Multi-agent urban transport simulations using OD matrices from mobile phone data

Journal Article

Author(s):
Anda, Cuauhtémoc; Ordoñez Medina, Sergio A.; Fourie, Pieter J.

Publication date:
2018

Permanent link:
https://doi.org/10.3929/ethz-b-000262223

Originally published in:
Procedia Computer Science 130, https://doi.org/10.1016/j.procs.2018.04.139

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 130 (2018) 803–809

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Conference on Ambient Systems, Networks and Technologies, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, 8-11 May, 2018, Porto, Portugal

Multi-agent urban transport simulations using OD matrices from mobile phone data

Cuauhtemoc Anda[a,*], Sergio A. Ordonez Medina[a], Pieter Fourie[a]

[a]*Singapore-ETH Centre, 1 CREATE Way, 06-01 CREATE Tower, 138602, Singapore*

Abstract

Although new available *big data* sources have revealed themselves to be extraordinarily useful for transport demand modelling, they have not come into widespread use due to the justifiable privacy concerns of data stewards. In this study, we step back and re-evaluate the way in which mobile phone *telco* data can be introduced for the task of transport and land-use policy evaluation, travel demand forecasting and transport infrastructure testing through large-scale transportation simulations. We investigated that question by deploying a multi-agent transport simulation driven primarily by hourly-aggregated telco Origin-Destination (OD) matrices. We address the principal four challenges: spatial and temporal disaggregation, mode imputation and route choice. For temporal disaggregation, we propose a convolution with an exponential kernel method. As for transport mode imputation, a supervised-learning framework is designed. The simulation results are compared against traffic count data and public transport smart card transactions, showing accurate patterns for private cars but overestimated public transport demand in the morning peak. Lastly, we set the future steps for the improvement of simulations driven by aggregated mobile phone data.

Keywords: Mobile phone data; MATSim; OD Matrices; Route choice; Disaggregation

1. Introduction

Even though big data sources have revealed themselves to be extraordinarily useful for transport demand modelling, they have not come into widespread use due to the justifiable privacy concerns of data stewards. The individual spatio-temporal traces the models are built on, are sensitive information that can reveal personal attributes and thus violate data privacy regulations. For instance,[1] showed that just with four spatio-temporal individual points aggregated by hours and at the spatial resolution of mobile phone antennas, one can identify 95% of the individuals in a given data set. Furthermore, other studies[2,3] have tried to implement data privacy obfuscation techniques like *Differential Privacy* on spatio-temporal traces but have failed to find an equilibrium between data utility and privacy preservation.

* Cuauhtemoc Anda. Tel.: +65-8732-3266
E-mail address: anda@arch.ethz.ch

1877-0509 © 2018 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the Conference Program Chairs.
10.1016/j.procs.2018.04.139

In this study, we step back and re-evaluate the way in which big data, explicitly mobile phone telco data, can be introduced for the task of transport and land-use policy evaluation, travel demand forecasting and transport infrastructure testing through large-scale transportation simulations. Different from other studies in which individual traces are used to calibrate different spatio-temporal models[4,5], we propose using mobile phone telco data at an aggregated level to set up micro-simulations for transport planning without compromising the users' privacy. Hence, the question becomes: is it possible to build realistic large-scale transport micro-simulations using only aggregated statistics from big data sources.

We investigated that question by deploying a multi-agent transport simulation driven primarily by hourly-aggregated telco Origin-Destination (OD) matrices at a spatial resolution of subzone planning boundaries, of approximately $2km^2$ in size (on average). We address the principal four challenges in doing so: spatial and temporal disaggregation, mode imputation and route choice. For the case of temporal disaggregation, we propose a convolution with an exponential kernel of the observed histogram to map from hourly-based departing times to a continuous function. As for transport mode imputation, a supervised-learning framework is designed by combining information from both a household travel survey and the mobile phone OD matrices. The simulation results are compared against traffic count data and public transport smart card transactions. Lastly, we set the future steps for the improvement of simulations driven by aggregated mobile phone data.

2. Mobile phone aggregated OD matrices

In this study, we build upon spatially and temporally aggregated OD matrices derived from mobile phone stay locations. The OD matrices were processed and provided by *DataSpark*[1], part of Singtel telecommunications group. This information represents around 50% of the total population in Singapore. The processing of the mobile phone traces into stay-point locations and aggregation into OD-matrices was performed by *DataSpark*. A stay-point is identified when a mobile phone connects to a mobile phone antenna for more than 20 minutes. However, for privacy reasons, the OD matrices provided were aggregated in hourly increments and with a spatial resolution of subzones as defined by the Urban Redevelopment Authority (URA)[2]. Hence, the OD matrices indicate the counts for users travelling from an origin subzone towards a destination subzone during a certain hour of the day.

2.1. Data cleaning

The first step was to choose one day for the simulation. We chose the 18th of April 2017, since it was a typical Tuesday without any special disruptions or events happening in the city. For the day chosen a total of 6,241,101 trips were detected. Next, after inspecting the OD-matrices we realised that 5% of the entries reflected travel times of more than 3 hours which, in Singapore, is unrealistic, since crossing the island from corner to corner by public transport takes only two and a half hours.
There are different reasons for such excessively long travel times. Taking into account that the stay-location algorithm performed by *DataSpark* has a temporal threshold of 20 minutes, the movements of bus drivers, taxi drivers, delivery people, who do not stop for more than 20 minutes over long periods of time, might be one of the reasons. Another reason is that there are still signalling errors within the mobile phone antennas that might corrupt the travel times in the OD matrix. For this reason, trips with more than 3 hours of travel time were filtered out.

3. Methodology

MATSim[6,7] is a computer program to simulate transport demand and supply interactions. Millions of agents represent the population of a city or region. Each agent has a transport demand represented by a sequence of activities it must perform during one day at different times in different places. All these agents are included in a mesoscopic

[1] https://www.datasparkanalytics.com/
[2] https://www.ura.gov.sg/uol/master-plan/Contacts/View-Planning-Boundaries

mobility simulation based on queues. They interact in capacity-limited transport networks generating dynamic congestion. The decisions on how to travel between places to perform activities are made before the mobility simulation and stored as a plan for every agent. Thus, a plan is a sequence of activities and trips. MATSim uses a coevolutionary algorithm to find an equilibrium of the transportation economic problem by repeating hundreds of times the mobility simulation.

In this work, the general strategy is to create one agent for each of the observable trips in the OD-matrices. Thus, agents' plans include two dummy activities at different locations and one trip between then. It is then necessary to provide a specific location for both the origin and destination of the trip along with the specific time, in seconds, at which the trip starts. In addition, one has to allocate the agent/trip with a transportation mode. For re-planning, only route mutations are allowed on agents' plans. Using the evolutionary algorithm of MATSim, agents will find optimal routes while maintaining initial departure times and travel modes.

3.1. Spatial disaggregation

In the first step, for each trip between two subzones, the origin and destination need to be mapped from subzones to a specific coordinate within the subzone. We use a facilities database that combines information on buildings and places of interest in Singapore to randomly sample an origin X-Y coordinate belonging to one facility and a destination X-Y coordinate. For the case of intra-subzone trips, the destination coordinates cannot be the same as the origin coordinates.

3.2. Mode choice

The aggregated OD-matrices lack information on how the trips were performed. To assign a travel mode for each of the observed trips, we designed a set of features for a supervised-learning framework that relates the mobile phone aggregated OD-matrices with the Singapore household travel survey 2012. The challenge in doing so is to find the same set of features in both the OD-mobile phone matrices and the household travel survey.
We simplify the travel mode categories reported in the household travel survey into 3 broad classes: *Car*, which includes trips performed by car and taxi; *PT*, which includes trips performed by bus, MRT, LRT and walking; and the mode *Teleport*, that accounts for trip modes that do not affect the main transport network such as being the passenger in a car trip, bicycle, or private buses (of which we do not possess any information on vehicles used and their associated routes and schedules). The trips/agents assigned with the *Teleport* mode will not participate in the transport network and will be 'teleported' from the origin to their destination.
Next, since we do not possess any type of personal information about the trip maker, we designed the feature set or covariates of the model according to the characteristics of the origin subzone, the characteristics of the destination, and the characteristics of the trip. The following table describes each of the features used:

Covariates related to the origin
Car_O : Variable indicating the *Car* mode share observed in the origin subzone
PT_O : Variable indicating the *PT* mode share observed in the origin subzone
Tel_O : Variable indicating the *Teleport* mode share observed in the origin subzone

Covariates related to the destination
Car_D : Variable indicating the *Car* mode share observed in the destination subzone
PT_D : Variable indicating the *PT* mode share observed in the destination subzone
Tel_D : Variable indicating the *Teleport* mode share observed in the destination subzone

Covariates related to the trip
$Distance$: Euclidean distance between origin subzone centroid and destination subzone centroid
$Minutes$: Time of the day of the trip in minute of the day

Once the features or covariates are chosen, the following step is to choose a class of functions within the supervised-learning collection to map the feature space into the probabilities of the three classes: *Car*, *PT* and *Teleport*. That is,

For a set of training examples,

$$\left\{ \mathbf{x}_i, y_i \right\}_{i=1}^{N} \qquad where \;\; \mathbf{x}_i \in \mathbb{R}^p \qquad and \;\; y_i \in \{car, pt, teleport\} \tag{1}$$

Find a function,

$$F\left[\, \mathbf{x};\, \boldsymbol{\beta} \,\right] \rightarrow \left[\, 0, 1 \,\right]^{|y|} \tag{2}$$

which parameters can be estimated through the optimisation problem,

$$\boldsymbol{\beta}_{estimate} = \;\; \boldsymbol{\beta} \;\rightarrow\; argmin \left\{ \sum_{i=1}^{N} Loss\left(F\left[\mathbf{x}; \boldsymbol{\beta}\right],\, y_i\right) + Regularisation \right\} \tag{3}$$

and where the loss function has the form,

$$-log\left(\frac{e^{F_{y_i}[\mathbf{x}_i]}}{\sum_{j=1}^{c} e^{F_j[x_i]}} \right) \tag{4}$$

Subject to the notation: N is the number of samples, $\mathbf{x}$ is the vector of features of size p, y is the mode of transport class, $\boldsymbol{\beta}$ is the vector of model parameters, $F[\cdot]$ is the prediction function, $Loss[\cdot]$ is the loss function, *Regularisation* is the term that penalizes model complexity to avoid overfitting (optional), and c is the number of classes.

We tested 3 different models that fit that description: Multinomial logistic regression with Lasso regularization, Random Forest and Gradient Boosting. The generalisation strategy was to train the models on 70% of the household travel survey and testing them on the 30% left for accuracy. Gradient Boosting outperform the other models with an accuracy of 65.74%, against 65.68% of Lasso logistic regression and 63.31% corresponding to Random Forest.
Table 1 shows the variable importance in the Gradient Boosting algorithm. We can see that the public transport shares of origin and destination are the most correlated features. We argue that this is a proxy for the socio-economic level of the subzone plus the level of accessibility of PT. Following the distance of the trip that comes in 3rd place.

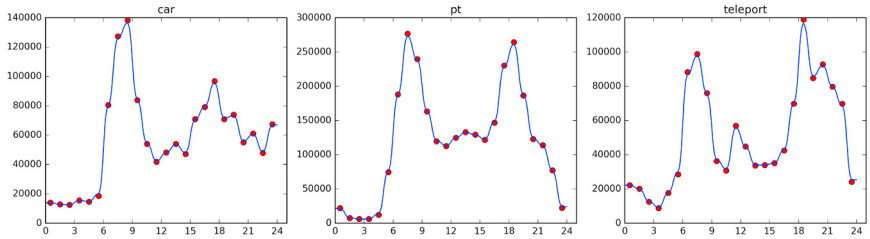
Table 1: Variable importance results for Gradient Boosting algorithm.

Rank	Var.	Rel. imp.
1	PT_O	23.29
2	PT_D	18.81
3	$Distance$	14.52
4	Tel_O	10.79
5	Car_D	9.60
6	Car_O	9.31
7	Tel_D	7.43
8	$Minutes$	6.24

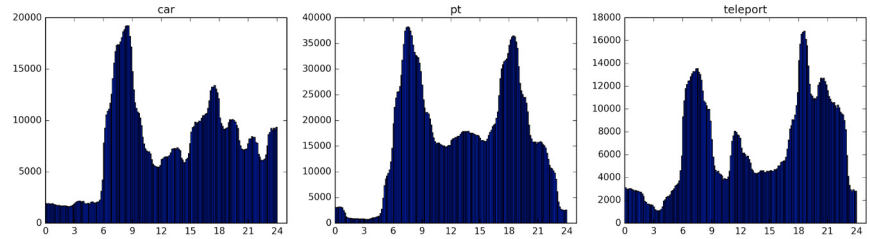
Lastly, we sample a travel mode for each of the observed trips in the OD matrices using the result probabilities of the Gradient Boosting. This strategy is used to achieve heterogeneity in the mode choice problem.

3.3. Temporal disaggregation

The last step of the travel demand input is to spread out trip start times across each hourly interval at a resolution of seconds, as the time step of the MATSim mobility simulation is one second. We propose a convolution with a



(a) Interpolation of observable departure times by travel mode



(b) Sampling results of departure times by travel mode after interpolation

Fig. 1: Departure time disaggregation with STH function

parametric exponential kernel: $g(t) = e^{-\lambda|t|}$

It is assumed that aggregated quantities are observed in the middle of each hourly interval (e.g. 9:30, 16:30, etc.). To find a value in between, the function of the kernel is to be a weight of the observed quantities. It gives more weight to observed values close in time to the unknown value and it gives negligible weight to farther observations. The parameter models how far should an observation still be taken into account to estimate a certain unknown value. Thus, the resulting continuous function of the start time histogram (STH) has the form:

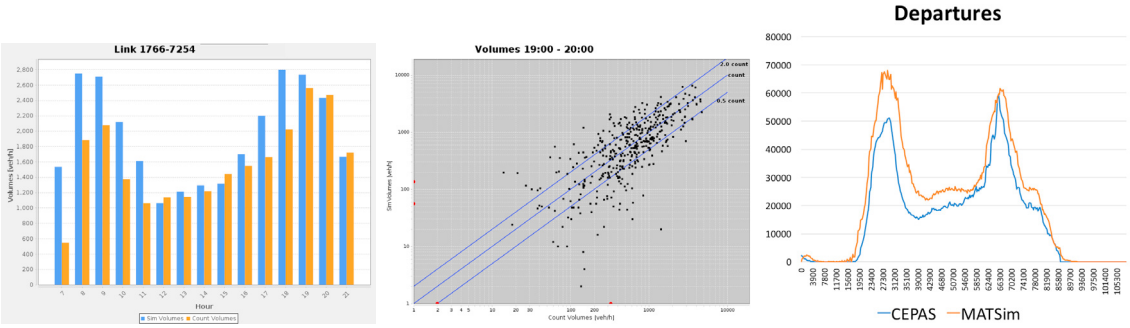
$$STH(t) = (f * g)(t) = \sum_{i=1}^{24} f(i) * g(t - i) \tag{5}$$

Where $f(t)$ is the given discrete histogram of start time and $g(t)$ is the exponential kernel. Fig.1a shows how the continuous start time histogram by mode (blue lines) were obtained with the given discrete distribution (red points) using an exponential kernel with a value of 0.001. Fig.1b presents the departure time distributions of disaggregated trips which were sampled using the described STH function.

3.4. MATSim for route choice and visualisation

The last piece of the general transport modelling framework is route choice. The strategy to follow is to let each agent decide which is its best route through the utility maximisation algorithm embedded in the evolutionary framework of MATSim.

A total of 5,558,642 agents representing approximately 50% of Singaporeans trips were included in the MATSim scenario. Road and public transport capacities were reduced to a half to account for this reduced demand. As mentioned before, the plans of each agent only include two activities and one trip. The utility function of MATSim was modified to only score the trip (negatively) and skip scoring activities. For re-planning, 20% of the agents change their routes every iteration in order to ease congestion (allowing larger shares of agent's to switch routes risks all of them jumping from one route to the same alternative). For public transport, the router proposed by [8] was used. It records experienced travel and waiting times to calculate routes of the next iteration. After 20 iterations of the mobility simulation, agents found optimal routes.



(a) Road counts against sim. results for a link across the day

(b) Road counts vs. sim. results for all links during afternoon peak hour

(c) Public transport demand: smart card (CEPAS) data vs sim. results

Fig. 2: Validation against traffic counts and public transport smart card data. Simulated car and pt trips were multiplicated by 2 to compensate mobile phone data market share.

4. Validation

To evaluate the model, we performed two types of validation. Simulated private vehicles were compared with Singapore traffic/road counts obtained in 2013. For public transport demand, we processed smart card data recorded in 2013 to compare with the simulation results.

Fig. 2a and Fig. 2b present the evaluation of the model using road counts. Only private cars are taken into account in this assessment. The number of vehicles of the simulation is multiplied by two because only 50% of the market was simulated. In Fig. 2a, one of the main links is compared, reporting similar tendency during the day. Fig. 2b shows the comparison of all links in the afternoon peak hour. Although both results look acceptable, there is an appreciable difference in the morning peak (represented by the link plot Fig. 2a). Links are systematically more congested in the simulation.

Fig. 2c shows a comparison of the public transport demand from the simulation and from recorded smart card ('CEPAS') data for the number of users starting a leg by time of the day. The simulated public transport demand is also duplicated, considering that a 50% scenario was simulated. Although the simulation in the afternoon peak looks accurate (including the small accumulation after the peak), there is a clear inconsistency in the morning peak. One possible reason is that train transfers are not visible in the tap-in tap-out card data, because passengers do not leave the system. Another possible reason of the problem is a special preference for other modes in the morning which was not included in the mode choice model.

5. Conclusion and future work

In this study, we introduced a new strategy to develop multi-agent transport simulations without using individual spatio-temporal traces. We demonstrated that it is possible to take advantage of emergent big data sources such as mobile phone network data without endangering the privacy of the users. Hourly aggregated OD matrices with a spatial resolution of subzone planning boundaries were used to drive the simulation. We solved the four main challenges: spatial disaggregation, temporal disaggregation, mode choice, and route choice. Specifically, we propose a convolution with exponential kernel of the observed histogram for temporal disaggregation, and a supervised-learning framework for mode choice using the household travel survey in a data fusion exercise. Results against traffic counts and public transport demand through smart card transactions show a sound match. However, several improvements are envisioned: the use of different aggregated statistics from mobile phone network data to develop a generative model of individual mobility patterns, a model for spatial disaggregation using different facilities attributes, and an extended mode choice model with additional aggregated mobile phone network data statistics. This study constitutes the first step in an effort to accurately simulate city-wide mobility patterns using privately-aggregated mobile phone data.

Acknowledgements

This research was conducted at the Future Cities Laboratory at the Singapore-ETH Centre, which was established collaboratively between ETH Zurich and Singapore's National Research Foundation (FI 370074016) under its Campus for Research Excellence and Technological Enterprise programme.

References

1. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* 2013;**3**:1376.
2. Primault V, Ben Mokhtar S, Lauradoux C, Brunie L. Differentially Private Location Privacy in Practice. *Most'14* 2014.
3. Hu X, Yuan M, Yao J, Deng Y, Chen L, Yang Q, Guan H, Zeng J. Differential privacy in telco big data platform. *Proceedings of the VLDB Endowment* 2015;**8**:1692-1703.
4. Jiang S, Yang Y, Gupta S, Veneziano D, Athavale S, González MC. The TimeGeo modeling framework for urban motility without travel surveys. *Proceedings of the National Academy of Sciences* 2016.
5. Lin Z, Yin M, Feygin S, Sheehan M, Paiement JF, Pozdnoukhov A. Deep Generative Models of Urban Mobility. *in review* 2017.
6. Horni A, Nagel K, Axhausen KW. *The Multi-Agent Transport Simulation MATSim*. London: Ubiquity Press; 2016.
7. Balmer M, Rieser M, Meister K, Charypar D, Lefebvre N, Nagel K, Axhausen KW. MATSim-T: Architecture and simulation times. *Multi-agent systems for traffic and transportation engineering* 2009;57-78.
8. Ordonez Medina SA, Erath A. New dynamic events-based public transport router for agent-based simulations. *ETH Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau* 2013;**921**.