikstelle-der-Landeshauptstadt-Hannover/Hannover-in-Zahlen/Bildung

Algorithm 1: Process of the applied MATSim initial demand generation

```
Result: MATSim Plan.xml
initialization();
data_collection_from_database();
data_processing() ;
foreach district do
    foreach populationType do
        while size(agents) < size(population in district) do
            create_agent();
            foreach Agent do
                set age; set gender; set homeCoordinate;
                set customAttributes = calculate(populationType, age, gender);
                set carAvailability = calculate(attributes);
                set mobility = calculate(type, carAvailability);
                if mobility = false then
                    activityPattern = Home ;
                else
                    if populationType = Commuter then
                        activityPattern = Home-Work-Home ;
                    if populationType = Commuter_Student then
                        activityPattern = Home-Education-Home ;
                    else
                        activityPattern = get_plan_pattern_from_survey(populationType);
                    foreach Activity do
                        foreach activityType do
                            while Coordinate not found do
                                coordinate = calculate(activityType);
                                if coordinate found then
                                    set coordinate;
                                else
                                    update(searcharea);
                            Plan = add(Activity);
                set times = calculate(Plan);
                Agent = add(Plan);
            Population = add(Agent);
```

Calibration of the MATSim model

The simulation was carried out over a simulation run with 500 iterations. Adopting a scoring function, that quantifies the individual benefits of activities and corresponding travel expenses, the plan of each agent was evaluated and iteratively optimized by the simulation framework. The applied modes were *car*, *public transport*, *walk*, *ride* and *bike*. In the first run, each agent having a car available was assigned to the mode *car*. However, if this was not the case, it was checked for each trip in the respective daily plan whether the activities were located within reach of a public transport stop. If this requirement was obeyed, the mode was set to *public transport (pt)*, otherwise it was switched to *ride*.

As the simulation progresses, agents change transport mode to find the most effective one for their trips. Table 1 shows the comparison between the surveyed values from the MiD and the simulation results. It turns out that

Table 1: Mode Share comparison between simulation and MiD

Transportation Mode	Mode Share (%)			
	MiD (region)	Simulation (region)	MiD (city area)	Simulation (city area)
Car	35	34	27	29
Ride	12	10	9	8
Public Transport	16	16	19	18
Bike	16	20	19	22
Walk	22	20	26	23

especially the bike share was slightly too high and the walk share too low. However, since these are teleported modes in the simulation model, the calibration to the mode share indicate that the model reproduced sufficient results.

The simulation was calibrated by comparing traffic counts with the simulated traffic volumes using the calibration module *Cadyts* (Calibration of dynamic traffic simulations) (Nagel et al., 2016). The simulation results indicated that traffic at peak hours was underestimated by approximately 20% (see figure 2). In the morning or evening hours, the relative error is higher because the sample size of counting stations is significantly smaller than during the rest of the day. As the model focuses on modelling the daily traffic, this error can be tolerated.

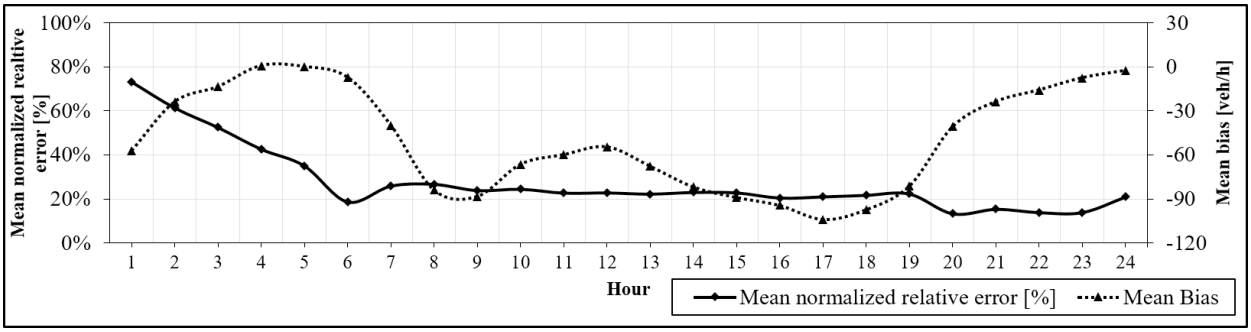

Fig. 2: Mean relative error and mean average bias comparing simulation results and traffic counts

4. Preliminary results and discussion

After calibration, the results of the model were compared to data from the MiD. In order to validate whether the daily plans of the agents and the location choice provided suitable results, we compared the relative frequency of distances and travel times (Figure 3). Especially the short distances were underestimated in the model. The mid-range distances were also less strongly represented than in reality, while long distances are more frequent. This observation is related to the way commuters are modeled. They are currently only represented by a simplified daily pattern. This leads to an under representation of the shorter distances. In addition, all commuters within a radius of 100 km of Hanover were modeled. In the commuter statistics on which these assumptions are based, commuters who held a secondary residence are also included. Thus, it is possible that e.g. a commuter lives in his second residence in Hanover during the week but is modeled as daily commuter. Looking at the trip duration, it is conspicuous that short travel times are underrepresented. The analysis of the simulation results indicated agents almost solely either walk or bike for these very short trips. However, the MiD shows that in large cities the car is used in 9% of the distances shorter than 500 m. This difference also shifts the distribution of travel times in the model. Despite this deviation, the cumulated relative frequency of trip lengths and trip durations between model and study fit very comparable.

Considering the fact, that the model currently only includes individual traffic and commercial traffic has not yet been implemented, the results are reasonable. Preliminary manual traffic counts as a part of our research have shown

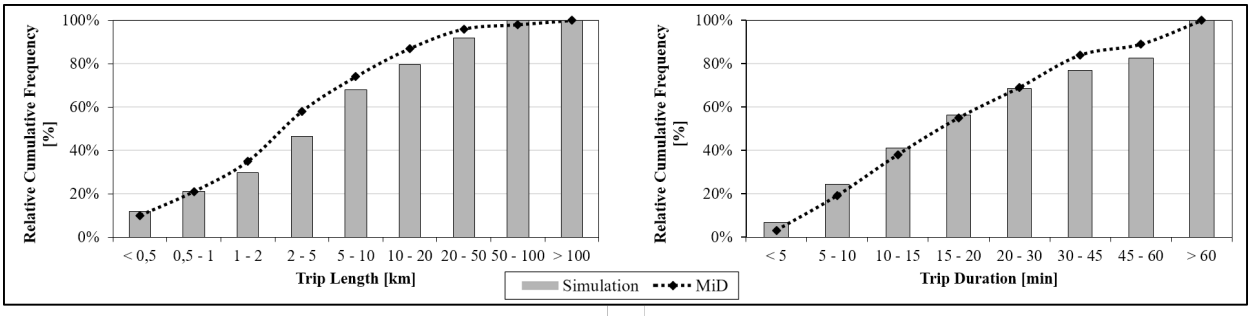

Fig. 3: Comparison between simulated trip length and trip duration frequency and similar values from the MiD

that the (visible and thus countable) share of commercial traffic in Hanover amounts to approx. 15 % of the total traffic. These results indicate that in addition of modelling the individual mobility of the inhabitants, commercial traffic must also be taken into account. Another factor influencing the calibration results is the accuracy of the counts used. These consisted partly of automated counts from inductions loops of traffic lights. Since counting vehicles is not the primary purpose of these loops, they tend in practice to under count (Briedis and Trueman, 2010).

In addition to the generation of the initial demand for the simulation model, the detailed extrapolation of activity locations was a major target of this work. The resulting distributions of activity locations for *home*, *work* and *shopping* activities are visualized by their spatial density in figure 4. The population density on the left indicates inhabitant hotspots in the residential districts adjacent to the centre with a lower density. In contrast, working place hotspots are among others in the city centre, the eastern business district and at the north-west edge of town (automotive factories). The shopping coordinates are concentrated on the inner city as well as on individual hotspots spreaded over the entire city area. At these spots shopping malls and supermarkets are situated. It is difficult to verify the accuracy of our extrapolation, especially for the workplace distribution, since no ground truth information is available. However, the model covers the characteristics and companies of the city of Hanover correctly.

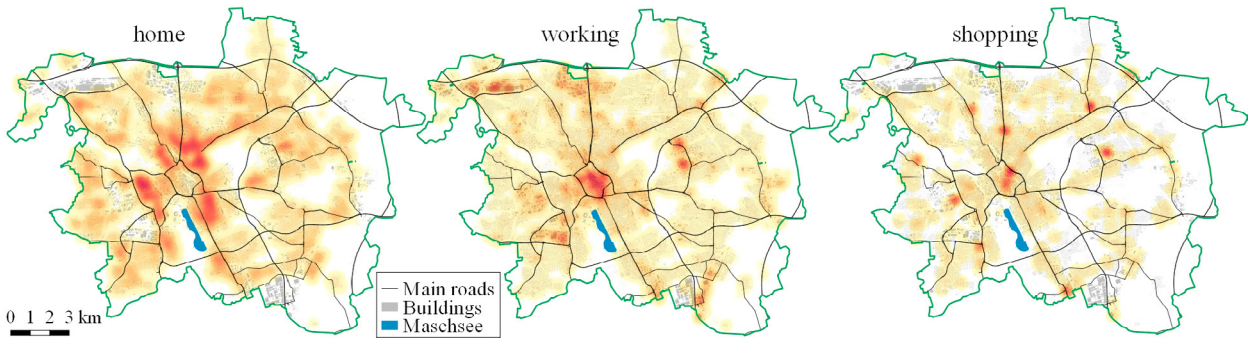

Fig. 4: Overview of the density distribution of home (left), working (centre) and shopping (right) activities over the city area. The densities are coloured from low/yellow to high/red and transparent for none existence.

5. Conclusion and outlook

As the calibration analysis showed, the model underestimates the daily traffic in Hanover by about 20%. The aim of future work is to include inner city commercial traffic to increase the accuracy of the model. It apparent, that a more accurate representation of the commuting agents in the model is necessary. The research area must be expanded in order to accurately represent the inhabitants of the Hanover region. As a result, the traffic passing through the city could be integrated into the model.

In the future, it will also be possible to use the raw data from MiD 2017 in order to integrate this wealth of information into the model generation process. The additional information will enable daily activity patterns to be allocated better based on different attributes of an agent. Thus, for example, an increased income of the household can lead to a completely different set of possible daily plans. A further improvement can be expected from the switch to a household based agent population. Since each agent is created individually, mutual dependencies cannot be represented at the moment. But in reality, there are strong relationships for instances between the daily plans of children and those of their parents. The methodology for extrapolating activity locations on a building level has to be further improved and validated. It is essential to investigate which degree of data density is useful and when further refinement will no longer be beneficial.

In addition to further refinements to the applied model generation methodology, the next step of the research is to calibrate and validate the model against further mobility and activity parameters. New technologies such as floating car data or GPS information from smartphones can be used to improve modeling accuracy and they can serve to validate agent-based models. By considering even inner-city inhomogeneities a more granular and data driven understanding of dependencies between peoples mobility and commercial traffic is expected. The development of this methodology should lead to a generalization of the approach, generating results not only for the city of Hanover.

Acknowledgement

The scientific research published in this article is granted by the Federal Ministry of Education and Research Germany for project USEfUL (grant ID 03SF0547). The authors cordially thank the partners and funding agency.

References

Anda, C., Medina, S.A.O., Fourie, P., 2018. Multi-agent urban transport simulations using OD matrices from mobile phone data. Procedia Computer Science 130, 803–809. doi:10.1016/j.procs.2018.04.139.

Bhat, C., Guo, J., Srinivasan, S., Sivakumar, A., 2008. CEMDAP User's Manual.

Briedis, P., Trueman, H., 2010. The accuracy of iductive loop detectors. 24th ARRB conference : building on 50 years of road and transport research : proceedings .

Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, 2018. ASR A1.2 Raumabmessungen und Bewegungsflächen.

Bundesministerium für Verkehr und digitale Infrastruktur, 2016. Deutsches Mobilitätspanel 2015/2016 (MOP), Dataset.

Bundesministerium für Verkehr und digitale Infrastruktur, 2018. Mobilität in Deutschland 2017.

Connect Fahrplanauskunft GmbH, 2019. Connect-OpenData-Pool. URL: http://www.connect-fahrplanauskunft.de/unsere-services/open-data.html.

Erol, K., Levy, R., Wentworth, J., 2000. Application of agent technology to traffic simulation .

FGSV, 2006. FGSV 147: Hinweise zur Schätzung des Vekehrsaufkommens von Gebietstypen .

Horni, A., Nagel, K., Axhausen, K.e., 2016. The Multi-Agent Transport Simulation MATSim. Ubiquity Press, London. doi:10.5334/baw.

Hörl, S., 2017. A matsim scenario for autonomous vehicles in la défense and Île-de-france doi:10.13140/rg.2.2.14946.12487.

Kickhöfer, B., Hosse, D., Turner, K., Tirachini, A., 2016. Creating an open matsim scenario from open data: The case of santiago de chile doi:10.13140/RG.2.2.25394.40649.

Loidl, M., Wallentin, G., Cyganski, R., Graser, A., Scholz, J., Haslauer, E., 2016. GIS and transport modeling—strengthening the spatial perspective. ISPRS International Journal of Geo-Information 5, 84. doi:10.3390/ijgi5060084.

Nagel, K., Zilske, M., Flötteröd, G., 2016. Cadyts: Calibration of dynamic traffic simulations doi:10.5334/baw.

OpenStreetMap, 2019. Openstreetmap contributors. URL: https://planet.openstreetmap.org.

OpenStreetMap Wiki contributors, 2018. Openstreetmap wiki. URL: https://wiki.openstreetmap.org.

Parsuvanathan, C., 2015. Big data and transport modelling: Opportunities and challenges. International Journal of Applied Engineering Research 10, 38038–38044.

Rieser, M., Horni, A., Nagel, K., 2016. Let's Get Started. Ubiquity Press. pp. 9–22. doi:http://dx.doi.org/10.5334/baw.

Statistisches Bundesamt, 2013. Zeitverwendungserhebung 2012/2013. doi:10.21242/63911.2013.00.00.4.1.0.

Vovsha, P., Petersen, E., Donnelly, R., 2002. Microsimulation in travel demand modeling: Lessons learned from the new york best practice model. Transportation Research Record: Journal of the Transportation Research Board 1805, 68–77. doi:10.3141/1805-09.

Ziemke, D., Kaddoura, I., Nagel, K., 2018. Entwicklung eines regionalen, agentenbasierten Verkehrssimulationsmodells zur Analyse zukünftiger Verkehrsszenarien für die Region Ruhr .

Ziemke, D., Kaddoura, I., Nagel, K., 2019. The matsim open berlin scenario: An openly available agent-based transport simulation scenario based on synthetic demand modeling and open data .

Ziemke, D., Nagel, K., Bhat, C., 2015. Integrating CEMDAP and MATSIM to increase the transferability of transport demand models. Transportation Research Record: Journal of the Transportation Research Board 2493, 117–125. doi:10.3141/2493-13.