Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 32 (2014) 802 – 807

Procedia Computer Science

The 3rd International Workshop on Agent-based Mobility, Traffic and Transportation Models, Methodologies and Applications (ABMTRANS)

Studying the accuracy of demand generation from mobile phone trajectories with synthetic data

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

[a] *Transport Systems Planning and Transport Telematics, TU Berlin, Berlin, Germany*
[b] *Industrial and Systems Engineering, University of Pretoria, Pretoria, South Africa*

Abstract

We investigate replacing travel diaries with sets of call detail records (CDRs) as input data for an agent-oriented traffic simulation. Synthetic CDRs are used in order to study the effect of this substitution in isolation. We introduce an experimental design where a detailed synthetic transportation scenario with individual simulated travellers is combined with a simple model of mobile phone usage to collect synthetic CDRs. This set of artificial CDRs is then considered as input for another instance of the same traffic model, disregarding all other information. We analyse to what degree the model reproduces the base case, depending on the frequency of the available CDRs.

Selection and Peer-review under responsibility of the Program Chairs.

Keywords: transport simulation; multi-agent simulation; mobile phone data

1. Introduction

Transport simulation scenarios usually make use of pen-and-paper trip diaries for their demand model. These are expensive to obtain. There is a substantial recent interest in using Call Detail Records (CDRs) as a data source for such simulations.

CDRs carry other information than travel diaries. They are mostly available for a longer timeframe, while travel diaries are typically obtained for a single study day. Conversely, travel diaries contain detailed data about the activities and trips conducted on the study day, while CDRs only witness the presence of the participant at a certain point in time in a certain mobile phone cell. Whether the person was travelling at that point in time, or conducting an activity, cannot be directly determined. Neither the mode of transport nor the activity type is available.

Privacy concerns inhibit the widespread use of such data. For research in an earlier stage where predictions in real-world scenarios are not yet attempted, it may be necessary and sufficient to work with abstractions from actual CDRs which are realistic in the sense that they reproduce statio-temporal properties of the behavior of people [1], but are unencumbered by the responsibility for the privacy of study participants.

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

1877-0509
Selection and Peer-review under responsibility of the Program Chairs.
doi:10.1016/j.procs.2014.05.494

In this work, we consider an agent-oriented transport simulation scenario and study the question of how much different the simulation outcome would be, if CDRs had been the only available input data for constructing the demand model.

2. Synthetic CDRs by simulation

We start with any full implementation of MATSim[2], our agent-oriented transport model. 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, annotated with mode of transport. We consider this a kind of ground truth of a hypothetical scenario. Note that due to the scale of the traffic model, many additional kinds of measurements can be taken from this output, in particular volumes on links at arbitrary time scales.

For this work, we developed a plug-in for MATSim 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, or for the situation where mobile internet usage produces CDRs and hence temporally very fine-grained data is available, or where cell handovers are recorded, but also allowing for experimental simplifications such as placing a call precisely on arrival at or departure from activity locations.

The output of this step is a set of CDRs

$$(p_i, t_i, c_i) \tag{1}$$

where p_i is a person identifier, t_i a timestamp, and c_i a cell. We consider this the available data for traffic demand modeling in the hypothetical scenario. This framework allows us to study methods for constructing demand models from CDRs, and how much information from CDRs is needed by these methods to re-approximate the state of the traffic system in the ground truth scenario to which degree. It isolates these questions from the different question of how good MATSim itself is at approximating reality.

3. Simulation driven by CDRs

We convert each CDR trajectory into a travel diary in a straightforward way. For every person identifier observed, a MATsim person is created. Every call is converted into an activity. Activities are connected by trips. Several calls in the same zone without a call in a different zone between them are fused, since there is no evidence of travel between them. Similarly, there is no evidence for detours, so no additional activities are inserted. The only degree of freedom is the departure time from each activity location, which must be no earlier than the time of the last sighting at the activity location, and no later than the time of the first sighting at the next activity location minus the required travel time. The simplest solution is to set the activity end time to the time of the last sighting at the activity location: the phone call is assumed to have taken place at the time the agent leaves the activity.

The resulting plans are simulated. The output of this step is of the same form as the ground truth scenario. The two scenarios can now be compared to assess the approximation quality.

4. Results

4.1. Test scenario

To test our software and illustrate its workings for a corner case, we use the following scenario:

- As a ground truth scenario, some agents take trips between some random activity locations on an uncongested network. In MATSim, this means that everybody uses fastest routes with respect to free speed travel time.

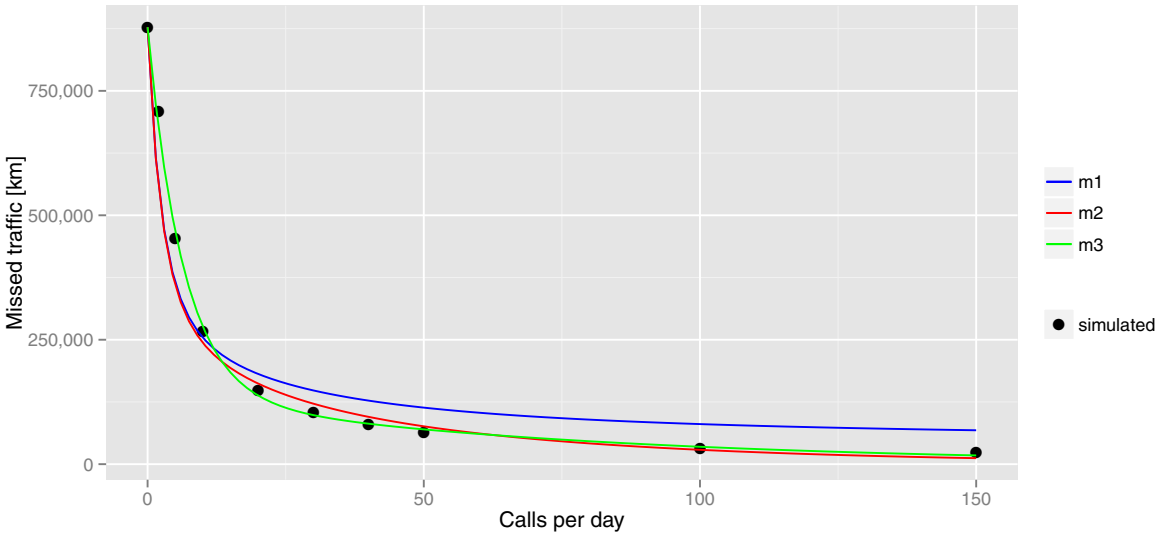

Fig. 1. Missed amount of traffic at different call rates. Simulation runs and models.

- Each link is its own mobile phone cell. Since MATSim works on the level of links, this means that CDRs are taken at the highest spacial resolution available.
- Agents make calls exactly when they start and end activities.

As expected, in this setup, traffic is reproduced exactly. The same happens when even more phone calls during activities or during travel are inserted, since they do not add additional information to this scenario. Also, they do not add artifacts to the simulation. This result is independent of the network or the demand.

4.2. Uncongested scenario

We consider a realistic travel demand model generated from empirical data. We used a 1994 household survey which contains complete trip diaries from one specific day of 2% of the Berlin population. 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. We select all individuals who only travel by car and obtain 18377 individuals. The network is extracted from OpenStreetMap (OSM) and contains 61920 links, with link speeds assigned according to defaults based on the value of the OSM highway category tag[3]. For this experiment, we assume no capacity constraints. With this setup, every agent chooses fastest routes with respect to free-speed travel time. We obtain a total travelled distance $L \approx 878000km$.

Agents place calls with uniform probability throughout the day, at a specified daily rate λ. Every agent has the same call rate. Multiple runs from the same ground truth scenario are run, with varying call rates. While average call rates of 50 calls per day or higher are certainly not realistic, these cases are still important to consider, because in practice, CDRs or similar data points need not be caused by actual phone calls, but might also appear as a consequence of, for instance, internet usage. We consider the terms CDR, call rate, and cell, to be interchangeable with corresponding concepts in other current or future technologies which produce trajectories.

We define the missed total travelled distance in the network compared to the base case as the error measure. As expected, the error drops with increasing call rate, on account of fewer trips to and from activities missed by the sampling (Fig. 1). Note that even with a high average call rate of 50 calls per day, the approach still misses about 10% of traffic. Thus, in order to make this model practically useful, some compensation method needs to be devised. As a first step in this direction, we investigate analytical models which could explain the data.

Table 1. Reproduction of total travel distance

	uncongested scenario		congested scenario	
rate	total travel distance [km]	relative to base	total travel distance [km]	relative to base
base	877934	1.0000	920928	1.0000
2	168672	0.1921	180407	0.1959
5	424627	0.4837	467898	0.5081
10	612295	0.6974	668082	0.7254
20	729089	0.8305	798749	0.8673
50	814241	0.9275	878649	0.9541
100	846466	0.9642	898857	0.9760
150	853450	0.9721	902353	0.9798
activity start/end	877934	1.0000	918223	0.9971

The probability of missing an activity of duration t with a call rate of λ is $e^{-\lambda t}$. Summing this over the empirical activity time sample and multiplying with the average trip length gives the expected missed traffic under the simplifying assumption that every activity accounts for the same amount of traffic.

$$m_1(\lambda; (t_1, \ldots, t_n)) = L\frac{1}{n}\sum_{i=1}^{n} e^{-\lambda t_i} \tag{2}$$

This model systematically predicts larger values of missed kilometers than the simulation at higher call rates. Following up on the intuition that this may in part be due to a correlation between the duration of an activity and the additional travel produced by it, we tried another analytical model, which incorporates trip distances, by considering the power sets of activity chains. For instance, an activity chain home-work-shopping-home has 16 ways of being sampled, among them home-shopping-home and home-work-home, but also shopping-home, since we do not consider home locations in a special way. We computed shortest-path travel distances for all partial plans of all agents, along with their probabilities depending on the call rate, and calculated the expected missed kilometers. This model fails to provide a better fit to the simulation runs. Its graph looks very similar to the much simpler model m_1, so it is omitted from the figure.

The only other way in which the simulation differs from the assumptions in m_1 is that in the simulation, agents also place calls while travelling. We tried to incorporate this effect into m_1 by capping the empirical activity durations at a given minimum duration t_{min}. The intuition behind this is that some time before and after the actual activity beginning and end, the agent is already or still near the activity location.

$$m_2(\lambda; (t_1, \ldots, t_n), t_{min}) = L\frac{1}{n}\sum_{i=1}^{n} e^{-\lambda \cdot max(t_i, t_{min})} \tag{3}$$

This model has a best fit to the simulation runs at a parameter value of $t_{min} = 23min$, where it is now quite close to the simulation result with some remaining errors at intermediate call rates (Fig. 1).

Following up on this, we now disregard the empirical activity duration sample and consider two typical activity durations, long (t_1) and short (t_2), with a relative frequency β of short activities, which yields a bi-exponential model.

$$m_3(\lambda; t_1, t_2, \beta) = L \cdot \left((1-\beta)e^{-\lambda t_1} + \beta e^{-\lambda t_2}\right) \tag{4}$$

At values of $t_1 = 3.7h$, $t_2 = 19.8min$ and $\beta = 0.16$, obtained by the nls solver of the R software package[4], this gives a good visual fit to the simulation runs. While these values do not approximate the empirical mean activity duration of $6.5h$, the model seems to pick up the effects of the empirical activity durations, the trip structures, and of calls placed while travelling.

4.3. Congested scenario

In a more realistic scenario, where roads are capacity-constrained, route choice in MATSim resembles a dynamic traffic assignment procedure. We generate a new ground-truth scenario based on the same set of travel diaries but

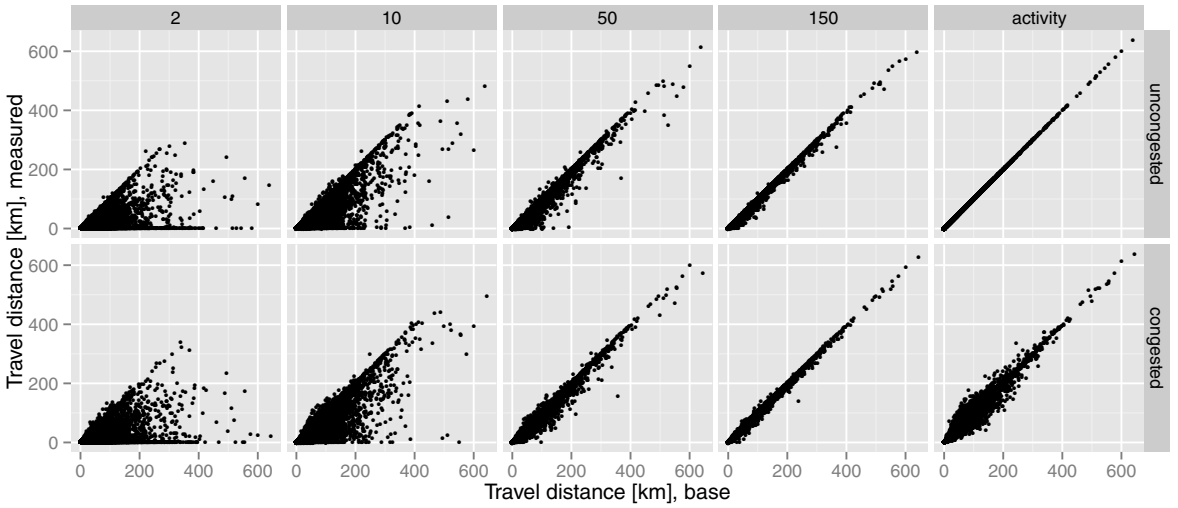

Fig. 2. Travelled distance per person, measured vs. base case, at different call rates

realistic link capacities, in a state resembling a dynamic user equilibrium with respect to travel times. Compared to the uncongested scenario, the total travelled distance increases by about 5% due to detours resulting from time-distance trade-offs. Applying the same process as above means that dynamic traffic assignment is now performed between sightings. Carrying out the same parametric runs as for the uncongested case reveals consistently less missed traffic in relative as well as absolute terms compared to the uncongested case (Table 1).

This can be explained as follows: In the uncongested scenario, the only effect of raising the call rate is that more activities are detected, so that more trips are taken. In the congested scenario, it has the additional effect that trajectories are traced more accurately, reproducing a larger share of the 5% travel distance increase which is due to detours.

In the congested scenario, all routes between sightings are fastest routes with respect to the congested traffic conditions. In every reconstructed case, however, there is overall less traffic, and in that situation, average fastest routes between any given pair of points tend to be shorter. For this reason, a higher sampling rate, and hence a more accurate tracing of routes, leads to a relatively higher reproduction of total travel compared to the uncongested scenario.

To illustrate the different behavior of the uncongested and the congested simulation, we consider the difference of travelled kilometers between base case and reconstructed case on a per person basis (Fig. 2). Each panel shows the reconstructed travel distance of each person in one of the scenarios versus the real travel distance in the base case. The rows are for the congested and uncongested scenario. The columns are call rates per day. The rightmost column shows a run with the artificial behavior of placing calls precisely at the beginnings and ends of activities. In the uncongested case, this reproduces travel distance per person exactly. More sightings would not add more information. In the congested case, we see a spread of missing or surplus travel distance per person. This reflects uncertainty about routes: The reconstructed scenario is in an equilibrium which is different from the base case. This means that under realistic traffic conditions, call rates which are higher than necessary to reproduce most activities still lead to an improvement in the reproduction of the traffic state, as individual routes are reproduced more faithfully.

The data points above the diagonal in the uncongested cases are a special case. They represent agents for which a missed activity results in a longer travel distance, because the activity is located on a tour of shorter length but longer duration. Missing the activity allows the agent to take a faster but longer tour through its remaining destinations.

5. Discussion and future work

The purpose of this work was to introduce the idea of simulating the acquisition of CDRs, which is a sampling process, as a framework to evaluate methods to create demand models from CDRs and to give a first estimate of the error introduced by this sampling. The method itself, as specified in section 3, is simplistic in that the resulting traffic will always be too little, because unsampled activities are simply missing. Adding additional trips to an individual, within the constraints of the potential path area induced by the CDRs, would be possible, but would require additional behavioral parameters. However, taking in some simple observable data, like the total travel distance or link volume counts, either the travel demand or the simulation output could be scaled to compensate. Our setup allows for the evaluation of such methods.

So far, we have only considered car traffic. We started out with agents using cars and recreated trajectories on the premise that they were from car users. However, there are multi-modal scenarios available which are open to similar methods: Synthetic CDRs can be generated from such a scenario without attaching the information whether the user is driving, using public transit, or walking, and these CDRs would then be used to reproduce a multi-modal scenario. This would need to incorporate a mode detection model, and this model would be validated by the degree to which the original modal split is reproduced. As far as public transit is concerned, it would be interesting to what degree line occupancy and line switch patterns could be reproduced.

We need to consider how to make use of the fact that real CDR datasets can span several days. This question arises both in synthesizing and in using CDRs. Our usual view of MATSim is that we simulate a typical work day. Synthetic CDRs spanning several days can be easily obtained by letting the same simulated day restart and continuing to apply the phone usage model. However, this would amount only to differently sampled concatenations of the same day, with zero temporal variability in the underlying travel behavior. Since MATSim is a stochastic model, it has variability between different runs with the same input data, but it is not clear if this translates well into temporal variability[5].

There are other modes of obtaining trajectories without user interaction. Services like Google Location History (formerly called Latitude) take location measurements directly from the handset and store them on a server, with explicit consent of the user, and without involving the phone operator. This acquisition mode has different characteristics from CDRs: The acquisition is not tied to the user using the phone, but runs in the background, at a rate which is determined by a trade-off between accuracy and energy usage. Our experimental setup can be easily adapted to synthesize such trajectories, with some of their special characteristics in mind (e.g. GPS does not work on underground trains.)

Acknowledgements

KN thanks the University of Pretoria for hospitality during a sabbatical.

References

1. S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, W. Willinger, Human mobility modeling at metropolitan scales, in: Proceedings of the 10th international conference on Mobile systems, applications, and services, MobiSys '12, ACM, New York, NY, USA, 2012, pp. 239–252. doi:10.1145/2307636.2307659.
2. M. Balmer, M. Rieser, K. Meister, D. Charypar, N. Lefebvre, K. Nagel, K. Axhausen, MATSim-T: Architecture and simulation times, in: A. Bazzan, F. Klügl (Eds.), Multi-Agent Systems for Traffic and Transportation, IGI Global, 2009, pp. 57–78.
3. M. Zilske, A. Neumann, K. Nagel, OpenStreetMap for traffic simulation, in: M. Schmidt, G. Gartner (Eds.), Proceedings of the 1st European State of the Map – OpenStreetMap conference, no. 11-10, Vienna, 2011, pp. 126–134.
URL `sotm-eu.org/userfiles/proceedings_sotmEU2011.pdf`
4. R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013).
URL `http://www.R-project.org`
5. A. Horni, Destination choice modeling of discretionary activities in transport microsimulations, Ph.D. thesis, Zürich (2013).