Privacy-by-design generative models of urban mobility

Author(s):
Anda, Cuauhtémoc; Ordonez Medina, Sergio Arturo

Publication date:
2019

Permanent link:
https://doi.org/10.3929/ethz-b-000357034

Originally published in:
Arbeitsberichte Verkehrs- und Raumplanung 1454

Privacy-by-Design Generative Models of Urban Mobility

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

[a] *Future Cities Laboratory, ETH Zurich, 1 Create Way #06-01, Singapore 138602*

Abstract

New streams of Location-based Services (LBS) Big data have risen society's concerns in regards to data privacy. Even though these type of data sets are anonymised and aggregated in space and time, the risk of a privacy breach by user's re-identification is still imminent. Still, LBS data has the potential to improve current travel demand models and transportation applications. We this in mind, we introduce a Privacy by Design framework that generates realistic disaggregated daily mobility patterns without the need for any personal information or access to individual-level LBS data. On the first step of the framework, we estimate the joint probability distribution of daily mobility patterns using modified Markov models, followed by an adaptation of the rejection sampling algorithm to improve the distribution of the daily tour types. We validate the synthetic mobility patterns against six different distributions and reach an average accuracy over 95%. With this, we hope to open the discussion in the transportation community in regards to data privacy and travel demand models.

Keywords: travel demand models, generative models, data privacy, Big data

1. Introduction

New streams of location-based Big data (LBS) allows us to observe and understand mobility behaviour on an unprecedented level of detail [1]. From the array of LBS data, mobile phone *telco* data has drawn special attention due to its pervasiveness, extensive coverage, and persistent collection. These

*Corresponding author.

Email address: anda@arch.ethz.ch (Cuauhtemoc Anda)

type of data corresponds to events on the phone, such as voice call, internet usage, periodical updates and location area changes. After these event are triggered, a timestamp is recorded along with the user id and the connected cell tower, generally the one closest to the device. The nature of this data collection process requires some processing steps to filter out systematic noise and be able to extract the trajectories [2] for further mobility analysis. Nonetheless, it has already improved current travel demand models for transport planning [3][4] and our knowledge on human mobility [5][6].

Conversely, the fact that daily mobility patterns can be reconstruct from these a series of LBS data points has awaken growing concerns in regards to data privacy [7]. People's patterns of movement in space and time are repetitive and predictable, making LBS data a potent quasi-identifier for single person [8]. For instance, in [9] was found that even for data with a temporal resolution of one hour and a spatial resolution equal to the cellular network's base tower cells, just four spatio-temporal points were sufficient to isolate and uniquely identify 95% of the individuals. This means that anonymising LBS data sets is by any means a solution to guarantee users' privacy.

Despite LBS data being particularly vulnerable to breaches in privacy, the challenge of balancing out the privacy concerns with the usefulness in travel demand models has attracted little attention in the transportation field. We argue that one of the reasons is that traditionally, transport planning data sets (e.g. household travel surveys) have by default people's consent, plus the data sets are generally owned by the same public agencies that calibrate the models. However, if we want to make LBS Big data useful for transport planning applications, we need to tackle first the growing concerns related to data privacy.

To this extent we introduce a new framework to reproduce realistic individual level mobility patterns by taking the Privacy by Design approach. This approach holds that data collection systems and practices should be designed from the ground up to include strong and irreversible pro-privacy measures [10]. This translates into taking privacy measures upfront when designing new travel demand models fuelled by LBS Big data. Specifically, our framework is designed in such way that it does not require any personal information, including any individual level trajectories. With this we hope to open the discussion in the transportation community in regards to data privacy and travel demand models.

The remainder of this paper is organised as follows. In Section 2, we review previous literature on travel demand models. Section 3 provides an overview of the general framework. In Section 4 we introduce a couple of models that capture the joint probability distribution of individual mobility patterns. In Section 5 we present the results for the different models and strategies. Finally, Section 6 and 7 contains further discussion about the proposed framework and the conclusions of the work respectively.

2. Literature Review

The traditional approach to model travel demand is with choice models. They are generally estimated using census and household travel surveys which collect personal information, information about the household and information about the journeys. They also generally use the utility maximisation paradigm where the different alternatives are weighted through parameters corresponding to the characteristics of the individual making the decision (e.g. socio-demographics), the characteristics of the alternatives and some context information. The realisation of the utility function for each individual dictates then the choice probabilities among the alternatives. Prominent examples are [11], where individual tours with activities and itineraries are constructed through a series of discrete choice models; and [12], where the choice of different daily plan aspects are modelled through a series of decision trees.

However, for the case of LBS big data, seldom times one has access to personal information such as socio-demographics, household structure, or trip purpose. In exchange, the sensing nature of LBS Big data allows for greater spatio-temporal granularity, wider population coverage, and a persistent data collection process. This has open opportunities to model travel demand with a human dynamics perspective. In [13] home and work activity locations are inferred from mobile phone data. These information along with spatial and temporal mechanisms inferred as well from mobile phone data are used to model flexible activity locations and schedules. In a similar way, [14] inferred primary activity locations from mobile phone data, and then trained a Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) to model the spatio-temporal aspect of flexible activities. In both frameworks access to individual level LBS data is required, as well as the the identification of home and work locations of mobile phone users.

The difference in the framework we are proposing is that we aim aim to generate not only the information related to flexible activities, but all the sequence of daily locations with schedules for a person. Furthermore, we designed our framework adopting the Privacy by Design approach, meaning that no individual level trajectory or personal information is used. To this extent, we employ generative models in the centre of our framework to accomplish our aim. These models have been already used in the field of transportation. Principally to produced synthetic populations [15] [16] [17], but also to generate mobility patterns [14] [18].

3. General Framework

In order to satisfy the Privacy by Design approach the objective of our framework is to reproduce a population of individual mobility patterns for one day by means of only user-aggregated mobile phone data from the *telco* operator, a data trust, or any other data steward. Such population should behave as close as possible to the real population in terms of the closeness to a series of target histograms related to temporal, spatial and individual aspects of mobility.

We start by assuming that there exist a true distribution that describes the population mobility patterns. This true distribution encodes the joint probability distribution of the series of places visited along with their temporal description (i.e. start times, durations).

$$f_X(x) = P(X_1, X_2, X_3, ..., X_N) \tag{1}$$

All the spatial and temporal information related to every stay-location throughout the day is encapsulated by X_i where $i = 1, 2, 3...N$ where N is the index for the last stay-location of the day. Every X_i corresponds then to the tuple $[L_i, St_i, D_i]$ which relates to the spatial stay-location, stay-location start time, and stay-location duration respectively.

The idea is then to approach as much as possible to the true distribution by constructing a proposal distribution $g(x)$ that encloses the real distribution $f(x)$. We then follow up with a adaptation of the rejection sampling algorithm to improve over the model deficiencies and ultimately get individual mobility patterns samples as close as possible to the real population.

3.1. Markov Models

Given the total number of different possible combinations of the random variables in the joint distribution $f(x)$, we require a model that can factorised

$f(x)$ into a set of marginal and conditional probability distributions. We use then Dynamic Bayesian Networks to build different Markov models that can approximate $f(x)$. This type of models inherit the 1st order Markov constraint which means that future states, or locations in our case, depend only on the current state. For the case of mobility patterns this represents an important constraint to model real tour structures. Thus, the different models proposed principally differed in the introduction of different strategies to mitigate this constraint. Eq. 2 introduces our general approximation model $g(X)$ as the factorisation given by the 1st order Markov property where the current state X_i only depends on the information of previous the state X_{i-1}

$$f(x) \approx g(x) = P(X_1) \prod_{i=2}^{n} P(X_i|X_{i-1}) \tag{2}$$

3.1.1. Privacy by Design via Maximum Likelihood Estimation

We estimate the model parameters of $g(X)$ using the Maximum Likelihood Estimation (MLE). The Dynamic Bayesian Network framework allows us to generalise the factorisation of the transition probabilities of the Markov property into a factorisation of conditional probabilities $P(X_{i,k}|U_{X_{i,k}})$, where $U_{X_{i,k}}$ refers to $X_{i,k}$ parents or dependants, and k is the iterator across the tuple $[L, St, D]$. Hence, given that we have a data set D with a list of samples $\{d_m\}_{m=1}^{M}$, we can construct the likelihood function as:

$$L_G(\Theta : D) = \prod_{m} \prod_{i} \prod_{k} P(X_{i,k}[m]|U_{X_{i,k}}[m] : \Theta) \tag{3}$$

A second restriction for the design of our Markov models is that the random variables involved should be of categorical nature and fully observable. This means that we can represent the different conditional probabilities $P(X_k|U_{X_k})$ as tables and the parameters $\theta_{k,x,u}$ being the entry values of those tables. Taking this into account the log likelihood can be express as:

$$l_G(\Theta : D) = \sum_{k} \sum_{x} \sum_{u} M[u,x] log(\theta_{k,x,u})) \tag{4}$$

Where $M[u,x]$ is the number of times that $X_k = x$ and $U_{X_k} = u$ happens in D. Hence, $x \in Val(X_k)$ and $u \in Val(U_{X_k})$.

After having constructed the log likelihood (Eq. 4) we can then proceed by formulating the optimisation problem to calculate $\hat{\Theta}$, as follows,

$$\hat{\Theta} = argmax_{\Theta} l_G(\Theta : D) \qquad s.t. \quad \sum_x \theta_{k,x,u} = 1 \forall (k,u) \tag{5}$$

And finally get the closed form solution of the optimisation problem:

$$\hat{\theta}_{k,x,u} = \frac{M[u,x]}{M[u]} \qquad \forall \quad (k,x,u) \tag{6}$$

Eq. 6 means that for the Markov models designed under the conditions of the random variables being categorical distributions and completely observable, the estimation of the parameters $\hat{\Theta}$ via MLE result in counting the frequencies of the different events as described by the conditional and marginal probabilities. Hence, only requiring histograms where the data is user-aggregated to estimate $g(x)$ and satisfy the Privacy by Design approach.

3.1.2. Sampling

Having estimated $g(x)$ we can proceed with the generation of the different individual locations and schedules throughout one day by using forward sampling. This method of sampling consists in assigning an outcome to the marginal distributions and then continue sampling following the order of the conditional probabilities. The sampling is stopped after the full day is completed.

3.2. Rejection Sampling

The second step of the framework takes into advantage the ability of generating any number of samples from g_X. We adapt the original idea of rejection sampling to further improve the daily tour type distribution in relation to the target f_{tour}. Since this daily tour type distribution is not directly encoded in g_X, we then estimate an empirical proposal distribution $\hat{g}_{tour}$ by drawing a large pool of samples from g_X. We then calculate the envelope factor $M = sup_x \frac{f(x)}{g(x)}, \quad x \in Val(X)$ and proceed with the rejection sampling algorithm:

1. Generate $\mathbf{Y} \sim g_X(x)$
2. Calculate $\mathbf{Y}_{tour}|\mathbf{Y}$
3. Generate $U \sim Uniform[0, M\hat{g}_{tour}(\mathbf{Y}_{tour})]$
4. If $U \leq f_{tour}(\mathbf{Y}_{tour})$, then accept: set $\mathbf{X}_{tour} = \mathbf{Y}_{tour}$ and stop. Otherwise, reject: return to step (1)

4. Modified Markov models for individual mobility patterns

The base idea for the two architectures proposed is to model the sequence of individual stay-zones, stay-zone start times, end times, and durations for one day, where a stay-zone is defined as the location where the individual performs an activity. From [19] this is factorised as:

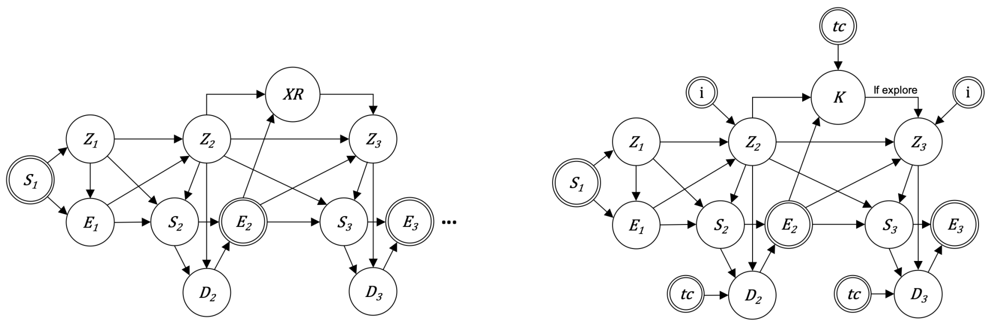
$$P(Z_{1:N}, S_{1:N}, E_{1:N}, D_{2:N}) = P(S_1)P(Z_1|S_1)P(E_1|Z_1, S_1)$$
$$\prod_{k=2}^{N} P(Z_k|Z_{k-1}, E_{k-1})P(S_k|Z_k, Z_{k-1}, E_{k-1})P(D_k|Z_k, S_k)P(E_k|S_k, D_k) \tag{7}$$

Where,
Z = Stay-zone
S = Stay-zone start time
E = Stay-zone end time
D = Stay-zone duration

This means that the next stay-location depends only on the previous stay-location and the previous end time. Another remark is that the first end time is model as a probability that refers to the first departure time of the day, while $E_k = S_k + D_k$ for $k = 2, ..., N$. As mentioned previously, the 1st order Markov constraint is an important restriction to generate realistic daily tours. To this end, we present two different variations on the base architecture to capture longer dependencies in an efficient way.

4.1. Explore & Return Model

Following the idea in [20] that exploration and preferential return are two mechanisms that describe human mobility, we added an Explore/Return (XR) random variable. This variable dictates whether the agent will explore a new stay-zone or will return to a previously visited one. It depends on the current stay-zone and the current end time, so as day develops, the agent will have a higher probability of returning to one of the previously visited places, specially if the agent is currently in a non-residential zone. The transition probability is now encoded as $P(Z_k|Z_{k-1}, E_{k-1}, XR_k)$. If the agent chooses to explore, then the previously visited zones are filtered out from the original $P(Z_k|Z_{k-1}, E_{k-1})$ and the probabilities are re-normalised. If the agent chooses to return, then only the already visited zones are considered in $P(Z_k|Z_{k-1}, E_{k-1})$ and the probabilities are as well re-normalised. Fig.1a shows the graphical representation of the model.



(a) Explore and Return model (b) Tour kernel model

Figure 1: Graphical representation of modified Markov models for individual mobility patterns. (a) Explore & Return model (b) Tour kernel model

4.2. Tour Kernel Model

Instead of having the Explore/Return variable that models indirectly the individual tour types, we can add a random variable that captures in a more direct way the construction of the daily tour chains. If we encode a tour chain as a sequence of digits, where every digit refers to a particular location, then the sequence 01020 might refer to someone that performs the activity chain: *Home, Work, Home, Shopping, Home*, where it is assumed that each type of activity is performed in a different location. For the tour kernel model we introduce a random variable K that models the next digit in the sequence given the current tour sequence or chain, the current time and the current zone. This is $P(K|Z_k, E_k, tc)$, where tc is the current tour chain. If K is present already in tc, then the transition is made directly to the linked zone. Otherwise, the transition is made through the probability $P(Z_{k+1}|Z_k, E_k, K, i)$, where i is the iterator of the state number. Fig. 1b shows the graphical representation of the model.

4.3. Types of urban travellers

Another strategy that we tested was the idea of having independent models for each type of traveller, instead of a general model for the full population. The intuition is that tour sequences can be more accurately constructed if the conditional and marginal distributions come from a series of homogeneous groups. In traditional travel demand models, this segmentation is taking into consideration through the demographics and social roles, however, in LBS Big data, seldom times we have access to these type of personal information. To this extent, in [21] a clustering framework based only on the series of individual stay-locations for one day was proposed. A set of five variables that reflect travel behaviour is designed, and different clustering algorithms are tested and validated. Adopting this framework, we tested the Explore and Return and the Tour kernel models for both cases: trained on the full population, and as independent models for each of the types of urban travellers.

5. Results

The framework was tested using mobile phone data from one the major *telco* operators in Singapore. All histograms relate to the 18th of April of 2017, a typical working Tuesday. For the spatial resolution, all histograms provided were aggregated into subzone planning boundaries[1]. Where these subzones are divisions within a planning area centred around a focal point such as a neighbourhood centre or an activity node. A total of 315 subzones which cover the extension of the main island were considered. As for the temporal resolution, the histograms were aggregated in an hourly basis. For the types of urban travellers part, we considered 16 different clusters as obtained in [21] for the case of Singapore.

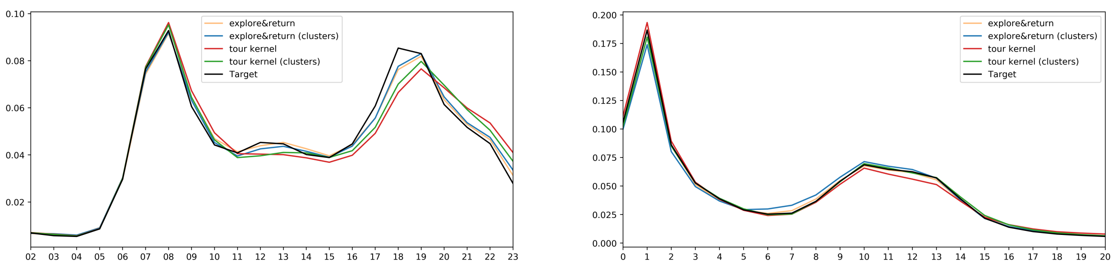
For the validation part, we considered 6 different target distributions: start time, duration, subzone, distance travelled, number of trips and tour type distribution. We assume that if our models are capable to match those target distributions, then we can conclude that the mobility patterns of the synthetic population behave similarly to the real population ones. We use the Root Sum of Squared Errors (RSSE) to measure the error between the distributions produced ($\hat{\pi}$) and the target ones (π), where RSSE is defined as:

$$RSSE(\hat{\pi}, \pi) = \sqrt{\sum_i (\hat{\pi}_i - \pi_i)^2} \tag{8}$$

[1]https://data.gov.sg/dataset?q=Subzone+Boundary

5.1. Temporal distributions

Fig. 2a presents the results for the start times distribution of every stay-zone. The x-axis represents the hour of the day, and the black colour plot represents the target distribution. Fig. 2b presents the results for the durations distributions. Here the x-axis is for the different durations from 0 hour duration to 20 hour duration. For both target distributions we can identify a close match.



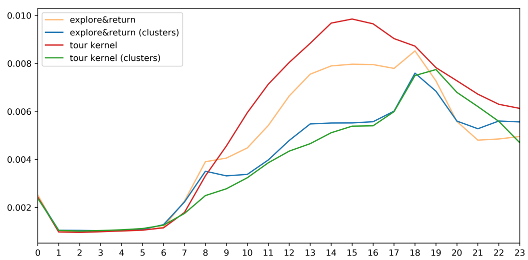
(a) Start times distribution

(b) Durations distribution

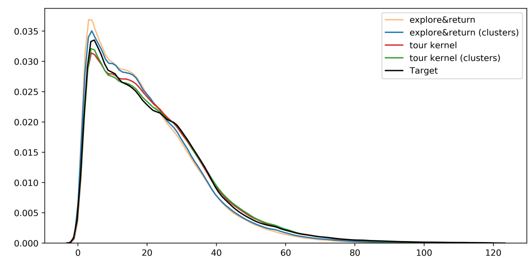
Figure 2: Temporal distributions validation. (a) Start time distributions (b) Durations distribution

5.2. Spatial distributions

For the case of the subzone distribution, we calculated the RSSE for each hour of the day. Fig. 3a shows how this error develops across the day for the different models proposed. One can notice that for all models and all hours of the day the error does not surpass the threshold of 0.1%. In Fig. 3b we can see a close match in the total distance travelled distribution by agent in a day. The units of the x-axis are given in km.



(a) Subzone error distribution

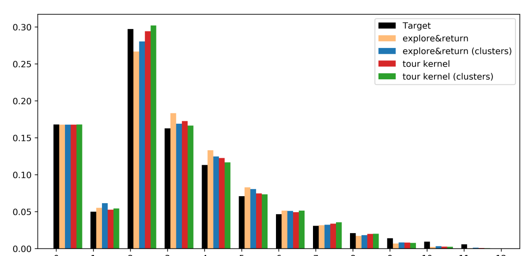


(b) Distance travelled distribution

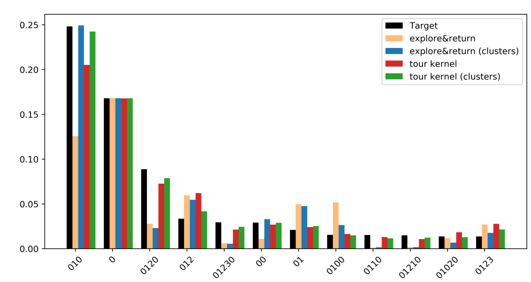
Figure 3: Spatial distributions validation. (a) Subzone error distribution (b) Distance travlled distribution

5.3. Individual related distributions

In Fig. 4a we present the distribution over the number of trips performed during the day by a single agent. The x-axis indicates the number of trips. Fig. 4b shows the daily tour chain distribution. Here, the x-axis indicates the target top 12 tour chains. We can notice that as compared to the temporal and spatial distributions, the tour chain distribution is more difficult to match, firstly because it is not directly encoded in the joint probability distribution, and secondly, because of the 1st order Markov property in the models.



(a) Number of trips per person distribution



(b) Daily tour chain distribution

Figure 4: Individual related distributions validation. (a) Number of trips per person distribution (b) Daily tour chain distribution

5.4. Rejection sampling efficiency

Another important metric for model comparison is the rejection sampling efficiency ($1/M$). It is a measurement of how far your proposal distribution is from the target distribution. The rejection sampling efficiency can also be interpret with its inverse M, which refers to the expected number of rejections needed in order to get one accepted sample. As explained in section 3.2, we have applied an adaptation of the rejection sampling algorithm to match the daily tour chain distribution. The calculation of M_{tour} then gives us a proxy of the distance between our models and the true distribution.

Fig. 5 shows the relationship between model performance, complexity and rejection sampling efficiency. Here the y-axis indicates the average model accuracy which is calculated as the complement of the average error for all target distributions, the x-axis indicates the number of model parameters, and the size of the dot relates to the expected number of rejections per sample. The first conclusion that we can draw is that there is an improvement in terms of model accuracy when clusters are considered. Another conclusion is that the Tour kernel model performs generally better than the Explore & Return one. The model that achieved the highest accuracy was the Tour kernel model with clusters, however, the number of parameters for this model is considerably larger as compared to the other ones. A balanced model is the Tour kernel (without clusters) since it still achieves over 90% accuracy, it has a good rejection efficiency (4.21 rejections per acceptance), and the number of parameters is not as large as the version with clusters.

Finally, Table 1 presents the full results on all the RSSE for every target distribution, as well as the RSSE average, the average accuracy, number of parameters and expected number of rejections per acceptance. We also present the results after doing rejection sampling on the Tour kernel model. As expected, the error of the top 100 daily tours drops down to virtually zero. What it is relevant to notice is that the change in this distribution does not substantially degrade the performance over the other distributions.

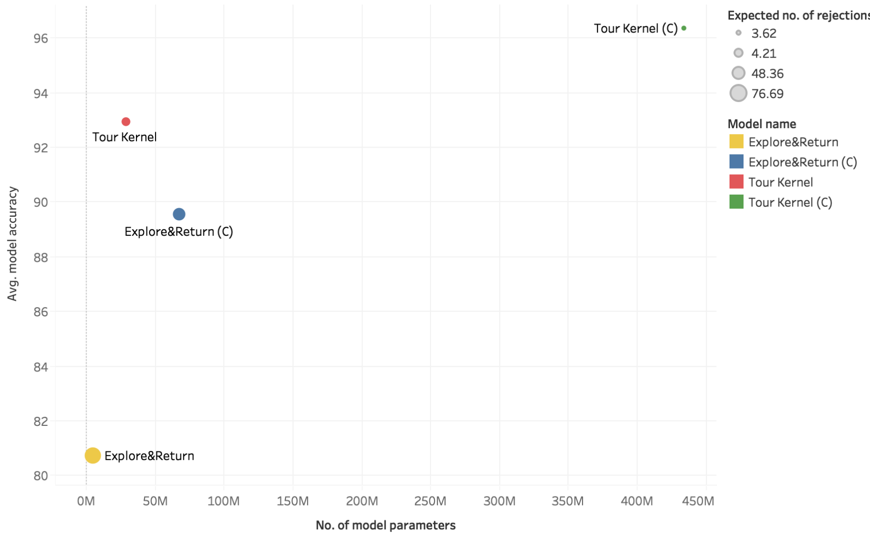


Figure 5: Model performance vs. complexity. Y-axis denotes the average accuracy performance of the model, X-axis denotes the total number of model parameters, the diameter size denotes the expected number of rejections for an acceptance.

Table 1: Table of results showing RSSE for target distributions, model performance, complexity and rejection sampling efficiency. E&R = Explore and Return model, E&R_C = Explore and Return model with clusters, TK = Tour kernel model, TK_C = Tour kernel model with clusters, TK_RS = Tour kernel model after rejection sampling

	RSSE (Root Sum of Squared Errors)							Model performance	Model complexity	Rejection sampling
Model name	Start time	Duration	Number of trips	Tours top 100	Subzone (24 hours mean)	Distance travelled	RSSE average	Average accuracy	Number of params.	Expected rejections / sample
E&R	1.30%	0.47%	11.00%	15.77%	0.46%	1.36%	19.28%	80.72%	4.50E+06	76.69
E&R_C	1.28%	1.95%	5.58%	8.52%	0.39%	1.00%	10.46%	89.54%	6.76E+07	48.36
TK	3.26%	1.50%	1.52%	5.88%	0.53%	0.58%	7.07%	92.93%	2.89E+07	4.21
TK_C	2.57%	0.87%	1.39%	2.00%	0.38%	0.39%	3.66%	96.34%	4.34E+08	3.62
TK_RS	3.11%	1.44%	1.37%	0.09%	0.51%	0.65%	1.20%	98.81%	2.89E+07	4.21

6. Discussion

As observed in Table 1 the average accuracy of the models proposed ranged from 80% to 96%. The principal variation in the models' accuracy comes from the error of the daily top 100 tours distribution. This translates in some models being able to overcome the 1st order Markov constraint better than others. However, the generative nature of the model, allows us to sample indefinite times and, as mentioned previously, use rejection sampling to improve the tour types distribution. It means that any of the models is useful as long as one has the computational power and time to produce the required number of expected rejections per acceptance needed. In theory, one could just sample from the random variables independently (i.e. without any model behind) and then use rejection sampling. However, given the dimensions of the variables and all the possible combinations it would not result in a practical solution. This is why the first step of the framework is to develop different model architectures to get as close as possible to the target distribution, and have a good rejection sampling efficiency for the second part.

Another point to discuss is the adoption of generative models through Dynamic Bayesian Networks instead of recent developments in deep learning generative models for sequences. Models such as Long Short Term Memory (LSTM) Recurrent Neural Networks (RNN) can encode in an efficient way the joint probability distribution over the whole sequence. However, adopting the deep learning approach would defeat in principle the Privacy by Design purpose since one would require access to individual level data to train these models. In contrast, for the case of our Explore & Return model, it is only 5 user-aggregated histograms that are required from the data provider: an initial zone histogram, the histogram of the time of the first departure | zone, dynamic origin and destination matrices, the histogram of duration | (time,zone) histogram, and the explore/return | (time,zone) histogram.

7. Conclusion

We introduced a new framework to harness LBS Big Data in transportation while mitigating privacy breach risks. The Privacy by Design Generative Models of Urban Mobility produce realistic daily mobility patterns without any personal information, including any individual level LBS data. The framework consists of two steps. The first step approximates the joint probability distribution over the different stay-locations and temporal attributes by modified Markov models. The second step applies rejection sampling to further improve the generation of daily tour sequences. For the different models and strategies the average accuracy spanned from 80% to 96% when applied to Singapore mobile *telco* data before rejection sampling. We also showed that rejection sampling on the daily tour types distribution further improves model performance.

There are several directions in which the current framework can be extended: an efficient adaption of the rejection sampling algorithm for several targets, a rigorous test on user re-identification, an extrapolation of the model for future scenarios, combination of other data sources to include mode of transport and socio-demographic information, and a study that measure the performance of synthetic mobility patterns against real mobility patterns in an agent-based simulation.

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.

The authors would also like to thank DataSpark for providing the aggregated mobile phone data in this study.

References

[1] C. Anda, A. Erath, P. J. Fourie, Transport modelling in the age of big data, International Journal of Urban Sciences 21 (2017) 19–42.

[2] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, M. C. González, A review of urban computing for mobile phone traces: current methods, challenges and opportunities, in: Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing, ACM, p. 2.

[3] L. Alexander, S. Jiang, M. Murga, M. C. González, Origin–destination trips by purpose and time of day inferred from mobile phone data, Transportation research part c: emerging technologies 58 (2015) 240–250.

[4] J. L. Toole, S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, M. C. González, The path most traveled: Travel demand estimation using big data resources, Transportation Research Part C: Emerging Technologies 58 (2015) 162–177.

[5] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, Understanding individual human mobility patterns, nature 453 (2008) 779.

[6] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in human mobility, Science 327 (2010) 1018–1021.

[7] J. Valentino-DeVries, N. Singer, M. H. Keller, A. Krolik, Your apps know where you were last night, and theyre not keeping it secret, New York Times 10 (2018).

[8] I. T. Forum, Big data and transport (2015).

[9] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, Unique in the crowd: The privacy bounds of human mobility, Scientific reports 3 (2013) 1376.

[10] M. Langheinrich, Privacy by designprinciples of privacy-aware ubiquitous systems, in: International conference on Ubiquitous Computing, Springer, pp. 273–291.

[11] J. L. Bowman, M. E. Ben-Akiva, Activity-based disaggregate travel demand model system with activity schedules, Transportation research part a: policy and practice 35 (2001) 1–28.

[12] T. Arentze, H. Timmermans, Albatross: a learning based transportation oriented simulation system, Citeseer, 2000.

[13] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, M. C. González, The timegeo modeling framework for urban mobility without travel surveys, Proceedings of the National Academy of Sciences 113 (2016) E5370–E5378.

[14] Z. Lin, M. Yin, S. Feygin, M. Sheehan, J.-F. Paiement, A. Pozdnoukhov, Deep generative models of urban mobility, IEEE Transactions on Intelligent Transportation Systems (2017).

[15] L. Sun, A. Erath, A bayesian network approach for population synthesis, Transportation Research Part C: Emerging Technologies 61 (2015) 49–62.

[16] L. Sun, A. Erath, M. Cai, A hierarchical mixture modeling framework for population synthesis, Transportation Research Part B: Methodological 114 (2018) 199–212.

[17] S. S. Borysov, J. Rich, F. C. Pereira, Scalable population synthesis with deep generative modeling, arXiv preprint arXiv:1808.06910 (2018).

[18] K. Ouyang, R. Shokri, D. S. Rosenblum, W. Yang, A non-parametric generative model for human trajectories., in: IJCAI, pp. 3812–3817.

[19] C. Anda, S. A. Ordoñez Medina, A time-space model of individual traces from aggregated telco data, in: 15th International Conference on Travel Behavior Research (IATBR 2018).

[20] C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling properties of human mobility, Nature Physics 6 (2010) 818.

[21] C. Anda, Archetypes of urban travelers: Clustering of mobile phone users in singapore, in: Mobile Tartu 2018, FCL, Singapore ETH Centre.