


# Privacy-by-design generative models of urban mobility

**Working Paper****Author(s):**

Anda, Cuauhtémoc; [Ordonez Medina, Sergio Arturo](#) 

**Publication date:**

2019

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000357034>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

**Originally published in:**

Arbeitsberichte Verkehrs- und Raumplanung 1454

# Privacy-by-Design Generative Models of Urban Mobility

Cuauhtemoc Anda<sup>a,\*</sup>, Sergio A. Ordonez Medina<sup>a</sup>

<sup>a</sup>*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

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

---

\*Corresponding author.

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

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

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

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

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

43 The remainder of this paper is organised as follows. In Section 2, we

44 review previous literature on travel demand models. Section 3 provides an  
45 overview of the general framework. In Section 4 we introduce a couple of  
46 models that capture the joint probability distribution of individual mobility  
47 patterns. In Section 5 we present the results for the different models and  
48 strategies. Finally, Section 6 and 7 contains further discussion about the  
49 proposed framework and the conclusions of the work respectively.

## 50 2. Literature Review

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

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

78 The difference in the framework we are proposing is that we aim to  
79 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).

$$f_X(x) = P(X_1, X_2, X_3, \dots, X_N) \quad (1)$$

All the spatial and temporal information related to every stay-location throughout the day is encapsulated by  $X_i$  where  $i = 1, 2, 3 \dots N$  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

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

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

### 123 3.1.1. Privacy by Design via Maximum Likelihood Estimation

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

$$L_G(\Theta : D) = \prod_m \prod_i \prod_k P(X_{i,k}[m]|U_{X_{i,k}}[m] : \Theta) \quad (3)$$

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

$$l_G(\Theta : D) = \sum_k \sum_x \sum_u M[u, x] \log(\theta_{k,x,u}) \quad (4)$$

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

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

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

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

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

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

### 147 3.1.2. Sampling

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

### 154 3.2. Rejection Sampling

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

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

#### 168 4. Modified Markov models for individual mobility patterns

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

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

173 Where,  
 174  $Z$  = Stay-zone  
 175  $S$  = Stay-zone start time  
 176  $E$  = Stay-zone end time  
 177  $D$  = Stay-zone duration

178 This means that the next stay-location depends only on the previous  
 179 stay-location and the previous end time. Another remark is that the first  
 180 end time is model as a probability that refers to the first departure time of  
 181 the day, while  $E_k = S_k + D_k$  for  $k = 2, \dots, N$ . As mentioned previously, the  
 182 1st order Markov constraint is an important restriction to generate realistic  
 183 daily tours. To this end, we present two different variations on the base  
 184 architecture to capture longer dependencies in an efficient way.

##### 185 4.1. Explore & Return Model

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



197 in  $P(Z_k|Z_{k-1}, E_{k-1})$  and the probabilities are as well re-normalised. Fig.1a  
 198 shows the graphical representation of the model.

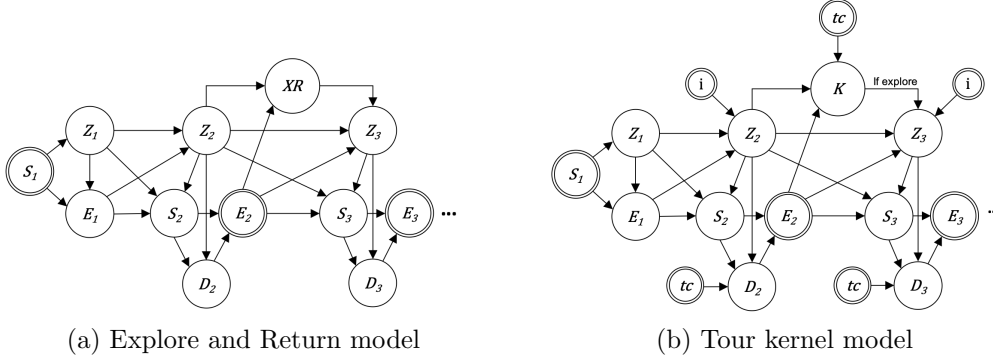


Figure 1: Graphical representation of modified Markov models for individual mobility patterns. (a) Explore & Return model (b) Tour kernel model

#### 199 4.2. Tour Kernel Model

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

#### 214 4.3. Types of urban travellers

215 Another strategy that we tested was the idea of having independent mod-  
 216 els for each type of traveller, instead of a general model for the full population.  
 217 The intuition is that tour sequences can be more accurately constructed if the  
 218 conditional and marginal distributions come from a series of homogeneous

219 groups. In traditional travel demand models, this segmentation is taking  
 220 into consideration through the demographics and social roles, however, in  
 221 LBS Big data, seldom times we have access to these type of personal infor-  
 222 mation. To this extent, in [21] a clustering framework based only on the  
 223 series of individual stay-locations for one day was proposed. A set of five  
 224 variables that reflect travel behaviour is designed, and different clustering  
 225 algorithms are tested and validated. Adopting this framework, we tested the  
 226 Explore and Return and the Tour kernel models for both cases: trained on  
 227 the full population, and as independent models for each of the types of urban  
 228 travellers.

## 229 5. Results

230 The framework was tested using mobile phone data from one the major  
 231 *telco* operators in Singapore. All histograms relate to the 18th of April of  
 232 2017, a typical working Tuesday. For the spatial resolution, all histograms  
 233 provided were aggregated into subzone planning boundaries<sup>1</sup>. Where these  
 234 subzones are divisions within a planning area centred around a focal point  
 235 such as a neighbourhood centre or an activity node. A total of 315 subzones  
 236 which cover the extension of the main island were considered. As for the  
 237 temporal resolution, the histograms were aggregated in an hourly basis. For  
 238 the types of urban travellers part, we considered 16 different clusters as  
 239 obtained in [21] for the case of Singapore.

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

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

---

<sup>1</sup><https://data.gov.sg/dataset?q=Subzone+Boundary>

### 248 5.1. Temporal distributions

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

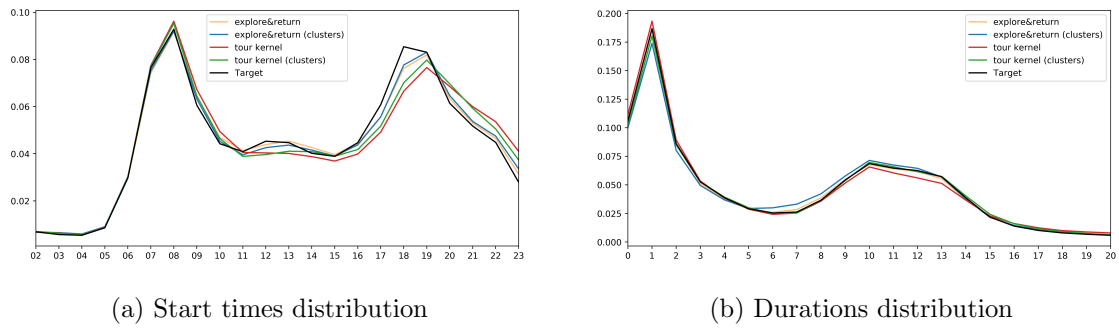


Figure 2: Temporal distributions validation. (a) Start time distributions (b) Durations distribution

### 255 5.2. Spatial distributions

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

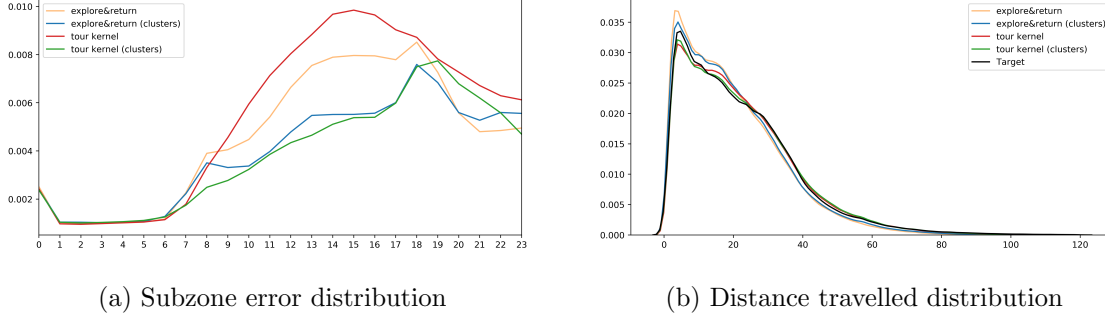


Figure 3: Spatial distributions validation. (a) Subzone error distribution (b) Distance travelled distribution

### 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.

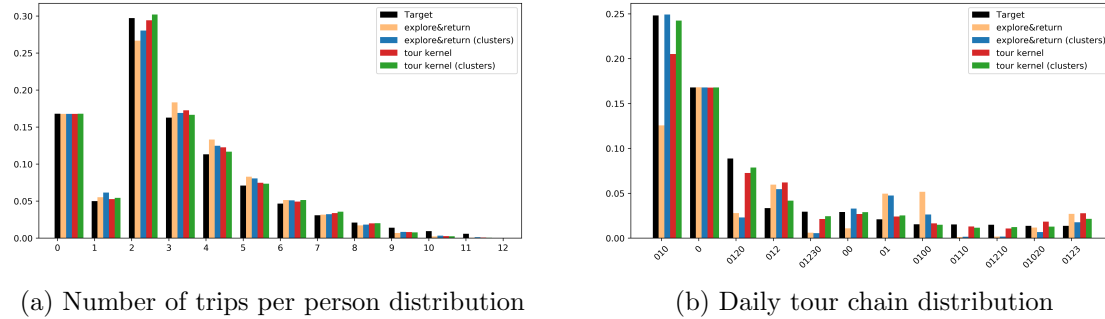


Figure 4: Individual related distributions validation. (a) Number of trips per person distribution (b) Daily tour chain distribution

#### 271 5.4. Rejection sampling efficiency

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

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

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

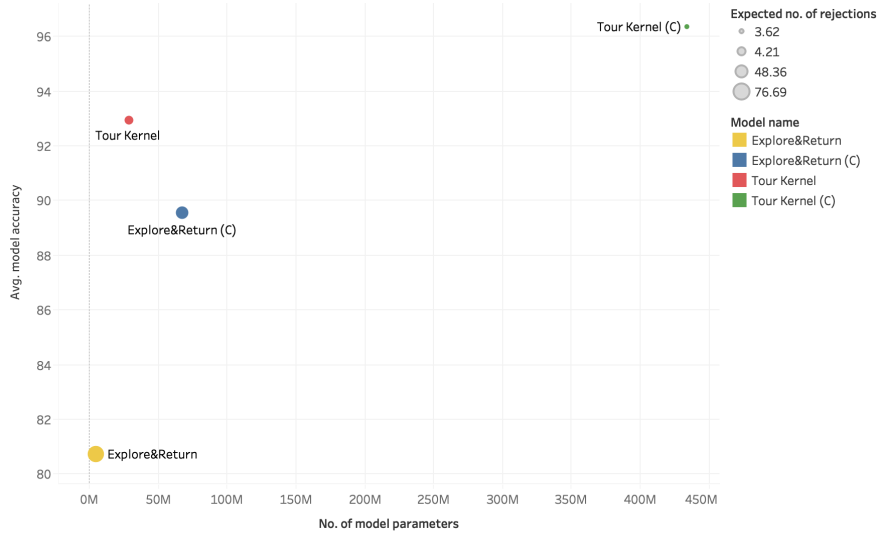


Figure 5: Model performance vs. complexity. Y-axis denotes the average accuracy performance of the model, X-axis denotes the total number of model parameters, the diameter size denotes the expected number of rejections for an acceptance.

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

## 301 6. Discussion

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

318 Another point to discuss is the adoption of generative models through  
319 Dynamic Bayesian Networks instead of recent developments in deep learning  
320 generative models for sequences. Models such as Long Short Term Memory  
321 (LSTM) Recurrent Neural Networks (RNN) can encode in an efficient way  
322 the joint probability distribution over the whole sequence. However, adopting  
323 the deep learning approach would defeat in principle the Privacy by Design  
324 purpose since one would require access to individual level data to train these  
325 models. In contrast, for the case of our Explore & Return model, it is only  
326 5 user-aggregated histograms that are required from the data provider: an  
327 initial zone histogram, the histogram of the time of the first departure |  
328 zone, dynamic origin and destination matrices, the histogram of duration |  
329 (time,zone) histogram, and the explore/return | (time,zone) histogram.

## 330 7. Conclusion

331 We introduced a new framework to harness LBS Big Data in transporta-  
332 tion while mitigating privacy breach risks. The Privacy by Design Generative  
333 Models of Urban Mobility produce realistic daily mobility patterns without  
334 any personal information, including any individual level LBS data. The  
335 framework consists of two steps. The first step approximates the joint prob-  
336 ability 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.



- 370 [5] M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, Understanding individual  
371 human mobility patterns, *nature* 453 (2008) 779.
- 372 [6] C. Song, Z. Qu, N. Blumm, A.-L. Barabási, Limits of predictability in  
373 human mobility, *Science* 327 (2010) 1018–1021.
- 374 [7] J. Valentino-DeVries, N. Singer, M. H. Keller, A. Krolik, Your apps  
375 know where you were last night, and theyre not keeping it secret, *New*  
376 *York Times* 10 (2018).
- 377 [8] I. T. Forum, Big data and transport (2015).
- 378 [9] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, V. D. Blondel, Unique  
379 in the crowd: The privacy bounds of human mobility, *Scientific reports*  
380 3 (2013) 1376.
- 381 [10] M. Langheinrich, Privacy by designprinciples of privacy-aware ubiqui-  
382 tous systems, in: *International conference on Ubiquitous Computing*,  
383 Springer, pp. 273–291.
- 384 [11] J. L. Bowman, M. E. Ben-Akiva, Activity-based disaggregate travel  
385 demand model system with activity schedules, *Transportation research*  
386 *part a: policy and practice* 35 (2001) 1–28.
- 387 [12] T. Arentze, H. Timmermans, Albatross: a learning based transportation  
388 oriented simulation system, Citeseer, 2000.
- 389 [13] S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, M. C. González,  
390 The timegeo modeling framework for urban mobility without travel sur-  
391 veys, *Proceedings of the National Academy of Sciences* 113 (2016)  
392 E5370–E5378.
- 393 [14] Z. Lin, M. Yin, S. Feygin, M. Sheehan, J.-F. Paiement, A. Pozdnoukhov,  
394 Deep generative models of urban mobility, *IEEE Transactions on Intel-*  
395 *ligent Transportation Systems* (2017).
- 396 [15] L. Sun, A. Erath, A bayesian network approach for population synthesis,  
397 *Transportation Research Part C: Emerging Technologies* 61 (2015) 49–  
398 62.

- 399 [16] L. Sun, A. Erath, M. Cai, A hierarchical mixture modeling framework for  
400 population synthesis, *Transportation Research Part B: Methodological*  
401 114 (2018) 199–212.
- 402 [17] S. S. Borysov, J. Rich, F. C. Pereira, Scalable population synthesis with  
403 deep generative modeling, *arXiv preprint arXiv:1808.06910* (2018).
- 404 [18] K. Ouyang, R. Shokri, D. S. Rosenblum, W. Yang, A non-parametric  
405 generative model for human trajectories., in: *IJCAI*, pp. 3812–3817.
- 406 [19] C. Anda, S. A. Ordoñez Medina, A time-space model of individual traces  
407 from aggregated telco data, in: *15th International Conference on Travel*  
408 *Behavior Research (IATBR 2018)*.
- 409 [20] C. Song, T. Koren, P. Wang, A.-L. Barabási, Modelling the scaling  
410 properties of human mobility, *Nature Physics* 6 (2010) 818.
- 411 [21] C. Anda, Archetypes of urban travelers: Clustering of mobile phone  
412 users in singapore, in: *Mobile Tartu 2018*, FCL, Singapore ETH Centre.