A Weather-Aware Framework for Population Mobility Modelling

Vanessa Brum-Bastos
Wroclaw University of Environmental and Life Sciences, Poland
University of Canterbury, Christchurch, New Zealand

Kamil Smolak
Wroclaw University of Environmental and Life Sciences, Poland

Witold Rohm
Wroclaw University of Environmental and Life Sciences, Poland

Katarzyna Sila-Nowicka
The University of Auckland, New Zealand
Wroclaw University of Environmental and Life Sciences, Poland
University of Glasgow, UK

Abstract

The widespread availability of GPS-enabled mobile devices has contributed towards an unprecedented volume of data on human movement. Human mobility data are the key input for developing accurate mobility models that can support decision-making in, for example, urban planning, transportation planning and disease spread. However, the increasing geoprivacy concerns have been limiting the use of and access to such data. For this reason, the WHO-WHERE-WHEN (3W) model, a privacy-protective model for generating synthetic mobility data, has been developed. However, human mobility is affected by multiple factors that must be accounted for to produce synthetic mobility trajectories that accurately simulate the fluctuations of population in a study area. The 3W model already considers four main factors affecting human mobility: size and shape of activity spaces, circadian rhythm, and home and work locations. Yet, meteorological factors are known to affect human mobility patterns but, to our knowledge, there is not a model that accounts for weather conditions. In this paper, we propose a theoretical framework to extend the 3W model to a 4W model: WHO-WHERE-WHEN-WEATHER. We hypothesise that accounting for weather conditions in human mobility predictions will increase the overall accuracy of predicted mobility patterns.

2012 ACM Subject Classification Applied computing; Human-centered computing

Keywords and phrases movement analytics, human movement, mobility models, context-awareness

Digital Object Identifier 10.4230/LIPIcs.COSIT.2022.17

Category Short Paper

Funding *Vanessa Brum-Bastos*: The Polish National Agency for Academic Exchange (NAWA) Ulam Fellowship PPN/ULM/2020/1/00207/DEC/1 and The Royal Society International Exchanges Grant IES/R1/201127.
Kamil Smolak: The research was financed under the National Science Centre, Poland research grant "Explanation and mitigation of the bias in human mobility predictions" no. 2021/41/N/HS4/03084.
Katarzyna Sila-Nowicka: Supported by the Marsden Fund Council from Government funding, managed by Royal Society Te Apārangi – UOA2037. Also supported by The Royal Society International Exchanges Grant IES/R1/201127.

1 Introduction

Mobility models are developed with data on human movements to identify and predict mobility patterns. However, human movements do not take place in a vacuum but are rather embedded and influenced by the surrounding environment, also known as movement context [6]. Thus, recent mobility studies have been linking movement data to contextual information to gather insights into commuting behaviour [24], tourist behaviour [11] and how weather influences human movement patterns [7]. Similarly, mobility models have also been incorporating movement context by considering work and home regions [16], commuting distance [17] or social media data [10].

Mobility models do not account for weather variables, yet human movement behaviour is impacted by meteorological conditions: [7] and [29] found that wind-speed and direction had an effect on the proportional distribution of time spent in different activities as well as promoted changes in the choice of transportation modes. [9] and [14] described an effect of rain on the proportion of vehicular and walking trips. [23] studied the effect of wind and rain on peoples shopping behaviour. [11] and [30] found a positive effect of increasing temperatures on walking. [5] discovered that increased rainfall led to higher use of public transportation in Bergen, Norway. [22] found that weather conditions influence urban mobility patterns.

Despite the increasing body of evidence supporting the role played by weather in human mobility patterns, up-to-date to our knowledge no mobility model accounts for meteorological variables when modelling nor predicting human movement patterns. Therefore, the goal of this paper is laying the foundations of a framework for weather-aware human mobility modelling at the population level. To consider weather conditions we chose to extend the WHO-WHERE-WHEN (3W) model, a privacy-protective model for generating synthetic mobility data [25], by incorporating ERA5 reanalysis data from the European Centre for Medium-Range Weather Forecasts (ECWMF).

2 The WHO-WHERE-WHEN (3W) model

The 3W is an agent-based model that generates synthetic trajectories by imitating real spatio-temporal characteristics of movement data [15, 19]. Such models can be used to simulate hypothetical scenarios, which are useful in disease spread prediction and urban planning. Additionally, in models such as 3W, the generation of synthetic trajectories from real data safeguards individual's privacy while also providing synthetic data that are still useful for analyses due to their realism [26, 13].

3 The WHO-WHERE-WHEN-WEATHER (4W) model concept

In this section we overview the conceptual framework for the 4W model (Figure 1) and describe the required input data, trajectory sampling process, synthetic trajectory generation process and respective output.

3.1 Inputs

3.1.1 Mobility data

We will use Near® unidentifiable mobility data, which is an aggregate of mobile location data collected from smartphone apps. When a user installs an app, they are asked to share their location data. Users can opt in or out to location sharing, but when they opt in, their phone collects location data and shares it with the app provider and Near®. Latitude-longitude coordinate pairs are collected by smartphones as they move through time and space. These lat-long pairs are associated with the specific device through a unique identifier also available in the dataset. The data are provided as a table containing latitude, longitude, timestamp and device id.

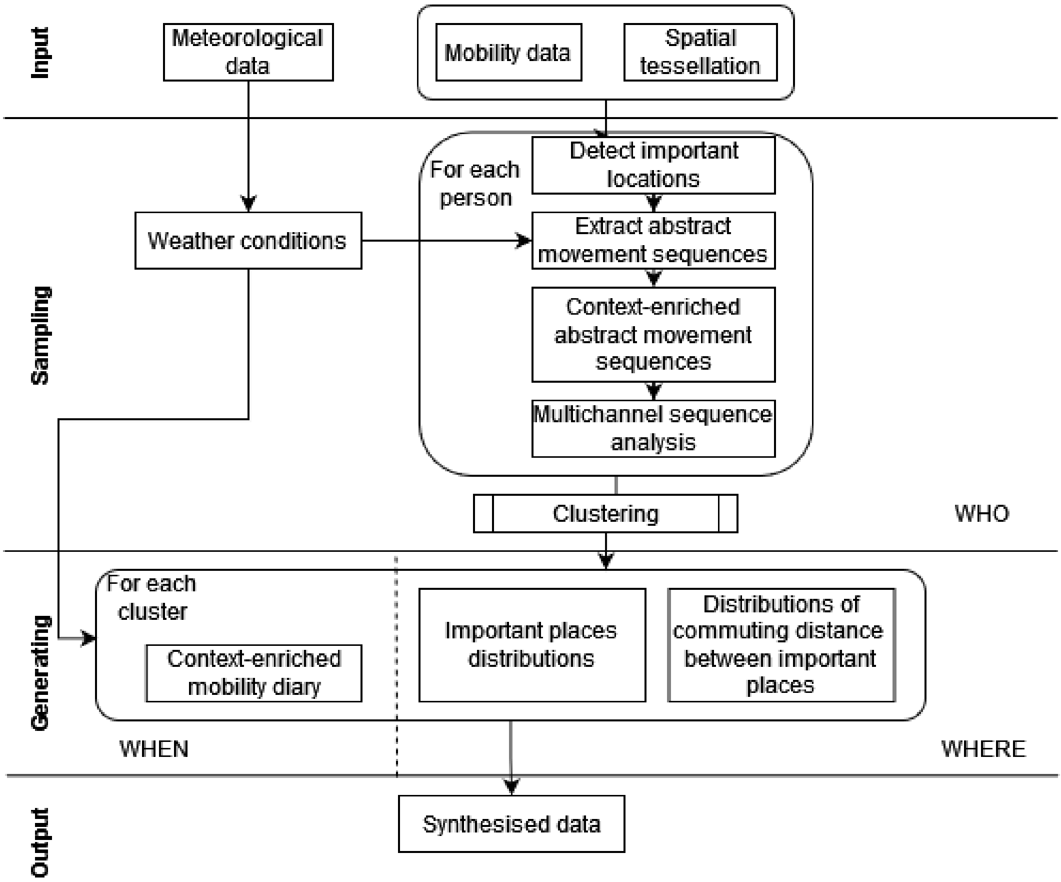

Figure 1 Concept overview for the WHO-WHERE-WHEN-WEATHER (4W) model.

The Near® mobility dataset we use was generated by about 4.7 million devices and it covers Rio de Janeiro, a city of approximately 6.7 million people in Brazil, between May 2019 and January 2021. The sampling rate of location data varies from milliseconds to hours or days depending on the device.

3.1.2 Weather data

The weather data we plan to use comes from the fifth generation ECMWF atmospheric reanalysis model (ERA5) of the global climate covering the period from January 1950 to present. ERA5 provides hourly global estimates of atmospheric, land and oceanic climate variables on a 30 km grid from the surface up to a height of 80km. The outputs from ERA 5 reanalysis have been validated and performed with high accuracy when compared to data from meteorological stations in Brazil [4]. Even though ERA5 reanalysis model provides a large number of atmospheric variables, we are only interested in using the ones that have potential direct effect on human mobility patterns. More specifically, we will look at wind speed components at ten meters height from the surface, temperature at two meters height from the surface, total cloud cover, total precipitation, rain rate and accumulated snowfall. We will also take into account the daylight conditions by also annotating the trajectories with sunrise and sunset information from the daylight Python package [1].

3.1.3 Spatial tessellation

The spatial tessellation is an aggregation layer that divides the simulation area, into a set of non-overlapping polygons, each with a unique identifier and centroid coordinates [20]. In our case, we will use a common approach, which is constructing a grid over the whole area of simulation [17, 8]. The resolution of the grid should not be higher than the resolution of the input data. We plan to test our model using various grid resolutions.

3.1.4 Data pre-processing

The raw data were not sorted, therefore we first split them into separate files for each unique device identifier. Furthermore, to facilitate sample selection for further analyses each device data file was also annotated with metadata on data incompleteness. The incompleteness q is expressed by the total number of missing observations in one-hour time intervals [27]. In many cases, we observed large gaps between consecutive records in movement trajectories, reaching up to a few months. Therefore, we calculated the highest possible q that could be achieved from a one-month-long sample of the movement trajectory for a given device ID.

At this stage, mobility data have a form of a set of raw movement trajectories P_i of i devices. Each P_i is a set of x, y coordinates and timestamps t ordered by time. In the next step, we will select a sample of at least 5000 movement trajectories with $q < 0.2$. In order to reduce noise in the data, we will extract stay-points from the movement trajectories by grouping data points from each P_i into clusters of spatio-temporally neighbouring points [25]. Stay-points are defined as locations where a device spends more than tau time within a range of $delta$. In our case, we plan to apply commonly used values of $tau = 10$ minutes and $delta = 300$ metres [18, 25]. and Then, we will assign all detected stay-points for each mobility trajectory P_i to the respective cells in the spatial tessellation, creating a movement trajectory where the coordinates are the centroids of the spatial tessellation cell. Finally, we will create movement sequences by aggregating data to one-hour time-bins. To achieve that, movement trajectories will be transformed into a sequence of cells visited every hour, where at each time interval the cell visited for the longest period will be selected as the location of a device for the one-hour time-bin .

3.2 Sampling

In the sampling stage, input data are used to estimate mobility-related probability distributions. This stage of the 4W model is called the WHO module. First, we use the mobility dataset to detect important places for each individual. Important places are defined as regularly visited locations that have particular importance or function to an individual [27]. In the 3W model, important places refer to work and home only, i.e. the first and second most significant locations. In the 4W model we will expand this to the n^{th} most important locations. Introducing a 3^{rd} or even 4^{th} most important location can help accounting for other regular trips, such as attending an evening course or going to a gym, which could improve the accuracy of the model [10].

We will assign a rank to detected important locations, where the highest rank indicates the most significant place. Based on that, we will replace locations recorded in a movement sequence of each individual by the important rank for this place. Locations that are not given a rank will be denoted as "other". This process creates an abstract movement sequence consisting of "abstract locations" [20], i.e., a temporal sequence of places without geographic coordinates and in which the spatial component is represented by places, such as "Home", "Work", "n^{th} most significant place" and "Other".

In the next step, we will enrich abstract movement sequences with the chosen meteorological data. Each record of the abstract movement sequence will be annotated with information on the weather conditions at the time and location where it was registered. Similarly to [7], we will translate weather conditions into sequences of qualitative states. The combination of the sequences describing weather conditions and the abstract mobility trajectory will create a contextualised abstract mobility trajectory that can be represented by multiple sequences of strings aligned in time (Fig 2). These are the so-called multi-channel sequences (MCSA), which is a bioinformatics technique for sequencing and analysing human genome [3]. However, MCSA has been extensively used for longitudinal studies in social sciences [2] and more recently in animal and human mobility studies [12, 7].

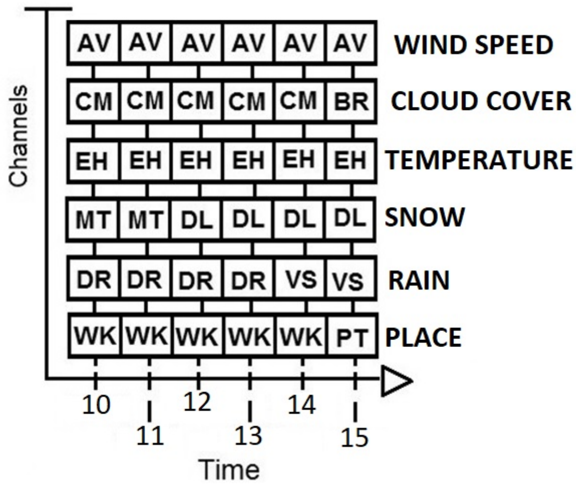

Figure 2 Schematic representation of one context-enriched abstract movement trajectory in the form of a multi-channel sequence. Each channel represents a property of movement or movement's surroundings. Here we have one channel for each meteorological variable and one channel representing places, which is our geographical dimension here. The strings in each sequence represent a specific state for that channel. For example, "AV" indicates average wind speed, "EH" indicates extremely high temperature, "DR" indicates dry weather.

All context-enriched abstract movement sequences will be used to find groups of individuals with similar mobility patterns while also taking into account the weather conditions affecting movement. We will use multi-channel sequence analysis (MSCA), optimal matching and hierarchical clustering to group contextualised abstract movement trajectories according to mobility behaviour and weather condition. The number of clusters of mobility behaviour will be estimated using Silhouette Coefficient criterion. The Silhouette Coefficient criterion is a partitioning metric that takes into account the mean intra-cluster distance and the mean distance to the nearest cluster for each sample [21]. We will use MCSA to calculate pairwise dissimilarity between all context-enriched abstract movement sequences. Dissimilarity is computed based on the minimal cost of replacing, deleting or inserting strings in the channels representing "place" and "weather conditions" so that one sequence is identical to another. This will generate a dissimilarity matrix that can then be used as a distance metric for clustering [7].

Clustered abstract trajectories will be used to calculate conditional probabilities expressing how likely it is for a person from one of the behavioural groups to visit a particular abstract location (identified by its rank) at a given time of the day and under certain weather conditions. The set of these probabilities is called a mobility diary. We will use these conditional probabilities in the generation stage of the model to create synthetic movement trajectories. Clustered abstract trajectories will also be used to determine the proportion of agents in each behavioural group for the generation process. The proportion will be established as equal to the one found in the original data.

Finally, we will calculate the distribution of significant locations and the distributions of commuting distances between important locations for each behavioural group. The distribution of significant places indicates how likely it is to find the important location of a particular rank in each cell of spatial tessellation. Such distribution is referred to as a weighed spatial tessellation [20]. Distributions of commuting distances indicate the median distance between locations of a particular rank, in all possible combinations without repetitions.

3.3 Generating

The generation process will start initializing agents. Agents will be assigned a mobility diary from clusters in a ratio corresponding to the ratio of individuals clustered in each group. After that, each agent will be assigned important locations based on sampled spatial distributions, following the procedure shown in Fig. 3. Important places will be generated for each agent, using the weighted spatial tessellation and distributions of commuting distance from the agent's cluster. The first significant location will be selected using the spatial distribution of the most significant locations. The next places, if considered within the model, will be selected using weighed spatial tessellation for a currently generated rank of places and all combinations of commuting distance distributions. For example, when generating the second place, the commuting distance between the first and the second place will be considered (see Fig. 3). The second location will be selected from an underlying weighed spatial tessellation within the commuting distance from the first place. When the third place is generated, its location will be selected from its weighted spatial tessellation in the area which lies within commuting distances from the first and the second place. The process will follow the same logic for the fourth, fifth and n^{th} most significant places. The location of the fourth place, for example, will be selected in the area which lies within commuting distances from the first, second and third places.

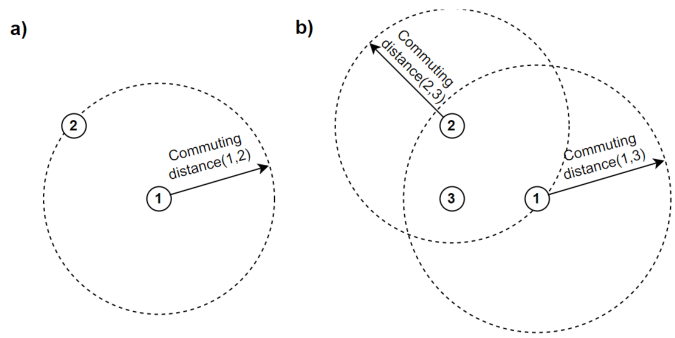

Figure 3 Important places generation process on the example of the second place (a) and the third place (b) selection. The second place is selected within commuting distance from the first place using underlying weighted spatial tessellation for the second place. The third place is selected within a commuting distance from the first and the second place using underlying spatial tessellation for the third places.

After assigning important places to every agent, the data generation process will start. That process will be controlled by the WHERE and WHEN modules, which separately simulate temporal and spatial aspects of mobility. This will be an iterative process, where each location for every agent will be generated using its mobility diary and considering current weather conditions. The WHEN module will select the next abstract location for an agent, which can be either one of the important places or another location. The WHERE module, depending on the selection, will generate the next data point in the location of a selected important place or will use the Exploration and Preferential Return (EPR) mechanism [28]to select other location. The EPR mechanism selects the next location to be an exploration or return to a previously visited place. This decision is based on the number of locations already visited by the agent and accounts for exploratory behaviour in human mobility. Locations for exploration will be selected using their attractiveness, which is defined as the population distribution in a given location divided by a distance friction function. The distance friction function determines that locations further away are less attractive than closer locations. The location of return will be selected using a probability distribution based on the frequency of previous visits.

The above process will be completed when the desired number of agents and data records are generated. The output from the model will have the form of a raw movement trajectory file, where each row consists of an identifier, timestamp and geographic coordinates. Generation accuracy will be assessed by comparing the results from the 4W model to real data.

4 Expected results and final considerations

We introduced a theoretical framework for taking weather into account when performing human mobility modelling. More specifically, we lay the foundations for extending the WHO-WHERE-WHEN (3W) model [26] into a WHO-WHERE-WHEN-WEATHER (4W) model. The 4W model will not only introduce weather information, but also experiment with varying the number of most important places. We hope that, by introducing weather information and expanding the number of most significant locations, we will be able to improve the accuracy of predictions, which we will test by comparing the results from the 4W model to the results from the original 3W model and real data.

References

1. Daylight python library. URL: `https://pypi.org/project/daylight/`.
2. Andrew Abbott. Sequence Analysis: New Methods for Old Ideas. *Source: Annual Review of Sociology Annu. Rev. Sociol*, 21(21), 1995. URL: `http://www.jstor.org/stable/2083405`.
3. Andrew Abbott and Angela Tsay. Sequence Analysis and Optimal Matching Methods in Sociology. *Sociological Methods & Research*, 29(1):3–33, August 2000. `doi:10.1177/0049124100029001001`.
4. Lucas Eduardo de Oliveira Aparecido, Glauco de Souza Rolim, Jose Reinaldo da Silva Cabral de Moraes, Guilherme Botega Torsoni, Kamila Cunha de Meneses, and Cicero Teixeira Silva Costa. Accuracy of ECMWF ERA-interim reanalysis and its application in the estimation of the water deficieny in paraná, Brazil. *Revista Brasileira de Meteorologia*, 34(4):515–528, 2019. `doi:10.1590/0102-7786344066`.
5. Kenneth Apeland. *Analysis Using Machine Learning*. PhD thesis, Western University of Applied Sciences, Bergen, June 2020. URL: `https://bora.uib.no/bora-xmlui/bitstream/handle/1956/23882/A-Weather-Mobility-Analysis-using-Machine-Learning.pdf?sequence=1`.
6. Vanessa Brum-Bastos, Marcelina Łoś, Jed A. Long, Trisalyn Nelson, and Urška Demšar. Context-aware movement analysis in ecology: a systematic review. *International Journal of Geographical Information Science*, 36(2):405–427, 2022. `doi:10.1080/13658816.2021.1962528`.

7 Vanessa S. Brum-Bastos, Jed A. Long, and Urška Demšar. Weather effects on human mobility: A study using multi-channel sequence analysis. *Computers, Environment and Urban Systems*, 70:1–17, September 2018. doi:10.3233/AIC-2008-0431.

8 F. Calabrese, G. Di Lorenzo, and C. Ratti. Human mobility prediction based on individual and collective geographical preferences. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 312–317, 2010. doi:10.1109/ITSC.2010.5625119.

9 Dandan Chen, Yong Zhang, Liangpeng Gao, Nana Geng, and Xuefeng Li. The impact of rainfall on the temporal and spatial distribution of taxi passengers. *PLOS ONE*, 12(9), September 2017. doi:10.1371/journal.pone.0183574.

10 Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090, 2011.

11 M. Cools, E. Moons, L. Creemers, and G. Wets. Changes in Travel Behavior in Response to Weather Conditions: Whether Type of Weather and Trip Purpose Matter? *Journal of the Transportation Research Board*, 2157:22–28, 2010. doi:10.3141/2157-03.

12 Johannes De Groeve, Nico Van de Weghe, Nathan Ranc, Tijs Neutens, Lino Ometto, Omar Rota-Stabelli, and Francesca Cagnacci. Extracting spatio-temporal patterns in animal trajectories: An ecological application of sequence analysis methods. *Methods in Ecology and Evolution*, 7(3):369–379, 2016. doi:10.1111/2041-210X.12453.

13 Marco Fiore, Panagiota Katsikouli, Elli Zavou, Mathieu Cunche, Françoise Fessant, Dominique Le Hello, Ulrich Matchi Aivodji, Baptiste Olivier, Tony Quertier, and Razvan Stanica. Privacy in trajectory micro-data publishing: A survey. *Transactions on Data Privacy*, 13(2):91–149, 2020. arXiv:1903.12211.

14 Zhan Guo, Nigel Wilson, and Adam Rahbee. Impact of Weather on Transit Ridership in Chicago, Illinois. *Transportation Research Record: Journal of the Transportation Research Board*, 2034:3–10, December 2007. doi:10.3141/2034-01.

15 Andrea Hess. A Data-driven Human Mobility Modeling: A Survey and Engineering Guidance for Mobile Networking. *ACM Computing Surveys (CSUR)*, 48(8), 2015. doi:10.1145/0000000.0000000.

16 Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6696 LNCS, pages 133–151. Springer, Berlin, Heidelberg, 2011. doi:10.1007/978-3-642-21726-5_9.

17 Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, and I Computing Methodologies Simulation. Human Mobility Modeling at Metropolitan Scales: Spatial and Temporal Parameters for Mobility Modeling. In *MobiSys '12 Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252, 2012.

18 Shan Jiang, Joseph Ferreira, Marta C González, Fei Wang, Hanghang Tong, Phillip Yu, Charu S Aggarwal Jiang, J Ferreira, and M C González. Clustering daily patterns of human activities in the city. *Data Min Knowl Disc*, 25:478–510, 2012. doi:10.1007/s10618-012-0264-z.

19 Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. Deep Learning for Human Mobility: a Survey on Data and Models. *CoRR*, 2020. arXiv:2012.02825.

20 Luca Pappalardo and Filippo Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32(3):787–829, 2018.

21 Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi:10.1016/0377-0427(87)90125-7.

22 G. Sánchez-Barroso, J. González-Domínguez, J. García-Sanz-Calcedo, and M. Sokol. Impact of weather-influenced urban mobility on carbon footprint of Spanish healthcare centres. *Journal of Transport and Health*, 20, March 2021. doi:10.1016/j.jth.2021.101017.

23 Katarzyna Siła-Nowicka and A. Stewart Fotheringham. Calibrating spatial interaction models from GPS tracking data: An example of retail behaviour. *Computers, Environment and Urban Systems*, 74:136–150, March 2019. doi:10.1016/j.compenvurbsys.2018.10.005.

24 Katarzyna Sila-Nowicka, Jan Vandrol, Taylor Oshan, Jed A. Long, Urška Demšar, and A. Stewart Fotheringham. Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30(5):881–906, May 2016. doi:10.1080/13658816.2015.1100731.

25 Kamil Smolak, Witold Rohm, Krzysztof Knop, and Katarzyna Siła-Nowicka. Population mobility modelling for mobility data simulation. *Computers, Environment and Urban Systems*, 84(January), 2020. doi:10.1016/j.compenvurbsys.2020.101526.

26 Kamil Smolak, Katarzyna Sila-Nowicka, and Witold Rohm. Towards anonymous mobility data through the modelling of spatiotemporal circadian rhythms. In *LBS 2019; Adjunct Proceedings of the 15th International Conference on Location-Based Services/Gartner, Georg; Huang, Haosheng*. Wien, 2019.

27 Chaoming Song, Tal Koren, Pu Wang, and A-L Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):1–6, 2010. doi:10.1038/NPHYS1760.

28 Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature physics*, 6(10):818–823, 2010.

29 Victor W Stover, Nygaard Consulting Associates, and Seattle D Edward McCormack. The Impact of Weather on Bus Ridership in Pierce County The Impact of Weather on Bus Ridership in Pierce County, Washington. *Journal of Public Transportation*, 15(1):95–110, 2012. URL: https://www.nctr.usf.edu/wp-content/uploads/2012/04/JPT15.1Stover.pdf.

30 P. Tucker and J. Gilliland. The effect of season and weather on physical activity: A systematic review. *Public Health*, 121(12):909–922, 2007. doi:10.1016/j.puhe.2007.04.009.