Towards Anonymous Mobility Data Through the Modelling of Spatiotemporal Circadian Rhythms

Kamil Smolak*, Katarzyna Siła-Nowicka**, Witold Rohm*

* Institute of Geodesy and Geoinformatics, Wrocław University of Environmental and Life Sciences, ul . Grunwaldzka 55, Wrocław, Poland
** School of Environment, University of Auckland, 23 Symonds St, Auckland, New Zealand

Abstract. Recent years have seen an extensive exploration of the potential of mobility data. Mobility data are gathered from personal devices and as technologies advance, so does the volume and variety of data that they gather. Such data, however, bring about increased concern over the potential for revealing sensitive information. Although there have been many methods proposed for protecting mobile data privacy, they come at a price of either limiting data utility or the level of privacy protection. In this work, we present an approach that preserves global mobility-related data properties and at the same time protects privacy. Our preliminary results show that we can enhance the usability of synthesized data while proving no breaches of privacy. Our contribution can be considered as a new mobility modelling method as well as a privacy-protecting algorithm.

Keywords. Human Mobility, Data Privacy, Mobility Modelling

1. Introduction

The availability of movement trajectories has influenced rapid development in human mobility studies. Location data are now extensively collected through ubiquitous devices, such as mobile phones, fitness bracelets and location loggers. The high temporal and spatial resolution of these data unlocks the potential for many applications where human movement is an important factor to consider. Mobility traces have proven their significance, for example, in traffic forecasting, city planning and utilities management.

High data utility comes at a price of privacy disclosure. Simply removing personal details such as a name from a released dataset does not preserve

Published in "Adjunct Proceedings of the 15th International Conference on Location Based Services (LBS 2019)", edited by Georg Gartner and Haosheng Huang, LBS 2019, 11–13 November 2019, Vienna, Austria.

This contribution underwent double-blind peer review based on the paper. https://doi.org/10.34726/lbs2019.45 |

privacy, as in many cases a person can still be re-identified from a combination of attributes such as postcode, gender and age. Moreover, due to high uniqueness of human mobility trajectories, aggregation does not improve the privacy of the data and at the same time causes a loss of precision and limits data utility (Fiore et al., 2019). Hence, data accessibility is limited in many countries by laws such as the General Data Protection Regulation (GDPR) (European Parliament, 2016).

The goal of anonymisation is to protect the privacy of individuals and retain the utility of human mobility traces (Fiore et al., 2019). The two most commonly used groups of anonymisation methods are based on 1) k-anonymity and 2) uninformativeness and differential privacy principle (Mir et al., 2013; Fiore et al., 2019). k-anonymity of traces is achieved when a subset of spatiotemporal points of each person is indistinguishable from at least k - 1 other subsets. Nevertheless, it is not clear what k value is considered sufficient for full privacy protection. The anonymisation methods based on the uninformativeness and differential privacy principle assume that data are stored in a database and are accessible only through the limited subset of queries, which are modified to produce noisy outputs. Differential privacy is satisfied when an observer cannot tell when a particular person's data were used to produce a result. However, because the data themselves are not modified, they can be stored in this database but cannot be published. One of the alternative approaches proposed in the literature is to synthesize traces using the original mobility dataset (Mir et al., 2013). Synthesized data preserve global mobility properties, while the individual traces do not contain true information. Therefore, such data can be freely published but analysed only at a collective level.

2. Modelling

In this paper, we present an ongoing work to extend the Differential Privacy - Work Home Extracted REgions (DP-WHERE) model proposed by Mir et al. (2013). It draws probability distributions of various statistical features of human mobility from a dataset. Next, uses them to generate synthetical trajectories which can be freely published. We extend the DP-WHERE model into the DP-WHO-WHEN-WHERE model which includes spatial and temporal aspects of human mobility and captures multiple mobility behaviours. As a result, we improve the modelling process to achieve higher accuracy of the global properties of the mobility-related dataset. Similarly to the DP-WHERE model, derived distributions are modified by a noise creating the DP-WHO-WHEN-WHERE, a differential privacy protection algorithm.

Three aspects of human mobility, further named components, are used as the foundation of the DP-WHO-WHERE-WHEN model (Fig. 2). The first component, WHO (Working HOurs shift groups) is responsible for finding individuals with similar temporal mobility behaviour. The WHERE component controls the spatial aspect of mobility. The last component, WHEN (Work-HomE circadiaN rhythm) controls the temporal aspect of mobility and determines the circadian rhythm of each synthesized group of people.

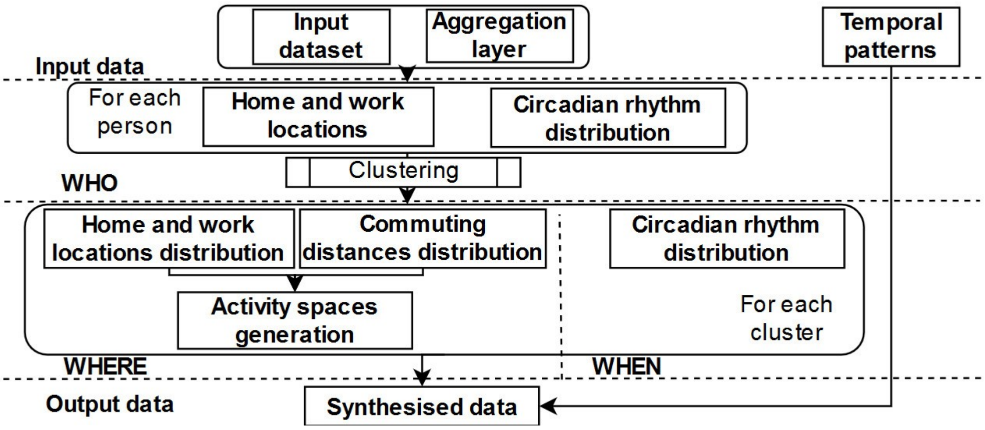

Figure 1. The proposed WHO-WHEN-WHERE modelling process.

The input data are composed of the input dataset, aggregation layer and temporal patterns. The input dataset consists of original mobility traces used to estimate the mobility-related distributions. The aggregation layer (i.e. grid-based partition) is used as a spatial reference for the data. The temporal patterns are used as a temporal reference for synthesised mobility traces. They determine the exact timestamp of a point in the trajectory and reflect the frequency of location updates.

The whole modelling process is repeated the number of times determined by the WHO component. First, home and work locations and the distance between them are determined for each person. These are used to construct his/her parametric activity space (also known as the use of space) in which a person spends most of his/her time. Any place apart from home and work locations inside the activity space is further considered as ‘another place’. Next, the person's circadian rhythm (empirical distribution) is drawn from his/her movement and divided into three categories: home, work and other places. These locations have each an assigned probability of a person appearance in a given time of day. It is calculated by counting and normalising the total number of appearances of each person in one of those locations throughout the whole period of the study. Using a clustering algorithm (for example Self-Organising Maps (Bianchi, Rizzi, Sadeghian & Moiso, 2016)) circadian rhythms are grouped to find people with similar temporal patterns of movement. In doing so, the population groups as well as their share in the whole population are determined.

For each cluster WHERE and WHEN components are calculated. First, the algorithm transforms the detected home and work locations into two spatial distributions. Next, a third spatial distribution of median commuting distance (median distance between home and work) is determined. The WHEN component calculates an average circadian rhythm in each cluster.

Mobility traces are synthesised using the distributions from each cluster. The spatial distribution of home locations is used to select the home location for each synthesised trace and the commuting distance is used to determine the distance in which the work location is selected. This step is identical to the original DP-WHERE model (for more information see Isaacman et al., 2012). From these, an activity space is constructed using one of the activity space approximation algorithms.

Each trace is synthesised accordingly to the temporal pattern from the input data. Single spatiotemporal point is generated in a repetitive process as follows: (1) a timestamp is read from the temporal pattern; (2) current location (either home, work or another place) is selected for a given time of day from circadian rhythm distribution; (3) coordinates of that place are read from the activity space. If another place is selected, it is randomly chosen inside the activity space. Coordinates and a timestamp are written as a single record with a random user identifier.

3. Preliminary Results and DIscussion

Number of clusters to synthesize increases computation complexity linearly. Therefore, at this stage of development, we evaluate the model without the clustering process, hence the average of distributions is calculated for the whole input dataset and it can be considered as WHEN-WHERE model. We compare its performance to the WHERE model in two cases (Test Case - TC1 and Test Case 2 - TC2).

TC1 is based on the 5000 mobility traces generated from the New York Taxi Cab and the US Census data for the city of New York. As an aggregation layer for the test case and further calculations, we use census tracts from the Census Bureau's geographic database. To preserve the real distribution of the New York City population, we use census data to calculate home and work locations. Also, we sample New York City cab traces to determine the commuting distances. We randomise each person's circadian rhythm. To eliminate the impact of the temporal aspects on the results, trajectories are synthesized with the same frequency as they were recorded.

For the TC2, we evaluate our model on the real data gathered by Global Positioning System (GPS) loggers from 173 people from the Kingdom of Fife in Scotland. This test case investigated the ability of the evaluated algorithms to capture and model real-life movement flows. To model population mobility, we aggregated the data into a regular grid of 81 x 66 km, divided into 1 x 1 km squares.

For the TCs, we compare hourly population distributions of the input and output datasets using the Earth Mover's Distance (EMD) measure (results shown in Fig. 2). When referenced to the same aggregation layer, quantified similarity of two spatial distributions can be interpreted directly in meters.

Two variants of the WHEN-WHERE model are evaluated, one full and one that does not extend the WHERE component with the activity space. In both cases, we note higher accuracy levels yielded by our model. For the TC1 there is more than 50% of improvement over the WHERE model. The WHEN-WHERE model is more than one kilometre more accurate on mean position error, which is 2272 m for the WHEN-WHERE and 3588 m for the WHERE. For the TC2 the accuracy varies from 2 km for the late evening hours, when most of the people are at home, up to 5 km in the morning. Our model continually performed better than the WHERE model. The WHEN-WHERE model produced at least 12% smaller position error than WHERE. High errors for the morning hours were caused by the gaps in the original GPS data.

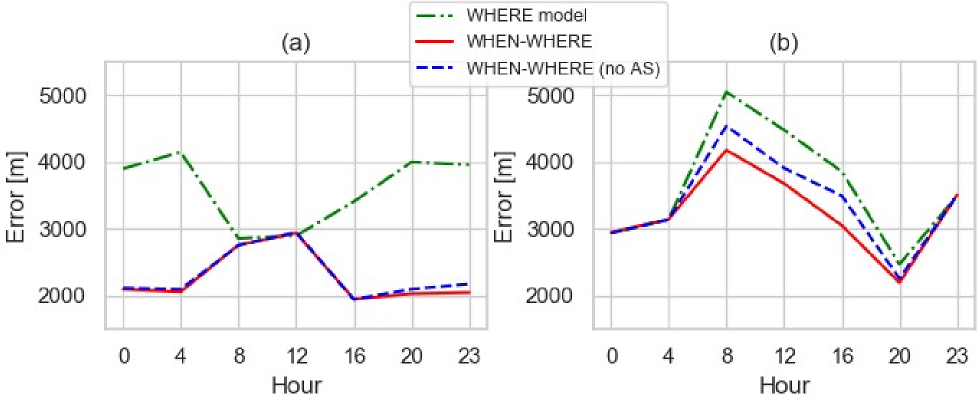

Figure 2. Comparison of the EMD error for datasets synthesised using (a) generated test case and (b) real data.

We measured the individuals' privacy protection by comparing the daily traces found in the original and synthesised data. It is expected that no trajectory would be matched in both datasets, which means that any real trace is included in the synthesised data. We considered trajectories to be identical when they had the same sequence of consecutively visited locations in an individual's daily itinerary. The longest matching sequence appearing in the compared datasets contained three locations for the first test case and two for the second, which stand for 10% and 5,4% of a daily trace, respectively. Furthermore, as a measure of activity spaces similarity, we calculated a number of people whose most frequently visited locations were identical. There were 340 (TC1) and 2 (TC2) people having the same set of the two most frequently visited locations and 5 (TC1) having the same set of the three most frequently visited places.

We compared the performance of the WHERE model and not yet fully developed WHO-WHEN-WHERE model for the synthetic and real data. We demonstrated that our model reaches a higher accuracy than the WHERE model and guarantees no breaches of privacy in published data.

In future work, we plan to introduce the WHO component into the model that will enable us to achieve an even larger accuracy improvement in all the cases. We will also adjust the model to satisfy the requirements of differential privacy. At the current stage, the synthesized mobility traces could be published, but the calculated probability distributions cannot as they breach the privacy.

References

Bianchi, F M, Rizzi, A, Sadeghian, A, & Moiso, C (2016). Identifying user habits through data mining on call data records. Engineering Applications of Artificial Intelligence, 54, 49-61.

European Parliament (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 Data Protection Regulation)

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

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012, June). Human mobility modeling at metropolitan scales. In Proceedings of the 10th international conference on Mobile systems, applications, and services (pp. 239-252). ACM.

Mir, D. J., Isaacman, S., Cáceres, R., Martonosi, M., & Wright, R. N. (2013, October). Dp-where: Differentially private modeling of human mobility. In 2013 IEEE international conference on big data (pp. 580-588). IEEE.