Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 52 (2015) 468 – 475

Procedia Computer Science

6th International Conference on Ambient Systems, Networks and Technologies, ANT 2015

A simulation-based approach for constructing all-day travel chains from mobile phone data

Michael Zilske[a,*], Kai Nagel[a]

[a] *Technische Universität Berlin, Transport Systems Planning and Transport Telematics, Salzufer 17-19, 10587 Berlin, Germany*

Abstract

The purpose of this work is to investigate replacing travel diaries with sets of call detail records (CDRs) as inputs for an agent-oriented traffic simulation. We propose constructing an agent population directly from a CDR dataset and fusing it with link volume counts to reduce spatio-temporal uncertainty and correct for underrepresented traffic segments. The problem of finding a set of travel plans which realizes a set of CDR trajectories and is consistent with a set of link volume counts is rephrased in terms of calibrating a choice model. This enables us to make use of an existing calibration scheme for agent-oriented simulations. We demonstrate our approach by an illustrative scenario with synthetic data.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Peer-review under responsibility of the Conference Program Chairs

Keywords: Agent-oriented simulation ; demand generation ; mobile phone data

1. Introduction

Building a traffic model for a city or region takes considerable effort. A typical approach is the four step process, consisting of the trip generation (sources and sinks of trips), trip distribution (destination choice), mode choice, and (static) route assignment.[17]

Newer approaches combine *activity-based demand generation* (ABDG)[1,3,6,18,21] with *dynamic traffic assignment* (DTA).[9,20,23] The output of the ABDG, and therefore the input to the DTA, can be hourly origin-destination (OD) matrices, a list of trips, or full daily activity chains, where each activity comes with a location, an end time, and the mode of transport that leads to the next activity.

Typical direct inputs to these modelling suites are digital road network data and digital land use data; other data, in particular revealed and/or stated choice data, are used to calibrate the behavioral models. All these data have been difficult to procure in the past. Fortunately, things are changing: Collaborative mapping projects such as OpenStreetMap (OSM) provide digital road network data whose completeness and accuracy is approaching or, occasionally, surpassing commercial sources[16]. For land use information,

* Corresponding author. Tel.: +49-30-314-28666 ; fax: +49-30-314-26269.
E-mail address: michael.zilske@tu-berlin.de

1877-0509 © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2015.05.017

OSM can already be considered a viable alternative data source in some locations. Globally, it is promising because it facilitates manual mapping of land use information from earth observation data by volunteers[2]. What is still more difficult to get is the demand data. Here, it has become clear over the past years that automatically and passively collected data will play an increasing role: Instead of taking a survey and calibrating trip generation/distribution or an activity based model with it, it seems attractive to use, for instance, mobile phone data, which often come in the form of call detail records (CDR).[8,7]

Inferring OD-matrices[13] or activity-trip-chains with activity types and possibly modes of transport[22] from CDRs are large research topics in their own right. Once created, such matrices or chains can be used as input to a traffic assignment model.[5] However, this initial modelling step typically takes several person-months of work, and often additional context-dependent information is invested, such as typical daily routines or activity opportunities in the area of interest. Considering this, it seems attractive to use the CDRs *directly* to drive a traffic model, and only then use additional data to make the model more realistic, if required.

In a first step, CDR datasets can be used directly as travel chains for a synthetic population. The CDRs of a single person are assigned to a simulated individual, which visits all call locations in order, conducting unlabelled dummy activities so that, at each point in time, the simulated location corresponds to the observed location. It need not even be attempted to distinguish between actual activity locations and transient locations passed while travelling. The latter turn up as very short activities, serving as sampling points for route generation.

This simple approach can already be useful, since it allows to identify bottlenecks in the simulated transport system, which may either also exist in reality and thus point to possible infrastructure improvements, or they may be artifacts pointing to problems with the road network data. In previous work[25], this approach was implemented using the MATSim transport microsimulation. Experimenting with a synthetic population with realistic activity-travel-patterns at different daily call rates, it was found that with 50 calls per day (uniformly), the resulting travel is underestimated by less than 5%, while 5 calls per day are clearly insufficient in that this leads to an under-estimation of travel by more than half.

In a second step, the model can then be enriched with additional data. A typical approach is to use the remaining freedom in a behavioral model to generate multiple alternatives, and then use the additional data to select between these alternatives. In particular, anonymous traffic counts have been used to select between multiple car routes,[11] between multiple public transit routes,[14] or between multiple daily activity patterns.[24]

Similarly, traffic counts can be used to enrich a model built from CDR data that are partially sparse in time. Illustrative of the general problem, this paper investigates a synthetic scenario where half of the population makes 50 calls per day while the other half makes only 5 calls per day. The temporal uncertainty in the sparse half of the data is used to direct the model to match the traffic counts, using the Cadyts calibration scheme[10]. In addition, the counts are used to compensate for the overall lost mileage by scaling up the overall demand, still making use of the spatial information that is contained in the sparse data. In order to address a likely problem with using mobile phone data directly, it is assumed that different calling behavior is associated with different activity behavior.

The rest of this paper is organized as follows: First, MATSim and its interaction with Cadyts are described, and how the two models together can be used to scale and reweigh an initial set of travel plans using link counts. Given this framework, we then discuss replacing travel plans with CDRs as the initial demand specification. Our experimental scenario is derived from a full activity-oriented assignment model for Berlin. We demonstrate in how far two segments which differ both in terms of travel behavior and in terms of calling behavior can be fused into a correct estimate of traffic state over time. The paper is concluded by a discussion and a summary.

2. MATSim and Cadyts

2.1. MATSim

MATSim combines a traffic demand model based on individual daily travel plans with a microscopic traffic flow simulation to iteratively calculate a dynamic user equilibrium. Its demand model consists of a population of agents

$$A_1, \ldots, A_N \tag{1}$$

Each agent has a mutable set of plans which can be understood as a choice set. The options are identical in the fixed dimensions (typically, the chain of activities with type and location), and vary in the open dimensions (typically, routes, modes of transport, and departure times). Every plan is assigned a mutable score, V_i, initialized to $+\infty$. Often, the score can be interpreted as utility.

Initial plans are auto-completed by the simulation as much as possible; for example, links are assigned to coordinates, and shortest path routes are computed if no routes are in the initial plans. Then, the following steps are iterated:

- Each agent chooses from its plan set according to a random utility model, where the choice distribution follows $P(i) = \exp(V_i)/\sum_j \exp(V_j)$.
- The chosen plans are loaded onto the network.
- For every chosen plan, V_i is re-calculated as a function of the plan's performance during the network loading (e.g. valuing travel time negatively) and assigned to that plan.
- Each agent in a random subset of the population adds a new plan to its plan set (identical to its other plans in the fixed choice dimensions, and distinct in the open dimensions) and removing an existing one if its plan set is now greater than a specified maximum.

The simulation is run until the variables on which the utility perception depends (e.g. dynamic link travel times) have converged to a steady state, and hence the choice distribution has become stationary. At that point, plan set mutation is ceased, so that the choice distribution now strictly follows the perceived utilities, and the simulation is continued until it converges a second time.

2.2. Cadyts

Cadyts is a calibration scheme which, when applied to MATSim and a vector of link traffic counts y, works by directing the plan choice probabilities of the whole agent population towards choices more consistent with the counts. This is achieved by calculating an offset to the score V_i of each chosen plan, iteration by iteration. Under certain additional assumptions, e.g. about the error distribution of the measurements, the adjusted choice distribution can be shown to approximate the posterior choice distribution given y[12,10]. It follows

$$P(i|y) = \frac{\exp\left(V_i + \sum_{ak\sim i} \frac{y_{ak}-q_{ak}}{\sigma^2_{ak}}\right)}{\sum_j \exp\left(V_j + \sum_{ak\sim j} \frac{y_{ak}-q_{ak}}{\sigma^2_{ak}}\right)} \tag{2}$$

where y_{ak} is the traffic count measurement on link a in time interval k, σ^2_{ak} is that measurement's error variance, and q_{ak} is the simulated value corresponding to that measurement. The condition $ak \sim i$ denotes that following plan i crosses link a in time window k.

Intuitively, the offset is calculated based on how much this choice of the plan contributes to the whole traffic system fitting to the traffic counts. Plans which traverse links where flow is underestimated are favored and vice versa, and σ denotes the trust level that is put into the measurement – high trust levels lead to small values of σ and thus to large correction terms.

This calibration can be seen as reducing uncertainty about travel behavior in the open choice dimensions, but it can also be applied to estimate the population size[12] if each agent is given an additional, synthetic plan to do nothing, disappearing from the scenario.

3. From call detail records to a population of agents

A CDR dataset consists of records of the form

$$T_n := [(p_n, t_1, c_1), \ldots, (p_n, t_K, c_K)] \tag{3}$$

where p_n is a person identifier, t_k are timestamps, and c_k are cell tower identifiers.

Each trace T_n is converted into a travel plan in a straightforward way: Calls are converted into activities. Several calls in the same cell without a call in a different cell between them are fused, that is, they are converted into a single activity that starts no later than the first call and ends no earlier than the last call in the same cell. No additional activities are added. Activities are connected by trips (only the car mode is considered here). Congestion is disregarded. It is assumed that fastest routes on the empty network are taken. The only degree of freedom under consideration is the departure time from each activity location, which can be chosen anywhere between the time of the last sighting at location i and the latest possible departure time to make it to the next sighting location $i + 1$ in time.

The full agent population is constructed by expanding the population generated from traces. Specifically, we create C agents $A_{n1}, \ldots, A_{nC}$ per trace T_n. The agents are initially equipped with one random realization of the trace T_n, and over the iterations (cf. section 2.1), they create new random realizations, varying in time structure. In addition, they are given a special plan which, if chosen, lets them stay at home. Agents choosing the stay-at-home option are considered to be removing themselves from the simulation.

The resulting agent population is

$$A_{11}, \ldots, A_{1C}, \ldots, A_{N1}, \ldots A_{NC} \tag{4}$$

This expanded population is used as a buffer, which the calibrator uses to steer the demand towards matching the known link volume counts. The utility function is constructed so that, for each agent, the probability of choosing one of its travel plans is $p^0_{nc} = 1/C$, and the probability of choosing the stay-at-home-plan is $1 - p^0_{nc}$. In consequence, the prior expected behavior of the simulation is that the population size is N, and on average one instance of each trace is realized.

By calculating offsets to this prior utility of plans, the calibrator simultaneously adjusts the population size, the weights assigned to the individual traces, and the temporal realization of the trajectories.

This results in a distribution of individual choices among possible trajectories and stay-at-home plans. In particular, we obtain posterior travel probabilities p_{nc}. The sum over the posterior travel probabilities of the agents associated with trace T_n, $w_n = \sum_{c=1}^{C} p_{nc}$, is the expected number of instances of trace T_n to appear in any iteration of the calibrated scenario after achieving stationarity, and $(w_1, \ldots, w_N)$ is a weight vector with which the CDR dataset has effectively been resampled, a common concept in synthetic population generation, where a survey population is adjusted to fit exogeneously given marginal sums[4,15], whose role is in the present case assumed by the traffic counts.

The population expansion described here is a particularly straight-forward way of implementing uncertainty about the CDR sample in the MATSim-Cadyts-ensemble, because it reduces the estimation of weights, as well as which temporal realization of a CDR trace to use, to individual agent decisions.

The expansion factor C is selected by the modeller. It needs to be large if highly underestimated demand segments are to be compensated for, so that there is a sufficient number of individuals in the population to draw from.

4. Experiment

4.1. Synthetic CDRs

In order to have full control over the ground truth, for the present study the CDR data is – as in the preceding study[25] – synthetically generated from a simulated scenario. A full implementation of MATSim is used as a synthetic ground-truth scenario. The output of this model is a set of complete descriptions of mobility behavior of an agent population with labeled activities and space-time trajectories on the level of network links. Note that additional kinds of measurements can be taken from this output, in particular link traffic counts.

A simple phone usage model is applied to the synthetic population: In every timestep, every agent gets to decide whether or not to make a phone call. When a phone call is made, the framework locates the agent and records a CDR. The first output of this step is a set of CDRs as specified in equation 3. The second output is a set of link traffic counts y_{ak}, the number of vehicles which have passed link a in time window k.

This is considered the available data for traffic modeling in the hypothetical scenario, and simulation runs are based only on this data.

The output of each iteration of the simulation is of the same form as the ground truth scenario. Any of its properties can be compared to the ground truth scenario to assess the approximation quality. In fact, since every iteration is a draw from the combined choice distributions of all agents, properties of the full statistical distribution of these draws can be used to compare with the ground truth.

This framework allows studying this and other methods for constructing demand models from CDRs, and how much information from CDRs and link traffic counts is needed to re-approximate the state of the traffic system over time in the ground truth scenario to which degree. It isolates these questions from the different question of how good the traffic simulation model itself is at approximating reality.

4.2. Scenario description

The experimental scenario is created from a 1998 household survey which contains complete trip diaries from one specific day of 2% of the Berlin population. The survey is not publicly available, but has been used before[19,14]. It contains activity locations, activity types, activity start and end times, and modes of transport for each trip. It does not contain any route information. For the present study, only individuals who only travel by car are considered, which produces 18 377 individuals. The network contains 61 920 links, of which a random 5% are chosen to collect volume counts in hourly time windows. Disregarding the spatial uncertainty of sightings, each link is associated with its own phone cell. We also disregard capacity constraints in the traffic network, i.e. for the present study there is no traffic congestion. Every agent chooses fastest routes with respect to free-speed travel time.

Agents place calls randomly at an individual daily call rate. Deliberately constructing a strong correlation between phone usage and travel behavior, we partition the agent population into two segments called workers and non-workers, where a worker is defined as an individual stating at least one work-related activity in the survey. The call rate of the workers is fixed at 50 calls per day (frequent callers), and that of the non-workers at 5 calls per day (infrequent callers).

Some travel chains obtained by the process are shown in Fig. 1a, while the true underlying behavior is shown in Fig. 1b. The orange plan contains a work activity, thus corresponding to a frequent caller. While the activity chain alone gives the traveller the freedom of many routes around and through the city, the sightings effectively pin one of the trips to the northern route. The two plans in blue do not contain a work activity, and are in consequence not sampled frequently. Several activities and related travel are missed. In fact, the light blue trace does not even result in a round trip any more.

4.3. Results

With any mobile phone data set in hand, the modeller has to decide on a threshold how many calls per day are necessary for a trace so that it can be meaningfully included in the model input.

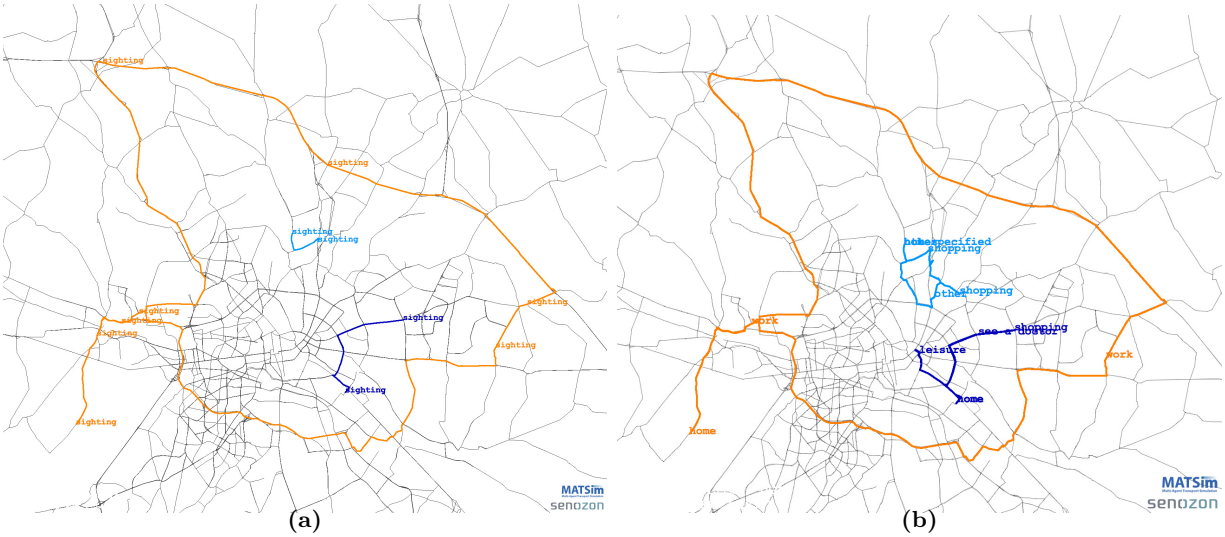


Fig. 1: Travel chains for three different travellers (a) and the underlying ground truth behavior (b).

Using the binary-distributed synthetic data, we compare two options:

- Leave the sparse traces out of the simulation. This effectively means accepting a lower sample size and possibly introducing a bias towards a traffic pattern associated with frequent callers.
- Include the sparse traces even though their spatio-temporal resolution is such that they contain only limited information.

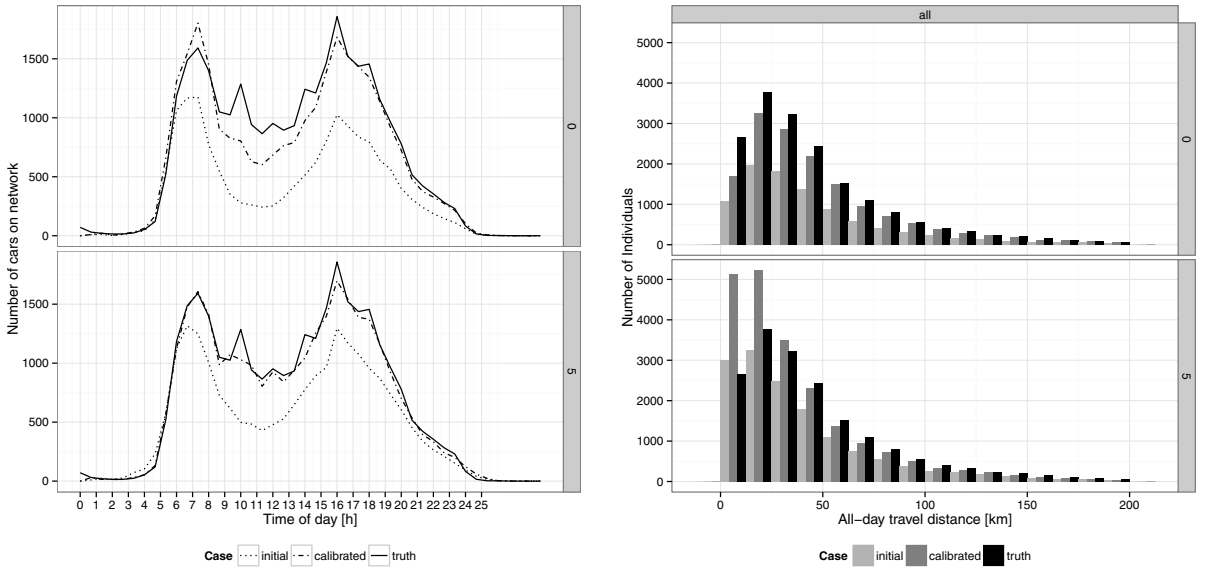
Fig. 2a shows network load over time for the initial situation where the population constructed from the available traces is simulated without adjusted weights, for the final estimation where the weights are adjusted towards fitting the link counts, and for the ground truth.

The first scenario shows the full effect of removing non-workers from the sample. In the initial estimation, there is too little traffic, but especially the load during mid-day is too small. In the final estimation, this gap is partly compensated for. In turn, the morning peak is overestimated, because there are only well-sampled traces of workers, which are mostly morning commuters, to draw from: In order to reduce the underprediction of mid-day load, the morning peak load has to be overestimated.

In the scenario where the traces of the non-workers, sampled at a low rate, are included, the final estimation has a closer fit to the ground truth (Fig. 2a bottom). In the initial estimation, the demand share generated from the undersampled non-worker traces is not only too low, but diffused over time: Possible trajectories through few sightings have more temporal freedom than those through many sightings. In the final estimation, while still too low, its time structure more closely resembles the ground truth: The temporal uncertainty of the CDR data is reduced by taking the link counts into account. Intuitively, the sparsely sampled trajectories are fitted to that share of the measured volumes which is not accounted for by well-sampled trajectories. The overall final demand estimation is better because it now contains this time-adjusted non-worker demand as a component.

Considering the all-day travel distance distribution (Fig. 2b) reveals that it is distorted in both cases.

In the first scenario, where the infrequent callers are excluded, the number of individuals travelling little is underestimated. There are at least two independent causes for this. The first is that workers travel more than non-workers, and traces of non-workers are missing by construction. Secondly, the estimation process itself is in this case biased towards far-travelling individuals: When the initial demand is too low overall, the contribution of most links to the Cadyts score correction (equation 2) is positive, so the utility offset of a plan is the larger the more links it crosses. In consequence, far-travelling agents will on average end up



(a) Network load over time of day

(b) All-day travel distance distribution

Fig. 2: Network load and travel distance distribution for the scenario where the non-worker demand segment is missing (top) or represented by undersampled trajectories (bottom).

with a higher probability of travelling. This effect is absent when the initial demand is a priori scaled to the known change in sample size. But an alternative interpretation of this experiment is that a population segment is missing from the sample altogether, without this fact or indeed the true size of the travelling population being known to the modeller, so the initial demand was left unchanged here.

In the second scenario, where the infrequent callers are now included, the number of individuals travelling little is overestimated. Since the initial overall travelled distance is much closer to the truth, the calibration signal and hence the bias towards longer trips introduced by the plan correction is not as strong. It is dominated by an effect in the opposite direction which is created by the plan creation itself: Since the travelled distance of each plan is by construction at the lower bound of what is consistent with the sightings, the distance distribution is shifted to the left.

5. Summary

We studied the problem of expanding a CDR dataset to a set of all-day travel chains by data fusion with link traffic counts. The proposed method is to construct, for each individual CDR trace, multiple chains which are consistent with the observations, and then to calibrate the vector of probabilities of each chain being in the final set using the link traffic counts.

Our experimental scenario illustrates two cases:

- When a large population segment is missing or removed from the CDR sample because of its low daily call rate, the remaining sample is scaled up and reweighed in the process to fit link counts.
- When the same population segment is kept in the sample, represented by sparse traces generated by only 5 calls per day, the process is able to reduce the resulting temporal diffusion by producing trajectories which are more consistent with the traffic counts. This case yields a better fit to the real traffic flow.

Overall, the results demonstrate that even a heavily biased cell phone dataset, together with anonymous traffic measurements, can be used to re-construct the traffic state over time quite well. Any algorithm which attaches behavioral interpretation to a CDR trace can be used in the plan generation step to enrich the model.

References

1. Arentze, T., and Timmermans, H., Eds. *ALBATROSS–Version 2.0 – A learning based transportation oriented simulation system*. EIRASS (European Institute of Retailing and Services Studies), TU Eindhoven, NL, 2005.
2. Arsanjani, J. J., Mooney, P., Zipf, A., and Schauss, A. Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. Tech. rep.
3. Axhausen, K. W. *Eine ereignisorientierte Simulation von Aktivitätenketten zur Parkstandswahl*. PhD thesis, Universität Karlsruhe, Germany, 1988.
4. Bar-Gera, H., Konduri, K. C., Sana, B., Ye, X., and Pendyala, R. M. Estimating survey weights with multiple constraints using entropy optimization methods. Tech. Rep. 09-1354, Transportation Research Board, Washington D.C., 2009.
5. Bekhor, S., Dobler, C., and Axhausen, K. W. Integration of activity-based and agent-based models. *Transportation Research Record: Journal of the Transportation Research Board 2255* (Dec 2011), 38–47.
6. Bhat, C., Guo, J., Srinivasan, S., and Sivakumar, A. *CEMDAP User's Manual*, 3.1 ed. University of Texas at Austin, Austin, TX, USA, 2008.
7. Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing 10*, 4 (Oct. 2011), 36–44.
8. Chen, C., Bian, L., and Ma, J. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies 46*, 0 (2014), 326 – 337.
9. Chiu, Y.-C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T., and Hicks, J. A primer for dynamic traffic assignment. Transportation Research Circular E-C153, Transportation Research Board, 2011.
10. Flötteröd, G., Bierlaire, M., and Nagel, K. Bayesian demand calibration for dynamic traffic simulations. *Transportation Science 45*, 4 (2011), 541–561.
11. Flötteröd, G., Chen, Y., and Nagel, K. Behavioral calibration and analysis of a large-scale travel microsimulation. *Networks and Spatial Economics 12*, 4 (2011), 481–502.
12. Flötteröd, G., and Liu, R. Disaggregate path flow estimation in an iterated DTA microsimulation. Tech. rep., 2010.
13. Iqbal, M. S., Choudhury, C. F., Wang, P., and González, M. C. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies 40*, 0 (2014), 63 – 74.
14. Moyo Oliveros, M., and Nagel, K. Automatic calibration of agent-based public transit assignment path choice to count data. In *Conference on Agent-Based Modeling in Transportation Planning and Operations* (Blacksburg, Virginia, USA, 2013). Also VSP WP 13-13, see http://www.vsp.tu-berlin.de/publications.
15. Müller, K., and Axhausen, K. W. *Population synthesis for microsimulation: State of the art*. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT), 2010.
16. Neis, P., and Zielstra, D. Recent developments and future trends in volunteered geographic information research: The case of openstreetmap. *Future Internet 6*, 1 (2014), 76–106.
17. Ortúzar, J. d. D., and Willumsen, L. G. *Modelling Transport*, 4th ed. Wiley, 2011.
18. Pendyala, R. Phased implementation of a multimodal activity-based travel demand modeling system in Florida. volume II: FAMOS users guide. Research report, Florida Department of Transportation, Tallahassee, 2004. See www.eng.usf.edu/~pendyala/publications.
19. Scheiner, J. Daily mobility in Berlin: On 'inner unity' and the explanation of travel behaviour. *European Journal of Transport and Infrastructure Research 5* (2005), 159–186.
20. Tampére, Ch. und Viti, F. Dynamic traffic assignment under equilibrium and non-equilibrium: Do we need a paradigm shift?
21. Vovsha, P., and Bradley, M. Advanced activity-based models in context of planning decisions. *Transportation Research Record 1981* (2006), 34–41.
22. Wang, H., Calabrese, F., Di Lorenzo, G., and Ratti, C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on* (2010), IEEE, pp. 318–323.
23. Zhou, X., Mahmassani, H., and Zhang, K. Dynamic micro-assignment modeling approach for integrated multimodal urban corridor management. *Transportation Research Part C 16*, 2 (2007), 167–186.
24. Ziemke, D., Nagel, K., and Bhat, C. Integrating CEMDAP and MATSim to increase the transferability of transport demand models. Annual Meeting Preprint 15-5516, Transportation Research Board, Washington D.C., 2015. Also VSP WP 14-15, see http://www.vsp.tu-berlin.de/publications.
25. Zilske, M., and Nagel, K. Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science 32* (2014), 802–807.