A time-space model of disaggregated urban mobility from aggregated mobile phone data

Other Conference Item

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

Publication date:
2018-07-16

Permanent link:
https://doi.org/10.3929/ethz-b-000268852

A time-space model of disaggregated urban mobility from aggregated mobile phone data

Cuauhtemoc Anda
ETH Zurich
Singapore-ETH Centre
anda@arch.ethz.ch

Sergio Arturo Ordoñez Medina
ETH Zurich
Singapore-ETH Centre
ordonez@ivt.baug.ethz.ch

Abstract

Mobile phone *telco* data represents very valuable information for transport planners. Its large-scale coverage together with its spatio-temporal resolution makes it compatible with agent-based simulations for transport planning and a promise to improve travel demand models. However, such data is particularly vulnerable to breaches in privacy. Even if anonymized, there is a risk that users could be re-identified. In this work, we propose a model capable of generating individual space and time traces without looking at any individual *telco* trace. We define a set of aggregated histograms needed from the telco company to generate a population of individual mobility patterns using a Dynamic Bayesian Network. We present the validation results against the original telco data, showing that the model proposed is a viable tool to exploit telco data for transport planning without compromising users' privacy.

1 Introduction

Mobile phone telco data refers to the geo-coded time-stamps collected by the mobile phone network operator. These user location updates happen as a *handshake* between the mobile phone devices and a close cell-tower or antenna. In previous mobile phone data studies, this type of data was called Call Detail Records (CDR) data, however, since now the origin of the *handshakes* is no more restricted to just calls or SMS messages, but triggered also by other mechanisms related to internet connection (e.g. email polling), we encapsulate them by the name of mobile phone telco data.

The spatio-temporal resolution of the individual telco traces, together with its large-scale coverage make telco data a promising data source to develop

travel demand models for transport planning. For the task of simulating individual spatio-temporal traces in an urban setting, two main studies have emerged in the recent years which harnessed mobile phone telco data. On the one hand the TimeGeo modelling framework by Jiang et al. (2016), which is capable of generating individual urban mobility patterns with a resolution of 10 min and hundreds of meters. Their temporal and spatial choice model parameters are obtained from processing individual telco data. For example, they need to infer the users' home and working location *a priori.* On the other hand, the work by Lin et al. (2016) focuses on generating activity sequences with location and time information using a generative recurrent neural network with Long Short-Term Memory cells. Similar to Jiang et al. (2016) the model is feed by processed individual telco data (e.g. the stay-locations of the user are classified in *home*, *work*, or *other*).

These type of models, require access to individual mobile phone telco data, which sometimes is not easy to be shared by telco operators. One of the reasons is that even if anonymized, there is a risk that users could be re-identified since spatio-temporal traces are pseudo-identifiers (OECD 2015). For instance, de Montjoye et al. (2013) showed that even when the location of an individual is specified hourly with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals. With this in mind, we propose a time-space model of individual traces from aggregated mobile phone telco data. The model is capable of generating a population of individual traces without looking at any individual telco trace, hence, making it a more viable option for the development of applications which require a population of agents with mobility patterns (e.g. agent-based simulations for transport planning).

2 Aggregated telco data

Instead of directly working with individual telco data, we asked the mobile phone telco company to provide us with 3 user-aggregated datasets. The 3 datasets have an hourly temporal resolution and a spatial resolution by sub-regional planning areas or subzones (e.g. for our study-case Singapore, the city-state is divided in 320 subzones).

1. Origin-Destination (OD) matrices: This table consists of entries with the origin subzone, the departure time, the destination subzone, the arriving time, and the number of counts associated to that dynamic OD-pair. All arriving and departure times are reported hourly.

2. Durations histogram: This table consists of entries with the subzone, the starting hour of the stay, the stay-duration in that subzone, and the number of counts observed. Again, the starting hours and durations are reported hourly.

3. Initial counts of number of users by subzone: An initial picture of how

many users were observed during the first-time of the generation process (e.g. at midnight) at each of the subzones.

3 Dynamic Bayesian Network

3.1 Representation

Fig. 1 shows the proposed Dynamic Bayesian Network for the task of generating individual spatio-temporal traces.

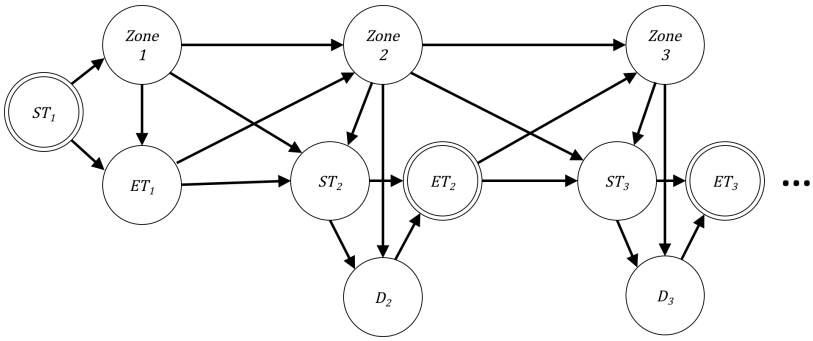


Fig. 1. Graphical representation of Dynamic Bayesian Network for individual spatio-temporal traces generation.

The proposed model encodes the joint probability distribution shown in Eq. 1.

$$P(z_{1:N}, d_{2:N}, st_{1:N}, et_{1:N}) = P(st_1)\, P(z_1|st_1)P(et_1|st_1, z_1)\prod_{k=1}^{N} P(z_k|z_{k-1}, et_{k-1})P(d_k|z_k, st_k)P(st_k|\, z_k, z_{k-1}, st_{k-1})P(et_k|st_k, d_k) \quad (1)$$

Where,

z = Sub-regional area or subzone
d = Duration of stay
st = Start time of stay
et = End time of stay
N = Number of stay-locations

The model is an extension of a Markov chain in which the next *subzone* visited by the agent depends on the previous *subzone* and the *end time* of the previous staying *subzone*. *start time* depends on the current *subzone*, the previous *subzone* and the last *end time*. The *duration* of the stay depends on the *subzone* and the *start time*. Finally, the *end time* of the *subzone* is a deterministic function corresponding to the sum of the *duration* of that *subzone* plus the *start time*.

3.2 Learning

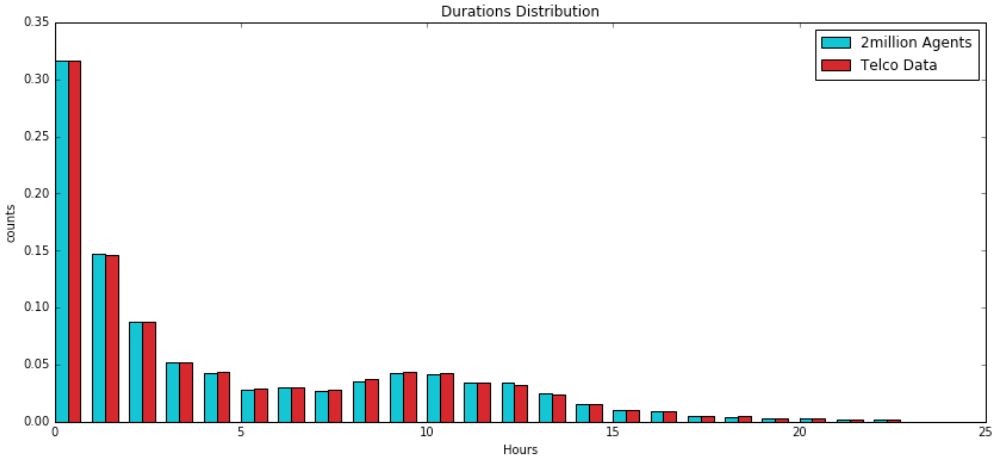
The parameters of the dynamic Bayesian network proposed can be estimated by Maximum Likelihood Estimation (MLE). Since all the variables are observable in the telco data and we are working with discretized variables, then the MLE solution of the model collapses to directly computing the different conditional and marginal probability distributions. This means that we only require the frequencies in which the combinations of the outcomes of the random variables occurred in the data. With this, we can achieve estimating the model parameters by only looking to certain aggregated histograms/distributions and avoid recurring to individual data.

3.3 Sampling

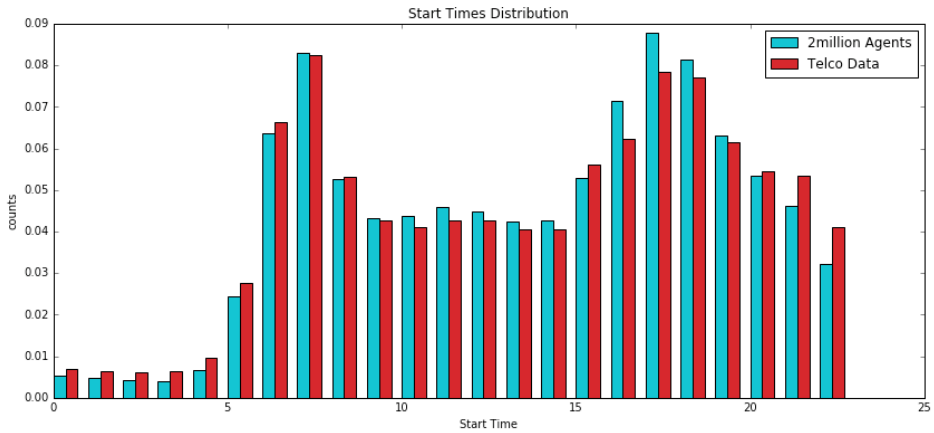
Forward sampling is used to generate individual spatio-temporal traces. This method of sampling starts by assigning an outcome for the marginal distributions of the model and then continues by following the order of the conditional probabilities.

4 Results

We used Singapore as a case-study. The aggregated mobile phone telco histograms were provided by one of the largest telco operators in Singapore. At any point individual data was consulted to calibrate the model. A sample of 2 million individual spatio-temporal traces were generated and compared to some statistic distributions. Fig. 2 shows the results so far.



(a)



(b)

Fig. 2 a) Comparison of durations distribution by the 2 million agents generated by the model (in blue) against the original information of the telco data (red). b) Comparison of the start times distribution of the 2 million agents generated (blue) against the original distribution of telco data (red).

Fig 2 shows a good match in the durations and start times distribution with the telco data. At this point, we are currently working on more validation results (e.g. distribution of number of trips by agent, spatial distribution of agents by time).

5 Conclusion

We propose a model that is capable of generating individual spatio-temporal traces from aggregated mobile phone data. Different than previous studies, our model has a privacy focus, in which, we avoid having model parameters that would need to be calibrated by observing individual data. This allows telco operators and transport planners to collaborate closely without compromising the privacy of the mobile phone users. Results by the generation process can be included as a first step towards developing agent-based simulations for Transport Planning.

Acknowledgments

This research has been conducted at the Singapore-ETH Centre for Global Environmental Sustainability (SEC), co-funded by the Singapore National Research Foundation (NRF) and ETH Zurich.

References

de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. 2013. "Unique in the Crowd: The Privacy Bounds of Human Mobility." *Scientific Reports* 3: 1376. doi:10.1038/srep01376.

Jiang, Shan, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. González. 2016. "The TimeGeo Modeling Framework for Urban Motility without Travel Surveys." *Proceedings of the National Academy of Sciences*, no. August: 201524261. doi:10.1073/pnas.1524261113.

Lin, Ziheng, Mogeng Yin, Sidney Feygin, Madeleine Sheehan, Jean-Francois Paiement, and Alexei Pozdnoukhov. 2016. "Deep Generative Models of Urban Mobility." In .

OECD. 2015. "Big Data and Transport: Understanding and Assessing Options." *International Transport Forum*. doi:http://dx.doi.org/10.1787/5jlwvzdb6r47-en.