Handling missing data in smartphone location logs

Boaz Sobrado

01 December 2017

Contents

1 Abstract

Using objective location data to infer the mobility measures of individuals is highly desirable, but methodologically difficult. Using commercially gathered location logs from smartphones holds great promise, as they have already been gathered, often span years and can be associated to individuals. However, due to technical constraints this data is more sparse and inaccurate than that produced by specialised equipment. In this paper we present a model which leverages the periodicity of human mobility in order to impute missing data values. Moreover, we will assess the performance of the model relative to currently used methods, such as linear interpolation.

2 Introduction

How people move about in their environment affects a wide range of outcomes, such as health, income and social capital (Goodchild and Janelle, 2010). A better understanding of mobility could lead to better health and urban-planning policies. Yet a large part of studies on human mobility are conducted with pen-and-paper travel diaries, despite well known methodological flaws. The high cost and burden to respondents limits the span of data collection. Short trips are frequently under-reported (Wolf et al., 2003) . The self-reported duration of commutes is often underestimated (Delclòs-Alió et al., 2017).

These obstacles can be overcome by using objective data on human mobility. Such data can now be obtained using the Global Positioning System (GPS). GPS uses the distance between a device and several satellites to determine location. GPS measurements can be used to infer a vast range of socioeconomic and behavioural measures, including where the individual lives, how much time he or she spends at home and where (and how) they travel.

Within behavioural sciences, researchers have used GPS data to investigate wideranging topics. For example, Zenk et al. (2009) investigate the effects of the food environment on eating patterns. Harari et al. (2016) look at the movement correlates of personality, finding that extrovers individuals spend less time at home than introverts. Wang et al. (2015) looks at how academic performance is affected by movement patterns. Palmius et al. (2017) use mobility patterns to predict bipolar depression.

In most studies participants receive a specialised GPS devices to track their movement. We call resulting logs *specialised logs*. Barnett and Onnela (2016) point out several methodological issues with these studies. Like studies using pen-and-paper travel diaries, collecting specialised logs is costly and places a high burden on participants. Besides, introducing a new device to the participant's life may bias their behaviour. Due to these drawbacks specialised logs usually span a short amount of time. Barnett and Onnela (2016) advocate installing a custom-made tracking app on user's phones (*custom logs*). Another solution is to take advantage of existing smartphone location logs (*secondary logs*) . For instance, Google Location History contains information on millions of users (Location History, 2017). Often, secondary logs span several years. By law, secondary logs are accessible to users for free (Commission, 2017). Yet, secondary logs also present methodological challenges. They were created for non-academic purposes under engineering constraints (detailed subsequently). These constraints mean that sensors do not track users continuously, meaning that the resulting logs can be sparse and inaccurate. Hence, two important challenges are dealing with measurement noise and missing data.

Missing data is a pervasive issue as it can arise due to several reasons. Technical reasons include signal loss, battery failure and device failure. Behavioural reasons include leaving the phone at home or switching the device off. As a result, secondary logs often contain wide temporal gaps with no measurements. For instance, several research groups studying mental health report missing data rates between 30% to 50% (Saeb et al., 2015; Grünerbl et al., 2015; Palmius et al., 2017). Other researchers report similar trends in different fields (e.g. Harari et al., 2016; Jankowska et al., 2015).

There is no golden standard for dealing with missing data in GPS logs (Barnett and Onnela, 2016). Importantly, spatiotemporal data measurements are often correlated in time and space. This means that common methods, such as mean imputation, are unsuitable. For example, imagine an individual who splits almost all her time between work and home. Suppose she spends a small amount of time commuting between the two along a circular path. Using mean imputation to estimate her missing coordinates, we impute her to be at the midpoint between home and work. She has never and will never be there! Worryingly, there is little transparency on how researchers deal with missing data (Jankowska et al., 2015).

The accuracy of smartphone location measurements is substantially lower than that of professional GPS trackers. Android phones collect location information through a variety of methods. Other than GPS measurements, Androids use less-accurate heuristics such as WiFi access points and cellphone triangulation. Different methods are used because of computational and battery constraints. GPS is the most energy consuming sensor on most smartphones (LaMarca et al., 2005; Chen et al., 2006). In professional GPS trackers less than 80% of measurements fall within 10 meters of the true location. GPS measures are most inaccurate in dense urban locations and indoors (Schipperijn et al., 2014; Duncan et al., 2013). Unfortunately for researchers, this is where people in the developed world spend most of their time.

Noisy data can lead to inaccurate conclusions if it is not accounted for. Suppose we wish to calculate an individual's movement in a day. A simple approach would be to calculate the sum of the distance between each measurement. But if there is noise, the coordinates will vary even though the individual is not moving. If the measurements are frequent and noisy, we will calculate a lot of movement, even if the individual did not move at all! This issue is also visualised in Figure 1. The problem is further complicated because missing data and noisy measurements are related. Methods used by researchers to reduce noise, such as throwing out inaccurate measurements (e.g. Palmius et al., 2017), can exacerbate the severity of the missing data problem.

In this paper we will explore in detail the problem of missing data and measurement error in secondary location logs. Moreover, we will compare methods used to deal with these problems.

2.1 A concrete example

There is little literature on dealing with missing data in custom or secondary logs. Thus it is worth illustrating the typical characteristics using an example data set. The example data set comes from the Google Location History of a single individual. It spans from January 2013 to January 2017 and contains 814 941 measurements. The data set contains a multitude of variables, including inferred activity and velocity. We will focus on measurements of latitude, longitude, accuracy (defined below). All measurements are paired with a timestamp.

2.1.1 Aggregate measures

Social scientists are most interested in aggregating spatiotemporal data to more socially relevant information, such as distance travelled. As we discussed earlier, aggregations without data processing can be highly biased. However, as an example we calculate time spent at home of the user for each day of the week in the month of February (1). Between individuals, time spent at home has been found to be a reliable predictor of extraversion (Harari et al., 2016).

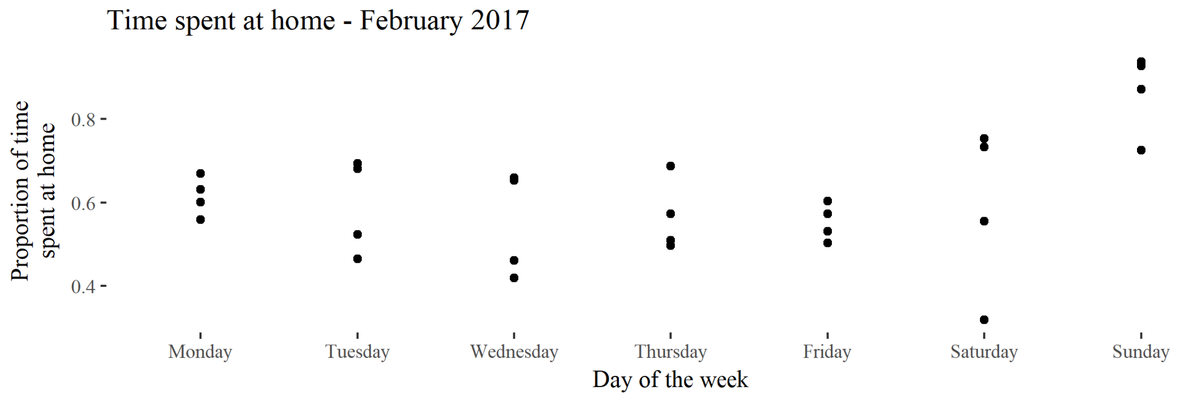
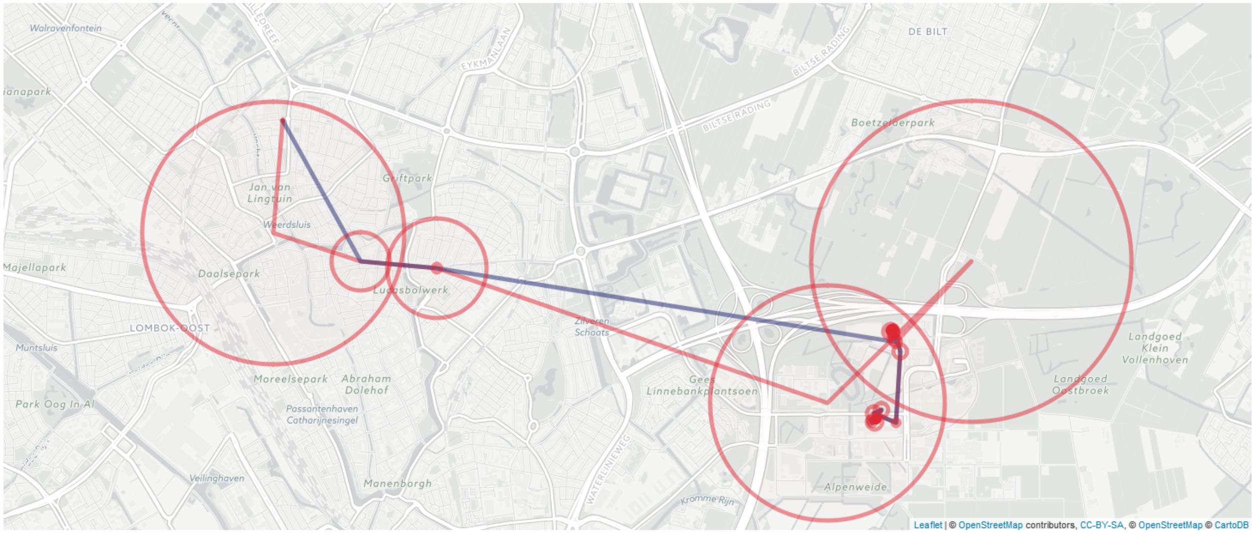


Figure 1: Proportion of time spent at home in February 2017. We estimate this by calculating the mean lattitude and longitude for every 5 minute time period in the month. Missing values are filled in by the previous observed value. Behavioural trends are evident, the user spends more time at home on weekends than on weekdays. Moreover, the variance is greater on weekends, due to travel.

2.1.2 Accuracy in location logs

Google location history provides a measure of accuracy that is given in meters such that it represents the radius of a 67% confidence circle. In the example data set the distribution of accuracy is highly right skewed, with a median of 28, $\mu = 127$ and the maximum value at 26 km. Palmius et al. (2017) note that in their Android based custom logs inaccurate location values are interspersed between more accurate location values at higher sample rates per hour. We observe similar patterns in secondary logs. 2 shows how accuracy can vary as a function of user behaviour, time and location. Inaccurate measures are often followed by more accurate measures. Most notably, low accuracy often (but not always) is associated with movement (3). Stationary accuracy varies depending on phone battery level, wifi connection and user phone use. There are several recurring low-accuracy points, possibly the result of cell-phone tower triangulation.



Leaflet | © OpenStreetMap contributors, CC-BY-SA, © OpenStreetMap © CartoDB

Figure 2: Measurement accuracy of each logged measurement of a morning journey on February 15th 2017. This includes all measurements from midnight to midday. The red circles denote the accuracy of all logged measurement points (the raw data). The points connected in time are connected by a line. The blue line shows the path without the most inaccurate (accuracy > 400 meters) points filtered out. The red line shows the path with all measurements included.

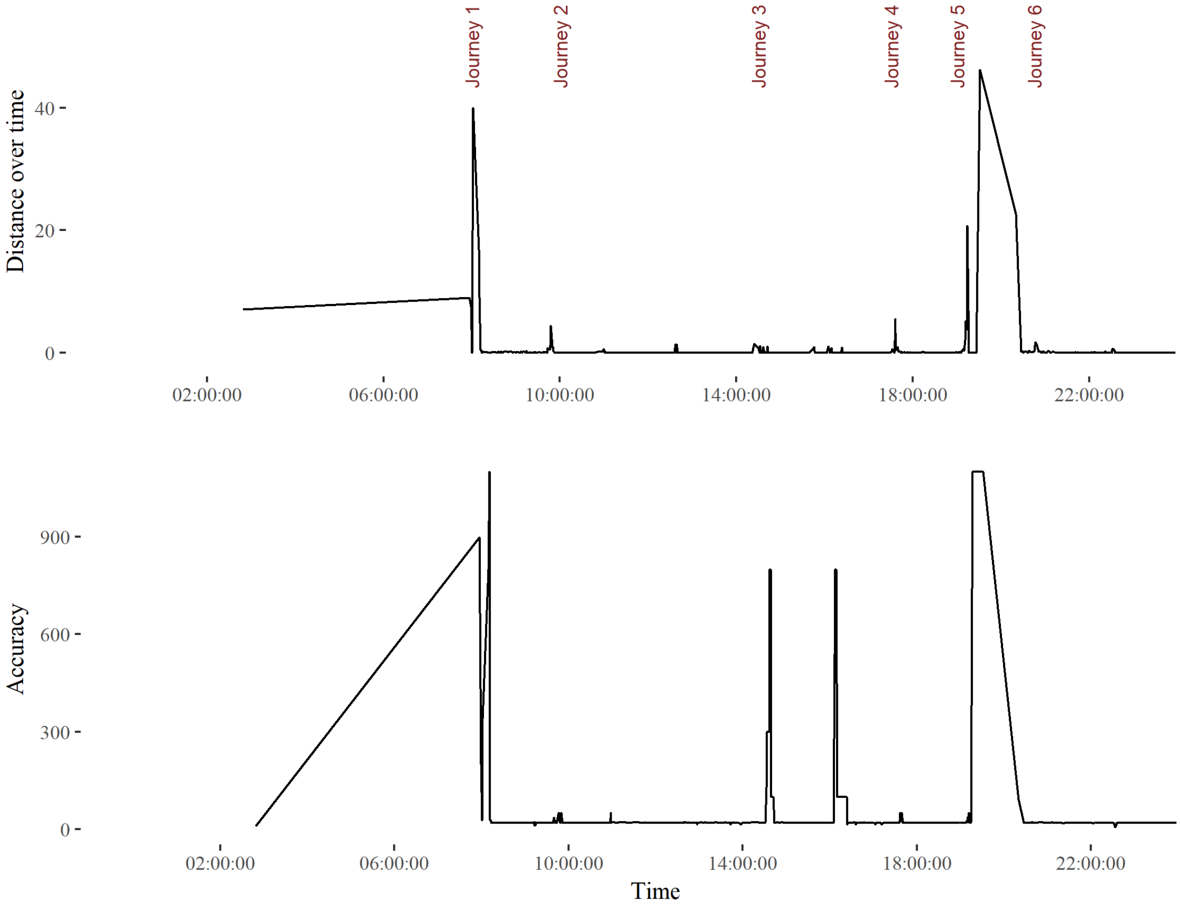


Figure 3: Measures of user activity and measurement accuracy on February 15th 2017.The upper chart shows the distance from the next measured point in meters over the course of the day. All journeys are marked with a red line. The first peak corresponds to the first journey from the user's home to a gym around 8am. The second, smaller peak before 10 reflects a journey from the gym to the nearby lecture theatre. Both journeys can be seen in Figure 1. All other journeys are not shown in Figure 1. The large jump between journey 5 and 6 is measurement error. The lower chart shows the accuracy over the course of the day. The figure shows that measurement inaccuracy is sometimes related to the movement of the individual.

2.1.3 Missingness

Over 54% of the data is missing for the entire duration of the log. This may be misleading as there are several long periods with no measurements whatsoever (see Figure 4). For days which were not entirely missing, approximately 22% of all five minute segments were missing. The structure of missingness of a day with measurements is shown in Figure 5. As you can see, there are several long periods over the course of the log for which there are no measurements. Moreover, even during a single day there are continuous periods where there is missing data, mostly during the late hours of the night in this case.

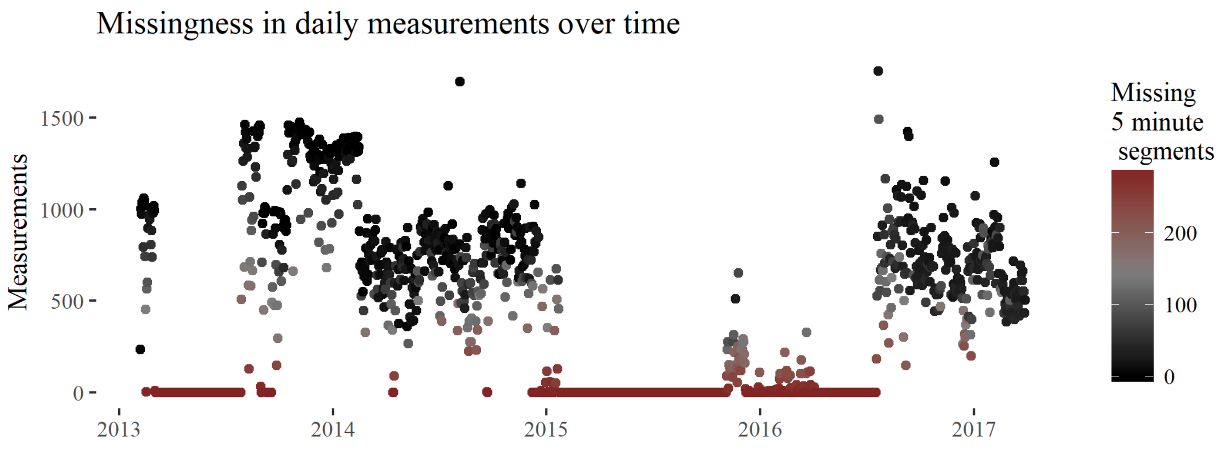


Figure 4: Example of missing data over the entire duration of the log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points are filled with red and lie on the x-axis.

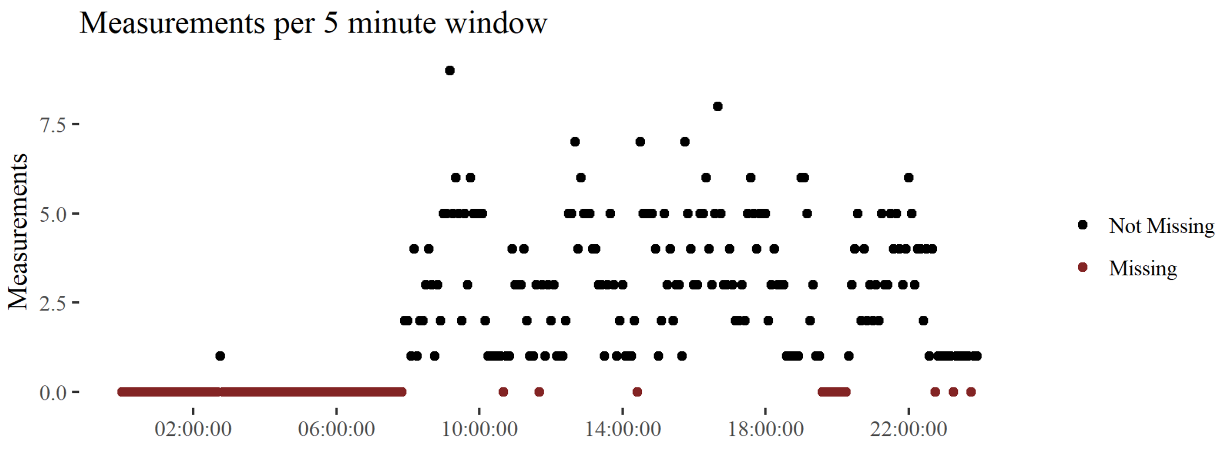


Figure 5: Example of missing data on February 15th 2017. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points are filled with red and lie on the x-axis.

3 Methods

3.1 Two-dimensional projections and notation

GPS measurements provide us with coordinates on the surface of the earth. Because most mobility metrics are computed for data in $\mathbb{R}^2$, we are interested in mapping these $\mathbb{R}^3$ measurements on a 2D Euclidean plane. Projecting three dimensional measurements onto a two dimensional plane results in distortion. To minimise errors we borrow an error minimising projection method from Barnett and Onnela (2016).

Having thus converted lattitude and longitude onto coordinates unique to each individual, let a person's true location on this two-dimensional plane be $G(t) = [G_x(t)G_y(t)]$ where $G_x(t)$ and $G_y(t)$ denote the location of the individual at time t on the x-axis and y-axis respectively. Moreover, let $D \in \mathbb{R}^2$ be the recorded data containing lattitude and longitude. In addition, let a denote the estimated accuracy of the recorded data. accuracy. $G(t)$, D and a are indexed by time labled by the set $T = t_1 < ... < t_{n+1}$. For simplicity, let each entry in the discrete index set T represent a 5 minute window. The measure of accuracy a_t is given in meters such that it represents the radius of a 67% confidence circle. If $D_t = \emptyset$ it is considered *missing* and it is not missing otherwise.

3.2 Selecting candidate models

There is no golden standard or established practice in how to deal with missing data in GPS logs. Researchers are generally vague about what practices they follow (Jankowska et al., 2015). Ostensibly this is because they are unaware of possible solutions. In an attempt to elucidate the topic, we explore potential solutions. We will argue that extensively used spatiotemporal methods, such as state space models (SSMs), are not well suited to deal with human mobility patterns. We also discuss in detail two approaches which deal explicitly with mobility patterns from custom or secondary logs (Palmius et al., 2017, ; Barnett and Onnela, 2016).

There is a vast literature on using SSMs in spatiotemporal statistics. For example, ecologists have used SSMs to explain how animals interact with their environment (Patterson et al., 2008). These models can be quite complex. Preisler et al. (2004) uses Markovian movement processes to characterise the effect of roads, food patches and streams on cyclical elk movements. The most well studied SSM is the Kalman filter, which is the optimal algorithm for inferring linear Gaussian systems. The extended Kalman filter is the de facto standard for GPS navigation (Chen and Brown, 2013). The advantage of state space models is that they are flexible, deal with measurement inaccuracy, include information from different sources and can be used in real time.

For us, the main limitation of SSMs is that they ignore regular movement routines. For instance, humans tend to go to work on weekdays and sleep at night. Because SSMs are based on the Markov property, they cannot incorporate this information. The estimated location $G(t)$ at timepoint t is often based only upon measurements D_t, D_{t-1} and ignores all $D_{t-i} | i \geq 2$. Hierarchical structuring and conditioning on a larger context have been suggested as ways to add periodicity to Markovian models. These solutions are often computationally intractable or unfeasible (Sadilek and Krumm, 2016). For this reason we do not consider SSMs to be useful for imputing missing data. Nonetheless, they could be of use in filtering noise.

In climate or geological research spatiotemporal imputation methods are often used. For instance, the CUTOFF method estimates missing values using the nearest observed neighbours (Feng et al., 2014,). The authors illustrate their example using rainfall data from gauging stations across Australia. Similarly, Zhang et al. (2017) use a variety of machine learning methods to impute missing values. The example provided relates to underground water data. Generally these models assume fixed measurement stations (such as rainfall gauging stations).

For this reason they cannot be easily applied to missing mobility tracks. Feng et al. (2014) claim their model could be used to establish mobility patterns. This may be possible by dividing the sample space into rasters. Each raster would be analogous to a measurement station. These artificial stations could "measure" the probability of the individual being there. To our knowledge such models have not been implemented for mobility traces and seems computationally inefficient.

On the other hand, a few researchers have explicitly attempted to impute missing data from human mobility patterns. Palmius et al. (2017) deal with the measurement inaccuracy of D in custom logs by removing from the data set all unique low-accuracy a data points that had $\frac{d}{dt}D > 100\frac{km}{h}$. Subsequently the researchers down sample the data to a sample rate of 12 per hour using a median filter. Moreover, Palmius et al. (2017) explain:

> "If the standard deviation of $[D]$ in both latitude and longitude within a 1 h epoch was less than 0.01 km, then all samples within the hour were set to the mean value of the recorded data, otherwise a 5 min median filter window was applied to the recorded latitude and longitude in the epoch".

Table 1: Table with descriptives about the data sets used to build the imputation methods. Missing data stands for the proportion of missing 5 minute windows within days that were not missing entirely.

Log duration	Logged days	Observations	Missing days	Missing data	Mean Accuracy
From 2013-02-06 to 2017-03-29	1512	646376	635	0.22	127.78
From 2016-07-14 to 2017-05-10	300	158382	3	0.41	1394.60
From 2014-01-22 to 2017-01-23	1097	814941	80	0.25	121.83

Missing data was imputed using the mean of measurements close in time if the participant was recorded within 500m of either end of a missing section and the missing section had a length of $\leq 2h$ or $\leq 12h$ after 9pm.

Barnett and Onnela (2016) follow a different approach which is, to the best of our knowledge, the only pricipled approach to dealing with missing data in human mobility data. Barnett and Onnela (2016) work with custom logs where location is measured for 2 minutes and subsequently not measured for 10 minutes. In the words of the authors, Barnett and Onnela (2016) handle missing data by:

> “simulat[ing] flights and pauses over the period of missingness where the direction, duration, and spatial length of each flight, the fraction of flights versus the fraction of pauses, and the duration of pauses are sampled from observed data.”

This method can be extended to imputing the data based on temporally, spatially or periodically close flights and pauses. In other words, for a given missing period, the individual’s mobility can be estimated based on measured movements in that area, at that point in time or movements in the last 24 hours (*circadian proximity*).

3.2.1 Datasets & Analyses

The data used to train the imputation methods was collected between 2013 and 2017 on different Android devices from several individuals (table 1).

In addition to the secondary logs, participants also volunteered to carry with them a specialised GPS tracker for a week. This specialised log was used to evaluate the models.

Analyses were performed using R and a multitude of other statistical packages (R Core Team, 2017; Wickham, 2009; Wickham and Francois, 2016; Pebesma and Bivand, 2005; Bivand et al., 2013).

3.2.2 Data pre-processing & filtering

The goal of filtering was to remove noise from the measurements and to aggregate multiple measurements into 12 per hour. Three different filtering methods were tested:

1. The filtered rolling-median downsampling method described by Palmius et al. (2017).
2. A weighted mean approach taking $f(a)$ as a weight.
3. A Kalman filter commonly used for GPS measurements (Doust, 2013).

The output of all of these methods was taken as the input of the imputation methods.

3.2.3 Imputation methods

Three imputation methods were selected in order to cover a wide range of techniques applied in the literature:

1. Mean imputation as described by Palmius et al. (2017).
2. The model developed by Barnett and Onnela (2016) using both spatial and temporal proximity.
3. Simple linear interpolation was used as a benchmark model.

3.2.4 Evaluation criteria

The entire length of the secondary logs were used as a training set. The specialised logs were used as a test set. The missing data imputation models were evaluated both directly, and on two computed measures: amount of trips made and distance traveled.

The direct evaluation involved calculating the error of each D_t compared to $G(t)$ approximated by the specialised log. The error measures used were root mean square error (RMSE) and mean absolute error (MAE).

The evaluation on computed measures involved calculating a mobility trace following the rectangular method of Rhee et al. (2007) for each imputed dataset. Like Barnett and Onnela (2016) we calculate bias by substracting the estimated measure under each approach for the same measure calculated on the full data. For simulation-based imputation approaches a mean value over 100 samples was taken.

Each imputation method used each of the three filtering methods as an input. Thus in the end we end up with eight methods to evaluate, three for each filtering method as well as four for each imputation method.

References

Barnett, I. and Onnela, J.-P. (2016). Inferring Mobility Measures from GPS Traces with Missing Data. *arXiv:1606.06328 [stat]*.

Bivand, R. S., Pebesma, E., and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Chen, M. Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., LaMarca, A., Potter, F., Smith, I., and Varshavsky, A. (2006). Practical Metropolitan-Scale Positioning for GSM Phones. In *UbiComp 2006: Ubiquitous Computing*, Lecture Notes in Computer Science, pages 225–242. Springer, Berlin, Heidelberg.

Chen, Z. and Brown, E. N. (2013). State space model. *Scholarpedia*, 8(3):30868.

Commission, E. (2017). Protecting your data: your rights - European Commission.

Delclòs-Alió, X., Marquet, O., and Miralles-Guasch, C. (2017). Keeping track of time: A Smartphone-based analysis of travel time perception in a suburban environment. *Travel Behaviour and Society*, 9(Supplement C):1–9.

Doust, P. (2013). *smoothing - Smooth GPS data - Stack Overflow*.

Duncan, S., Stewart, T. I., Oliver, M., Mavoa, S., MacRae, D., Badland, H. M., and Duncan, M. J. (2013). Portable global positioning system receivers: static validity and environmental conditions. *American Journal of Preventive Medicine*, 44(2):e19–29.

Feng, L., Nowak, G., O'Neill, T., and Welsh, A. (2014). CUTOFF: A spatio-temporal imputation method. *Journal of Hydrology*, 519:3591–3605.

Goodchild, M. F. and Janelle, D. G. (2010). Toward critical spatial thinking in the social sciences and humanities. *GeoJournal*, 75(1):3–13.

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Tröster, G., Mayora, O., Haring, C., and Lukowicz, P. (2015). Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE journal of biomedical and health informatics*, 19(1):140–148.

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., and Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science*, 11(6):838–854.

Jankowska, M. M., Schipperijn, J., and Kerr, J. (2015). A Framework For Using GPS Data In Physical Activity And Sedentary Behavior Studies. *Exercise and sport sciences reviews*, 43(1):48–56.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., and Schilit, B. (2005). Place Lab: Device Positioning Using Radio Beacons in the Wild. In *Pervasive Computing*, Lecture Notes in Computer Science, pages 116–133. Springer, Berlin, Heidelberg.

Location History, G. (2017). Timeline.

Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., and Vos, M. D. (2017). Detecting Bipolar Depression From Geographic Location Data. *IEEE Transactions on Biomedical Engineering*, 64(8):1761–1771.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology & Evolution*, 23(2):87–94.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*, 5(2):9–13.

Preisler, H. K., Ager, A. A., Johnson, B. K., and Kie, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics 15: p. 643-657.*

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rhee, I., Shin, M., Hong, S., Lee, K., and Chong, S. (2007). Human Mobility Patterns and Their Impact on Routing in Human-Driven Mobile Networks. *ACM HotNets 2007.*

Sadilek, A. and Krumm, J. (2016). Far Out: Predicting Long-Term Human Mobility. *Microsoft Research.*

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., and Mohr, D. C. (2015). Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, 17(7):e175.

Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C. D., and Troelsen, J. (2014). Dynamic Accuracy of GPS Receivers for Use in Health Research: A Novel Method to Assess GPS Accuracy in Real-World Settings. *Frontiers in Public Health*, 2:21.

Wang, R., Harari, G., Hao, P., Zhou, X., and Campbell, A. T. (2015). SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp ’15, pages 295–306, New York, NY, USA. ACM.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wickham, H. and Francois, R. (2016). *dplyr: A Grammar of Data Manipulation.* R package version 0.5.0.

Wolf, J., Oliveira, M., and Thompson, M. (2003). Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey. *Transportation Research Record: Journal of the Transportation Research Board*, 1854:189–198.

Zenk, S. N., Schulz, A. J., and Odoms-Young, A. (2009). How Neighborhood Environments Contribute to Obesity. *The American journal of nursing*, 109(7):61–64.

Zhang, Z., Yang, X., Li, H., Li, W., Yan, H., and Shi, F. (2017). Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level. *Journal of Hydrology*, 553(Supplement C):384–397.