Synthesising digital twin travellers
Individual travel demand from aggregated mobile phone data

Journal Article

Author(s): Anda, Cuauhtémoc; Ordonez Medina, Sergio Arturo; Axhausen, Kay W.

Publication date: 2021-07

Permanent link: https://doi.org/10.3929/ethz-b-000477530

Rights / license: Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International


This page was generated automatically upon download from the ETH Zurich Research Collection. For more information, please consult the Terms of use.
Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data

Cuauhtemoc Anda\textsuperscript{a,*}, Sergio A. Ordonez Medina\textsuperscript{a}, Kay W. Axhausen\textsuperscript{b}

\textsuperscript{a} Future Cities Laboratory, ETH Zurich, 1 Create Way #06-01, Singapore 138602, Singapore
\textsuperscript{b} IVT, ETH, Wolfgang-Pauli-Strasse 15, HIL F31.1, 8093 Zurich, Switzerland

Keywords:
Travel demand models
Mobile phone data
Generative models
Data privacy

ABSTRACT

Mobile phone data generated in mobile communication networks has the potential to improve current travel demand models and in general, how we plan for better urban transportation systems. However, due to its high-dimensionality, even if anonymised there still exists the possibility to re-identify the users behind the mobile phone traces. This risk makes its usage outside the telecommunication network incompatible with recent data privacy regulations, hampering its adoption in transportation-related applications. To address this issue, we propose a framework designed only with user-aggregated mobile phone data to synthesise realistic daily individual mobility — Digital Twin Travellers. We explore different strategies built around modified Markov models and an adaption of the Rejection Sampling algorithm to recreate realistic daily schedules and locations. We also define a one-day mobility population score to measure the similarity between the population of generated agents and the real mobile phone user population. Ultimately, we show how with a series of histograms provided by the telecommunication service provider (TSP) it is possible and plausible to disaggregate them into new synthetic and useful individual-level information, building in this way a big data travel demand framework that is designed in accordance with current data privacy regulations.

1. Introduction

New streams of geo-coded \textit{Big Data} allows us to observe and understand mobility behaviour on an unprecedented level of detail (Anda et al., 2017). From this array of data sources, mobile phone network data from the telecommunications service providers (TSP) has drawn special attention in the transportation field due to its pervasiveness, extensive coverage, and persistent collection. These records are generated in the telecommunications network as a result of user-related events such as voice calls and internet usage, and network-related events such as location area changes and periodical network updates. Once an event is triggered, a timestamp and the mobile device id are recorded in a cell tower, normally the closest to the mobile device. In such way that if the timestamps were filtered by mobile id, sorted by time of the day and plotted on a map, we would obtain a view of visited locations and trajectories of a mobile phone user. This enables a wide range of research towards using mobile phone data to improve travel demand models (Iqbal et al., 2014; Alexander et al., 2015; Toole et al., 2015; Molloy and Moeckel, 2017; Bwambale et al., 2019) and uncover insights into human mobility (Gonzalez et al., 2008; Song et al., 2010).

Conversely, the fact that individual daily mobility patterns can be reconstructed from a series of mobile phone data points has...
awakened growing concerns in regards to data privacy (Valentino-DeVries et al., 2018). People’s patterns of movement in space and time are highly dimensional, making mobile phone data a potent quasi-identifier for a single person (International Transport Forum, 2015). For instance, De Montjoye et al. (2013) found that even for data with a temporal resolution of an hour and a spatial resolution equal to the density of cell towers, just four spatio-temporal points were sufficient to isolate and uniquely identify 95% of the individuals.

The emergence of Big Data has pushed forward recent updates on data protection and privacy regulations. New provisions and requirements that extend the responsibilities related to personal data processing have been included in different regulations. For example, recital 26 of the European General Data Protection Regulation (GDPR) (European Commission, 2018) states that the principles of data protection should apply to any information concerning an identified or identifiable natural person or to personal data which have undergone pseudonymisation which could be attributed to a natural person by using additional information. This means that ‘anonymising’ mobile phone network data is not enough to comply with the GDPR (similarly with other privacy regulations like the Singapore Personal Data Protection Act (PDPA) (Personal Data Protection Commission Singapore, 2012)). The result is an additional barrier-to-adopting mobile phone data into transport planning since it becomes a challenge to strike a balance between data privacy and data utility.

One possibility to comply with these regulations is to group data in such a way that individual records no longer exist and cannot be distinguished from other records in the same grouping. In this scenario, we are only able to request user-aggregated histograms from the TSPs. Therefore, what we present here are some ideas on how mobile phone data utility can be maximised under this circumstance. With the premise of disaggregating aggregated data into an alternative but representative mobility population, we have developed a framework built around modified Markov models and an adaption of the rejection sampling algorithm to produce realistic synthetic individual level mobility data at an urban scale. The output of this work can be used, for example, to set up agent-based transport simulations (Anda et al., 2018; Bassolas et al., 2019).

The remainder of this paper is organised as follows. In Section 2, we review the related work on travel demand and generative models. Section 3 provides an overview of the general framework. In Section 4, we introduce the modified Markov models along with a baseline Markov model. Section 5 shows the clustering strategy used to segment mobile phone users. Section 6 explains how the models and strategies were evaluated and shows the results obtained. Finally, Section 7 and 8 contain further discussion and the conclusions, respectively.

2. Literature review

The traditional approach to model travel demand is by means of the four-step travel model (de Dios Ortúzar and Willumsen, 2011). It is comprised by a trip generation step, where land-use related factors are employed to calculate the incoming and outgoing number of trips per aggregate zone level; trip distribution, which allocates those trips into origin–destination pairs following a gravity model function; mode choice, to compute the proportion of trips by transportation mode, generally by using a utility-maximisation choice model; and traffic assignment, which allocates the routes taken following the Wardrop’s principle of user-equilibrium. The final output is an estimate of travel demand aggregated at the zone level.

In recent years, a new transport forecasting approach has emerged based on agent-based simulations. Where instead of focusing on zones, the main axis of the analysis is the individuals and their travel and activity plans (Axhausen and Garling, 1992). For this disaggregated paradigm, travel demand models have been developed accordingly. Prominent examples are Bowman and Ben-Akiva (2001), where individual tours with activities and itineraries are constructed through a series of discrete choice models; and Arenzte and Timmermans (2000), where the choice of different daily plan aspects are modelled through a series of decision trees. In both cases the models are developed around household travel surveys, which include socio-demographic information, and fit within the class of utility-based travel demand models designed to respond to what-if scenario analysis due to policy, infrastructure, socio-demographical or behavioural changes.

Mobile phone data is a natural fit into the disaggregated modelling paradigm, however, rarely one has access to socio-demographic information which is required in travel survey-centric travel demand models. In addition, mobile phone data embeds other type of characteristics that can potentially improve the current transport planning capabilities. Two general approaches have been taken in regards to the use of mobile phone data in travel demand models, one is to propose ways to calibrate utility-based travel demand models with mobile phone data (Bwambale et al., 2019), the other is to develop new data-driven travel demand models. Following the latter approach, Jiang et al. (2016) proposed TimeGeo, a travel demand framework without travel surveys. In TimeGeo, home and work activity locations are inferred from mobile phone data first, then locations and schedules of flexible (i.e. secondary) activities are simulated using spatial and temporal mechanisms of human dynamics. Lin et al. (2017) similarly inferred primary activity locations from mobile phone data, but trained instead a Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) to model flexible activities. In both studies, access to individual-level mobile phone data was required, as well as the identification of home and work locations of mobile phone users. This can become problematic when transferred into practice, specifically in cases where stakeholders are required to comply with data privacy regulations.

A possible solution to the privacy problem is to use synthetic data as an alternative to real users’ data. This idea has emerged in the computer science domain as part of a line of research that focuses on different methods to ensure data privacy in user trajectory data. There are several notable examples with similar objectives to this work but with important differences. Isaacman et al. (2012) introduced WHERE, a probabilistic modelling approach to produce synthetic Call Detail Records (CDRs). Later, Mir et al. (2013) extended the approach to add a differential privacy mechanism (DP-WHERE) as a formal privacy-preservation guarantee. For both cases, synthetic CDRs are generated from a series of distributions related to home/work locations and call patterns. One of the key
differences with this study is that in mobile phone network data full location trajectories can be usually derived, whereas, in CDRs the location is only known at the time when a call is made. Hence, this paper can be considered an updated version of Isaacman et al. (2012).

Another relevant example is the work by Bindschaedler and Shokri (2016). Here, a Markov model is trained on a set of seed traces to synthesise plausible location traces. The model incorporates current location (i.e. the state) and a period of time random variable (morning, afternoon, evening and night) to predict/generate the next location following a semantic seed structure. Additionally, geographic and semantic similarity metrics were also defined to evaluate the performance of the synthetic traces. It is important to note that this framework was designed to generate realistic but independent location traces, thus, no guarantees or validation exist for a population of urban-wide mobility traces. Moreover, given how individual traces are sampled from seeds (i.e. adding noise to the Viterbi reconstruction step), if a population of traces were generated, it would most probably lose the statistical properties of the real population.

More recently, the work by Pappalardo and Simini (2018) looked into generating samples of realistic spatio-temporal trajectories while remaining as close as possible to several real population statistics. They proposed a two-step framework. First, a Markov chain (non-parametric) to transform a base diary with only home locations into a mobility diary containing additional abstract locations. Second, the use of an exploration and preferential return (parametric) model to allocate the geographical information of the abstract locations. Although the framework outputs individual mobility trajectories, as mentioned by the authors, these mobility trajectories exclude travel times. Furthermore, their validation excludes relevant aspects in transportation, like the distribution of tours/mobility motifs Schneider et al. (2013), making the accuracy of their results uncertain for transport-specific applications.

From the literature, it emerges the importance to continue developing new urban-scale travel demand models based on mobile phone network data. In particular, models that are capable of generating realistic and individual mobility trajectories, including locations, activity/stay times and durations, and travel times. All of these, while considering the implications of deploying them into real-world transport-specific applications, namely, transport-oriented validation results that confirm the data-utility for applications in this field, and the data-sharing constraints between practitioners and TSPs due to data privacy regulations.

**Position of our work.** As an effort to close the research gap towards the aforementioned direction, in this paper, we introduce the Digital Twin Travellers framework to synthesise individual travel demand from aggregated mobile phone data. Different to Isaacman et al. (2012) and Mir et al. (2013), our proposed framework is based on trajectories derived from mobile phone data where location updates are not exclusive to calls (i.e. CDRs) but also come from other user-intrinsic events (e.g. internet usage) as well as other user-extrinsic mechanisms (e.g. transition between cell-location areas). Also, as opposed to Jiang et al. (2016) and Pappalardo and Simini (2018), our approach is a fully data-driven non-parametric model that does not require the inference of home and work locations. It is also an integrated generative framework, meaning that the spatial and temporal components are not split (Pappalardo and Simini, 2018), and where the objective is to estimate a joint probability distribution of mobility patterns in a population. This allows, for instance, to recreate the number of trips per agent as part of the Markov process rather than from a predefined base itinerary or seed (Bindschaedler and Shokri, 2016; Pappalardo and Simini, 2018). It also ensures the statistical match among several distributions directly modelled. Finally, another important difference is the restriction of user-aggregated data as input. This is to align the framework to recent data privacy regulations and make it readily-available for real-world applications. Another benefit of relying uniquely on aggregated data is that the proposed framework serves as the foundation for a future iteration where formal privacy-guarantees can be added, similar to Mir et al. (2013). Bindschaedler and Shokri (2016) noted the drawbacks on using only aggregates of data to synthesise new mobility traces, in particular the result of unlikely traces that do not adhere to realistic mobility constraints. In this work, we explore different strategies in the generative modelling framework and sampling methods to compensate for this. These strategies aim to maximise data accuracy for transport applications while keeping efficient model designs with generalisation capabilities. It is also important to mention that generative modelling techniques have already being explored in the field of transportation, principally to produce synthetic populations (Sun and Erath, 2015) (Sun et al., 2018) (Borysov et al., 2019), simulate activity chains (Axhausen and Herz, 1989), and generate mobility patterns (Lin et al., 2017) (Ouyang et al., 2018).

Thus, the contribution of this work is threefold. First, we propose the non-parametric and fully data-driven Digital Twin Travellers framework to synthesise realistic spatio-temporal trajectories from aggregated mobile phone data. Second, we define a validation metric to compare synthetic populations of disaggregated mobility data with real ones, the one-day mobility population accuracy score, which is oriented to transportation-specific applications and which considers relevant distributions often not validated, such as the distribution of tours. Third, the exploration between model complexity and accuracy between different strategies within a generative modelling framework to synthesise realistic spatio-temporal trajectories. All of these, with the ultimate objective of enabling the usage of mobile phone data in disaggregated transport applications (e.g. agent-based simulations) while complying with recent data privacy regulations.
3. General framework

The objective is to synthesise realistic and individual daily travel demand in the form of stay-point trajectories, or as defined by Pappalardo and Simini (2018), mobility diaries containing the time and duration of the visits in the various locations without explicitly specifying the type of activity. More specifically, we define a stay-point as a location (i.e. the stay-zone) where a user remains for at least 15 min, along with its start time and end time (see Fig. 1). In terms of the four step model, the proposed work solves the trip generation and trip distribution steps in a disaggregated manner. All these under the premise that we have access to mobile phone data representative of the population, but this access is restricted to only user-aggregated queries. Hence, the output could be used to:

1. Generate the complementary individual travel demand data of users subscribed to a different TSP.
2. Generate a full digital twin population with complete one-day tours (i.e. Digital Twin Travellers).

We start by assuming that there exists a true distribution that describes the mobility patterns of an entire urban population. This true distribution (Eq. 1) encodes the joint probability distribution of the sequence of stay-points visited by an individual during one day.

\[ f_X(x) = P(X_1 = x_1, X_2 = x_2, ..., X_N = x_N) \]  

On Eq. 1, each \( X_i \) corresponds to a set of random variables that describe the spatial and temporal aspects of the \( i \)th stay-point. For example, \( X_i \) can be the tuple of random variables \([Z_i, St_i, D_i]\): the \( i \)th stay-zone, start time and duration respectively. \( x_i \) is the set of realisations of the random variables in \( X_i \). Each individual has a total of \( N \) stay locations throughout a day.

The main strategy is to approach as much as possible to this true distribution by constructing a proposal distribution \( g_X(x) \) that encloses the real distribution \( f_X(x) \). Then, follow up with an adaptation of the rejection sampling algorithm to improve over the model deficiencies and ultimately get individual mobility demand samples that have similar statistics as the ones in the real population.

3.1. Markov models

Given the complexity in Eq. 1, we require a modelling framework that is computationally feasible and simplifies the joint distribution by factorising it into a set of marginal and conditional probability distributions. We employ Dynamic Bayesian Networks to build different Markov models that can approximate \( f_X(x) \). These type of models are commonly used to describe the different states in a dynamical system. We interpret individual travel demand as a dynamical system, where the different stay-zones represent the different states of the individual. The essence of Markov models is that the immediate future state of a system depends only in the current state, thus, disregarding any information on previous states, this is called the first-order Markov property. For our case of interest, the first-
order Markov property represents an important constraint if we desire to model realistic daily tours. We represent daily tours as mobility networks derived from the stay-zones visited by an individual on a single day. Each node in the network represents a different stay-zone, and the links indicate the trips done between stay-zones. This is similar to the definition of mobility motifs in Schneider et al. (2013). As an example, if individual A wakes up at home (first state) and goes to work (second state), then the choice of the next state will be only dependent on the information of the current state (work state), making it unlikely to return to the exact same home location at any point of time, since this information has already being forgotten by the model.

Therefore, the models proposed differed in the introduction of strategies to mitigate the first-order Markov constraint and capture long-term dependencies, while keeping the models efficient and with generalisation capabilities (i.e. without increasing the Markov order). Eq. 2 introduces the general approximation Markov model \( g_X(x) \), composed by the marginal probability of the initial state \( X_1 \) and the product denoting the conditional probability of transitions, where the current state \( X_i \) only depends on the previous state \( X_{i-1} \). While all the proposed models follow this general form, each model has a particular definition of the Markov state and different added mechanisms for the state transitions \( X_i|X_{i-1} \) (specifics explained in Section 4).

\[
    f_X(x) \approx g_X(x) = P(X_1) \prod_{i=2}^{n} P(X_i|X_{i-1})
\]

3.1.1. User-aggregated data restriction via Maximum Likelihood Estimation

We meet the user-aggregated data requirement by setting a couple of conditions in the design of the random variables and estimating the models using the Maximum Likelihood Estimation (MLE) procedure. Random variables are required to be categorical and fully observable. This means that temporal and spatial continuous variables need to be discretised, and that the models proposed cannot include any type of latent variables. It is also worth mentioning that using MLE fits well with the extensive coverage and pervasiveness properties in mobile network data, assuming that the aggregates come from an extensive and unbiased sample, which is generally the case for at least one of the TSPs.

The first step in the MLE procedure is to construct the likelihood function. We can rewrite the marginal and transition probabilities in Eq. 2 as general conditional probabilities \( P(X_i|X_{i-1}) \) following the convention in Bayesian Networks, where \( U_{X_i} \) refers to \( X_k \) parents or dependants, \( i \) is the iterator for the stay-points and \( k \) is the iterator across the tuple of random variables that describe each stay-point. 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 over a set of parameters \( \Theta \) as:

\[
    L_\theta(\Theta : D) = \prod_{m=1}^{M} \prod_{i=1}^{n} P(X_i|m|U_{X_i}|m : \Theta)
\]

The second step is to make use of the categorical nature of the random variables to simplify the likelihood function. The different conditional probabilities \( P(X_i|U_{X_i}) \) can be represented as a series of tables with the parameters \( \theta_{k,x,u} \) being the entry values of those tables, where \( x \in \text{Val}(X_k) \) and \( u \in \text{Val}(U_{X_k}) \).

\[
    L_\theta(\Theta : D) = \sum_{X} \sum_{S} \sum_{U} \nu[u,x]\log(\theta_{k,x,u})
\]

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

Once we have constructed the log likelihood (Eq. 4) we can proceed with the formulation of the MLE optimisation problem to calculate the parameters \( \Theta \), as follows:

\[
    \hat{\Theta} = \arg\max_{\Theta} L_\theta(\Theta : D) \quad s.t. \quad \sum_k \theta_{k,x,u} = 1 \quad \forall \quad (k,u)
\]

This yields the following closed form solution:

\[
    \tilde{\theta}_{k,x,u} = \frac{\nu[u,x]}{\nu[u]} \quad \forall \quad (k,x,u)
\]

Eq. 6 means that for the Markov models designed under the conditions of the random variables being categorical and fully observable, the estimation of the parameters \( \hat{\Theta} \) via MLE results in counting the frequencies of the different events as described by the conditional and marginal probabilities. Hence, only requiring histograms or distributions where the data is user-aggregated.

3.1.2. Sampling

Having estimated \( g_X \) we can proceed with the generation of stay-point data by using forward sampling. This method of sampling consists in assigning an outcome to the marginal distributions and then continue sampling following the topological order in the

---

1. \( \text{Val}(X) \) corresponds to the set of possible values that the random variable \( X \) can take
2. Also note how we have changed the iterators on the number of samples \( m \) and the number of stay-points \( i \) in Eq. 3 for an iteration over all possible values of \( x \) and the parents \( u \) in Eq. 4
conditional probabilities. The sampling is stopped after one full day for an individual agent is simulated.

3.2. Rejection Sampling

The second step in the framework takes into account the ability of the joint probability distribution \( g_x \) to generate any number of samples, making it feasible to select only the ones that improve the distribution of daily tour networks. This can be achieved by using rejection sampling, where the goal is to draw samples from a complicated target distribution, in our case \( f_{\text{tour}} \), using a simpler proposal distribution, \( \tilde{g}_{\text{tour}} \), that envelops the target one. However, we do not explicitly know the density function of \( \tilde{g}_{\text{tour}} \) since it is encoded within \( g_x \). To this end, the original rejection sampling is adapted by using instead the empirical proposal distribution \( \tilde{g}_{\text{tour}} \) which is obtained from drawing a large pool of samples from \( g_x \). We calculate the envelope factor as:

\[
\lambda = \sup_x \frac{f_{\text{tour}}(x)}{\tilde{g}_{\text{tour}}(x)}
\]

(7)

where, \( \sup \) is the supremum operator or least upper bound and \( f_{\text{tour}} \) is the target distribution of daily tour networks, and proceed with the rejection sampling algorithm:

Algorithm 1. Rejection sampling algorithm

4. Modified Markov models for individual travel demand

4.1. Baseline Markov model

We define our baseline Markov model as introduced in our previous work (Anda and Ordoñez Medina, 2018b). It is an expansion of the time-dependent first-order Markov chain mobility model in Bindschaedler and Shokri (2016) with the additional features of a stay-point duration random variables dependent on the stay-zone and the hour of the day and an hourly discretisation of the temporal variables. Different to Bindschaedler and Shokri (2016), we do not aim to generate location traces based on seed samples, but directly sample from the joint probability distribution of the generative model. Other random variables included in our baseline Markov model are stay-zones, their start times, and the time of the first departure of the day (i.e. the first stay-point end time; consecutive end times are calculated in a deterministic way as the sum of of the stay-point start time with its duration). The selection of these random variables constitute the minimum required to recreate the sequence of locations and times of a single agent, including its travel times. The baseline Markov model is the joint probability distribution of these random variables factorised as follows:

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

(8)

with the auxiliary function,

\[
E_k = S_k + D_k, \quad k \geq 2
\]

(9)

where,

- \( Z \): Stay-zone
- \( S \): Stay-point start time
- \( E \): Stay-point end time
- \( D \): Stay-duration

The Markov property can be detected in the probability of stay-zone conditioned only on the previous stay-zone and previous end time. As mentioned before, this becomes an important restriction in the generation of realistic daily tour structures. Following we present two different model variations to capture longer dependencies in an efficient way without the need of increasing the Markov order.

4.2. Explore and Return Model

In Song et al. (2010) two principles to accurately model human trajectories were introduced: exploration and preferential return. The authors noted that in random walk models of human mobility the visitation probability should not be random and uniform in space but that there exists a tendency of decreasing the exploration of additional locations with time. Following this idea, we add an Explore/Return (XR) random variable to the baseline Markov model. This new random variable dictates whether the agent being sampled will explore a new stay-zone or will return to a previously visited one. It is dependent on the current stay-zone and the current end time, so as the 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 stay-zone related to a non-residential area). For this model, the transition probability is encoded as \( P(Z_k|Z_{k-1}, E_{k-1}, XR_k) \) and its implementation is as follows: if the agent chooses to explore (i.e. explore is sampled from XR), then the previously visited stay-
Set number of target samples
Initialise list of accepted samples

while accepted samples < target samples do
  Simulate $Y \sim g_X$; // obtain one sample from the proposed model
  Calculate $Y_{tour} | Y$; // extract the tour network from the sample drawn
  Simulate $U \sim Uniform[0, \lambda g_{tour}(Y_{tour})]$ if $U \leq f_{tour}(Y_{tour})$ then
    Append $Y$ to accepted samples ; // accept sample
  else
    Dismiss $Y$ ; // reject sample
  end
end
zones are filtered out from the original $P(Z_{k}|Z_{k-1}, E_{k-1})$ and the probabilities re-normalised; but if the agent chooses to return (i.e. return is sampled from XR), then only the already visited stay-zones are considered in $P(Z_{k}|Z_{k-1}, E_{k-1})$ and its probabilities re-normalised. The case that the agent stays in the same stay-zone is interpreted an exploration step since it is assumed that the agent finished his/her activity at the present location and will continue to a different location within the same stay-zone. Fig. 2a shows the graphical representation of the model.

4.3. Tour Explicit Model

The Explore/Return (XR) variable models indirectly the daily individual tour networks. Alternatively, we can introduce a random variable that directly reconstructs tour networks. This is the essence behind the Tour explicit (TX) model. We first define a way to encode tour networks as sequences of digits, where each digit corresponds to a stay-zone. In this way, different stay-zones are mapped to different digits while same stay-zones are mapped to the same digit. For instance, the sequence 01020 refers to an individual that started the day at location 0, then moved to location 1, back to location 0, then to location 2, and finalised the day back at the initial location 0. In this example, the semantics behind this type of network might refer to the activity chain [Home-Work-Home-Shopping-Home], where it is assumed that each type of activity is done at a different location. Second, we introduce the random variable $T$ that models the next digit in the sequence given the current tour-code sequence $tc$, the current time and the current stay-zone. This is $P(T|Z_k, E_k, tc)$. The implementation works as follows, a digit is sampled from $T$. This digit can be either any of the digits already in the current $tc$ or the next digit to the highest one in $tc$. If the digit sampled is not in $tc$ then it is interpreted as an explore transition step (similar to the Explore and Return model); else, the transition is done directly to the stay-zone corresponding to the digit sampled. There are two other changes in the conditional probabilities. One is to make the explore transition probability dependent additionally on $i$, which is a counter for the number of stay-points. The other one is to make stay-zone duration ($D$) dependent on the current tour sequence $tc$ as well. These augmented conditional distributions are $P(Z_{k+1}|Z_k, E_k, T, i)$ and $P(D_k|Z_k, S_k, tc)$ respectively. Fig. 2b shows the graphical representation of the Tour explicit model.

Table 1

Types of histograms required (BM: Baseline Markov model, XR: Explore and Return, TX: Tour explicit).

<table>
<thead>
<tr>
<th>Model</th>
<th>CPD</th>
<th>Histogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM, XR, TX</td>
<td>$P(Z_k</td>
<td>S_1)$</td>
</tr>
<tr>
<td>BM, XR, TX</td>
<td>$P(E_1</td>
<td>Z_1, S_1)$</td>
</tr>
<tr>
<td>BM, XR</td>
<td>$P(Z_k</td>
<td>Z_{k-1}, E_{k-1})$</td>
</tr>
<tr>
<td>TX</td>
<td>$P(Z_k</td>
<td>Z_{k-1}, E_{k-1}, i)$</td>
</tr>
<tr>
<td>BM, XR, TX</td>
<td>$P(S_k</td>
<td>Z_{k-1}, E_{k-1})$</td>
</tr>
<tr>
<td>BM, XR</td>
<td>$P(D_k</td>
<td>Z_k, S_k)$</td>
</tr>
<tr>
<td>TX</td>
<td>$P(D_k</td>
<td>Z_k, S_k, tc)$</td>
</tr>
<tr>
<td>XR</td>
<td>$P(XR_k</td>
<td>Z_k, S_k)$</td>
</tr>
<tr>
<td>TX</td>
<td>$P(T</td>
<td>Z_k, E_k, tc)$</td>
</tr>
</tbody>
</table>
6. Types of urban travellers

A complementary strategy that we investigated was the idea of learning a generative model for each type of traveller rather than a general model for the entire population. The intuition behind is to keep the randomness of the sampling process constraint to the space of a more homogeneous group of travellers, avoiding in this way the generation of synthetic travellers that are a mixture of different types.

To obtain different types of travellers, traditionally, travel demand is segmented according to socio-demographic information (e.g. worker, student and non-worker) (Kutter, 1972) and in more recent studies, according to certain personal attributes (e.g. licence ownership, rail discount card ownership) (Schlich and Axhausen, 2005) and activity chains (Schlich and Axhausen, 2005; Jiang et al., 2012). Unfortunately, neither personal attributes nor activity labels are often available in mobile phone datasets. The travel demand segmentation, hence, requires to be done using only the intrinsic mobility patterns in mobile phone data.

A straightforward option is to segment travellers based on their mobility tour networks. However, this approach carries two main drawbacks. First, there might be too many different types of tour networks (e.g. +100), resulting in not enough data to populate the required distributions. Second, tour networks do not necessarily distinguish different types of travellers. Take for instance three different individuals: Alice, which starts her day at home, goes to work at 9 am, then at 6 pm goes to the gym, followed by dinner at 8 pm and returning home at 10 pm. Alice’s tour network code would be “01230”, given that each activity has its unique location; Bob, also starts his day at home, goes to work at 9 am, does some shopping at 7 pm, and returns home at 9 pm. Bob’s tour network would be “0120”; and Carlos, who starts his schedule at home, goes to the park at noon, has lunch at a cafe at 2 pm and comes back home at 4 pm. Carlos’ tour network would be “01230”, which is the same as Alice. If we would manually cluster them into two groups, a more sensible choice would be to put Alice and Bob together (since they are both commuters with extra activities after work), and Carlos in the other group (a non-worker group).

In order to achieve a more meaningful segmentation of travellers based only on their mobile phone traces, we have previously proposed in Anda and Ordoñez Medina (2018a) clustering a set of variables that capture different traits of travel behaviour to obtain different types of travellers:

- **Stay-points mean duration.** The average duration across all stay-points of a user during one day. It is a proxy of the number of trips done by a person.

- **Stay-points standard deviation duration.** The standard deviation of the duration across all stay-points of a user. It differentiates users with homogeneous activity durations from users with non-homogeneous activity durations.

- **Bias morning-night.** The difference between the average of stay-points durations before noon and the average of stay-points durations after noon. It describes whether a person is an early traveller or a late traveller.

- **First departure.** Time of the day in which the user performs his/her first trip departure.

- **Last arrival.** Time of the day in which the user made his/her last arrival.

In Anda and Ordoñez Medina (2018a), we also tested a couple of clustering algorithms from which HDBSCAN (Campello et al., 2013) resulted in more representative clusters. HDBSCAN is a hierarchical extension to DBSCAN, a density-based clustering algorithm successfully being used in many real-world applications (Schubert et al., 2017). The hierarchical extension allows density-varying clusters, a more efficient implementation, and trades the neighbourhood distance parameter $\varepsilon$ in DBSCAN for the more intuitive minimum size of a cluster parameter.

In terms of a real-world application, a per-cluster version of the proposed methodology has the implication that the clusters need to be calculated by the data provider to remain congruent with the user-aggregated data constraint of the proposed framework. This means that data provider has to extract the set of features described, run the clustering algorithm, and provide the aggregated histograms by cluster. A less troublesome option for the data provider would be to produce the required histograms given existing costumer segmentation profiles (e.g. using available socio-demographical information). In that case, the results presented here, can be use as a reference of the level of improvement that can be achieved when following a per type of traveller implementation.

6. Evaluation

6.1. Dataset

The proposed methodology was tested using user-aggregated data from one of the major mobile network operators in Singapore. Raw mobile phone data is processed and converted to individual stay-point data by the operator. The histograms required for this study are aggregations on top of the stay-point data. Table 1 specifies each of the histograms required, their corresponding conditional probability distribution (CPD) and the model(s) they correspond to. All histograms relate to the 18th April 2017, a typical working Tuesday. They account for 50% percent of the total population (around 2.8 million users). In terms of the spatial resolution, they are aggregated into subzone planning boundaries\(^3\), which are divisions within a planning area centred around a focal point such as a neighbourhood centre or an activity node and with an average coverage of 2.3 squared km. A total of 315 subzones which cover the extension of the main island were considered. As for the temporal resolution, all histograms were aggregated on an hourly basis.

Histograms depending on the type of traveller were aggregated considering the 16 different clusters (7 related to commuters and 9

---

to non-commuters) reported in Anda and Ordonez Medina (2018a).

6.2. Experiments

The models proposed were trained on aggregates corresponding to 70% of the data and used to generate one million samples (i.e., synthetic travellers). We compared a series of distributions of the generated samples against the distributions of the 30% left out sample. We replicated each experiment setting five times in a Monte Carlo cross-validation (MCCV) fashion to report on the sensitivity of different data partitions. We conclude that the synthetic travellers behave similarly to the real ones if the model can replicate accurately the target distributions.

Following are the different configurations tested:

1. Baseline Markov model (BM)
2. Explore and Return model (XR)
3. Tour explicit model (TX)
4. Explore and Return for each cluster of travellers (XR-C)
5. Tour explicit for each cluster of travellers (TX-C)

Additionally, we have also included the DITRAS model (Pappalardo and Simini, 2018) as another baseline method. For the implementation of DITRAS we assumed the users’ home location to be the location of the first stay-point observed during the study day. The weighted spatial tessellation was defined as the subzone planning boundaries described previously, and the relevance of a subzone was calculated as the total number of connections to mobile towers within that subzone during the study day. As suggested, the preferential return constants were selected as \( \rho = 0.6 \) and \( \gamma = 0.21 \).

6.3. Validation metric

We use the \( W_1 \) Wasserstein metric (or Earth Mover’s Distance) to measure the distances between a series of target distributions \( q \) and the obtained ones \( p \). The \( W_1 \) metric can be interpreted as the minimum amount of work needed to transform one pile of dirt (i.e., one distribution) into another one. Hence, this minimum amount of work can be regarded as a metric to compare how much two distributions differ. In a more formal way, and considering the Kantorovich formulation from the optimal transport problem (Levina and Bickel, 2001), given two random variables \( X \) and \( Y \) with probability distributions \( p \) and \( q \) respectively, we want to find the optimal transport map \( f \), which is a joint distribution of \( X \) and \( Y \), that minimises the expectation under \( f \) of an arbitrary distance function between \( X \) and \( Y \). This is,

\[
W_1(p, q) = \min_f \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} d(x_i, y_j) \right] \tag{10}
\]

As our target and generated distributions come from histograms, we can consider the discrete case for the Wasserstein metric and express the expectation in Eq. 10 as,

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} d(x_i, y_j) \tag{11}
\]

where we assume that the support of \( X \) and \( Y \) is of same size \( n \). It is also more clear now how the Wasserstein metric relates to an energy function, where \( f_{ij} \) is the mass we want to transport and \( d(x_i, y_j) \) is the cost or distance required to move that mass. Furthermore, if we consider the constraints:

\[
\sum_{j=1}^{m} f_{ij} = 1, \quad f_{ij} \geq 0 \quad \forall i, j, \quad \sum_{i=1}^{n} f_{ij} = q_j, \quad \sum_{j=1}^{m} f_{ij} = p_i \tag{12}
\]

then Eq. 11 can be expressed as a linear program (Rubner et al., 2000).

In contrast to other ways of comparing two distributions (e.g. Jensen-Shannon divergence, Kolmogorov–Smirnov test), the Wasserstein metric takes into consideration an arbitrary distance function for the support of the distribution. This distance function can be chosen freely depending on the type of distributions to compare. In our case, we are handling different types of support: categorical (e.g. type of tour network), numerical acyclical (e.g. hourly duration) and numerical cyclical (e.g. hour of the day). As a way to provide one distance metric for all, and following Bindschaedler and Shokri (2016), we choose \( d(x_i, y_j) = 1_{|\mu|} \). This choice yields the following closed form solution for the optimisation problem in Eq. 10,

\[
W_1(p, q) = 1 - \sum_i \min\{p_i, q_i\} \tag{13}
\]

In order to compare the different experiments using a single score, we define the one-day mobility population accuracy score as,
\[ \text{accscore}(\text{pop}) = \left(1 - \frac{1}{K} \sum_{k=1}^{K} W_i(p^{(k)}, q^{(k)}) \right) \times 100 \] (14)

Fig. 3. Validation results. (a) Activity start time distribution (b) Activity duration distribution (c) Daily number of trips per agent distribution (d) Daily tour network distribution (e) Spatial error by hour of the day distribution (f) Total distance travelled (g) Activity space area distribution (h) Mobility entropy distribution.
where $K$ is the total number of distributions to compare, $p^{(k)}$ is the $k$th obtained distribution from the generative model and $q^{(k)}$ its corresponding target distribution. Since $W_1$ is bounded $[0,1]$, the range of Eq. 14 is $[0,100]$, where 100 is a perfect score for the mobility population. We simplify the one-day mobility population accuracy score by substituting Eq. 13 as follows,

$$\text{accscore}(\text{pop}) = 100 \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{n} \min(p^{(k)}_i, q^{(k)}_i)$$

(15)

Fig. 4. Spatial distribution error at 16:00 h.

Additionally, we were also interested in validating a notion of semantic similarity following Bindschaedler and Shokri (2016) and Ouyang et al. (2018). Semantic similarity arises from the idea that even though two different mobility traces might differ geographically, they could be semantically similar if the purpose of the visits and the way they happen are similar. However, the difficulty in doing so is that mobile phone data commonly lacks labels about the type of activities or the type of travellers. To this end, we make use again of the five temporal variables in Section 5 used to characterise different types of travellers. The objective is to compare how similar is the distribution of points in the 5-dimensional space of the generated population to the real one. The closer the models are to the target, the closer they are able to generate semantically similar mobility traces.

We then define semantic similarity (SS) as:

$$SS = (1 - W_1(p^{\text{clus}}, q^{\text{clus}})) \times 100$$

(16)

Where, $p^{\text{clus}}$ and $q^{\text{clus}}$ are the cluster membership distribution of the 5-dimensional space of the target population and the model distribution respectively. Cluster membership for the generated populations can be found in a supervised-learning way (e.g. using K-nearest neighbours) by using the centroids of the clusters as labels.
### Table 2
Validation results with 95% confidence interval.

<table>
<thead>
<tr>
<th>Model</th>
<th>Start time</th>
<th>Duration</th>
<th>Number of trips (12)</th>
<th>Tours top 100</th>
<th>Spatial error (24 h avg)</th>
<th>Distance travelled</th>
<th>Activity space</th>
<th>Mobility entropy</th>
<th>Accuracy Score</th>
<th>Semantic Similarity (%)</th>
<th>Model complexity</th>
<th>No. of parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM</td>
<td>0.0120 ± 2e-4</td>
<td>0.0090 ± 6e-4</td>
<td>0.1448 ± 8e-4</td>
<td>0.6532 ± 6e-4</td>
<td>0.0154 ± 2e-4</td>
<td>0.0905 ± 7e-4</td>
<td>0.1843 ± 1.3e-3</td>
<td>0.5552 ± 1.3e-3</td>
<td>79.19 ± 0.07</td>
<td>86.14 ± 0.06</td>
<td>1.40e6</td>
<td></td>
</tr>
<tr>
<td>XR</td>
<td>0.0111 ± 3e-4</td>
<td>0.0081 ± 7e-4</td>
<td>0.1289 ± 1.5e-3</td>
<td>0.3587 ± 2.1e-3</td>
<td>0.0270 ± 2e-4</td>
<td>0.0471 ± 1.3e-3</td>
<td>0.0559 ± 8e-4</td>
<td>0.1803 ± 9e-4</td>
<td>89.79 ± 0.1</td>
<td>86.85 ± 0.01</td>
<td>1.41e6</td>
<td></td>
</tr>
<tr>
<td>XR-C</td>
<td>0.0092 ± 3e-4</td>
<td>0.0185 ± 6e-4</td>
<td>0.0625 ± 8e-4</td>
<td>0.2198 ± 1.2e-3</td>
<td>0.0217 ± 5e-4</td>
<td>0.0296 ± 1e-3</td>
<td>0.0827 ± 1e-3</td>
<td>0.0907 ± 1.3e-3</td>
<td>93.32 ± 0.08</td>
<td>88.96 ± 0.04</td>
<td>2.44e6</td>
<td></td>
</tr>
<tr>
<td>TX</td>
<td>0.0081 ± 6e-4</td>
<td>0.0197 ± 9e-4</td>
<td>0.0074 ± 7e-4</td>
<td>0.0136 ± 8e-4</td>
<td>0.0207 ± 2e-4</td>
<td>0.0366 ± 1e-3</td>
<td>0.0742 ± 9e-4</td>
<td>0.0080 ± 1.2e-3</td>
<td>97.65 ± 0.08</td>
<td>94.95 ± 0.02</td>
<td>3.25e6</td>
<td></td>
</tr>
<tr>
<td>TX-C</td>
<td>0.0059 ± 4e-4</td>
<td>0.0134 ± 6e-4</td>
<td>0.0032 ± 3e-4</td>
<td>0.0085 ± 7e-4</td>
<td>0.0179 ± 3e-4</td>
<td>0.0355 ± 1.3e-3</td>
<td>0.0745 ± 1.2e-3</td>
<td>0.0036 ± 7e-4</td>
<td>97.97 ± 0.07</td>
<td>95.90 ± 0.01</td>
<td>4.17e6</td>
<td></td>
</tr>
<tr>
<td>DITRAS</td>
<td>0.0498 ± 9e-4</td>
<td>0.1574 ± 8e-4</td>
<td>0.1359 ± 7e-4</td>
<td>0.1961 ± 5e-4</td>
<td>0.1439 ± 4e-4</td>
<td>0.2940 ± 1e-3</td>
<td>0.2079 ± 1.1e-3</td>
<td>0.1709 ± 9e-4</td>
<td>83.05 ± 0.08</td>
<td>84.44 ± 0.02</td>
<td>1e5</td>
<td></td>
</tr>
</tbody>
</table>
Table 3
Validation results after Rejection Sampling.

<table>
<thead>
<tr>
<th>Model (RS)</th>
<th>Start time</th>
<th>Duration</th>
<th>Number of trips (12)</th>
<th>Tours top 100</th>
<th>Spatial error (24 h avg)</th>
<th>Distance travelled</th>
<th>Activity space</th>
<th>Mobility entropy</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>XR</td>
<td>0.0238</td>
<td>0.0458</td>
<td>0.0467</td>
<td>0.0034</td>
<td>0.0269</td>
<td>0.0254</td>
<td>0.0550</td>
<td>0.0124</td>
<td>97.01</td>
</tr>
<tr>
<td>XR-C</td>
<td>0.0160</td>
<td>0.0436</td>
<td>0.0214</td>
<td>0.0030</td>
<td>0.0269</td>
<td>0.0285</td>
<td>0.0712</td>
<td>0.0088</td>
<td>97.26</td>
</tr>
<tr>
<td>TX</td>
<td>0.0103</td>
<td>0.0186</td>
<td>0.0143</td>
<td>0.0078</td>
<td>0.0212</td>
<td>0.0390</td>
<td>0.0768</td>
<td>0.0058</td>
<td>98.03</td>
</tr>
<tr>
<td>TX-C</td>
<td>0.0064</td>
<td>0.0148</td>
<td>0.0045</td>
<td>0.0021</td>
<td>0.0188</td>
<td>0.0342</td>
<td>0.0742</td>
<td>0.0030</td>
<td>98.03</td>
</tr>
</tbody>
</table>
6.4. Results

Activity start time. Activity start time refers to the initial time of a stay-point—the time in which a person starts a new activity. This can also be interpreted as the trip arrival time if we assume that a stay-point happens immediately after the end of a trip. Fig. 3a shows the activity start time distribution, where the x-axis represents the hour of the day and the black-coloured plot represents the target distribution. We can observe that all models follow the target distribution closely.

Activity duration. Activity duration is the amount of time an individual spends at a particular location. Fig. 3b shows the activity duration distribution, where the x-axis is the time spent in hours, from 0-h duration (i.e. less than 1-h duration) to 20-h duration, and the black plot refers to the target distribution. We can observe that all proposed models have a close match to the target distribution. Together with the activity start time results, this means that the fundamental structure of the BM model, on which the other proposed models are built on top, effectively captures the sequence of temporal transitions. For the case of DITRAS, we can observe an underestimation of short-term activities, which are normally captured by mobile phone data.

Number of trips. Number of trips is the total number of trips performed by an individual during a day. Fig. 3c shows the distribution of number of trips for the different models tested as well as the target distribution (in black colour). We can observe that both the BM and the XR model are unable to emulate the 2-trips peak and instead overestimate agents with three to five trips when compared to the target distribution. The XR-C model, on the other hand, slightly underestimates a–2 trips and misses to generate enough agents with 3–4 trips. Both TX and TX-C have a closer match to the target distribution, while DITRAS has a 2-trips overshoot and an underestimation of agents with more than five trips (similar results are reported in the original paper for CDR data).

Tour network. As mentioned previously, one of the disadvantages of using Markov models is the lack or inefficiency in handling long term dependencies, thus, resulting in non-realistic daily tour networks for every agent. In Fig. 3d, we show how our models perform in terms of the top 12 daily tour network distributions. Here, the x-axis indicates the numerical code for each of the top 12 tour networks. As expected, the BM model shows a poor performance without being able even to generate the typical home-work-home (i.e. 010) network. The XR model starts capturing the return-to-home behaviour for the tour networks ending in 0. As we analyse the results for the rest of the proposed models, we can see that the XR-C outperforms the version without clusters, TX yields better results than the XR-C, and finally, the TX-C shows the best results out of all the proposed models. For the case of DITRAS, we can observe an inclination towards more typical tour networks (e.g. 010, 0120), but falls short when replicating more complex tour networks or when an inter-zone trip occurs (e.g. 0100, 0110). For a quantitative comparison, we present the W1 metric results in Table 1. On which, we have extended the evaluation of the daily tour network distribution to include the top 100 networks which account for more than 85% of all tour networks presented in the hold out 30% sample.

Spatial error. The spatial error is the W1 measure of the distribution of agents in each subzone, for every hour of the day. It indicates whether the agents generated are in the correct locations at the correct times as compared to the real population. Fig. 3e shows how this error develops across the day for the different models proposed (comparison figure with DITRAS can be found in A). We can observe that the BM model outperforms the other ones with a lower hourly error throughout the day. This result reflects one of the trade-offs between the BM model and the other models proposed; the BM model results in more accurate spatial distributions of locations in exchange for unrealistic individual tour networks. The XR model, on the other hand, improves on the tour network distribution at the cost of lower accuracy in the spatial error distribution (up to = 4% error). For the other proposed models, as the complexity increases, the spatial error distribution starts converging on average to the BM error, while keeping more accurate tour network distributions. Additionally, we have also included the BM spatial error when compared to the training data (grey plot) to further analyse how the spatial error is decomposed in the proposed models. The validation of the BM model against the same training data shows the part of the spatial error that comes from the dynamics of the Markov model and the discretisation of the time variables—the model error (grey area). With the model error we can proceed to identify the part of the spatial error that comes from validating against new data, regardless of the models proposed (yellow area). The third and last component of the spatial error is the

<table>
<thead>
<tr>
<th>Model</th>
<th>Rejection Sampling top 100 tours</th>
<th>Rejection Sampling all tours (14610)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tour networks coverage (%)</td>
<td>λ</td>
</tr>
<tr>
<td>BM</td>
<td>100</td>
<td>3094.91</td>
</tr>
<tr>
<td>XR</td>
<td>100</td>
<td>52.88</td>
</tr>
<tr>
<td>XR-C</td>
<td>100</td>
<td>35.87</td>
</tr>
<tr>
<td>TX</td>
<td>100</td>
<td>2.92</td>
</tr>
<tr>
<td>TX-C</td>
<td>100</td>
<td>1.26</td>
</tr>
</tbody>
</table>

The resulting graphs and analysis were done from a single Monte Carlo experiment since we found almost no meaningful variation across the different data partitions. Nonetheless, the average results of all Monte Carlo experiments along with a 95% confidence interval are reported in Table 2.
Transportation Research Part C 128 (2021) 103118

16

Fig. 4. One introduced by the added mechanisms to mitigate the first-order Markov property (measured from the BM error to the model of interest error). In Fig. 4 we present how the spatial error is distributed geographically (by planning subzones) at 16:00 h, which is the hour of the day with the highest spatial error. From the colour gradients, red means more individuals in the subzone than in the validation data, while blue means less individuals in the subzone than in the validation data. The first thing to notice is the significant spatial error in DITRAS; this is due to the gravity model being the mechanism driving the exploration of new locations. As for the proposed models, we can observe that at 16:00 h the XR and XR-C models locate more agents in some residential areas than required and less in the downtown area. This is improved in TX and TX-C. Table 1 presents the 24-h average of the spatial error for each of the models.

Distance travelled. Distance travelled is the total distance travelled by every individual. It is calculated as the sum of the euclidean distances between each of the centroids of the subzones visited. Fig. 3 f shows the distribution results, where the x-axis refers to the total km travelled. We can observe that both the BM model and DITRAS perform poorly, the former with a tendency of large distances travelled and the latter with a significant overshoot in short distances travelled. The rest of the models show a closer match to the target distribution. It is worth pointing out that a similar metric often used to characterise trip distances by an individual is the radius of gyration (Gonzalez et al., 2008; Pappalardo and Simini, 2018). However, the one-day generation scope of this study makes the radius of gyration highly correlated to the total distance travelled, making it a redundant metric if included. Nonetheless, we have calculated it and presented its results in B.

Activity space. Activity space is the spatial area of activity opportunities known by the individual, which includes locations which have been registered and seen, but not necessarily visited yet (Schönfelder and Axhausen, 2003). We calculate the activity space area of an individual $u$ as the two-dimensional confidence ellipse area described in Schönfelder and Axhausen (2003), given by,

$$ A(u) = 6\pi |S|^{1/2} $$

$$ S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} $$

$$ S_{xy} = S_{yx} = \frac{1}{N-2} \sum_{i=1}^{N} (x_i - x_m)(y_i - y_m) $$

$$ (17) $$

Where, $x_i$ and $y_i$ are the coordinates for location $i$, $x_m$ and $y_m$ are the centre of mass coordinates for individual $u$, and $N$ is the total number of locations visited by individual $u$. Fig. 3g shows the activity space area distribution for individuals/agents with more than two distinct locations. We can observe that BM tends to generate agents with larger activity space, while DITRAS generates agents with smaller activity space. The rest of the proposed models follow the target distribution more accurately, with a slight bias towards agents with larger activity space areas.

Mobility entropy. Mobility entropy captures the degree of predictability of an individual’s whereabouts; it characterises the heterogeneity of visitation patterns (Song et al., 2010). We calculate the mobility entropy of an individual $u$ as the temporal-uncorrelated entropy in Pappalardo and Simini (2018),

$$ S^{mc}(u) = -\sum_{i \in L^u} p_i \log(p_i) $$

$$ \frac{\log|L^u|}{log|L^u|} $$

$$ (18) $$

Fig. 5. Model performance vs. complexity. Y-axis denotes the accuracy score of the model, X-axis denotes the total number of model parameters, the diameter size denotes the log transform of the expected number of rejections per acceptance, and the cross markers refer to the increment in accuracy after rejection sampling is applied.
Where, $p_u$ is the probability that individual $u$ visits location $i$ and $|L^{(u)}|$ is the cardinality of the set of locations visited by individual $u$. 

Fig. 3h shows the mobility entropy distribution results. We can observe that BM is the model that produces agents with the highest mobility entropy, while DITRAS is unable to replicate users with lower mobility entropies (e.g. users that only perform inner-subzone trips). The tour explicit models follow similar mobility entropy distribution as in the target.

**Accuracy score and semantic similarity.** Aside from the W1 metric for all distributions, Table 1 also presents the one-day mobility population accuracy score, the semantic similarity metric and the number of parameters for each of the models. The accuracy score was computed taking into account activity start time, activity duration, top 12 number of trips, top 100 tour networks, 24-h average spatial error, distance travelled, activity space, and mobility entropy distributions. The TX-C model achieves the maximum accuracy score of 97.97% and the maximum semantic similarity of 95.90%. As expected, the worse performing model is BM due to the first order Markov property. For XR, XR-C, and TX models we can see gradual improvements as the model complexity grows. DITRAS scores 83.05% in accuracy and 84.44% in semantic similarity, the main problem being the gravity model unable to discover realistic new locations, and its lack of travel times and inner-zonal trips. Still, we acknowledge that it is a good alternative when there is not enough data to populate some of the data-intensive distributions from the proposed models.

6.5. Rejection sampling

Another metric for model comparison, exclusive to generative models, is the rejection sampling efficiency. We report this efficiency using the envelope factor $\lambda$, which can also be interpreted as the expected number of samples needed in order to get one accepted sample. The larger $\lambda$, the more samples needed, meaning the proposal distribution is farther away from the target. We implemented the rejection sampling algorithm as explained in Section 3.2. We chose the target to be the distribution of the top 100 tour networks in the holdout sample. The proposal distribution was calculated empirically from pools of 10 million samples per model. The result of each model $\lambda$ is reported in Table 4.

We can observe that the BM model requires around 3,095 samples to get an accepted one. This makes it computationally expensive and not viable in many situations. The XR models obtained $\lambda$ values of 52.88 (no clusters) and 35.87 (with clusters), making them more realistic options for rejection sampling. However, the most efficient models in terms of rejections sampling were the TX models with 2.92 (no clusters) and 1.26 (with clusters) values. Aside from reporting the $\lambda$ values, it was also relevant to show whether the models were capable of generating at least one instance of each of the top 100 tour networks in the proposal distributions. We define this property as the tour network coverage. As observed in Table 4, all models had a tour network coverage of 100%.

We have also included an analysis of rejection sampling if we would consider all tour networks available in the hold out sample (14610 in total). This allows us to understand and compare the generalisation power of each model, as some of the tour networks in the test set might not be included in the training set. We observe that the tour network coverage under this setting varies from 93.96% (BM) to 97.43% (TX-C).

Table 3 shows the final results when applying rejection sampling on the top 100 tour networks for all models except the BM model. As expected, the W1 metric for the top 100 tours distribution dropped down to virtually zero for every model. As for the other distributions, there was either some further improvement (e.g. number of trips) or no relevant change. In terms of the models, we can see substantial improvements in the accuracy score of the XR ones, but not so much for the TX ones. The reason is simply that the TX models had already good performance in the Top 100 tours distribution before rejection sampling.

6.6. Model complexity vs performance

Fig. 5 shows the relationship between accuracy score, complexity, and rejection sampling efficiency of all the models. Here the y-axis indicates the accuracy score of the models before rejection sampling, the x-axis indicates the number of model parameters, the area of the circle marker relates to the ratio of number of rejections per acceptance (i.e. $\lambda$). For XR and XR-C we also included the change in accuracy after rejection sampling, indicated by the cross markers. In general, we can see that models improve in accuracy and rejection sampling efficiency as the complexity grows larger. However, it is important to note that the largest improvement in accuracy given the least increased in number of parameters and without doing rejection sampling happens from the BM to the XR model. Furthermore, the XR model can substantially enhanced its accuracy score without adding complexity by doing rejection sampling. This makes the XR model an appealing option when considering both accuracy and model complexity.

7. Discussion

As observed in Table 2 and Fig. 5, there is a direct correlation between accuracy score and model complexity. However, when analysed by individual metric, we can find certain trade-offs between the models proposed. For instance, BM has the least spatial error, while having the worst accuracy score. We can also observe that the tour network distribution defines mostly the model performance ranking and that by itself, it is a good metric to evaluate whether the model is able to mitigate the first-order Markov constraint. For these reasons, we select it as the target distribution in the rejection sampling process.

The idea of rejection sampling comes from the generative nature of the models being able to sample indefinite times, opening the possibility of taking only the ‘good’ samples. This means that virtually any model is useful as long as one has the computational power and time to produce the required number of accepted samples. 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 aimed to develop different model architectures to get as close as possible to the target distribution and have plausible \( \lambda \) values.

The models proposed can be interpreted as non-parametric models. In this type of models, the number of parameters is not fixed since they not only depend on the model’s definition but also on the data set. It is clear to see how this flexible number of parameters can happen in our cases. If we consider for instance the TX models, for a specific data set we might have a particular tour network that might not happen for a different data set (think about two different data sets for different cities). Hence, depending on which tour networks can be found in the data set is the different parameters needed to define certain probability distributions. Non-parametric models are also more flexible since there is no need to make assumptions about the family of distribution that maps the data, but in exchange, require much more data than parametric models. This makes them suitable when Big Data is available. In Table 1, we have reported the number of parameters for each of the models, considering the experimental training data set used.

Another point to discuss is the possibility of overfitting, especially given that the number of parameters per model ranges from around 1 to 4 million. The first question to consider is whether we have enough input data. A quick approximation tells us we do. For a dataset of one million users in a single day we can extract 9 million data points. That is 9 data points per user considering a three-stay-point day (1 for the 1st subzone, 1 for the 1st departure, 2 subzone-transitions, 2 start times, 2 durations, and 1 for either the XR variable or the TX one). If the ratio number-of-datapoints/number-of-parameters may need to be higher, for instance, to cover the participation of single individuals, one could add up more days of data. The second question is how to avoid overfitting in models with a large number of parameters. In light of newer models (i.e. deep neural networks), it is a common approach to reduce generalisation error with larger models coupled with regularisation techniques (Kukacka et al., 2017). In fact, these types of models have shown their effectiveness in practice despite the number of parameters being generally even higher than the number of available training samples (Zhang et al., 2016). In our case, we argue that the regularisation effect comes from the simplified overall model architecture definition. Dynamic Bayesian Networks allow simplifying the joint distribution through the series of 1st order Markov transitions. This simplification strikes a balance with the non-parametric definition of the local probability distributions, making them able to generate realistic samples that are not included in the training set and thus, preventing overfitting. Evidence to this is our largest model TX-C, which despite the total number of parameters, still held more generalisation power (expressed as tour network coverage for all tours in Table 4).

Finally, we would like to address 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 efficiently encode the joint probability distribution over the whole sequence. However, adopting the deep learning approach makes it harder to develop open data pipelines following recent data privacy regulations, since access to individual data would be required to train such models. In contrast, the models proposed here are able to generate realistic and individual travel demand with only a series of user-aggregated information (i.e. a series of histograms) from the telecommunications provider, resulting in a less burdensome approach from the data sharing perspective. For instance, for the case of XR model, only seven histograms are required: the six histograms mentioned in Table 1 plus the target tour network histogram to be used in rejection sampling.

8. Conclusion

We effectively introduced a framework to synthesise realistic and individual travel demand form aggregates of mobile phone data — the Digital Twin Travellers. The idea of just having access to user-aggregated information goes along with the current data protection and privacy regulations. In a two-step framework, we first introduced two main architectures (Explore and Return: XR; and Tour explicit: TX) that by enriching the state information can efficiently mitigate the first-order Markov constraint and generate realistic daily tour networks. We showed that as the proposed models get more complex, they improve over the one-day mobility population accuracy score, which we also defined in this work. Nonetheless, we also show that if model complexity is an issue, the XR models can increase substantially its accuracy score by an adaptation of the rejection sampling algorithm. This makes the proposed approach useful in practice since one can be able to generate realistic individual travel demand by only requesting a series of user-aggregated histograms from the mobile phone operators. There are several directions in which the current framework can be extended: similar to Mir et al. (2013), the implementation of Differential Privacy in the current methodology to ensure that user re-identification is not possible and as a formal privacy-guarantee; extrapolation of the model for future scenarios; the combination of other data sources or supplementary mobile phone data to include mode of transport and socio-demographic information; and a study to measure the performance of the Digital Twin Travellers in different transport policy and infrastructure intervention scenarios in an agent-based simulation platform.

Acknowledgments

This research was conducted at the Singapore-ETH Centre for Global Environmental Sustainability (SEC), 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. The authors would like to thank DataSpark for providing the anonymised and aggregated mobile phone data in this study.

Appendix A. Spatial error comparison with DITRAS

Fig. A.6.
Appendix B. Radius of gyration

The radius of gyration indicates the characteristic distance travelled by a person during a period of observation (Gonzalez et al., 2008; Pappalardo and Simini, 2018). It is calculated as follows:

\[
r_g(u) = \sqrt{\frac{\sum_{i=1}^{N} p_i (l_i - l_{cm})^2}{\sum_{i=1}^{N} p_i}}\]

(B.1)

Where, \( r_g(u) \) is the radius of gyration of individual \( u \), \( l_i \) is a vector of coordinates for location \( i \), \( l_{cm} \) is a vector of coordinates for the individual’s centre of mass, \( p_i \) is the probability of the individual visiting location \( i \), and \( N \) is the number of locations visited by the individual. Fig. B.7 shows the radius of gyration distribution for all models.
References


