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

Michael Zilske, Kai Nagel
Technische Universität Berlin

October 26, 2014

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 illustrative scenarios with synthetic data.

1 Introduction

Traffic simulations build a virtual model for the traffic system. These models can reach from simple sketch planning tools to highly complex simulation systems. One class of complex simulation systems are microscopic simulation models, where all elements of the simulation such as travelers, vehicles, links, intersections, signals, etc. are resolved individually.

Such simulation systems have two important inputs: A description of the transport network (sometimes called the supply side), and a description of the demand. The traditional demand description is a – possibly time-dependent – origin-destination (OD) matrix. Some approaches use trip tables, i.e. lists of triples, each one consisting of starting time, starting location, and destination location (e.g. DynusT, 2014). Yet again others – and these are the ones that will be considered in this paper – use full daily travel plans.

The arguably most straightforward way to generate initial daily travel plans is to take them from a trip diary survey. Trip diaries typically record, for a given day, all trips of a specific individual, including locations, starting and ending times, modes of transport, and purposes of the activities between the trips. This can be used directly as an input for simulations which are based on travel plans. The process will typically look as follows (Balmer et al., 2006):

1. Convert all trip information of each person into a record of the following type:

Listing 1: Structure of a plan

```
<plan>
   <activity type="home" location="..." endTime="07:00" />
   <leg mode="car" />
   <activity type="work" ... />
   ...
</plan>
```

The `location` can either be given as a coordinate or as a reference to a network link.

2. Have the simulation system fill in routes, e.g. routes that are fastest in the empty network.

3. Perform a network loading by executing all plans simultaneously in a microscopic traffic flow simulation.

This gives an initial set of plans together with an initial network loading, from where on the simulation can iteratively evolve. A **choice set** of alternative plans is dynamically generated over the iterations, each one by mutating a previous plan in one or more choice dimensions, such as departure time or route. In consequence, all aspects of the initial plan that are *not* modified over the iterations need to be realistic from the start. Agents then perform a **choice** between their plans, typically according to a logit model. The initial plan, the free choice dimensions, and the parameters of the utility model which determines the choice probabilities together describe the choice distribution.

However, trip diaries are not always available. In such a situation, one can, for example, use behavioral models to generate initial plans (e.g. Kitamura, 1988; Bowman et al., 1998; Pendyala, 2004; Arentze and Timmermans, 2005; Vovsha and Bradley, 2006; Bhat et al., 2008; Balmer, 2007; Ziemke et al., 2014), or derive them from trip based models (e.g. Balmer et al., 2005; Neumann et al., 2014). An alternative approach, in line with "big data" or "smart city" considerations, is to use cell phone datasets, in particular call detail records (CDRs).

Many investigations have used cell phone data in studies of human mobility (e.g. González et al., 2008; Candia et al., 2008). A frequent approach is to estimate origin/destination flows (Iqbal et al., 2014; Calabrese et al., 2011; Gur et al., 2009), but it is also promising to reconstruct locations, activity types, and transport modes from the data (Dash et al., 2014; Wang et al., 2010; Chen et al., 2014), i.e. to estimate a set of annotated trajectories from a set of raw phone traces. A driver for such an approach is that the result can be used to replace in part the traditional trip diary survey, thus either saving money or extending the sample size. The resulting activity plans can then be used in the same way as the traditional trip diaries as input to a travel plan based simulation.

This two-step method is, however, not the only possible approach to the problem of initial plan generation from CDRs. In particular, the reconstruction of locations, activity types and transport modes in general comes with uncertainties, implying that the constructed activity chains are not the only ones consistent with the data. Furthermore, calling behavior varies among individuals, and may correlate with movement behavior (Wesolowski et al., 2013). This indicates that it may be more appropriate to carry these uncertainties into the downstream processes, for example by constructing multiple activity chains which are all consistent with a CDR trace.

Zilske and Nagel (2013) investigate an early version of such an approach, where it was simply assumed that callers leave an activity location exactly at the time when the last call at some location occurs, and travel directly to the location where the next call is registered. The approach is attractive, since one can build a traffic model based on travel chains using easily available road network data (e.g. from OpenStreetMap) together with CDRs, which are also easily available in certain situations. In particular, the approach promises to build initial chain-based models in areas where no other data is available, e.g. in developing countries.

However, for that investigation no additional data to either verify nor further calibrate the approach was available. For verification, Zilske and Nagel (2014) take a calibrated activity-oriented traffic model of the Berlin region, extract synthetic CDR data under various assumed calling patterns, and investigate the difference between the resulting synthetic traffic and the ground truth. The main result is that even under generous assumptions about the frequency of calls and even assuming a full sample, this "lower bound" approach loses so much car mileage that it must be compensated for.

The present paper investigates in how far additional data, here in the form of anonymous traffic counts, can be used to bring such a simulation closer to reality. The motivation is that anonymous traffic counts either already exist or are fairly easy to procure even in adverse situations.

The approach here will be based on the MATSim transport microsimulation and the Cadyts calibration scheme (Flötteröd, 2009; Flötteröd et al., 2011). The rest of this paper is organized as follows: First, MATSim is introduced. Then Cadyts and its interaction with MATSim is described, and how the two models together can be used to scale and reweigh an initial set of travel plans using link travel counts. Given this framework, we then discuss replacing travel plans with CDRs as the initial demand specification. Two scenarios are used to generate results: one is a simple illustrative loop scenario, and one is derived from a full activity-oriented assignment model for Berlin. The experimental studies are concerned in particular with the question 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 (Flötteröd and Liu, 2010; Flötteröd et al., 2011). 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 behavior in the open choice dimensions, but it can also be applied to adjust overall travel demand (Flötteröd and Liu, 2010), 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. Fig. 1 shows some examples of the spatial information that is available at this point.

It is now assumed that this is the only available data for initial demand generation.

For the present study, 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

Figure 1: Sightings for three different travellers.

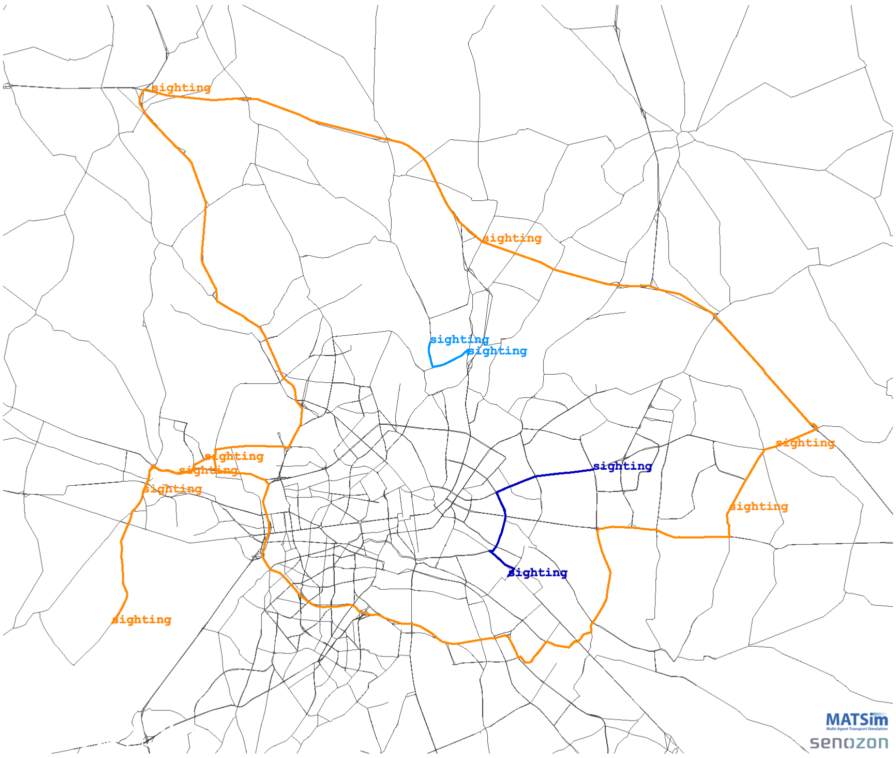

Figure 2: Initial plans for three different travellers.

on the empty network are taken. The only degree of freedom considered here 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.

A spatial visualization of the result can be found in Fig. 2.

Structurally, the plan at this point looks like

Listing 2: Structure of a plan derived from a phone trace

```
<plan>
   <activity type="sighting" location="..." endTime="..." />
   <leg mode="car"/>
   <activity type="sighting" ... />
   ...
</plan>
```

Compared to Listing 1, there are the following differences:

- There are no activity types (`"sighting"` is used as a generic label).
- The activity end time is randomly drawn within the time constraints.
- Sightings recorded while travelling will result in additional activities on the way.
- The location corresponds to the phone cell. At this point, phone cells are identified with links.

Clearly, this is not a behavioral plan, but rather a possible trajectory generated from a phone trace. In the same vein, its activities are rather waypoints, which can, but do not have to, be annotated with true behavioral activity types.

The term plan is not used here to denote a behavioral concept, but the same as the genotype in evolutionary computation, and as such it just needs to contain a description that can be interpreted by the downstream modules (Goldberg, 1989; Russel and Norvig, 2010), in this case by the traffic simulation.

In the following, the terms “activity” and “plan” will be used like this.

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 a 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

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

The 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 (e.g. Bar-Gera et al. (2009), for a survey see Müller and Axhausen (2010)), 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 Experiments

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 (Zilske and Nagel, 2014) – 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.

For this work, a plug-in for MATSim was developed for the purpose of obtaining synthetic CDRs from such a scenario. The software takes two additional inputs:

- A cell coverage, which partitions the simulated geographic area into mobile phone cells.
- A mobile phone usage model. The software exploits the benefits of an agent-oriented simulation framework, allowing for different population segments with different calling habits.

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 within the cell coverage, 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 Illustrative loop scenario

4.2.1 Scenario description

Consider a simple network consisting of only one home facility, one work facility, only one route connecting each location with the other, and a population which is divided into two segments of 1000 individuals each. One segment departs for work at 7am, and one at 9am. The entire population leaves work and heads home at 5pm. All individuals make a phone call and produce a CDR precisely at the time they leave and arrive at their home location. Most individuals also use their phone at work and place calls when they arrive and when they leave, but members of the early-rising population segment do so only with a probability of 70%. This condition is designed to reflect the real-world case where a certain calling behavior is associated with certain kinds of travel behavior.

In the traffic demand reconstructed directly from the resulting CDRs, the non-calling sub-segment of the early-rising population will effectively stay at home, because their travel plan is constructed from an undersampled trace without a sighting at the work location. It does not contain a trip. This leads to an initial underestimation of the traffic demand from the home location to the work location at 7am to 700 travelling individuals, and from the work location to the home location to 1700 individuals.

4.2.2 Results

Once adding a traffic measurement with the reference volume of $y = 1000$ during hour 8 (ranging from 7:00:00 to 7:59:59), the observed population segment which leaves at 7am is scaled up by the calibrator to fit that number, compensating for those unobserved early-risers who do not use their phone at work (Fig. 3 top left). The validation measurement in the opposite direction at hour 18 follows (Fig. 3 top right): The approach is capable of improving the simulation away from the measurement because of the all-day time structure in the phone data.

If $y = 2000$ at hour 18 (and no measurement at hour 8) is chosen as the calibration measurement instead (Fig. 3 bottom row), meaning that only the total number of travellers is known but nothing from which relative population weights could follow, both population segments are scaled up proportionally.

4.3 Berlin scenario

4.3.1 Scenario description

As a more realistic scenario, a travel demand model generated from real data is used. It 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 (Scheiner, 2005; Moyo Oliveros and Nagel, 2012, 2013). 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. A total travelled distance of about 878 000 km is obtained.

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

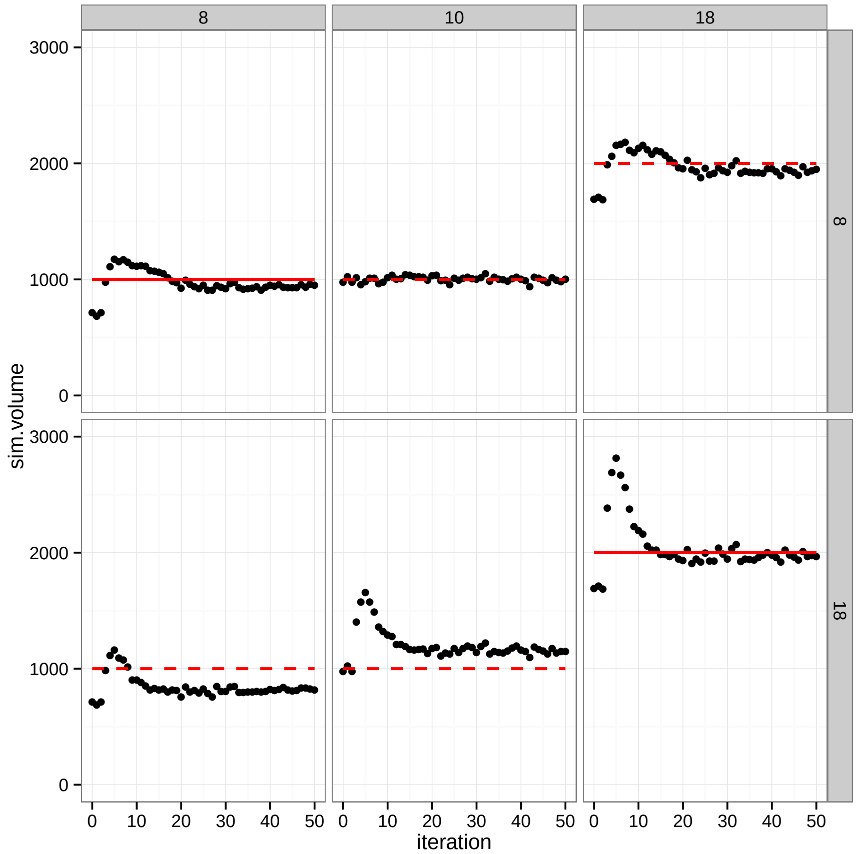

Figure 3: Simulated link volume over iterations, at hours 8, 10 and 18. The red lines, dashed and solid, denote the real value. The calibration target (solid red lines) is the measurement at hour 8 (top) or hour 18 (bottom).

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 traffic demand from these population segments is markedly distinct (Fig. 6 top vs. bottom, solid lines). 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).

The original plans underlying Fig. 1 are shown in Fig. 4. As one can see, the orange plan is a plan that contains a work activity, thus corresponding to a frequent caller (see Fig. 1). While the original plan gives the traveller the freedom of many routes around and through the city, the sightings (Fig. 1) 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. Many activities and related travel are missed (compare Fig. 4 with Fig. 1). In fact, the light blue CDR trace does not even result in a round trip any more.

4.3.2 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.

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 associ-

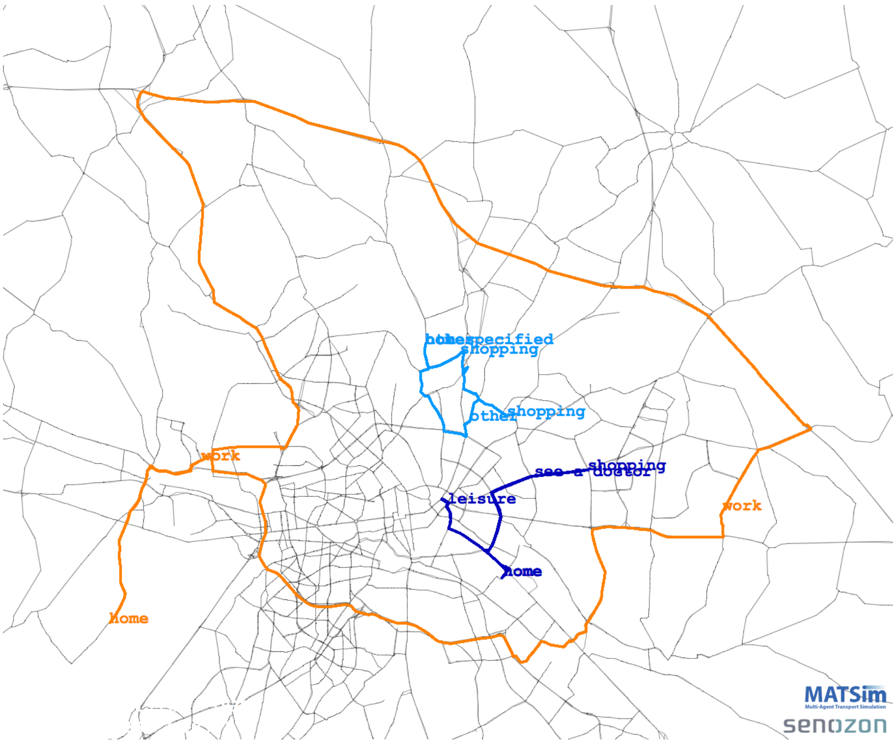

Figure 4: The original plans underlying Fig. 1.

ated with frequent callers.

- Include the sparse traces even though their spatio-temporal resolution is such that they contain only limited information.

Fig. 5 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. 5 bottom). In the initial estimation, the demand share generated from the undersampled non-worker traces is not only too low, but diffused over time (Fig. 6 bottom): 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. 7) reveals that it is distorted in both cases.

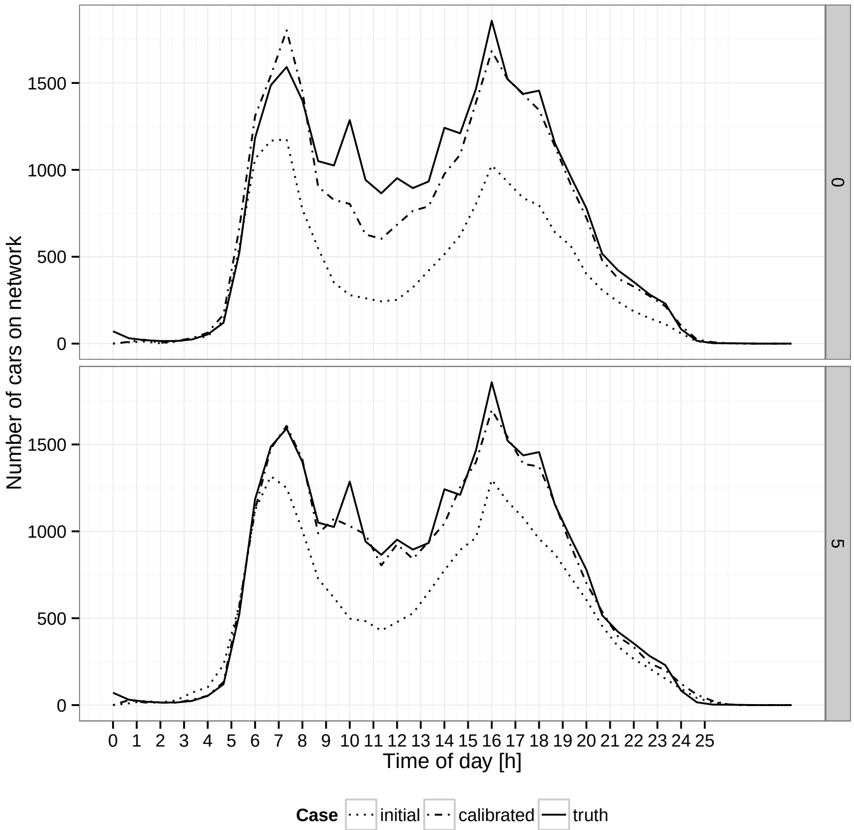

Figure 5: Network load over time of day where one demand segment ("non-workers") is missing (Scenario 1, top) or represented by undersampled trajectories (Scenario 2, bottom).

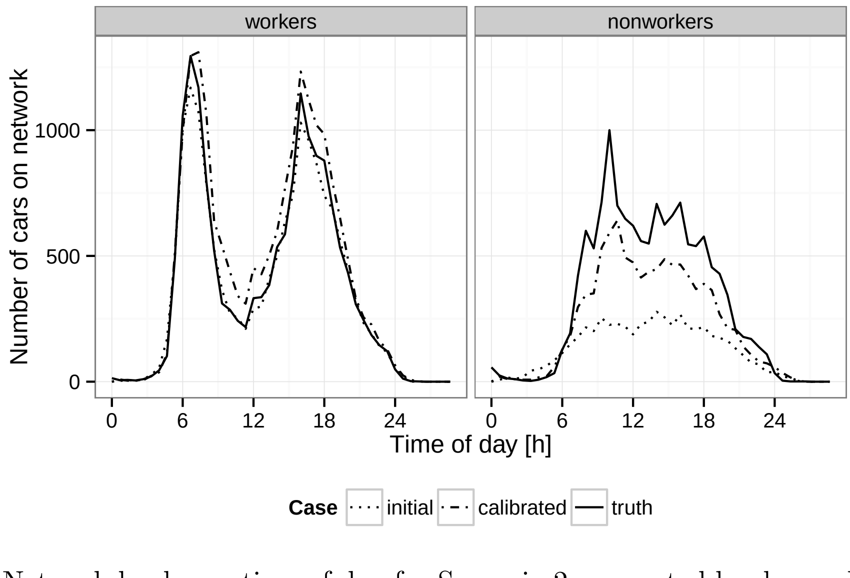

Figure 6: Network load over time of day for Scenario 2, separated by demand segments.

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

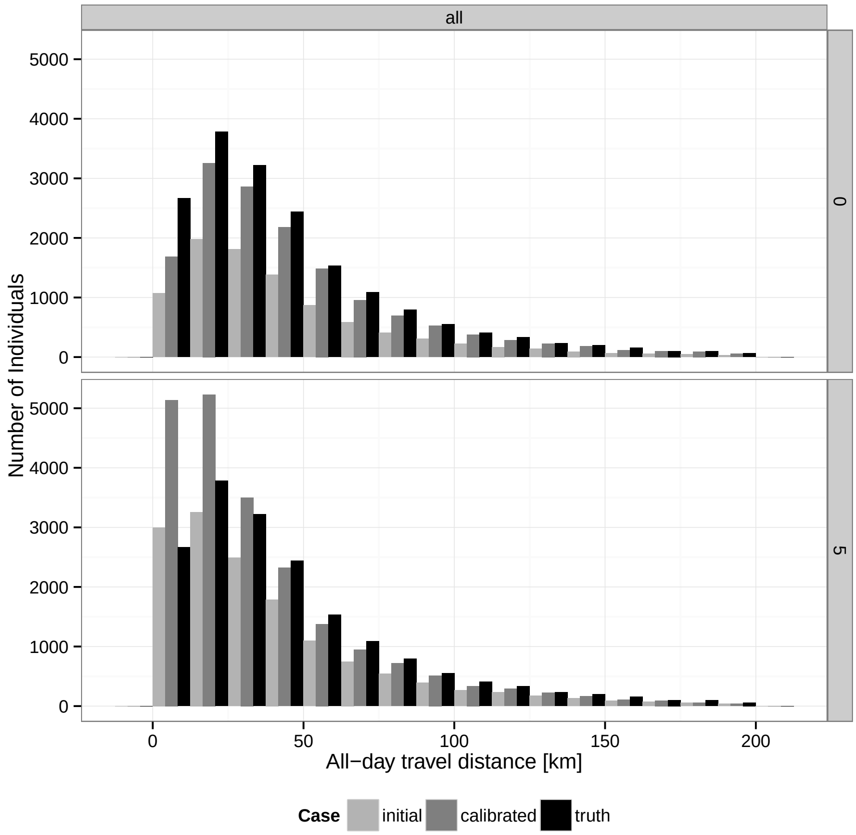

Figure 7: All-day travel distance distribution where sparse traces were removed (Scenario 1, top) or kept (Scenario 2, bottom).

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 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 Discussion

The starting point for this paper was the assumption that within CDR data, some traces may have a sufficient number of data points for full trajectory reconstruction, while others may not. In this situation, the statistically worst case is that the frequent callers belong to a different demand segment than the infrequent callers. The computational experiments show that even in this situation, a data fusion with anonymous traffic counts enriches the information in such a way that the resulting traffic is much closer to reality. That is, even trajectories with too few calls for reconstruction are useful as building material for a data fusion procedure.

Call rates While average phone call rates of 50 calls per day are certainly not realistic, these cases are still worth considering even outside of illustrative scenarios, because in practice, data points similar to CDRs need not be caused by actual phone calls, but can also appear as a consequence of, for instance, internet usage or recorded cell handovers. We consider the terms CDR, call rate, and cell, to be interchangeable with corresponding concepts in other current or future technologies which produce trajectories.

Additional or other measurements Cadyts is a quite flexible tool, allowing to adjust against arbitrary measurements that can be extracted from the simulation. This works since it takes each plan's contribution to each individual measurement from the simulation and builds an internal model around this. One alternative data source that comes to mind are link speed measurements, often also provided from cell phone data, but, because of fewer privacy restrictions, often available in much larger quantities. If available, it is also possible to add aggregate data to the process, such as the distribution of daily travel distance. These can be directly fused into the model, taking the same role as the link counts.

Activity types The approach discussed in the present paper does not add activity types to the trajectories. It is clear that this would be desirable, e.g. for planning purposes. Much work exists to attach activity types to trajectories (e.g. Chen et al., 2014). If such work is available for a certain scenario, its output can just be directly used as input to MATSim, including its Cadyts calibration approach (Flötteröd et al., 2012). If, however, such work is not available for a given scenario, it is our experience that such information reconstruction algorithms need further adaptation to a specific scenario. With our Ivory Coast scenario (Zilske and Nagel, 2013) in mind, we target scenarios where such additional information is not available.

Uncertainty in interpretation Also, our philosophy here is to retain the uncertainty that is in the data as long as possible throughout the process. For that reason, we just keep the actual sightings as fixed, while everything else in a plan is open to adjustment. An upstream method that assigns activity types or transport modes to sightings could be put into the simulation loop, enhancing the simple plan generation algorithm described in section 3, to generate possible activity chains consistent with the trace (cf. Ziemke et al., 2014, for a similar approach). Optimally, these would come together with levels of certainty or probabilities per activity chain from the perspective of the upstream algorithm, which can be used as initial plan choice probabilities. Cadyts would then concentrate on the most probable combination of plans consistent with the measurements.

Sightings "en route" Such an approach would make better use of sightings recorded during travel. Recall that in the present approach, sightings are identified with possible activity locations. If, say, a traveller made phone calls right at the end of the previous and at the beginning of the following true activity, this leaves no time for the additional activity corresponding to that en route sighting, and it will just serve as an additional constraint in the sense that the routing has to go through it. On the other hand, if there are no tight constraints caused by the previous and following sightings, then without additional information in fact we do not know if a certain sighting was generated en route or not. Again, an upstream method could generate multiple options here, possibly again with prior weights attached, and Cadyts would select the one most consistent with the measurements. Additionally, one could, if available, feed aggregated distributions such as the number of trips per person, into Cadyts as additional measurement.

Spatial uncertainty The present paper assumes that each CDR can be unequivocally assigned to a link. Clearly, this it not true in reality; first, phone cells are larger than this, and second, CDRs may wander between cells without the phone actually physically moving (Chen et al., 2014). Our intention is to address this in future work in the same way as the other uncertainties, i.e. to assume that we actually do *not* know the exact position of each call. Again, optimally an upstream algorithm would provide us with multiple plans which are all consistent with the data, and the Cadyts approach could then be used to select between them according to additional measurements such as traffic counts.

Behavioral priors In general, also behavioral priors can be added. In fact, the original formulation of Cadyts (Flötteröd et al., 2011) does exactly that: It assumes that there is a behavioral prior which results in prior choice probabilities, and Cadyts computes posterior probabilities after the measurements (also cf. Eq. (2)). For the present paper, the weight of the behavioral prior was essentially set to zero. Once it will be possible to have activity types, as discussed earlier, then those behavioral priors, in the shape of all-day scoring or utility functions, can also be used, assuming that sufficient data is available to estimate such utility functions for the scenario under consideration. This could then even include the effect of, say, joint activities (Dubernet and Axhausen, 2013) or car sharing (Ciari et al., 2013).

Sensitivity to policy The output of the described process is the estimation of a traffic state over time. It could be used, for instance, to identify users of a certain link or intersection, to compute emissions (Kickhöfer and Nagel, 2011), or as embedding scenario for a human-in-the-loop simulation. It is, at this point, clearly not useful as an input to policy analysis. The only behavioral investment is that drivers use fastest paths between sightings, and even that cannot be used as a choice dimension since some of the routes are pinned to certain links by sightings on these links obtained while driving. A step towards a behavioral model, reactive to changes in the environment and thus to policy measures, would be, again, to make draws from a larger space of feasible activity-trip-chains when realizing a CDR trace. This would work towards the goal in two ways at once: Agents with many calls would no longer be pinned to their routes by sightings while travelling, allowing them to re-route around disturbances, and the properties of the expanded population would not automatically be biased by the call rate distribution of the CDR input towards less travel activity than in reality.

6 Summary

We formulated the problem of fusing CDRs with traffic counts as a reduction to the calibration of individual travel choice probabilities in an iterated dynamic travel assignment scheme. The approach thus inherits known properties from the mobility simulation and from the calibrator.

A simple loop scenario illustrates our main argument for using an agent-based demand model even in the absence of activity diaries, with CDRs as an alternative input. CDR traces have an all-day structure, which a trip-based demand model does not capture. In the illustrative scenario, only one link count is needed to influence traffic in both directions.

The Berlin 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

Arentze, T. and Timmermans, H., editors (2005). *ALBATROSS–Version 2.0 – A learning based transportation oriented simulation system.* EIRASS (European Institute of Retailing and Services Studies), TU Eindhoven, NL.

Balmer, M. (2007). *Travel demand modeling for multi-agent transport simulations: Algorithms and systems.* PhD thesis, Swiss Federal Institute of Technology (ETH) Zürich, Switzerland.

Balmer, M., Axhausen, K., and Nagel, K. (2006). A demand generation framework for large scale micro simulations. *Transportation Research Record*, 1985:125–134.

Balmer, M., Rieser, M., Vogel, A., Axhausen, K., and Nagel, K. (2005). Generating day plans using hourly origin-destination matrices. In Bieger, T., Laesser, C., and Maggi, R., editors, *Jahrbuch 2004/05 Schweizerische Verkehrswirtschaft*, pages 5–36. Schweizer Verkehrswissenschaftliche Gesellschaft.

Bar-Gera, H., Konduri, K. C., Sana, B., Ye, X., and Pendyala, R. M. (2009). Estimating survey weights with multiple constraints using entropy optimization methods. Technical Report 09-1354, Transportation Research Board, Washington D.C.

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

Bowman, J., Bradley, M., Shiftan, Y., Lawton, T., and Ben-Akiva, M. (1998). Demonstration of an activity-based model for Portland. In *World Transport Research: Selected Proceedings of the 8th World Conference on Transport Research 1998*, volume 3, pages 171–184. Elsevier, Oxford.

Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015.

Chen, C., Bian, L., and Ma, J. (2014). From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, 46(0):326 – 337.

Ciari, F., Schuessler, N., and Axhausen, K. W. (2013). Estimation of carsharing demand using an activity-based microsimulation approach: Model discussion and some results. *International Journal of Sustainable Transportation*, 7(1):70–84.

Dash, M., Nguyen, H. L., Hong, C., Yap, G. E., Nguyen, M. N., Li, X., Krishnaswamy, S. P., Decraene, J., Antonatos, S., Wang, Y., et al. (2014). Home and work place prediction for urban planning using mobile network data. In *Mobile Data Management (MDM), 2014 IEEE 15th International Conference on*, volume 2, pages 37–42. IEEE.

Dubernet, T. and Axhausen, K. W. (2013). Including joint decision mechanisms in a multiagent transport simulation. *Transportation Letters*, 5(4):175–183.

DynusT (accessed August 2014). Dynamic urban systems for Transportation. `http://dynust.net`.

Flötteröd, G. (2009). Cadyts - A free calibration tool for dynamic traffic simulations. In *Swiss Transport Research Conference*. `http://www.strc.ch/conferences/2009/Floetteroed.pdf`.

Flötteröd, G., Bierlaire, M., and Nagel, K. (2011). Bayesian demand calibration for dynamic traffic simulations. *Transportation Science*, 45(4):541–561.

Flötteröd, G., Chen, Y., and Nagel, K. (2012). Behavioral calibration and analysis of a large-scale travel microsimulation.

Flötteröd, G. and Liu, R. (2010). Disaggregate path flow estimation in an iterated DTA microsimulation. Technical report.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley.

González, M., Hidalgo, C., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453:779–782.

Gur, Y. J., Bekhor, S., Solomon, C., and Kheifits, L. (2009). Intercity person trip tables for nationwide transportation planning in israel obtained from massive cell phone data. *Transportation Research Record: Journal of the Transportation Research Board*, 2121(1):145–151.

Iqbal, M. S., Choudhury, C. F., Wang, P., and González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40(0):63 – 74.

Kickhöfer, B. and Nagel, K. (2011). Mapping emissions to individuals – new insights with multi-agent transport simulations. In *Proceedings of the Conference on Computers in Urban Planning and Urban Management (CUPUM)*. Also VSP WP 11-02, see `www.vsp.tu-berlin.de/publications`.

Kitamura, R. (1988). An evaluation of activity-based travel analysis. *Transportation*, 15:9 – 34.

Moyo Oliveros, M. and Nagel, K. (2012). Automatic calibration of microscopic, activity-based demand for a public transit line. Annual Meeting Preprint 12-3279, Transportation Research Board, Washington D.C. Also VSP WP 11-13, see `www.vsp.tu-berlin.de/publications`.

Moyo Oliveros, M. and Nagel, K. (2013). 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. Also VSP WP 13-13, see `www.vsp.tu-berlin.de/publications`.

Müller, K. and Axhausen, K. W. (2010). *Population synthesis for microsimulation: State of the art*. ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau (IVT).

Neumann, A., Balmer, M., and Rieser, M. (2014). Converting a static trip-based model into a dynamic activity-based model to analyze public transport demand in Berlin. In Roorda, M. and Miller, E., editors, *Travel Behaviour Research: Current Foundations, Future Prospects*, chapter 7, pages 151–176. International Association for Travel Behaviour Research (IATBR).

Pendyala, R. (2004). 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. See www.eng.usf.edu/~pendyala/publications.

Russel, S. and Norvig, P. (2010). *Artificial Intelligence – A Modern Approach*. Pearson Education, Upper Saddle River, New Jersey 07458, 3 edition.

Scheiner, J. (2005). Daily mobility in Berlin: On 'inner unity' and the explanation of travel behaviour. *European Journal of Transport and Infrastructure Research*, 5:159–186.

Vovsha, P. and Bradley, M. (2006). Advanced activity-based models in context of planning decisions. *Transportation Research Record*, 1981:34–41.

Wang, H., Calabrese, F., Di Lorenzo, G., and Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. IEEE.

Wesolowski, A., Eagle, N., Noor, A., Snow, R., and Buckee, C. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society, Interface*, 10:20120986.

Ziemke, D., Nagel, K., and Bhat, C. (2014). Integrating CEMDAP and MATSim to increase the transferability of transport demand models. Annual meeting preprint, Transportation Research Board, Washington D.C. submitted.

Zilske, M. and Nagel, K. (2013). Building a minimal traffic model from mobile phone data. Technical report, MIT (Cambridge, MA). See perso.uclouvain.be/vincent.blondel/netmob/2013/NetMob2013-program-v1.pdf. Also VSP WP 13-03, see `www.vsp.tu-berlin.de/publications`.

Zilske, M. and Nagel, K. (2014). Studying the accuracy of demand generation from mobile phone trajectories with synthetic data. *Procedia Computer Science*, 32:802–807.