Computers, Environment and Urban Systems 84 (2020) 101526

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Population mobility modelling for mobility data simulation

Kamil Smolak[a,*], Witold Rohm[a], Krzysztof Knop[a], Katarzyna Siła-Nowicka[b,c]

[a] Institute of Geodesy and Geoinformatics, Wrocław University of Environmental and Life Sciences, Wrocław, Poland
[b] Urban Big Data Centre, University of Glasgow, Glasgow, UK
[c] School of Environment, University of Auckland, Auckland, New Zealand

ARTICLE INFO

Keywords:
Human mobility modelling
Privacy preservation
Movement trajectories
Mobility data

ABSTRACT

Mobility models have a broad range of applications in areas related to human movements, such as urban planning, transportation, and simulations of diseases spread. In the last decade, the extensive geolocated user trajectories collected from mobile devices allowed for more realistic mobility modelling, improving its accuracy. However, mobility data sharing raises privacy concerns, which in turn limits accessibility to the data.

In this paper, we propose a WHO-WHERE-WHEN (3W) model, an improved privacy-protective mobility modelling method for synthetic mobility data generation. Based on real trajectories, it produces artificial user mobility trajectories that simulate population fluctuations in a study area, and thus preserves the individual's privacy. The model simulates the individual spatiotemporal aspects of lives accurately, representing real population flows and distributions.

The proposed method was inspired by the Work and Home Extracted REgions (WHERE) algorithm, but we have extended it by considering the activity space and circadian rhythm of people. Furthermore, we propose a clustering approach to capture and reproduce the heterogeneous characteristic of mobility. We evaluate our model and compare its performance to the WHERE algorithm on the synthetic and real data test cases. Use of the 3W model improved the accuracy of population distribution reproduction by 35% measured using Earth Mover's Distance. The travel distances and the spatial distribution of the flows reproduced by the 3W model match input data with high accuracy. We also evaluate the level of privacy protection by comparing synthesised and input datasets. We find that no daily trajectory can be matched between input and synthesised datasets and the average length of the matching sequence of visited locations to contain only two locations.

1. Introduction

Understanding population mobility and behaviour are the basis for sustainable planning and resources management, which help to transform cities into more cost- and time-effective places. Several different data sources show potential for human mobility and behavioural studies. These are global navigation satellite system (GNSS) trackers, credit card transactions and geolocated data from social media and mobile phones.

Novel data sources can provide information about whereabouts of single individuals in a form of movement trajectories collected for a long period. This raised interest in analysing human mobility in the most detail and completeness, mining the complete picture of individual mobility. For the purpose of this work, we refer to such type of data as individual movement trajectories. These are time-ordered sequences of coordinates corresponding to the locations visited by single individuals (Giannotti et al., 2011). Depending on the used tracking technology, harvested datasets differ by size, sampling frequency, spatial accuracy, bias and associated additional information (Fiore et al., 2019). Recently, mobile phone data have reached high popularity in mobility studies (Jiang et al., 2013). The main reason of their reputation is the ubiquitousness of mobile phones, hence the ability to track the whole populations at large span of time and relatively high spatial accuracy (Calabrese, Ferrari, & Blondel, 2014; Deville et al., 2014). However, some researchers have pointed to the biases of these data, caused by the even-triggered nature of mobile phone data (Zhao et al., 2016).

Mobility data have been successfully applied to many studies relating to human mobility at the collective and individual levels. The collective level involves analyses of population flows, focusing on groups of travelling individuals, their interactions and impact on local environments, and includes studies of population distribution mapping (Deville et al., 2014), land-use classification (Ros & Muñoz, 2017), dynamic traffic analysis (Calabrese et al., 2014), analysis of mobility

* Corresponding author.
E-mail address: kamil.smolak@upwr.edu.pl (K. Smolak).

https://doi.org/10.1016/j.compenvurbsys.2020.101526
Received 24 January 2020; Received in revised form 15 July 2020; Accepted 27 July 2020
Available online 13 August 2020

patterns (Siła-Nowicka et al., 2016), simulations of the spread of disease (Bengtsson et al., 2015) and community detection (Ratti et al., 2010). The individual mobility level studies include behavioural profiling (Furletti, Gabrielli, Rinzivillo, & Renso, 2012) and interaction analysis (Calabrese, Smoreda, Blondel, & Ratti, 2011).

Access to individual movement trajectories raises serious privacy concerns (Ahas, Silm, Järv, Saluveer, & Tiru, 2010). Therefore, data availability, especially mobile phone trajectories, is regulated in many countries by laws such as the European Union directive (European Commision, 2016), state laws in the United States (Snape, 2016; State of California, 2015) or China's Cybersecurity Law (Greenleaf & Livingston, 2016). In the European Union, where strict rules are in place, location data are considered personal even if they do not contain any personal information.

In fact, removing personal data does not fully preserve privacy because individuals can be re-identified using the uniqueness of their trajectories (De Montjoye, Hidalgo, Verleysen, & Blondel, 2013). It is therefore likely that new laws will be introduced in the coming years that will be even more restrictive, such as obligatory data processing and giving mobile phone users the right to refuse to share their data, including their trajectories (European Commision, 2017). Because many privacy-protection methods require data to be highly aggregated, or truncated to a short period, it limits their usefulness, often making it impossible to infer human mobility patterns (Fiore et al., 2019; Zang & Bolot, 2011). It also discourages researchers from conducting further studies, thereby reducing the number of published works in this field (Ahas et al., 2010).

We believe, however, that it is possible to retain the full potential of individual movement trajectories for collective population flow analysis without violating an individual's privacy. Our solution is to base a single user's mobility patterns on real trajectories but to represent them as artificial movement trajectories.

The complexity of such processing requires the development of a consistent modelling framework. This work introduces a new WHO-WHERE-WHEN (3W) mobility modelling method, which has two main features:

1. The production of a synthetic population, reflecting real flows and mobility statistics, and therefore protecting the individual's privacy; and
2. The flexibility of feeding the model with freely available information about a study area (census, taxi pick-ups and drop-offs) to derive easy-to-use human mobility in a form of movement trajectories.

We evaluated the 3W model using two datasets. First, we applied our algorithm to the large-scale synthetic data. Then, we used a set of real individual movement trajectories from Global Positioning System (GPS) trackers to verify the algorithm's ability to reproduce real-life population flows.

The rest of the paper is organised as follows. In Section 2, relevant research is described. The concept of the proposed model is presented in Section 3. In Section 4 and 5, the prepared test cases and evaluation methodology are introduced. Section 6 presents the results of the validation. A discussion of the potential of the 3W and future directions are discussed in Section 7. Finally, in Section 8, we provide conclusions from our work.

2. Relevant research

The goal of current studies on mobility data privacy is to satisfy the principle of privacy-preserving data publishing (PPDP), that is to process the trajectories before publishing in a way that they retain full usefulness and protects individual's privacy at the same time (Fung, Wang, Chen, & Yu, 2010). Many approaches to the privacy protection of mobility data have been made but none of the research provided a solution satisfying PPDP criteria (Fiore et al., 2019). It is clear that simple processing, such as reducing the spatial and temporal resolution of the data, does not preserve privacy and significantly affects data utility as well. The most promising approach, proposed first by (Rui Chen, Gergely Acs, and Claude Castelluccia, 2012), is to generate artificial trajectories using real individual movement trajectories. The general idea behind synthetic trajectories generation is the creation of some representation of original data, which is further used to generate artificial trajectories. The main advantages of this approach are ease of adoption of the most strict privacy criteria through noise introduction into the representations and the form of output data which is almost indistinguishable from real individual movement trajectories. Proposed algorithms can be divided by the trajectories representation used, which is either tree-based (Rui Chen, Gergely Acs, and Claude Castelluccia, 2012) or comprise of a set of distributions (Gursoy, Liu, Truex, & Yu, 2018; Roy, Kantarcioglu, & Sweeney, 2016). Privacy-protection level of these methods is sufficient, as the presence of particular individuals cannot be inferred from the produced output, however, synthesised data retain only a limited set of global properties and individual characteristics of movement are not replicated (Fiore et al., 2019).

(Isaacman et al., 2012; Mir, Isaacman, Caceres, Martonosi, & Wright, 2013) went a step further and proposed the Work Home Extracted REgions (WHERE), a privacy-protective human mobility model. It is based on the idea of spatiotemporal trajectories generation through the probability-distribution-based representation of original data. WHERE is an agent-based model of mobility, based on human-related aspects of mobility, such as home location and commuting distance. With that, it can be classified as a human mobility model, attempting to simulate human mobility and not only to replicate input data characteristics, which distinguish it from other privacy-protection methods proposed. However, WHERE can replicate only a limited set of collective mobility statistics, such as hourly population distributions and the daily range of distance covered by each agent and was designed to replicate mobile phone data only. Its practical realisation simulates flows among a preset number of places. The two- and three-place variants are named WHERE2 and WHERE3, respectively.

The task of human mobility models is to replicate and extrapolate the spatiotemporal characteristics of mobility trajectories. Models are used to simulate human mobility at various scales under imposed conditions to study the impact of human movement on various phenomena. Mobility models have many potential applications in areas where it is crucial to understand mobility characteristics, such as disease spreading (Bengtsson et al., 2015), traffic analyses (Calabrese et al., 2014) and utility demand forecasting (Smolak et al., 2020).

WHERE model is the only up-to-date attempt to the creation of a privacy-preserving mobility model. In this work, we extend and modify assumptions of WHERE model to increase its capability of spatiotemporal mobility characteristics of replication. Importantly, we capture additional spatiotemporal features of mobility and extend its application beyond mobile phone data. The 3W model is designed to sample and synthesise any kind of individual movement trajectories.

3. Model concept

Three aspects of human mobility were used as the foundation for the WHO-WHERE-WHEN algorithm. We refer to them as model components:

- Working HOurs shift groups (WHO) extracts groups of people having similar temporal mobility behaviour,
- Work Home Extracted REgions (WHERE) controls the spatial aspect of mobility,
- Work-HomE circadiaN rhythm (WHEN) controls the temporal aspect of mobility.

Below, we describe all of the model input data and parameters,

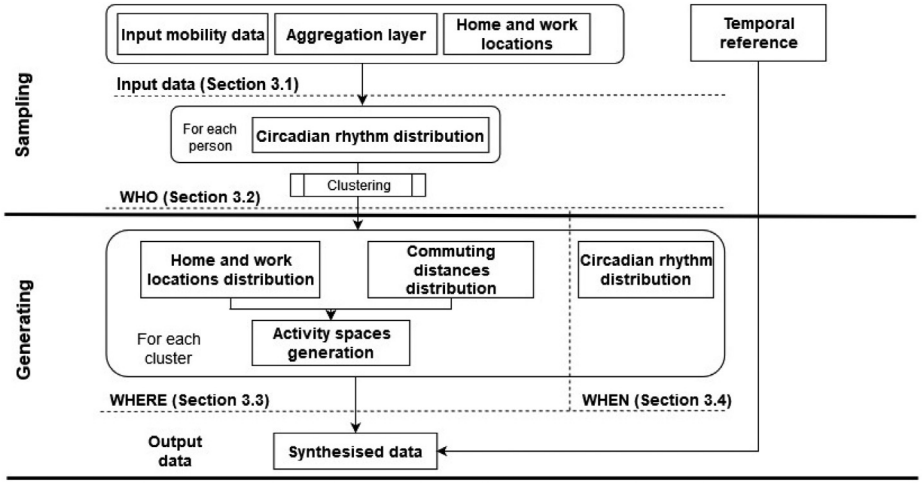

Fig. 1. Scheme of the 3W generation process.

including the data preprocessing steps. A general overview of the 3W generation process is presented in Fig. 1. The model can be divided into two phases - sampling, when all the distributions are calculated and generating when trajectories are synthesised. The WHO component is responsible for sampling, while the WHERE and WHEN are generating trajectories.

3.1. Input data

Before the generation process begins, the input data are used to estimate mobility-related probability distributions and parameters. In this section, we describe the required data, along with the preprocessing algorithms. These are a temporal reference file, spatial aggregation layer, input movement trajectories and home-work pairs detected for each person.

The temporal reference consists of timestamps required by the algorithm to embed data temporally. They determine dates and times when the location of the currently synthesised person is generated. This is an external file provided to the model. It also determines the generation time range and the temporal distribution of position logs.

Input mobility data files are individual movement trajectories which are sampled to deliver probability distributions on human mobility. It is possible to feed the model directly with probability distributions calculated from other sources, such as census data or taxi trajectories.

Input mobility data are chosen with respect to the aggregation layer, which is a spatial reference. This can be a regular grid of rectangles, but can also vary in shape to form, for example, hexagons or administrative units. The same aggregation layer has to be used in the sampling and generation steps. The resolution of a layer should correspond to the spatial accuracy of the input mobility data or the replacement data.

Meaningful locations for every human are known as regularly visited places that have a particular meaning for them, with a clear gradual pattern of visitation frequency (Song, Koren, Wang, & Barabási, 2010). The 3W model 'meaningful locations' refer to the top two most commonly visited places which are usually home and work locations (Ahas et al., 2007; Siła-Nowicka et al., 2016). Although, there are many methods to detect these locations (Ahas et al., 2010), none of them were proved to perform infallibly. Their accuracy may vary depending on the data and their scale. To eliminate the impact of home and work location detection algorithm on the results, we decided not to incorporate this algorithm into our model. The detected home and work pairs for each person in the input mobility data are given explicitly to the model at the beginning.

3.2. WHO component

Human movement patterns are associated with many variables such as socioeconomic status, social relations and trip purpose (Gabrielli, Furletti, Giannotti, & Nanni, 2015; Wesolowski, Eagle, Noor, Snow, & Buckee, 2013; Xu, Belyi, Bojic, & Ratti, 2018). The idea of the WHO component is to capture groups of similar mobility behaviour and their share in the whole population. We extract these groups by finding people having a similar circadian rhythm of movement.

The circadian rhythm is expressed by an empirical distribution *HWO*, divided into three categories: Home (H), Work (W) and Other place (O). Each of these three locations has an assigned probability of user appearance for each time window during the day. To derive the *HWO* distribution, input mobility data are analysed, as shown in Algorithm 1. First, the three empty vectors for home, work and other place are created. The length of each vector is determined by the preset temporal resolution. Next, the home and work locations are read from the provided information (see Section 3.1). The algorithm iterates through the user's trajectory, checking the time and place of appearance to account to a time slot in the one of the three vectors. When the end of the trajectory is reached, the common vector's time slots are stacked and normalised to represent the probability of appearance for each aggregation period.

Algorithm 1: HWO vectors calculation.

Data: Input mobility data $Input$, time resolution R, three empty vectors of R dimension H, W and O, home and work location in the aggregation layer $home, work$.
Result: HWO distribution filled with an appearance probability
for *user in Input* **do**
  for *position in user* **do**
    if $position = home$ **then**
      read $position[time]$;
      $H[round[position[time]]] += 1$;
    else if $position = work$ **then**
      read $position[time]$;
      $W[round[position[time]]] += 1$;
    else
      read $position[time]$;
      $O[round[position[time]]] += 1$;
for *time slot in R* **do**
  $HWO[timeslot] =$ normalise$([H[timeslot], W[timeslot], O[timeslot]])$
return HWO

Algorithm 1. HW Ovectors calculation.

The *HWO* is calculated for each person in the dataset. Next, all the distributions are fed to the K-means algorithm, which divides them into groups of similar circadian rhythms. The number of clusters to be

produced is estimated using Silhouette Coefficient criterion. This approach was used to cluster temporal patterns of human mobility from individual movement trajectories (Jiang, Ferreira, & González, 2012; Thuillier, Moalic, Lamrous, & Caminada, 2017). With the a being the mean intra-cluster distance and b being the mean nearest-cluster distance for each sample, the Silhouette Coefficient is (Rousseeuw, 1987).

$$SC = \frac{1}{n}\sum_{i=1}^{n}\frac{b(i)-a(i)}{max(a(i),\, b(i)}, \tag{1}$$

where n is the number of samples. Circadian rhythms are averaged in each cluster. The size of each cluster expressed as the number of *HWO* distributions in each of them is normalised and represents the share of each cluster in the population.

After clustering, for each extracted group a pair of spatial probability distributions of important places *HomeDistribution* and *WorkDistribution* is prepared. They determine the probability of finding a home or work place of each cluster in a particular location. Home and work locations of each person in each of the groups are aggregated into the previously derived reference layer. Next, the home and work location distributions are calculated and normalised across the entire aggregation layer, which results in a pair of spatial probability distributions. Using the approach presented by (Isaacman et al., 2012), a third spatial probability distribution of commuting distance is computed. This expresses the median distance that people from a particular home location are travelling to work. It is calculated and separately assigned to a home location of each individual. The median of the commuting distances is calculated for each aggregation cell, creating a *CommutingDistance* distribution for each of the groups. For each person, an ellipse is fitted to all the recorded activities. The length ratio of the semi-axes is taken as a *Spread* parameter, which is averaged for each cluster. It is required due to the adopted activity space simulation methodology (see Section 3.3). The process is presented by Algorithm 2.

Algorithm 2: Distributions calculations.

Data: Clustered input mobility data $InputClustered$, Aggregation layer $layer$, work and home locations $user_{work}, user_{home}$.
Result: $HomeDistribution$, $WorkDistribution, CommutingDistance, Spread$

for $group$ *in* $Input_{clustered}$ **do**
  for $user$ *in* $group$ **do**
    $worklayer[group] \longleftarrow layer.\text{count}[user_{work}]$
    $homelayer[group] \longleftarrow layer.\text{count}[user_{home}]$
    $CommutingDistance[group] \longleftarrow$ calculateDistance$(user_{work}, user_{home})$;
    for $position$ *in* $user$ **do**
      $AllActivities.\text{append}(user_{position})$;
    Find semi-axes length ratio $Ratio$ by fitting an ellipse E to $AllActivites$;
    $Spread[group].\text{append}(Ratio)$)
  $Spread[group] = \text{mean}(Spread)$

  for $cell$ *in* $worklayer[group]$ **do**
    $worklayer[group[cell[density]]] \longleftarrow cell[count]/cell.area$
  $WorkDistribution[group] = \text{normalise}(worklayer[group])$
  for $cell$ *in* $homelayer[group]$ **do**
    $homelayer[group[cell[density]]] \longleftarrow cell[count]/cell.area$
  $HomeDistribution[group] = \text{normalise}(homelayer)$
  for $cell$ *in* $CommutingDistance[group]$ **do**
    $CommutingDistance[group[cell[distance]]] \longleftarrow median(cell)$

return *Collection of* $WorkDistribution, HomeDistribution, CommutingDistance, Spread$

Algorithm 2. Distributions calculation.

3.3. WHERE component

The WHERE component starts a generation process. It is responsible for placing a person in the space of the aggregation layer. The total number of persons to be synthesised has to be determined at this step and is used to rescale the share of different mobility groups into the number of people to be generated for each cluster. For each generated person WHERE module assigns a pair of home and work locations and an activity space. Meaningful places are set using previously created distributions assigned to the currently synthesised group. First, the home location is selected using the *HomeDistribution*. Then, the average commuting distance, d, in the chosen location is taken from the *CommutingDistance*. Potential work locations are selected from the *WorkDistribution*, from inside a ring of a radius d with an origin in a home location. The ring width is determined by the mean aggregation cell size, creating an annulus. Chosen home and work locations are used to construct the activity space.

There are many possible approaches for computing activity spaces presented in the literature (Patterson & Farber, 2015). They can be divided into five categories: ellipses, minimum convex-hull geometries, kernel density approaches, network-based approaches and activity locations (Siła-Nowicka, 2016). The activity space computation method for a human mobility generation should be easily transferable and applicable to a common case. However, most of the methods are currently based on non-parametric approaches fit to an individual's trajectories (Patterson & Farber, 2015). In the 3W model, we decided to use the ellipse approach because it is a parametric and easily scalable method.

In the 3W, an activity space is constructed as an ellipse with the home location at its centre and the workplace at the edge (Schönfelder & Axhausen, 2003). Apart from the important places, all the locations inside the activity space are considered when predicting individual's position. The major semi-axis of the ellipse connects the two most important places and represents the commuting distance. The minor semi-axis length is set using a mean *Spread* ratio between the major and minor semi-axes of the ellipses of the current cluster (see Section 3.2 for explanation). The whole process is presented by Algorithm 3.

Algorithm 3: Creating users and assigning activity spaces to them.

Data: Collection of
$WorkDistribution, HomeDistribution, CommutingDistance,$
$Spread$, f factor, clustered input mobility data $InputClustered$
Result: $UsersSynthesised$
for *group in InputClustered* **do**
  for *user in group* **do**
    $user[home] \longleftarrow$ select home from $HomeDistribution[group]$;
    select d from $CommutingDistance[group][home]$;
      create $SearchRadius$ as a ring of $(1-f)*d$ inner radius and $(1+f)*d$ outer radius;
    create possible work place subset $WorkDistribution[group]_{subset}$ by selecting $WorkDistribution[group]$ within $SearchRadius$;
    $user[work] \longleftarrow$ select work from $WorkDistribution_{subset}$;
    $user[major\ semi-axis] \longleftarrow d$
    $user[minor\ semi-axis] \longleftarrow d * Spread[group]$
    construct $user[activity\ ellipse]$ from $user[work]$, $user[home]$, $user[major\ semi-axis]$, $user[minor\ semi-axis]$
return *Collections of users with assigned acitivity spaces* $UsersSynthesised$

Algorithm 3. Creating users and assigning activity spaces to them.

3.4. WHEN component

An activity space determines all the possible locations where a generated user can appear. In this step, movement between those positions has to be simulated, therefore, information about people's mobility routines is required. The entire generation process is represented by Algorithm 4 and depicted in Fig. 2. The algorithm iterates through the temporal reference (Fig. 2 at the top) and matches each given date and time with a proper time slot in the *HWO* (Fig. 2 in the middle) of a currently generated cluster, choosing a current position to be either home, work or other (Fig. 2 at the bottom). When 'other' is selected, a position is randomly chosen from the activity space. The random selection of location supports privacy protection for the individual. Lastly, the selected location is written down along with a current timestamp. When the end of the temporal reference for one user is reached, the algorithm moves to the next individual and repeats the process. At the end of the generation process, the output file contains synthesised individual movement trajectories where each row contains a unique identifier of a trajectory, a timestamp and assigned coordinates (see the top of the Fig. 2 for an example of the output file).

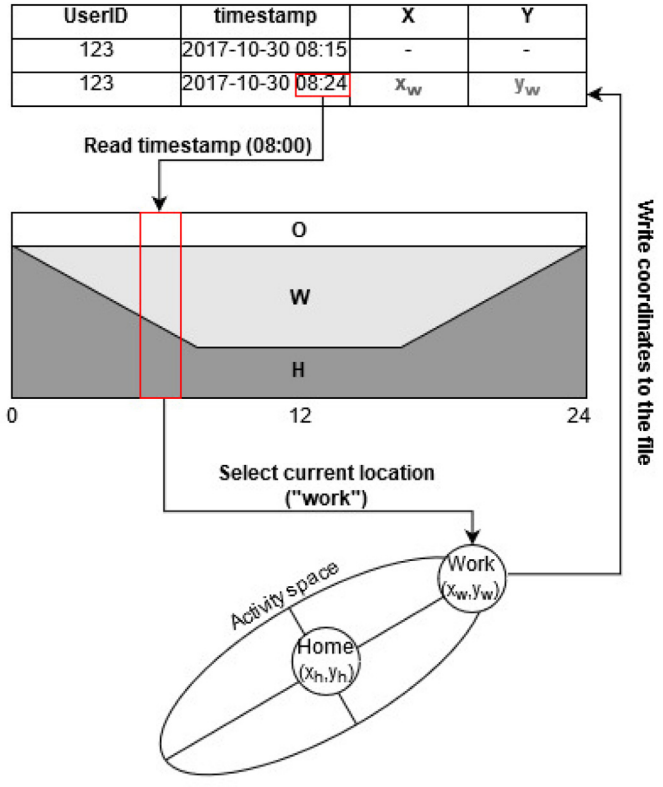

Fig. 2. Scheme of the 3W mobility simulation algorithm. The time from the temporal reference file is used to select the time slot in the *HWO* for a currently synthesised cluster. According to the selected probability, a current location is chosen. Its coordinates are written to the output file.

Algorithm 4: The mobility simulation algorithm.

Data: Collections of users with assigned activity spaces
$UsersSynthesised$, temporal reference $TemporalReference$,
HWO
for *group in UsersSynthesised* **do**
  for *user in group* **do**
    read $user[home], user[work]$;
    $others \longleftarrow$ select a set of all other places inside activity space;
    for *event in TemporalReference* **do**
      select position from $HWO[group[event[timestamp]]]$ in a given time slot;
      if *position = home* **then**
        $event[position] = user[home]$;
      else if *position = work* **then**
        $event[position] = user[work]$;
      else if *position = other* **then**
        $event[position] =$ random.choice$(other)$;
return *Output with assigned positions*

Algorithm 4. The mobility simulation algorithm.

4. Evaluation methodology

To compare the similarity of population mobility, we use information about the temporal and spatial distribution of the people. For that, we extract and summarise hourly peoples' positions from synthesised data and a reference dataset and convert them into hourly population distributions. The similarity between the two distributions can be quantified using one of the statistical distance measures. To maintain consistency with previous findings, we used the Earth Mover's Distance (EMD) method (Rubner, Tomasi, & Guibas, 2000).

Given two distributions, P, Q, with n and m clusters, respectively, the EMD is based on the minimal cost of transformation between P and Q on a given metric space. If d_{ij} is a ground distance and f_{ij} is a flow between an *ith* element of P and a *jth* element of Q, then we want to find the minimal cost:

$$J = min \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}. \tag{2}$$

When optimal flows (having minimal costs) are found, the EMD is defined as:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}. \tag{3}$$

For the calculations, we use the Fast EMD from (Pele & Werman, 2009). Since we compare the distributions with the same distance matrix, determined by an aggregation layer, the distributions are already normalised, and the results can be given directly in metres.

We also calculate the following collective mobility-related characteristics: 1) travel distances $P(d)$ distribution and 2) the pairwise comparison of the predicted and observed flows between units of the aggregation layer. These measures have been used as the mobility model validation measures at the population level before (Wang, Kong, Xia, & Sun, 2019; Yan, Wang, Gao, & Lai, 2017).

To calculate the distribution of travel distances, we extract a trip length (step distance) between consecutive records for each person in

each dataset. The similarity of these distributions to the same statistics drawn from the input mobility data demonstrates the ability of the model to reproduce movement at various spatial scales, from small trips to long journeys. To quantify the dissimilarity of the travel distances distribution and to keep consistency with current research in the area (Wang et al., 2019) we use Kullback-Leibler divergence measure. The divergence of probability distributions P and Q is defined as (Kullback & Leibler, 1951):

$$D_{KL}(P\|Q) = \sum_{x\in\chi} P(x)\, log\, \frac{P(x)}{Q(x)} \tag{4}$$

The second measure of mobility is the number of flows between units of the aggregation layer. We compare the number of incoming or outgoing trips with respect to geographic units. The similarity of these values indicates the ability of a model to estimate flows properly.

We measure the individuals' privacy protection by comparing the daily trajectories found in the original and synthesised data. We expect that no trajectory match both datasets, which would mean that any real movement trajectory is not included in the synthesised data. We consider the trajectories to be identical when they have the same sequence of consecutively visited locations in an individual's daily itinerary. We verify the maximum lengths of the matching sequences that can be found in both datasets. Furthermore, as a measure of similarity in activity spaces, we calculated the number of people whose most frequently visited locations are identical.

5. Empirical case studies

We evaluate our method by comparing its performance to the WHERE algorithm in two test cases at different scales. Firstly, we use the publicly available Census Tracks and New York Taxi Cab trajectories to generate a large-scale human mobility dataset. Secondly, we evaluate our algorithms using real individual movement trajectories, originally collected for the work of (Siła-Nowicka et al., 2016) in the form of GPS trajectories, and test the model's performance on a real and small-scale dataset. To evaluate the impact of each modifications introduced to the initial WHERE model, we apply three variants of the 3W algorithm, as presented in Table 1. The 2W variant is the simplest and extends WHERE model by the circadian rhythm (WHEN component) and activity space. The 3W-NS variant adds WHO component but it does not cluster spatial distributions. The 3W-Full variant additionally clusters spatial distributions and therefore, includes all the algorithms described in Section 3.

5.1. Synthetic data test case

Using publicly available data, we synthesise a dataset for the first test case. This approach provides a few advantages over using real mobility data for such tests: 1) knowledge of spatial and temporal distributions, such as home and work locations and circadian rhythms, originally hidden in the data; 2) an overview of the algorithm performance in capturing and reproducing random movement, giving a clear view of possible tendencies in the mobility imitation.

5.1.1. Generating mobility input from public data

We create an original dataset: home and work locations, commuting distance, circadian rhythm, activity spread and aggregation layer. As an aggregation layer for the test case and further calculations, we use census tracts from the Census Bureau's geographic database (Census Bureau, 2016). We select 1815 tracts in the area of the City of New York. To preserve the real distribution of the New York City population, we use census data to calculate the home and work locations. Also, we sample New York City cab trajectories to determine the commuting distances, which were used in the past to describe intra-urban mobility (Liu, Kang, Gao, Xiao, & Tian, 2012). At this point, we assume the commuting distance to be equal to a median trip length between the origin and the destination for each aggregation cell. To evaluate model's ability to capture and reproduce multiple mobility groups, we created four clusters of circadian rhythms, representing four distinct mobility behaviours. Each group of circadian rhythms is defined as a weighted mixture of Gaussian distributions from which a rhythm for each person is drawn. To increase the randomness of the synthetic dataset, we randomly select a *Spread* parameter value for each user. These data are used to generate synthetic dataset test case in a form of individual movement trajectories. Every user is assigned a one-month-long temporal reference with one record every hour.

5.1.2. Test case

Using the generated mobility dataset, we synthesise one month of data with the 3W (see Table 1) and WHERE2 algorithms for the same period and aggregation layer. We use the same temporal pattern for a generation as it was used when generating input mobility dataset. Using the same temporal pattern eliminates the impact of the temporal aspects on the results. Every generated test case contains 5000 users, which is consistent with other works on mobility models (Calabrese, Di Lorenzo, & Ratti, 2010; Isaacman et al., 2012).

5.2. Real data test case

In the absence of the real large-scale mobility data we use individual GPS trajectories, collected from 173 people from the Kingdom of Fife in Scotland, UK. In order to reduce potential bias, the participants were selected randomly from the overall population. Data were collected using i-Blue 747 ProS GPS loggers. The location of the individuals was stored every 5 s and the data consist of ID, latitude, longitude, elevation, date and time. The sub-sample selected for this study comprised 3,867,918 records, collected during a one-week-long observation campaign. We select a small subset of 28 people living in the area of Dunfermline and Edinburgh having at least four consecutive days of data. We down-sample the data to one-hour time-bins and fill the gaps in the trajectories using the last observed location. The data include trajectories of people living and travelling between towns in the area, and therefore represent inter-urban mobility behaviour.

In order to model population mobility, we aggregate data into a regular grid of 81 × 66 km to cover all the data we use in the study and divided it into 1 × 1 km squares. Using GPS trajectories and the aggregation layer, we synthesise four days worth of data for 28 people using the 3W (see Table 1) and WHERE2 algorithms. We calculate home and work location probability distributions, commuting distances, the average spread of activity space and a circadian rhythms using each participant's 'important places', as previously defined by (Siła-Nowicka et al., 2016). We use the same temporal pattern as in the synthetic test case to eliminate the impact of temporal aspects on the results.

Table 1
Capabilities of the tested 3W algorithm variants.

Variant	Capabilities
3W-Full	Uses WHERE and WHEN components, clusters circadian rhythms of population and spatial distributions
3W-NS	Uses WHERE and WHEN components, clusters circadian rhythms of population
2W	WHERE and WHEN components are used

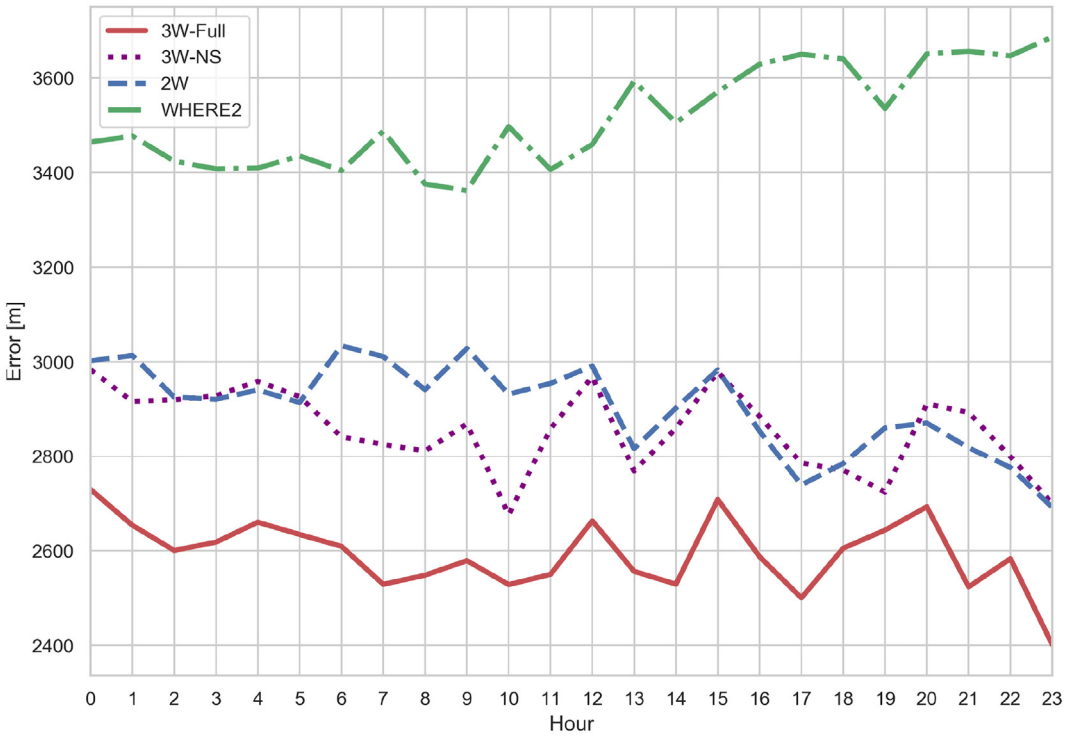

Fig. 3. Comparison of EMD distances for datasets synthesised using synthetic mobility data.

6. Results

In this section, we present the evaluation results based on the two experimental test cases. For each, we compare four different datasets, synthesised using three variants of the 3W (described in Table 1) and the WHERE2 model. All of the datasets are verified against original data by assessing the spatiotemporal distributions and descriptive statistics for similarity, and the privacy protection level.

6.1. Results: synthetic data test case

In this subsection, we report the results obtained from the synthesised data based on the synthetic test case. Due to the identical aggregation layer and temporal reference used for each dataset, including input data, the results can be compared directly.

6.1.1. Similarity of spatiotemporal distribution

We calculate the EMD to the population distribution of the input mobility data according to Section 5 and Eqs. (2) and (3) for each of the four synthesised datasets. Fig. 3 shows that, for the given test case, all the variants of the 3W model outperform the WHERE2 method, providing a more than 35% improvement in average accuracy. The 3W-Full dataset is almost one kilometre more accurate on mean position error, which is 2593 m for the 3W-Full dataset and 3515 m for the WHERE2 dataset. The gain of the 3W-Full dataset over datasets generated with 3W-NS and 2W variants corresponds to the impact of mobility groups extraction. Circadian rhythms and their share in the whole population recovered by the 3W-Full variant is identical to the input data used for artificial data generating. The datasets generated with 2W and 3W-NS variants have similar EMD across all the hours, with the 3W-NS dataset performing slightly better (average distances 2W: 2904 m, 3W-NS: 2857 m).

6.1.2. Similarity in collective mobility characteristics

Considering travel distances distribution $P(d)$ similarity measured with the Kullback-Leibler divergence, the data synthesised with the 3W-Full (0.0020) model outperforms other models (3W-NS: 0.0025, 2W: 0.0024, WHERE2: 0.0087). Simplified variants of the 3W model are only slightly worse than the 3W-Full model and significantly better than the WHERE2. The values of Kullback-Leibler divergence are small which can be seen in absolute differences $\Delta P(d)$ of travel distances distributions of the input and synthesised data in Fig. 4. Trips shorter than 5 km are better reflected by the 2W variant. Longer journeys are better reflected in the 3W-Full dataset. All the models usually overestimate the number of short trips and underestimate the number of long trips, however, the 3W-Full model is in the best agreement with the input mobility data.

The 3W-Full model corresponds to the input mobility data in terms of the number of incoming and outgoing flows to each unit of the aggregation layer (Fig. 5 a). The model reproduces low flows well and tends to slightly overestimate the high flows. The performance of simplified variants of the 3W model is slightly worse because of the larger overestimation of high flows present in these datasets (Fig. 5 b, c). The WHERE2 is unable to synthesise flows in the same locations as they are in the input data, resulting in a large underestimation of trips (Fig. 5 d).

6.1.3. Evaluation of privacy protection of individuals

We compare each combination of individual users daily trajectories between the input and the 3W-Full datasets. We find that the datasets do not contain any identical daily trajectories. The median length of the matching sequence appearing in the compared datasets contains two locations, which stand for 8.3% of a daily trajectory.

Dissimilarity of trajectories results from variations in the most frequently visited locations by the synthesised users. The number of users having this same set of locations is presented in Table 2. When considering two locations, these are selected from home and work locations distributions and hence, there is a higher probability of recurrence of the same combinations in both datasets. Further locations are selected at random and therefore, the number of the same most frequently visited locations drops significantly.

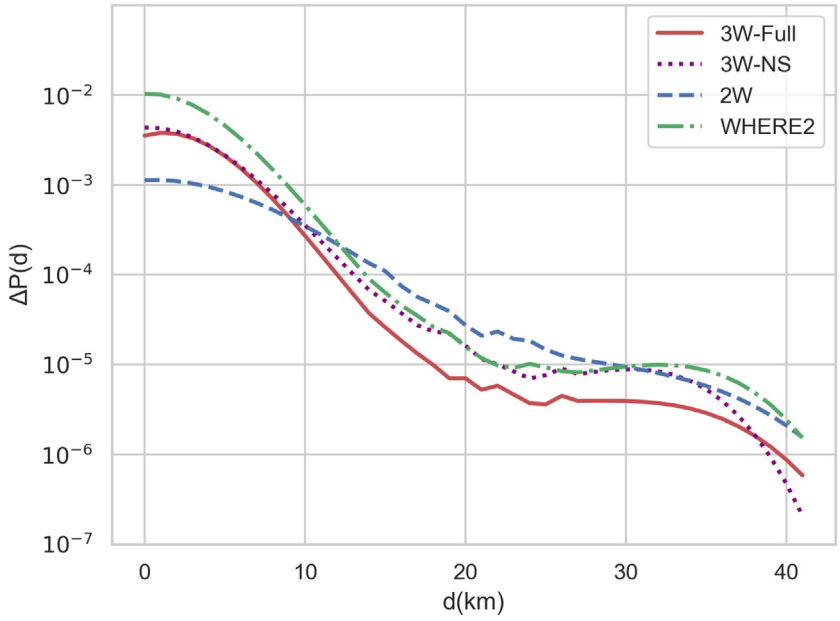

Fig. 4. The absolute difference between travel distances distribution of the input mobility data for the synthetic test case and four synthesised datasets.

6.2. Results: real data test case

In this subsection, we report the results based on the real GPS trajectories. They have been preprocessed as described in Section 5.2 and hence, synthesised datasets may be compared directly to the input mobility data.

6.2.1. Similarity in spatiotemporal distribution

As in Section 6.1.1, we calculate the EMD to the original distribution, according to Section 5 and Eqs. (2) and (3) for each of the four datasets. The EMD of the 3W-Full dataset varies from around 2 km for the night hours, when most of the people are at home, up to 6 km in the morning (see Fig. 6). Interestingly, the 3W-Full dataset has significantly

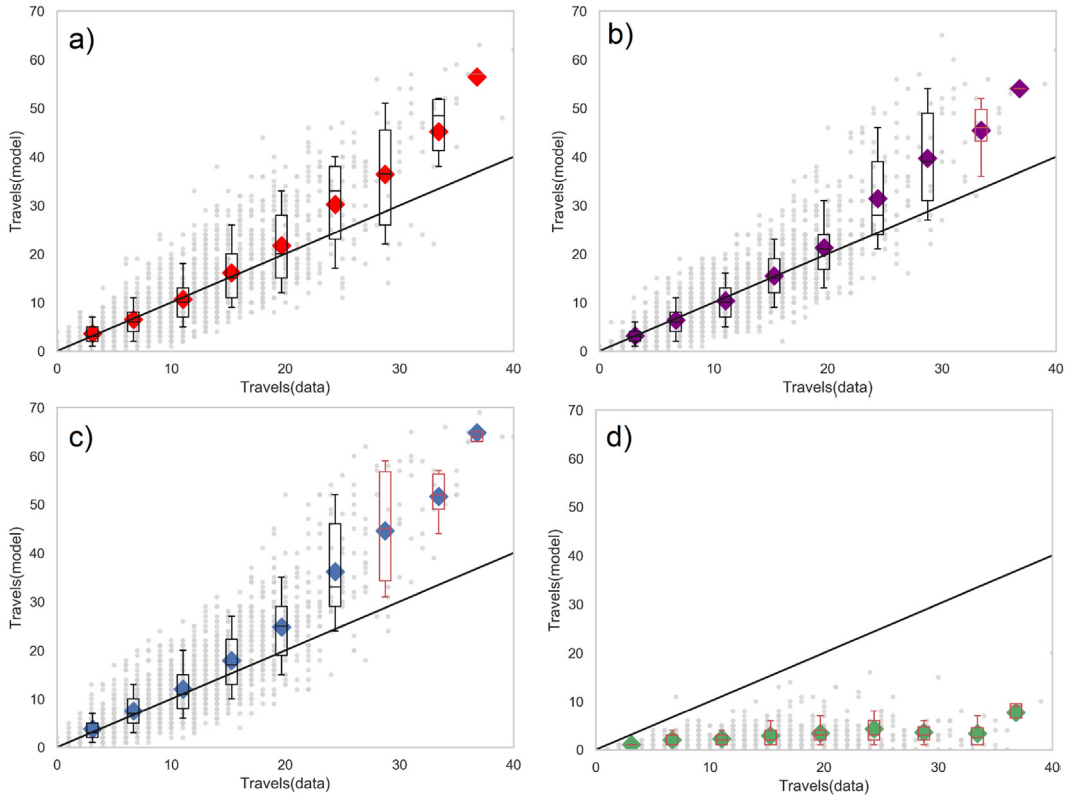

Fig. 5. Pairwise comparison of the flows observed in the input mobility data and datasets synthesised by 3W-Full (a), 3W-NS (b), 2W (c), and WHERE2 (d). Each grey point represents the number of flows calculated in the two datasets for a single location. The black line is x = y. The boxes depict the distribution of synthesised flows and the marker shows the average value of synthesised flows in that aggregation bin. If the line lies outside the range of 9th and 91st percentile (To keep consistency with recent research in the area we adopted the same criteria for model verification (Wang et al., 2019), the box is coloured red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
The number of people having identical sets of the most frequently visited places in the original and the 3W-Full datasets.

Places considered	People having the same set of most frequently visited locations (share of population)
2	264 (5.24%)
3	27 (0.54%)
4	4 (0.08%)
5	1 (0.02%)
6	0 (0%)

lower EMD for afternoon and evening hours than datasets produced by other 3W model variants and the WHERE2. All the variants of the 3W model perform better than the WHERE2. The 3W-Full model yielded, on average, more than 31% less position error than the WHERE2 (average errors: 3W-Full 3246 m, WHERE2: 4699 m). However, during the night hours when most of the people are staying in the most frequently visited location the performance of the WHERE2 is comparable to the 3W models and sometimes even slightly better. On average EMD of the 2W model (4512 m) and 3W-NS (4347 m) are only marginally lower than EMD of the WHERE2 model.

6.2.2. Similarity in collective mobility characteristics

The Kullback-Leibler divergence values for mobility-related characteristics confirm the superiority of the 3W-Full model (0.0855) in travel distances distribution reproduction, being 89% less distant from the target distribution than the WHERE2 (0.7375). According to this measure, the 3W-NS (0.1368) is the second and the 2W (0.1478) is the third-best performing model, which aligns with results obtained from EMD metric.

The differences of travel distances distributions $\Delta P(d)$ are presented in Fig. 7. The $P(d)$ of the 3W-Full is the most similar to the original data, very closely following the $P(d)$ of input mobility data in a range of 0 km to 30 km. The share of long-distance travels over 30 km is underestimated in the 3W-Full dataset, which results in the larger error. It is caused mainly by the presence of the travels of distance larger than 50 km, which are not present in the input data.

The 3W-NS dataset contains trips made on larger distances than those present in the input data and hence, the share of travels is underestimated for all the distances. The travel distances of 2W have the identical range as the input mobility data but underestimate the number of trips for all distances. However, the $P(d)$ of 3W-NS and 2W diverge less from the input data for large distances over 30 km. The $P(d)$ of WHERE2 is substantially different from the target. Moreover, trips longer than 45 km are not reproduced in this dataset.

According to (Fig. 8 a,b) the 3W-Full and 3W-NS variants are the best in reproducing the number of flows observed in the input mobility data. However, the number of flows in the locations with high flows in the input data is underestimated in the both model variants. The 2W model variant is performing slightly worse than other 3W variants (Fig. 8 c). Similarly to the synthetic test case, the WHERE2 model is unable to reproduce the flows in the same locations as they occur in the input mobility data, which results in a large underestimation of flows in the locations with medium and high numbers of flows (Fig. 8 d).

6.2.3. Evaluation of privacy protection of individuals

We find that no identical sequence of a minimum of two locations is present in the 3W-Full dataset and input data in the real data test case. We also calculate the number of users having the same set of most frequently visited locations and find that even considering two locations, no people have the same sets.

7. Discussion and further works

We compare the performance of the WHERE2 and three variants of the 3W model (3W-Full, 3W-NS, 2W) for synthetic (New York case) and real data (the Kingdom of Fife case). Using the EMD metric for the synthetic test case (Fig. 3), we find that the distance between the spatiotemporal distribution of the original and synthesised data is much smaller for all the variants of the 3W method throughout all the hours of the day. We find that the 3W-full method consisting of all the proposed components reaches 35% smaller mean position error than the WHERE2 model. This is related to the use a different approach to mobility simulation (Section 3). The 3W algorithm selects a current position inside the individual activity space of a user using *HWO* (Fig. 2), which simulates a circadian rhythm of the movement, while the

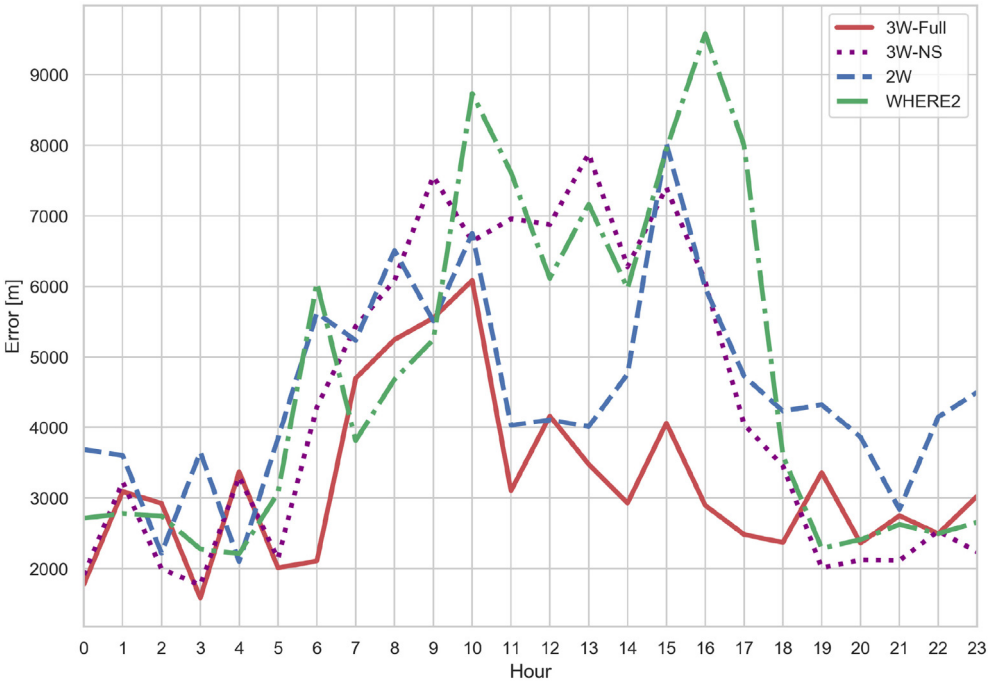

Fig. 6. Comparison of the EMD distances for datasets synthesised using real mobility data.

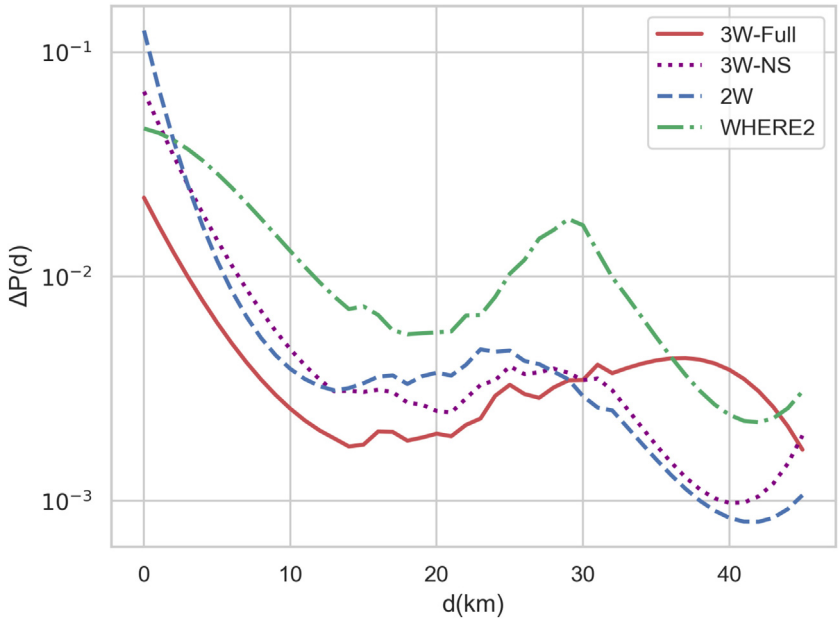

Fig. 7. The absolute difference between travel distances distribution of the input data for the real test case and four synthesised datasets.

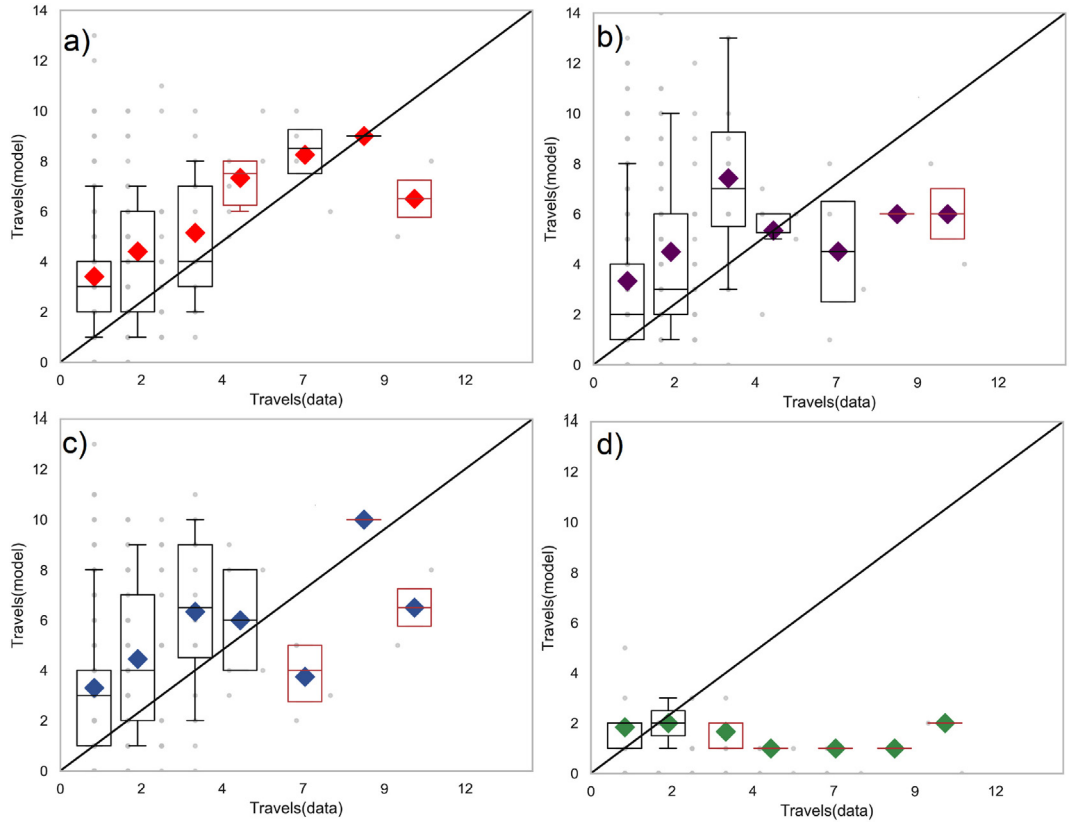

Fig. 8. Pairwise comparison of the flows observed in the input mobility data and datasets synthesised by 3W-Full (a), 3W-NS (b), 2W (c), and WHERE2 (d). Each grey point represents the number of flows calculated in the two datasets for a single location. The black line is x = y. The boxes depict the distribution of synthesised flows and the marker shows the average value of synthesised flows in that aggregation bin. If the line lies outside the range of 9th and 91st percentile (To keep consistency with recent research in the area we adopted the same criteria for model verification (Wang et al., 2019), the box is coloured red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

WHERE2 picks locations from the global population densities. The latter approach tends to choose places in more crowded areas that are not necessarily realistic locations for particular individuals. It is important to note, that individual activity space is simulated as an ellipse. This is a simple approach but it has limited flexibility, as it may cover areas where sampled movement trajectories have never appeared and on the other hand, may not cover areas where some positions were recorded.

The gain of the 3W-Full model over other variants stems also from applied clustering methodology. The EMD significantly drops when

both, spatial and temporal data are clustered. Using only temporal clustering provides very little improvement (compare 2W and 3W-NS in Fig. 3). It is important to note that four groups of the similar movement were created for the synthetic test case by providing only four different circadian rhythms and no separate spatial distributions. Yet, the EMD measure shows that spatial distributions are important in the modelling process, even if in that case they were randomly generated.

In the case of real mobility data modelling (Fig. 6), we gain more than 31% of improvement comparing 3W-Full and WHERE2. As expected for periods of lowest mobility, accuracy is highest during the night hours. However, using the 3W-Full model variant we observe the greatest accuracy gain for the middle daytime hours when people engage in various activities, other than staying in their home and work locations. Similarly to the synthetic test case, the importance of clustering home and work distributions according to detected circadian rhythms' groups is observable through the differences of EMD values of the 3W-Full and 3W-NS models.

In general, the average EMD values of the model variants align well with the incremental complexity of evaluated variants. The WHERE2 model can be considered the simplest and has the largest error. The 3W is built upon the WHERE2 idea and raises its complexity through additional components and so does the accuracy.

The EMD measure was used in (Isaacman et al., 2012; Mir et al., 2013), where the WHERE model was evaluated based on the mobile phone data dataset of 10,000 mobile phone users, using the WHERE2 and WHERE3 variations. The average EMD value for both models varied between 2 and 3 miles (around 3.22 and 4.82 km, respectively), which is similar to the result obtained in this work. An additional third place considered in the WHERE3 model slightly improved the accuracy. In this work, we extend the set of possible locations where a person can move by considering the activity space idea (Section 3.3). With that, the number of those places is not fixed and fluctuates depending on home and work locations.

We find that the 3W method reproduce a population whose collective mobility metrics are more similar to the original data than the WHERE2 output. Similarly to the EMD metric, the 3W-Full model reaches the highest similarity to the input mobility data distributions in the synthetic and real data test case. The travel distances distribution produced by the WHERE2 model highly differs from other distributions. Presumably, considering fixed locations forces mobility to occur at a limited set of distances imposed by the spatial distributions of home and work locations. In the real data test case, it can be observed as a significantly low share of medium distance travels. Adding an activity space introduces movement between many locations, which leads to many possible travel origin-destination combinations at a wide range of distances. All the evaluated 3W variants use activity space, therefore distributions reproduced by the 3W are significantly closer to the original one than the one obtained from WHERE2. The positive impact of mobility clustering is observable.

The pairwise comparison of the flows verifies the models' ability to synthesise the same amount of flows in the exact same location as in the input mobility data. Again, the 3W-Full model is performing the best. Other 3W variants are performing only slightly worse in all the test cases, with the more complex variants performing better. The WHERE2 model is not able to reproduce these flows, resulting in poor performance. This may be caused by the high number of self-transitions observable in the data (synthesised person stays in the same location for the next time bin), which are not considered as a trip.

The studies of (Wang et al., 2019; Yan et al., 2017) also used travel distances distributions and pairwise comparison of the flows to verify the model performance at the collective level of mobility. However, these works proposed a mobility modelling method based on a different concept, where human mobility is driven by an exploration and preferential return mechanism (EPR), that is a person movement occurs to a previously unvisited (exploration) or visited location (preferential return) with certain probabilities. Both works used different datasets for evaluation and while the Kullback-Leibler divergence values cannot be directly compared to each other the abilities of the EPR-based and 3W model to reproduce collective mobility characteristic can be compared.

In terms of travel distances distributions EPR-based models overestimate the number of short trips and long journeys. Similarly, the 3W model slightly overestimates the number of short trips but underestimates the number of long journeys. Also, the number of flows in locations where the low number of flows is observed in the input data is overestimated in the EPR-based models. In the case of the 3W model, these numbers are close (synthetic test case) or slightly overestimated (real data test case).

We also verify the privacy protection for individuals. We find that synthesised trajectories do not replicate the original data. Small parts of sequences are found to be identical in both test cases, but these are shorter than four consecutively visited locations. To compare the similarity of activity spaces in the original and synthesised datasets, we calculate the number of people having the same set of most frequently visited places. In the case of the synthetic data simulation, where 5000 users are simulated, 264 people (5.24% of users in the dataset) have two identical most frequently visited locations, and that number drops to four people (0.08% of users in the dataset) when considering four locations. Using the real invididual movement trajectories, we consider only 28 people, and we do not find two people having the same set of two most frequently visited places. Those values indicate a low similarity of user activity spaces between both datasets. This is the result of a random ‘other’ location selection, which introduce noise into the synthesised mobility data.

The 3W is an important step in the creation of privacy-protective mobility model. In comparison to previous works, its capabilities in mobility simulation are significantly improved and privacy protection abilities are retained.

The most important challenge for further development of the 3W model is to reproduce also individual characteristics of human mobility, which is necessary to truly retain the usefulness of reproduced data. To satisfy the requirements of PPDP the privacy-preserving mobility model should be able to replicate collective and individual mobility characteristics at all spatial scales. The integration of mobility mechanisms incorporated into recently presented unified mobility models (Wang et al., 2019; Yan et al., 2017), able to model human mobility at diverse spatial scales, may be a breakthrough in the pursue for mobility data anonymisation solution satisfying the PPDP principle.

Although in this work we decided to exclude home and work locations extraction process from the 3W model, it can be integrated with one of the known methods for detection of these locations. Such a method has to be at least able to detect meaningful locations in the trajectory and to rank them by their importance. There are many potential algorithms which can be used, based on various assumptions, such as statistical models (Ahas et al., 2010), machine-learning-based algorithms, from which a large portion of algorithms uses clustering methods (Isaacman et al., 2011; Nurmi & Koolwaaij, 2006), and mixed models (Siła-Nowicka et al., 2016). The two most important locations obtained from these algorithms can be considered home and work. The chosen method would probably have an impact on the accuracy of the model. In our work, home and work locations were already known in the dataset and therefore no home-work location extraction algorithm was used. It is important to note that some people have more than one home and work locations (Ahas et al., 2010) and at this moment the model is incapable of simulating such a scenario. Therefore, additional experiments should be run in the future.

8. Conclusions

In this work, we have proposed the WHO-WHERE-WHEN method, an improved privacy-protective population mobility model. Our model can synthesise artificial trajectories, representing a population of very similar movement behaviours to those of a real community. The

algorithm is designed to sample and reproduce data on a large scale.

The proposed model is a modification of the WHERE method, in which we mostly focused on the spatiotemporal aspect of urban-scale human mobility modelling. We have proposed a different approach that increases flexibility and improves the accuracy. It describes user mobility spatially using their activity space and temporally by adding the circadian rhythm. Furthermore, we cluster people by their mobility behaviour relying on their everyday movement routine to capture heterogeneous characteristics of the movement.

We validated our method using previously created test cases, with large-scale synthetic and real individual movement trajectories. Our results were compared to the results of the WHERE2 algorithm. We noted that the 3W imitates spatiotemporal population densities and flows with greater accuracy. Moreover, the results show that our algorithm better preserves collective mobility statistics.

This method contributes to the research devoted to overcome privacy issues, which limits the accessibility of mobility data. Its simplicity and capability of working on large sets of data allow for the synthesis of artificial user trajectories by imitating the mobility parameters of the real population. These can be further used in studies in place of real data, thus protecting the privacy of mobile phone users.

The 3W model was implemented in the Python programming language as a plugin for the QGIS platform. We have shared this tool in the GitHub repository (10.5281/zenodo.1240952).

Acknowledgments

The authors wish to thank the Editor and Reviewer for their effort and comments. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compenvurbsys.2020.101526.

References

Ahas, R., Aasa, A., Silm, S., Aunap, R., Kalle, H., & Mark, Ü. (2007). Mobile positioning in space—Time behaviour studies: Social positioning method experiments in Estonia. *Cartography and Geographic Information Science, 34*(4), 259–273.

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology, 17*(1), 3–27.

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., ... Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports, 5*.

Calabrese, F., Di Lorenzo, G., & Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences. *13th international IEEE conference on intelligent transportation systems* (pp. 312–317). IEEE.

Calabrese, F., Ferrari, L., & Blondel, V. D. (2014). Urban sensing using Mobile phone network data: A survey of research. *ACM Computing Surveys, 47*(2), 1–20.

Calabrese, F., Smoreda, Z., Blondel, V. D., & Ratti, C. (2011). Interplay between telecommunications and face-to- face interactions : A study using mobile phone data. *PLoS One, 6*(7).

Census Bureau (2016). *Census data*.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports, 3*, Article 1376.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*(45), 15888–15893.

European Commision (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general Da)*.

European Commision (2017). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL concerning the respect for private life and the protection of personal data in electronic communications and repealing directive 2002/58/EC (Regulation on privacy and electronic Commu)*.

Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Le Hello, D., ... Stanica, R. (2019). *Privacy of trajectory micro-data: A survey. arXiv preprint arXiv:1903.12211*.

Fung, B. C. M., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys (Csur), 42*(4), 1–53.

Furletti, B., Gabrielli, L., Rinzivillo, S., & Renso, C. (2012). Identifying users profiles from mobile calls habits. *Proceedings of the ACM SIGKDD international workshop on urban computing* (pp. 17–24). ACM.

Gabrielli, L., Furletti, B., Giannotti, F., & Nanni, M. (2015). Use of mobile phone data to estimate visitors mobility flows. *International conference on software engineering and formal methods* (pp. 1–13). .

Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., & Trasarti, R. (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB Journal, 20*(5), 695–719.

Greenleaf, G., & Livingston, S. (2016). *China's new cybersecurity law {−} also a data privacy law? Technical report 144*. University of New South Wales.

Gursoy, M. E., Liu, L., Truex, S., & Yu, L. (2018). Differentially private and utility preserving publication of trajectory data. *IEEE Transactions on Mobile Computing, 18*(10), 2315–2329.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., & Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. *Pervasive computing* (pp. 133–151). Berlin, Heidelberg: Springer Berlin Heidelberg.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012). Human mobility modeling at metropolitan scales. *Proceedings of the 10th international conference on mobile systems, applications, and services - MobiSys '12* (pp. 239). .

Jiang, S., Ferreira, J., & González, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery, 25*(3), 478–510.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E., & González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. *UrbComp '13 proceedings of the 2nd ACM SIGKDD international workshop on urban computing* (pp. 1–9). ACM.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79–86.

Liu, Y., Kang, C., Gao, S., Xiao, Y., & Tian, Y. (2012). Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems, 14*(4), 463–483.

Mir, D. J., Isaacman, S., Caceres, R., Martonosi, M., & Wright, R. N. (2013). DP-WHERE: Differentially private modeling of human mobility. *Proceedings - 2013 IEEE international conference on Big Data, Big Data 2013* (pp. 580–588). .

Nurmi, P., & Koolwaaij, J. (2006). Identifying meaningful locations. *2006 3rd annual international conference on Mobile and ubiquitous systems: Networking and services*MobiQuitous (May).

Patterson, Z., & Farber, S. (2015). Potential path areas and activity spaces in application: A review. *Transport Reviews, 35*(6), 679–700.

Pele, O., & Werman, M. (2009). Fast and robust earth mover's distances. *2009 IEEE 12th international conference on computer vision* (pp. 460–467). IEEE.

Ratti, C., Sobolevsky, S., Calabrese, F., Andris, C., Reades, J., Claxton, R., & Strogatz, S. H. (2010). Redrawing the map of Great Britain from a network of human interactions. *PLoS One, 5*(12).

Ros, S. A., & Muñoz, R. (2017). Land use detection with cell phone data using topic models: Case Santiago, Chile. *Computers, Environment and Urban Systems, 61*, 39–48.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65.

Roy, H., Kantarcioglu, M., & Sweeney, L. (2016). Practical differentially private modeling of human movement data. *IFIP annual conference on data and applications security and privacy* (pp. 170–178). Springer.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth Mover's distance as a metric for image retrieval. *International Journal of Computer Vision, 40*(2), 99–121.

Rui Chen, Gergely Acs, and Claude Castelluccia (2012). Differentially private sequential data publication via variable-length n-grams. *Proceedings of the 2012 ACM conference on computer and communications security* (pp. 638–649). .

Schönfelder, S., & Axhausen, K. W. (2003). Activity spaces: Measures of social exclusion? *Transport Policy, 10*(4), 273–286.

Siła-Nowicka, K. (2016). *Using GPS trajectories for further understanding of spatial behaviour*. (Unpublished results).

Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from gps trajectories and contextual information. *International Journal of Geographical Information Science, 30*(5), 881–906.

Smolak, K., Kasieczka, B., Fialkiewicz, W., Rohm, W., Siła-Nowicka, K., & Kopańczyk, K. (2020). Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models. *Urban Water Journal*, 1–11.

Snape, J. (2016). *California penal code 2016 book 1 of 2*. Lulu Press.

Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics, 6*(10), 818–823.

State of California (2015). *An act to add Chapter 3.6 (commencing with Section 1546) to Title 12 of Part 2 of the Penal Code, relating to privacy*.

Thuillier, E., Moalic, L., Lamrous, S., & Caminada, A. (2017). Clustering weekly patterns of human mobility through mobile phone data. *IEEE Transactions on Mobile Computing, 17*(4), 817–830.

Wang, J., Kong, X., Xia, F., & Sun, L. (2019). Urban Human Mobility. *ACM SIGKDD Explorations Newsletter, 21*(1), 1–19.

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface, 10*.

Xu, Y., Belyi, A., Bojic, I., & Ratti, C. (2018). Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems, 72*, 51–67.

Yan, X. Y., Wang, W. X., Gao, Z. Y., & Lai, Y. C. (2017). Universal model of individual and population mobility on diverse spatial scales. *Nature Communications, 8*(1), 1–9.

Zang, H., & Bolot, J. (2011). Anonymization of location data does not work: A large-scale measurement study. *Proceedings of the 17th annual international conference on mobile computing and networking* (pp. 145–156). ACM.

Zhao, Z., Shaw, S. L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science, 30*(9), 1738–1762.