Development of a fully synthetic and open scenario for agent-based transport simulations – The MATSim Open Berlin Scenario

Dominik Ziemke, Kai Nagel
Transport Systems Planning and Transport Telematics
Technische Universität Berlin

July 22, 2017

Abstract

This paper documents the development of a new Berlin scenario that is based on open data. Because of its use of standardized data, the generation procedure of the scenario is comparatively easy to be applied in other regional contexts.

1 Introduction

The purpose of this work is to create a scenario for agent-based transport simulation studies of the metropolitan region of Berlin and Brandenburg in Germany. The main characteristics of the scenario are that it

- contains full day plans of all agents who represent the people living in the region,
- considers all relevant modes of transport,
- relies on openly available and unrestrictedly usable data ("Open Data"),
- relies on standardized data to ensure spatial transferability of the scenario generation procedure,
- is policy-sensitive in terms of route and mode choice.

The development procedure is based on a person-specific activity scheduling model and a traffic-count-based calibration procedure. The procedure to create the scenario described in this paper can be regarded as an extension of the work described by Ziemke et al. (2015). In contrast to the preceding work, the scenario has been improved with regard to the following aspects:

- The representativeness of the synthetic population has been improved. The current scenario contains a representation of all inhabitants of the German states of Berlin and Brandenburg aged 18 years and older, whereas the previous work contained only a population scaled down to the share of tripmakers that use the car.
- More modes are considered (car, two different forms of public transport, cycling, walking), whereas the previous work only contained car drivers, neglecting trips and tripmakers in other modes of transport.
- In addition to route choice, which was already part of the preceding work, the agents in the current scenario are able to choose modes. The scenario has been calibrated regarding both choice dimension (route choice, mode choice) and can be regarded policy-sensitive with regard to these choice dimensions.

This scenario is similar to a scenario for Santiago de Chile (Kickhöfer et al., 2016) in that it is also open to use. In contrast to the Santiago scenario, which relies on a region-specific travel survey, this scenario is also open on the input side, i.e. only data sources that can be easily obtained in most places of the world are used. As such, the approach to create this scenario is readily portable to other regions and can serve as a blueprint for more easy and quicker scenario setup.

An approach to set up a scenario based on the same development trajectory for the Ruhr region in Germany is currently underway.

2 Data

The following data sources are used to create the described scenario:

- Zensus 2011 (a Germany-wide census, Statistische Ämter des Bundes und der Länder, accessed 21 July 2017).
- Pendlerstatistik (commuter statistics, Bundesagentur für Arbeit, 2010).
- OpenStreetMap (OpenStreetMap, accessed 30 Jan 2017).
- Traffic Counts.
- Shapefiles that describe the geometries of municipalities in Brandenburg and LORs[1] in Berlin.

3 Methodology

3.1 Generation of a synthetic population

The population of the scenario consists of all persons aged 18 and above who reside in the German states of Berlin and Brandenburg according to the "Zensus 2011" (Statistische Ämter des Bundes und der Länder, accessed 21 July 2017). The census file contains, amongst further information, for each municipality

- the total population, differentiated by gender and eleven age classes,
- the number of employees, differentiated by gender, and
- the number of students.

Workplaces for employees are informed by the "Pendlerstatistik 2009" (commuter statistics, Bundesagentur für Arbeit, 2010). The commuter statistics files contain the number of socially-secured employees, differentiated by gender, for each municipality-to-municipality pair. Note that Berlin is only one municipality, which makes this approach requiring refinement inside Berlin. In areas with smaller municipalities, like mainly found in Brandenburg, no further refinement is necessary. Also note that the census, in contrast to the commuter statistics, does not state the number of socially-secured employees, but all employees (based on their self-report).

Next to the census and the commuter statistics, the synthetic population generation procedure uses shapefiles that describe the geometries of municipalities in Brandenburg and LORs in Berlin as input.

First, the procedure (implemented in a Java class called `DemandGeneratorCensus`) reads in the aforementioned attributes of all municipalities from the census file and the number of commuter relations for each municipality-to-municipality pair of the commuter statistic files.

Second, a scaling procedure is applied, which scales municipality-to-municipality-based commuter values such that the number of all commuters departing from one zone (given by the commuter statistics, which is solely based on socially-secured employees) meets the number of employees residing in that zone (given by the census, which is based on all employees). This scaling is done separately for each gender-and-age-specific group of the population.

Third, the procedure iterates over all municipalities and creates in each municipality persons, assigning them age and gender such that the numbers of males and females in each age group of the census are met. If the municipality is Berlin, a random LOR in Berlin is chosen. Since LORs are defined such that the population of a LOR does not fall below or exceed a certain minimum or maximum (Bömermann et al., 2006), this procedure approximates population densities in Berlin. Based on the previous step, agents are assigned to be employed or not employed and to be a student or not a student in accordance with the relevant numbers from the census. If an agent is an employee or a student, they are assigned with a work or school location zone according to the commuter information. If the destination is Berlin, a random LOR is chosen.

Finally, this information is written out in tab-separated text files according to the format of CEMDAP (cf. section 3.3) input files as described by Bhat et al. (2008). In the current implementation, the output file carries values for the attributes household ID, person ID, being employed, being a student, having a driver's license, work zone, school zone, gender, age, and being a parent. All other attributes are set to zero and therefore not taken into account in subsequent modeling steps. In the same step, one household is created for each agent, i.e. household structures are not taken into account. Children (people under the age of 18 years) are not included.

[1] LOR stands for "Lebensweltliche orientierte Räume" and is a neighborhood-oriented zone system intended to become the standard spatial reference for Berlin.

3.2 Generation of other CEMDAP input information

Next to the synthetic population (person and household files), CEMDAP requires input files regarding land use and levels of service. A so-called `zone2zone` file carries information regarding the distance of any two zones and whether they are adjacent. Adjacency is computed based on methods from the `GeoTools` Java library and, similarly, distances are computed between zone centroids. The `zones` file can contain for each zone information regarding numbers of different types of employees in the zone, certain accessibility measures etc. In the current implementation, no such information is used, i.e. the model considers all zones equally attractive. The `los` (level of service) files contain travel times and travel costs between any two zones. Their values are computed based on distances between zone centroids.

3.3 Assignment of daily activity-travel plans

To assign daily activity-travel patterns for each member of the synthetic population, CEMDAP (Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns, Bhat et al., 2004) is used. CEMDAP is a software implementation of a system of random-utility-based models that represent the decision-making behavior of individuals. CEMDAP's output consists in the complete daily activity-travel patterns of each individual of the synthetic population and outlines the sequence of activities and intervening trips that a person undertakes during the day.

The input data files mentioned in sections 3.1 and 3.2 are read into a PostgreSQL database[2]. After starting the CEMDAP software, the database holding the input information and a model-specification file are specified. The model-specification file holds all parameters of all CEMDAP models. The model specification used here is taken from CEMDAP's latest implementation for the Los Angeles metropolitan areas in the United States. This is associated with the assumption that persons of the model estimation context (Los Angeles) with certain demographic attributes have to some extent similar day plans as persons with the same demographic attributes in the model application context (Berlin).

Note that this does *not* mean that the approach does not account for differences between the two regions. First, the fact that the synthetic population and the geography of the two regions are different has already been accounted for as described in sections 3.1 and 3.2. Second, despite the fact that CEMDAP is run for Berlin with the parameter set estimated for Los Angeles, the overall procedure must *not* be regarded an *un*modified application of the Los Angeles model. In contrast to the usual application of CEMDAP, when applied in a regional context for which an explicit model estimation has been carried out, here, the modeling results are not directly fed into the transport system. Instead, CEMDAP is run multiple times (in the current scenario five times) to generate a selection of potential activity-travel patterns for each agent. All these patterns are fed into the MATSim transport simulation (Horni et al., 2016), which, in correspondence with a calibration procedure (cf. section 3.4), sorts out those activity-travel patterns that do not contribute well to a reproduction of real-world traffic patterns as they are given on the basis of traffic counts.

Ziemke et al. (2015) describe different methods of model transfer and mention that an update of model parameters is generally advisable and found to lead to better results than approaches without updating. While in other model transfer studies updating is mostly done with regard to parts of model parameters, in the present approach, model updating operates on initial full daily activity plans, while the CEMDAP model parameters themselves remain unchanged.

3.4 Location Choice Calibration

As a result of the previous step (application of CEMDAP, cf. 3.3), five (potential) daily activity-travel patterns are obtained for each member of the synthetic population. The task of this modeling step is to choose those of the multiple initial activity-travel patterns that best represent real-world travel observations as given by traffic counts. Since the main difference among the different initial activity-travel patterns consists in different activity locations, this modeling step constitutes a process of location choice.

Using a class called `CemdapStops2MatsimPlansConverter`, the array of different activity-travel patterns modeled by CEMDAP are translated into the set of MATSim plans of the corresponding agent. Home and work locations are chosen by a random draw within the geography of the zone, where the home or work activity takes place. The home locations are kept constant over the various plans of a given agent. For computational resource efficiency, a 10% sample of the full population is created.

[2] The precise procedure of reading in the input files into a PostgreSQL database and its technical setup are described by Bhat et al. (2008).

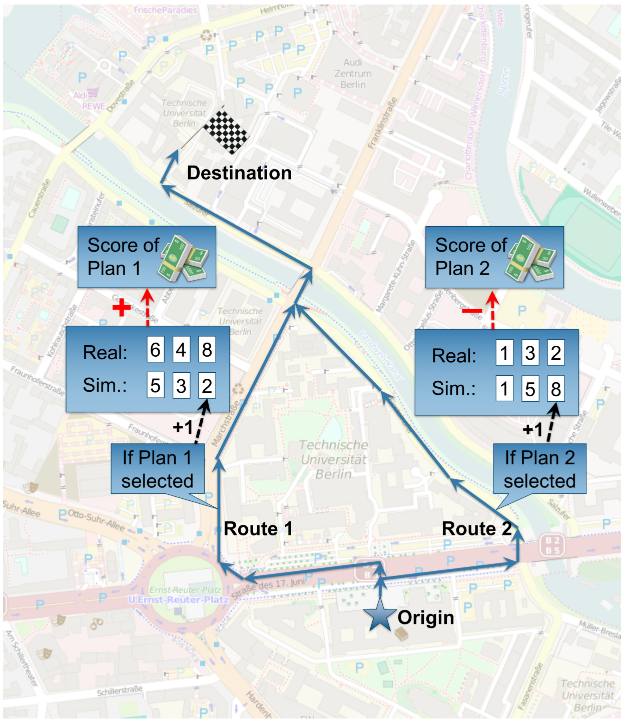


Figure 1: Simplified illustration of Cadyts calibration procedure.

The network for the transport simulation was created on the basis of data from OpenStreetMap (accessed 30 Jan 2017). After simplification, the network consists of 11,345 nodes and 24,335 single-direction car-only links.

Real-world travel observations are given by 8,304 hourly count values for 346 count stations. 250 of these count stations are operated by the Berlin Traffic Management Center (Verkehrsmanagementzentrale), while the remaining 96 stations belong to the motorway administration. In these values, no distinction is made between vehicles of different types (e.g. cars and trucks). Note that these data constitute the only component of all input data, which are not directly openly available, but only obtainable upon request. As such data are, however, available in the majority of regions, even in less-developed regions where such data can be provided by means of manual counts, the use of this dataset is not regarded as a major hurdle to model transferability.

To calibrate an agent's plan choice set against traffic counts, MATSim is run in conjunction with Cadyts (Calibration of dynamic traffic simulations, Flötteröd, 2010) as described in more details by Ziemke et al. (2015). In contrast to that work, here, travel demand is *not* scaled down to the share of car travelers since it is intended to create a scenario that considers travelers of all modes.

Instead, transport network properties are scaled up by the inverse of the Berlin-specific modal share for cars. Note that in this modeling step, the choice of transport modes has not been taken into account. All travelers travel on the car network in this modeling step. Because of the up-scaling of the network, the traffic dynamics of all tripmakers travelling on the car network are approximately the same as if only the car travelers moved on the original network (i.e. the network without up-scaling). This allows the calibration procedure to exert influence on the choice of plans of all agents, and not just car travelers as it was done previously (cf. Ziemke et al. (2015)). This involves the implicit assumption that spatial patterns of travelers in different modes of transport are similar, which is likely not the case. This initial inconsistency is, however, remedied in the subsequent modeling step (cf. section 3.5), where agents are allowed to change their modes of transport.

The calibration effect of Cadyts is incorporated into MATSim via the scoring of plans (cf. Nagel et al. (2016a) and Nagel et al. (2016b)). In simple terms, the standard MATSim scoring evaluates how much time agents spend performing their activities (according to their plan), which increases their score, and how much time they spend traveling and waiting in traffic jams, which decreases their score. Cadyts' calibration effect acts as an additional scoring component, which evaluates how well the plans (and the travel included in it), which agents perform during the simulation, match with real-world travel observations. If a plan is conducive to reproducing real-world traffic patterns, Cadyts will reward this plan with a positive score offset (as illustrated in figure 1), while a plan that produces trips that are not in line with real-world observations will receive a negative score offset. This offset should be interpreted as a plan-specific constant of the choice model, where Cadyts is the method to determine its value. As this procedure is integrated into MATSim's iterative transport demand adaptation process, travel demand becomes gradually more realistic. Figure 2 illustrates the relation between simulated and real-world traffic volumes averaged over all count stations.

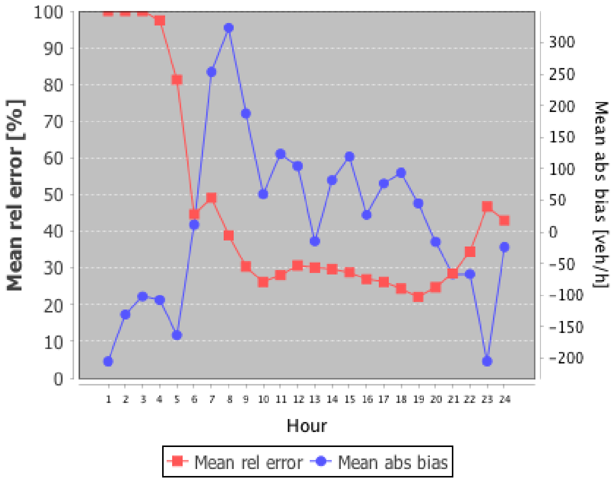


Figure 2: Evaluation of match of traffic counts

The following parameters are used in this modeling step:

- flowCapacityFactor = 0.47.
- storageCapacityFactor = 0.57.
- stuckTime = $30s$.
- countsScaleFactor = 3.2.
- Weight of strategies: ChangeExpBeta = 1.0, ReRoute = 0.2.
- fracOfIterationToDisableInnovation = 0.7.
- cadytsScoringWeight = 15.0.
- runId = be_203.

3.5 Mode Choice Calibration

After each agent has been assigned with locations by the previous step (cf. section 3.4), a mode choice calibration is carried out. By this process, agents can switch from the car to other modes of transport. Besides ”car”, these modes include ”pt”, ”ptSlow”, ”bicycle”, and ”walk”, and are associated with the following parameters:

- “Car” with alternative-specific constant (ASC) = −2.0 and monetary distance rate = $-0.00035 EUR/km$.
- “Pt” with ASC = −7.0 and teleported mode speed = $6.410 m/s$.
- “PtSlow” with ASC = −4.0 and teleported mode speed = $4.274 m/s$.
- “Bicycle” with ASC = −3.0 and teleported mode speed = $3.205 m/s$.
- “Walk” with ASC = 0.0 and teleported mode speed = $1.068 m/s$.

In the current scenario, all modes besides car are so-called teleported modes, i.e. trips made by these modes are handled by putting the agent to the location of the next activity with a time lag that corresponds with the distance from the previous activity and the specified speed of that mode. For ”walk”, for instance, a speed of $5.0 km/h$ is assumed. Accounting for the fact that in reality a tripmaker cannot walk on beeline between two activities and the real distance will be somewhat longer, speeds are divided by 1.3 leading to a speed for the ”walk” mode of $1.068 m/s$ ($= 5.0 km/h$ / $3.6 (km/h)/(m/s)$ / 1.3), which leads to the same result as multiplying the distance by a beeline factor of 1.3.

Two public-transport modes are used to account for the significant structural differences between urban public transport (e.g. bus, tram) and regional public transport (e.g. S-Bahn (commuter train), regional train). While the former has a lower travel speed and lower access costs (considered as one determinant or interpretation of the value of the mode-specific constant (ASC)), the latter has a higher travel speed, but also higher access costs.

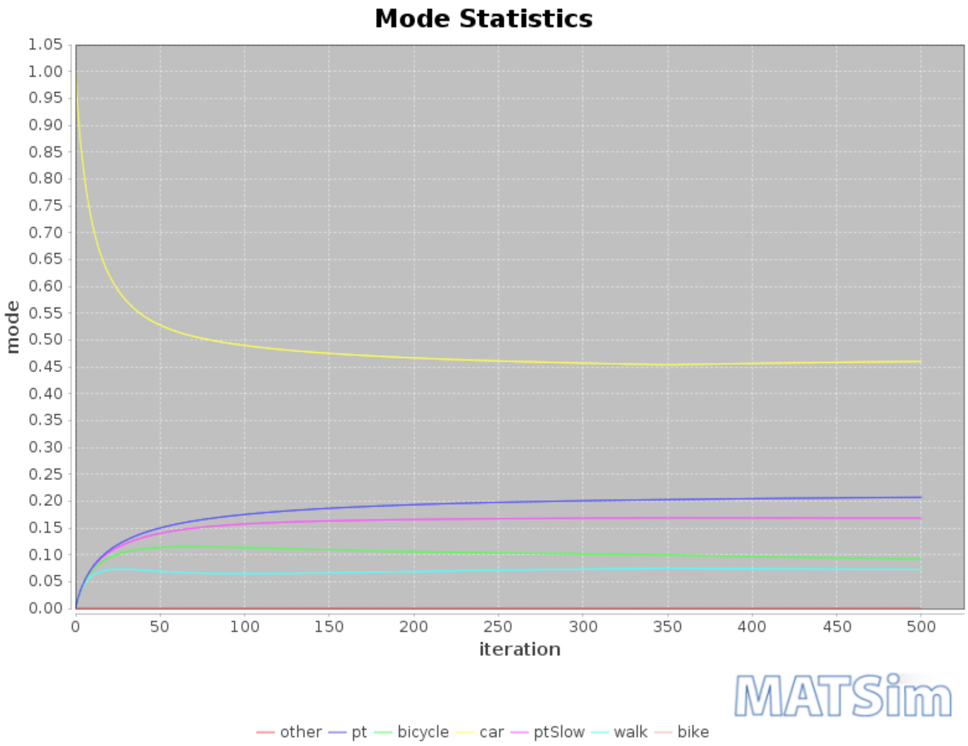


Figure 3: Development of choice of different modes.

Figure 3 illustrates how agents switch from car to other modes of transport during this modeling step. Note that in this modeling step, no (Cadyts) calibration is active. As such, this step must at the same time be regarded a stability test in the sense of Ziemke et al. (2015), i.e. a simulation run that confirms that choices made initially under the influence of Cadyts stay largely the same after this influence is removed. This is depicted in figure 4.

The following parameters were used in this modeling step:

- flowCapacityFactor = 0.12.
- storageCapacityFactor = 0.24.
- stuckTime = $30s$.
- Weights of strategies: ChangeExpBeta = 1.0, ReRoute = 0.2, ChangeTripMode = 0.2.
- fracOfIterationToDisableInnovation = 0.7.
- runId = be_251.

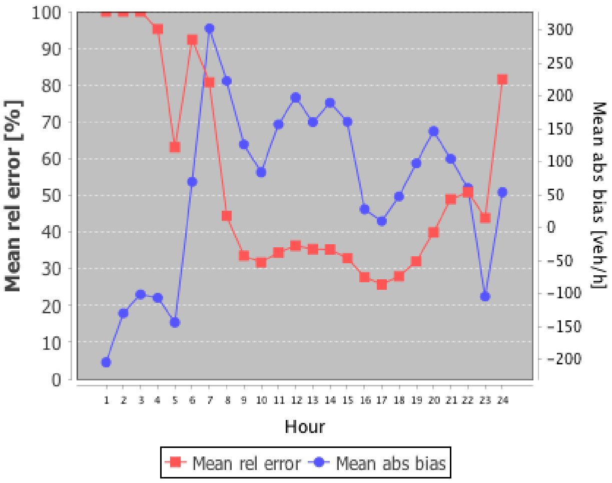
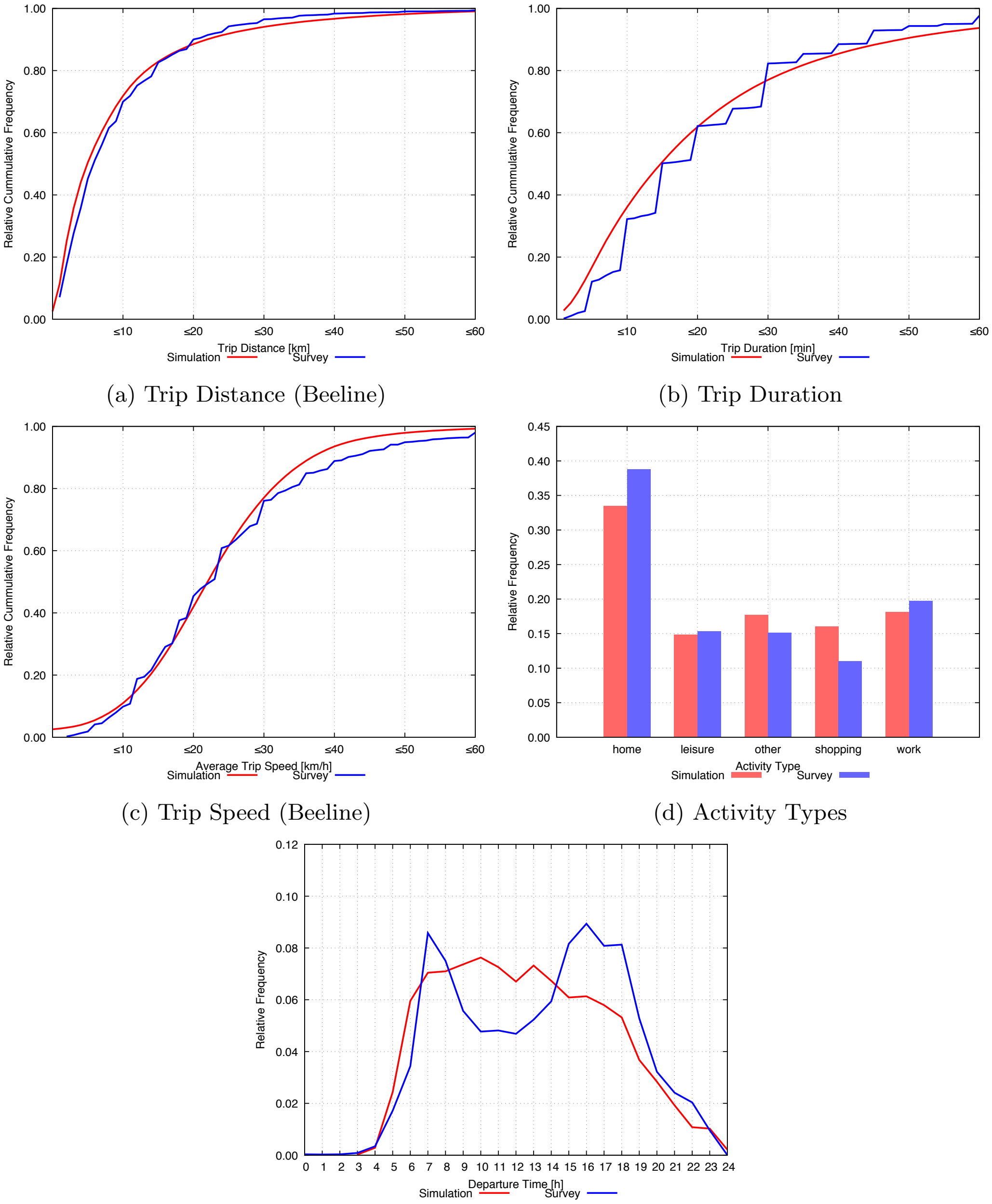


Figure 4: Evaluation of match of traffic counts.

4 Results

Figure 5 depicts properties of trips made by agents traveling by car compared to corresponding values from the SrV 2008 travel survey (Ahrens, 2009). For this analyses, only trips starting or ending in Berlin are taken into account to be comparable with the travel survey. Furthermore, only trips shorter than $100km$ are considered.



(a) Trip Distance (Beeline)

(b) Trip Duration

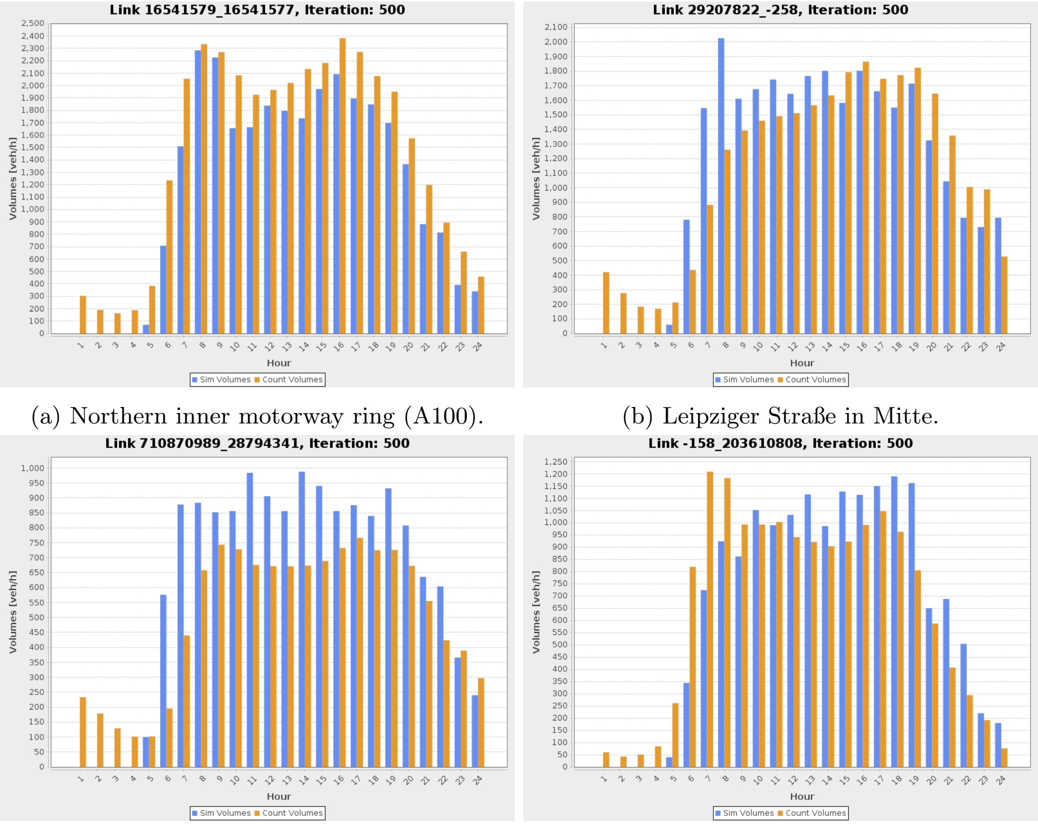
(c) Trip Speed (Beeline)

(d) Activity Types

(e) Departure Time

Figure 5: Car trips

Travel distances, durations, and, consequently, travel speeds match very well. Activities are also reproduced largely well, while the simulation produces more midday traffic than stated by the survey. These results are largely in line with the results of the car-traffic-only scenario that has been created in this study's predecessor (Ziemke et al., 2015). There, it was argued that the difference in midday traffic is likely a result of different demand segments between survey and traffic counts rather than a shortcoming of the calibration procedure. While the survey only contains personal traffic with the typical morning and afternoon peaks, the counts include heavy-goods traffic, which balances the midday drop of personal traffic to some extent. Accordingly, this difference would most likely vanish if counts were used that distinguish between personal and heavy-goods traffic. As explained in section 3.4, counts that contain such a distinction were not available for the present scenario. The steps in the survey graph of the trip duration diagram are a result of self-reported travel times in the survey as respondents tend to report "round" numbers rather than exact values by the minute.



(a) Northern inner motorway ring (A100).

(b) Leipziger Straße in Mitte.

(c) Schönhauser Allee in Prenzlauer Berg.

(d) Märkische Allee (L158) in Marzahn.

Figure 6: Comparison of simulated and counted traffic volumes of four individual count stations.

Figure 6 depicts a comparison between simulated and counted traffic volumes at four arbitrarily selected individual counts stations. While differences in the quality of the match for these four count stations are discernible, all four count stations have in common that they clearly do not show a marked midday drop as found in the survey data (cf. figure 5), reaffirming the above reasoning.

Figure 7 illustrates the overall match in simulated and counted traffic volumes as well as the spatial distribution of count stations in Berlin. The stations that form a line along the western and southern edge of the inner city trace route of the urban motorway A100.

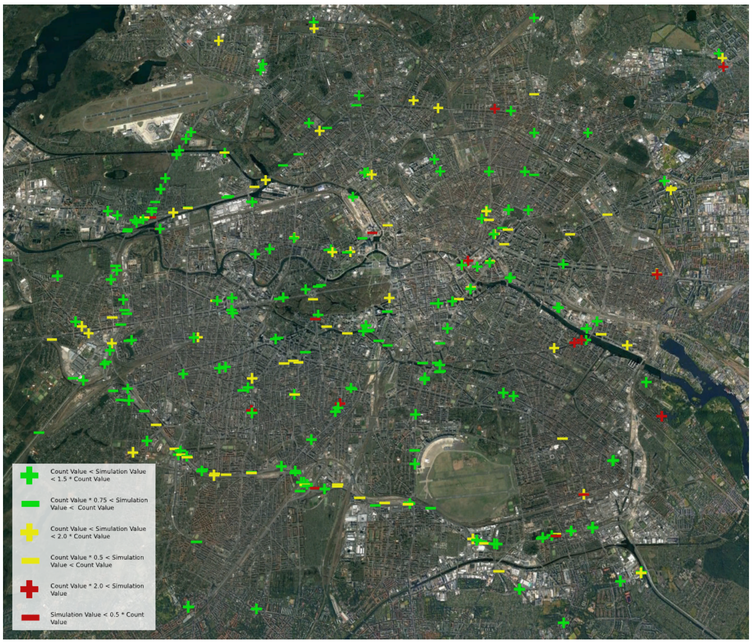
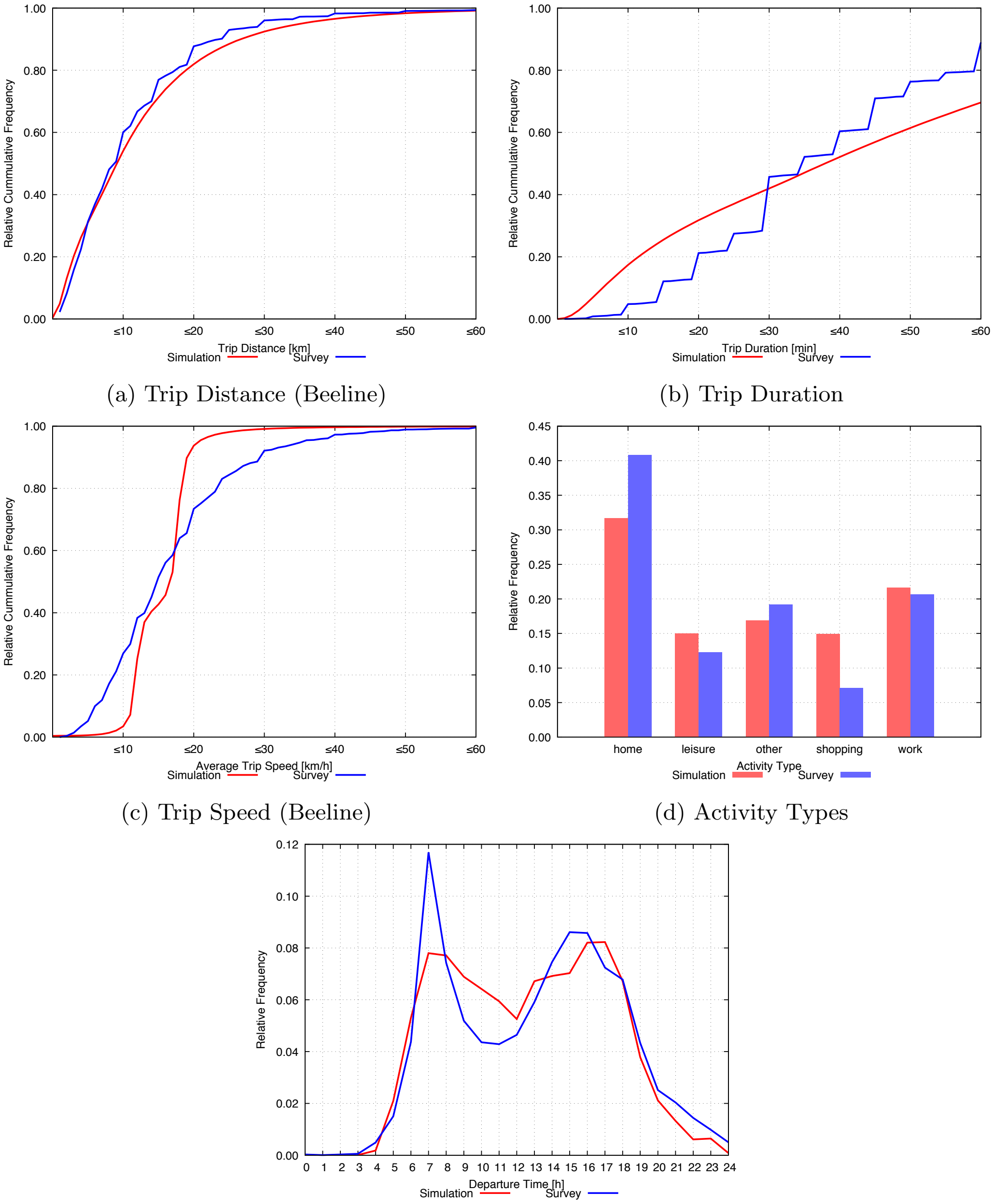

Count Value < Simulation Value < 1.5 * Count Value
Count Value * 0.75 < Simulation Value < Count Value
Count Value < Simulation Value < 2.0 * Count Value
Count Value * 0.5 < Simulation Value < Count Value
Count Value * 2.0 < Simulation Value
Simulation Value < 0.5 * Count Value

Figure 7: Comparison of simulated and counted traffic volumes (Background Map: Google Earth).

Figure 8 compares properties of trips made by agents traveling by public transport compared to corresponding values from the SrV 2008 travel survey. On the simulation side, trips made by both ”pt” and ”ptSlow” are considered, which is reflected by the inflection point in the trip speed diagram at a speed of approximately $15.0 km/h$.



(a) Trip Distance (Beeline)

(b) Trip Duration

(c) Trip Speed (Beeline)

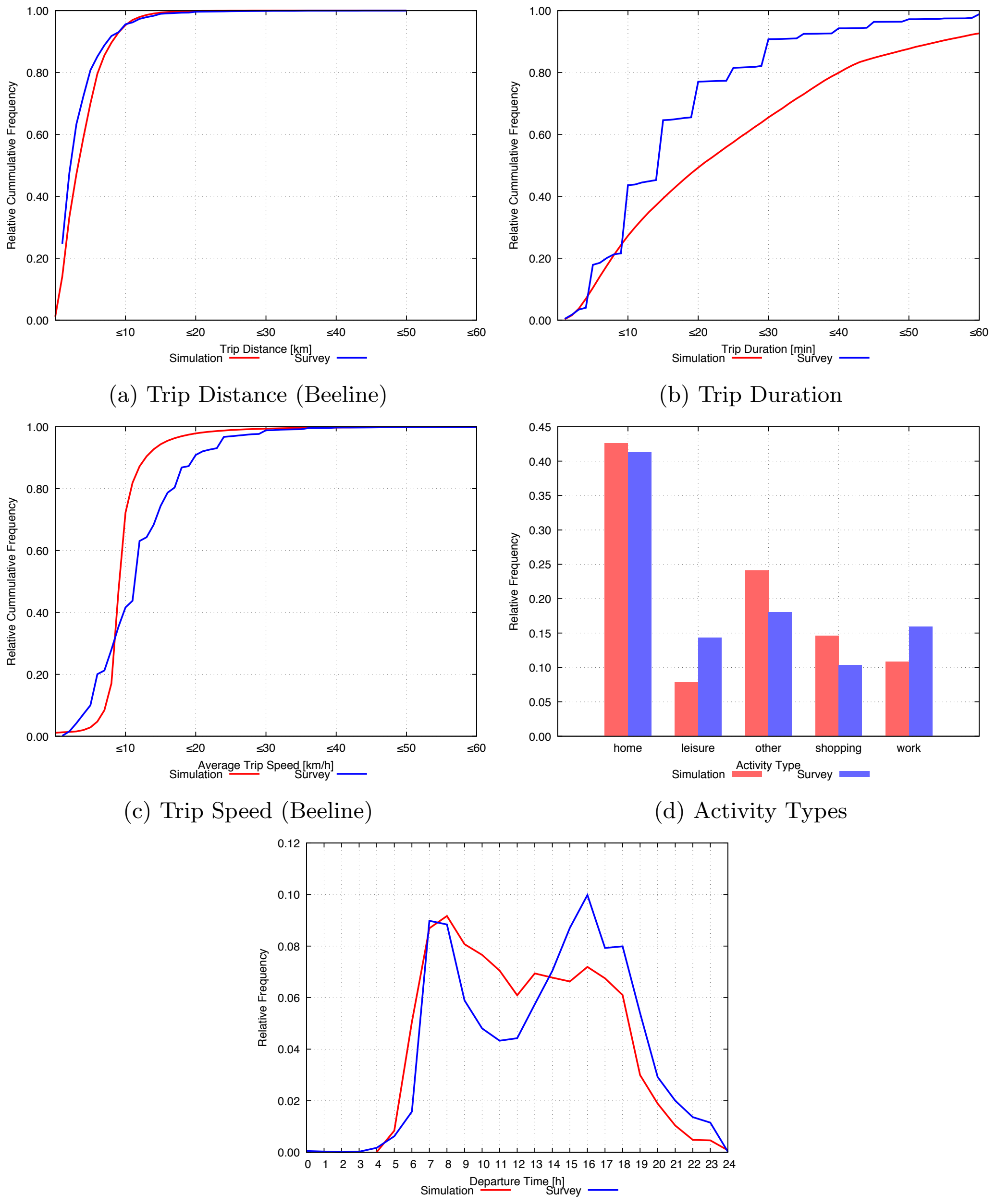
(d) Activity Types

(e) Departure Time

Figure 8: Public transport trips

Figure 9 depicts an analogous comparison for trips made by bicycle. It is visible that speeds have been chosen somewhat too low, which also leads to a mismatch in trip durations. An improvement to this shortcoming is underway.

A comparison of average values of trip distances and trip durations between simulation and survey for all modes of transport is given in table 1. While trip distances are generally met very well, trip durations for modes other than car tend to be somewhat overestimated. As already seen above, this is in particular true for the mode ”bicycle”.



(a) Trip Distance (Beeline)

(b) Trip Duration

(c) Trip Speed (Beeline)

(d) Activity Types

(e) Departure Time

Figure 9: Bicycle trips

Table 1: Comparison between simulation and SrV 2008 survey

Mode of Transport	**Trip Distance**		**Trip Duration**	
	Simulation	Survey	Simulation	Survey
Car	9.0km	9.6km	21.6min	22.3min
Public transport	12.2km	11.7km	47.5min	40.2min
Bicycle	3.9km	3.6km	25.7min	17.6min
Walk	1.0km	1.0km	19.8min	14.0min

5 Access to the Scenario

All input data to use the scenario and to perform own analyses are obtainable via `https://svn.vsp.tu-berlin.de/repos/public-svn/matsim/scenarios/countries/de/berlin/2017-07-20_car_pt_ptSlow_bicycle_walk_10pct/`. The MATSim software can be downloaded from `https://github.com/matsim-org/matsim/releases`.

6 Acknowledgment

The authors wish to thank Gabriel Thunig for carrying out parts of the required technical work and performing analyses. The rest of the VSP team, in particular Joschka Bischoff, Ihab Kaddoura, Amit Agarwal, and Theresa Thunig, provided feedback throughout the development of the scenario.

References

Gerd-Axel Ahrens. Endbericht zur Verkehrserhebung Mobilität in Städten – SrV 2008 in Berlin. Technical report, Technische Universität Dresden, 2009. http://www.stadtentwicklung.berlin.de/verkehr/politik_planung/zahlen_fakten/download/2_SrV_endbericht_tudresden_2008_berlin.pdf.

C.R. Bhat, J.Y. Guo, S. Srinivasan, and A. Sivakumar. A comprehensive econometric microsimulator for daily activity-travel patterns. Transportation Research Record, 1894:57–66, 2004. ISSN 0361-1981. doi: 10.3141/1894-07.

C.R. Bhat, J. Guo, S. Srinivasan, and A. Sivakumar. CEMDAP User's Manual. University of Texas at Austin, Austin, TX, USA, 3.1 edition, 2008.

H. Bömermann, S. Jahn, and K. Nelius. Lebensweltlich orientierte Räume im Regionalen Bezugssystem (Teil 1). Berliner Statistik, 8:366–371, 2006. http://www.stadtentwicklung.berlin.de/planen/basisdaten_stadtentwicklung/lor/download/BerlinerStatistik0608.pdf.

Bundesagentur für Arbeit. Pendlerstatistik 2010. CD-ROM, 2010.

G. Flötteröd. Cadyts – Calibration of dynamic traffic simulations – Version 1.1.0 manual. Transport and Mobility Laboratory, École Polytechnique Fédérale de Lausanne, November 2010. URL http://home.abe.kth.se/~gunnarfl/files/cadyts/Cadyts_manual_1-1-0.pdf.

A. Horni, K. Nagel, and K. W. Axhausen, editors. The Multi-Agent Transport Simulation MATSim. Ubiquity, London, 2016. doi: 10.5334/baw. URL http://matsim.org/the-book.

B. Kickhöfer, D. Hosse, K. Turner, and A. Tirachini. Creating an open MATSim scenario from open data: The case of Santiago de Chile. VSP Working Paper 16-02, TU Berlin, Transport Systems Planning and Transport Telematics, 2016. See http://www.vsp.tu-berlin.de/publications.

K. Nagel, B. Kickhöfer, A. Horni, and D. Charypar. A closer look at scoring. In Horni et al. (2016), chapter 3. doi: 10.5334/baw. URL http://matsim.org/the-book.

K. Nagel, M. Zilske, and G. Flötteröd. CaDyTS: Calibration of Dynamic Traffic Simulations. In Horni et al. (2016), chapter 32. doi: 10.5334/baw. URL http://matsim.org/the-book.

OpenStreetMap, accessed 30 Jan 2017. www.openstreetmap.org.

Statistische Ämter des Bundes und der Länder. Zensus 2011, accessed 21 July 2017. www.zensus2011.de.

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