Personalised Map Matched Imputation: imputing missing data from smartphone location logs

Boaz Sobrado[1]

[1] Utrecht University

Author Note

Department of Methodology & Statistics

Submitted as a thesis manuscript conforming to APA manuscript guidelines (6th edition).

Correspondence concerning this article should be addressed to Boaz Sobrado, . E-mail: boaz@boazsobrado.com

Abstract

Personal mobility, or how people move in their environment, is associated with a vast range of behavioural traits and outcomes, such as socioeconomic status, personality and health. The widespread adoption of location-sensor equipped smartphones has generated a wealth of objective personal mobility data. Nonetheless, smartphone collected personal mobility data has remained underused in behavioural research, partly due to the practical difficulties associated with obtaining the data and partly because of the methodological complexity associated with analysing it. Recent changes in European regulation have made it easier for researchers to obtain this data, but the methodological difficulties remain. The difficulty lies in that smartphone location data is irregularly sampled, sparse and often inaccurate. This results in a high proportion of missing data. In this paper we present a method called Personal Map Matched Imputation (PPMI) to deal with missing data and noise in smartphone location logs. The main innovation of PPMI is that it creates a personalised spatial map for each individual based on all the available data. In doing so PPMI leverages the regularity of human mobility in order to smoothen noisy measurements and impute missing data values. By simulating missing periods in real data we find that a simple implementation of PPMI performs as well as existing methods for short (5 minute) missing intervals and substantially better for longer (1 day) missing intervals. When imputing a subset of real missing data where travel logs are available as a reference points, we find that PPMI performs substantially better than existing models.

Keywords: Missing Data, Measurement Bias, GPS, Human Mobility

Word count: 8319

Personalised Map Matched Imputation: imputing missing data from smartphone location logs

Introduction

Why is human mobility important? Human mobility affects a wide range of outcomes, such as health, income and social capital (Goodchild & Janelle, 2010). Human mobility measures are quantified metrics of how people move about in their environment. The most widely administered personality questionnaires ask individuals to what extent they agree with statements such as "I love large parties", "I prefer going to the movies to watching videos at home" and "I love to travel to places that I have never been before" (Goldberg et al., 2006). At their root these questions are mobility measures. In economic research, the postal code of an individual's home address is often used as a proxy for socioeconomic status (e.g. Villanueva & Aggarwal, 2013). Economists are interested not only in where people live, but also where they go to work: geographic labour mobility is related to the income of individuals, the well-being of a country's economy and even informs the policy of bodies such as the European Comission (Tatsiramos, 2009). The extensive use of measures like these across different domains in social science strongly suggests that mobility metrics are linked to important real world outcomes. Behavioural researchers have found that mobility measures can be used to predict academic performance (R. Wang, Harari, Hao, Zhou, & Campbell, 2015), the incidence of obesity (Zenk, Schulz, & Odoms-Young, 2009) or even the onset of a depressive episode in bipolar depression patients (Palmius et al., 2017). To fully understand behaviour not only must we understand how behaviour can vary within and between individuals, but also across environments.

Despite the importance of mobility measures, the majority of metrics are obtained through the use of questionnaires (such as the aforementioned personality questionnaire) or specifically through pen-and-paper travel diaries. As a thought experiment we encourage the reader to try to remember where exactly he or she was five weeks ago on Saturday at midday. Most people find this sort of task difficult (if not impossible), which illustrates the main

methodological flaws of questionnaires and travel diaries. Travel diaries are burdensome to collect because participants must be explicitly asked to write down on their movement patterns at frequent intervals (lest they forget them). The burden to the participants makes data collection expensive, which in turn limits the duration of data collection in practice. In addition, the frequent reporting duties of the participant may bias the participant's behaviour. The limits of human cognition also limit the accuracy of self-reported measures. There is evidence that statistics derived from self-reported mobility measures are systematically biased. For instance, participants under-report the frequency of short trips (Wolf, Oliveira, & Thompson, 2003) and underestimate the duration of regular commutes (Delclòs-Alió, Marquet, & Miralles-Guasch, 2017). These obstacles can only be overcome by using objective data on human mobility.

Social scientists now have the unprecedented opportunity to easily obtain objective data on human mobility from smartphones. Before smartphones the only way to collect objective data on human mobility involved giving participants an expensive professional-grade location sensor and convincing them to take it with themselves at all times. Barnett and Onnela (2016) points out that introducing a new device to the participant's life may bias their behaviour. Moreover, collecting data in such a way is costly, places a high burden on participants and therefore the logs do not span a long time (Barnett & Onnela, 2016). Today millions of individuals carry smartphones with themselves every day and do not need to be encouraged to do so by researchers. Smartphones are equipped with a range of sensors that can be used to track the location of the device at all times. These smartphones can collect and store hundreds of location measurements a day. For instance, Google Location History contains movement information on millions of users, often spanning years (*Timeline*, 2017). Moreover, recent changes in EU regulations with regard to consumer data-portability rights ensure that a willing participant should be able to easily share this information with researchers at no cost to the either the participant or the researchers (European Commission, 2017). Researchers now have the ability to easily access the

objective movement records of millions of individuals spanning several years at little cost and without a significant burden to the participants.

This paper wishes to achieve four objectives: First, we have argued that understanding human mobility is important. Secondly, we argued that social scientists should leverage data logs from smartphones to study human mobility, instead of relying on out-dated pen-and-paper questionnaires. Now we will explore the practical difficulties in using smartphone location logs. Finally we will introduce Personal Map Matched Imputation (PPMI), a method for surmounting these difficulties. We will compare PPMI to existing methods in the literature.

Background

Smartphone location measures are obtained primarily (but not exclusively) by Global Positioning System (GPS) measurements. A GPS sensor uses the distance between a device and several satellites to determine the location of the device. Although using a GPS sensor is the most accurate way to establish location on a smartphone, the GPS sensor is also the most energy consuming sensor on most smartphones (M. Y. Chen et al., 2006; LaMarca et al., 2005). In order to avoid battery depletion and to overcome computational constraints smartphones also use less-accurate heuristics such as WiFi access points and cellphone tower triangulation. Smartphone location logs contain measurements from all of these sources, usually in the form of time-stamped latitude, longitude and accuracy values. The accuracy a of any given measurement is given in meters such that it represents the radius of a 67% confidence circle (*Timeline*, 2017). In other words, the true location of a device should be within the radius a of the measurement 67% of the time.

Researchers often develop custom-made tracking applications which participants are instructed to download on their phone. Alternatively participants are given a phone to use for a given period of time with the custom-made tracking app pre-installed. We call location logs resulting from these custom-made apps *custom logs*. The advantage of custom logs is

that the researchers can adjust tracking parameters, such as the the frequency of measurements and the sensor used to make the measurement. The disadvantage of this approach is that researchers have to develop or adapt a custom-made tracking application (which is not easy given hundreds of different types of smartphone models), distribute it among research participants and enforce participation. Participants may dislike tracking apps because they view them as more intrusive and these apps regularly drain the battery of the device (Harari et al., 2016).

We focus on another solution, which is to take advantage of existing smartphone location logs (*secondary logs*). The advantages are clear: repositories such as Google's Location History contain years of information on millions of users (*Timeline*, 2017) and participants can share the data by clicking of a button. The disadvantage is that researchers have no control over the tracking parameters, often resulting in logs with sparse and inaccurate measurements. Hence, two important challenges in using secondary logs are how to deal with missing data and measurement noise.

In order to work with secondary logs, researchers need to be able to handle the data sparsity that leads to missing data. Missing data is a pervasive issue in secondary logs as it can arise due to several reasons. Technical reasons include signal loss, battery failure and device failure. Behavioural reasons include switching the device or its location services 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% (Grünerbl et al., 2015; Palmius et al., 2017; Saeb et al., 2015). Other researchers report similar trends in different fields (e.g. Harari et al., 2016; Jankowska, Schipperijn, & Kerr, 2015). Figure 1 shows a secondary log spanning years where despite its long duration the sparsity of measurements is also evident.

There is no golden standard for dealing with missing data in GPS logs (Barnett & Onnela, 2016). Importantly, spatiotemporal data measurements are auto-correlated in both time and space. This means that common practices with other types of data, 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, even though she has never been there. Worryingly, there is little transparency on how researchers deal with missing data (Jankowska et al., 2015).

Another methodological problem is related to the noise in the measurements. The accuracy of smartphone location measurements is substantially lower than that of professional GPS location trackers because smartphones often use less accurate sensors. 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 (S. Duncan et al., 2013; Schipperijn et al., 2014). Unfortunately for researchers, this is where people in the developed world spend most of their time. Figure 2 shows how accuracy can vary as a function of user behaviour, time and location. Most notably, low accuracy is often (but not always) associated with movement (see Figure 3).

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 distances 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! 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 propose PPMI as a method for dealing with missing data and measurement error in secondary location logs. PPMI creates a discrete personalised map based on the individuals mobility history. Then it assigns each measurement to a discrete location based on the individuals entire mobility history. Finally, it uses a classification

method to impute missing periods.

Related Work

How have researchers dealt with missing data in human mobility logs thus far? Unfortunately there is no golden standard in how to deal with this type of missing data. Researchers are generally vague about what practices they follow (Jankowska et al., 2015). This vagueness is worrisome as it invites solutions which contain significant researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). Most researchers simply down-sample temporally and remove missing observations or use some sort of rule-based common sense imputations (e.g. Palmius et al., 2017). The only principled approach (that we know of) aiming to solve the issue of missing data in this context (i.e. location logs as they relate to human mobility) is that of Barnett and Onnela (2016). We will explore the methods of Barnett and Onnela (2016) and Palmius et al. (2017) in detail subsequently, after introducing other spatiotemporal methods.

A lack in methods for missing data imputation for human mobility patterns does not imply there is not a vast literature on modelling movement. The most widespread models are SSMs, therefore we shall detail a few examples and subsequently argue that they are nonetheless unsuited for long term human mobility logs. Ecologists have used SSMs to explain how animals interact with their environment (T. A. Patterson, Thomas, Wilcox, Ovaskainen, & Matthiopoulos, 2008). Preisler, Ager, Johnson, and Kie (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 (Z. Chen & 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 secondary logs the main limitation of SSM implementations is that they ignore

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. In other words, the estimated location at a time-point t is often based only upon measurements at timepoints t and $t-1$. 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 & Krumm, 2016). Moreover, these models often assume time and space invariance (i.e. location is not a direct function of time or space). These mathematical assumptions are violated in the case of human movement patterns. For this reason we do not consider existing SSMs to be useful for imputing missing data in smartphone mobility logs.

In the wider realm of spatiotemporal statistics there are numerous missing data imputation methods. These often come from climate or geological research and rely on spatiotemporal auto-correlations. For instance, Feng, Nowak, O'Neill, and Welsh (2014) estimate missing values by incorporating similar observed temporal information from the value's nearest spatial neighbors. The authors illustrate their example using rainfall data from gauging stations across Australia. Similarly, Z. 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 without significant pre-processing.

On the other hand, a few researchers have explicitly attempted to impute missing data from human mobility patterns (Barnett & Onnela, 2016; Palmius et al., 2017 ; Wu, Fu, Wang, Xiao, & Fu, 2017). Importantly, none of them worked with secondary logs, but rather with custom logs that have higher sampling frequency. Nonetheless we will detail what they did as informative examples. Palmius et al. (2017) deal with the measurement inaccuracy in custom logs by removing from the data set all unique low-accuracy data points with a speed greater than 100 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 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".

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 $\leq 12h$ if this occurred after 9pm. In cases where the previous conditions are not met no values are imputed.

Barnett and Onnela (2016) follow a different approach which is, to the best of our knowledge, the only principled 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. Barnett and Onnela (2016) handle missing data by first converting data to mobility traces, which are defined as a sequence of flights and pauses. Flights are segments of linear movements and pauses corresponding to periods of time where a person does not move. Subsequently, the authors impute 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.

It is also worth mentioning that there is a body of work that incorporates road features and other background information to generate predictions or imputations. For instance, Liao,

Patterson, Fox, and Kautz (2007) use hierarchical Markov models in combination with knowledge about the transportation system to bridge the gap between raw GPS sensor measurements and high level information such as a user's destination and mode of transportation. This can be then used to impute or predict missing steps. Along similar lines Wu et al. (2017) use what they call a Spatial Temporal Semantic Neural Network (STS-NN) to predict future human movement. While the authors are concerned with prediction and not imputation, they devised a method called the Spatial Temporal Semantic (STS) algorithm which converts raw measurements to machine learning friendly discrete bins. Working with high-frequency measurements, Wu et al. (2017)'s method down-samples the raw data temporally and map-matches the resulting bins to discrete points along pre-established geographical features such as roads and highways. This minimises measurement error and paves the way for applying machine learning methods to human mobility problems. Subsequently we will focus on methods which do not incorporate external information as they are more generally applicable.

In this section we have argued that there is a lack of established practices to follow with respect to missing data in human mobility logs. Moreover, extensively used spatiotemporal methods, such as state space models (SSMs), are not well suited to deal with human mobility patterns in secondary logs. Finally we discussed in detail three approaches which deal explicitly with mobility patterns from custom or secondary logs (Barnett & Onnela, 2016; Palmius et al., 2017 ; Wu et al., 2017).

Methodology

Notation

Location measurements, such as those produced by GPS sensors, provide us with coordinates (latitude and longitude) on the surface of the earth, which is ellipsoid shaped. Projecting three dimensional measurements in $\mathbb{R}^3$ onto a two dimensional plane in $\mathbb{R}^2$ results in distortion. For clarity, when we use the term distance we refer to the geodesic distances

on an ellipsoid using the WGS84 ellipsoid parameters.

Subsequently let us simplify by assuming that a persons location is on a two-dimensional Euclidean plane. 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. For simplicity, we can assume that the x-axis is the longitude and the y-axis is the latitude. Moreover, let $D \in \mathbb{R}^2$ be the recorded data containing the latitude and longitude. In addition, let a denote the estimated accuracy of the recorded data. Furthermore, $G(t)$, D and a are indexed by time labeled by the countable set $t = t_1 < ... < t_{n+1}$. If $D_t = \emptyset$ it is considered *missing*. The measure of accuracy a_t is given in meters such that it represents the radius of a 67% confidence circle. The reader should keep in mind that a high accuracy value a_t for a measurement D_t means that the measurement is inaccurate (this can be rather confusing).

Personalised Map Matched Imputation

Our algorithm is designed to leverage the periodic nature of human movement along with the long span of secondary to deal with measurement sparsity and inaccuracy.

Modelling assumptions. First, following Barnett and Onnela (2016) we categorise all time-points t as either belonging to the set P (pause) or set F (flight). Conceptually pauses can be understood as periods of time where an individual spends significant amount of continuous time without moving. Flights are the times where the individual is moving. Let t_a be a pause of length n.

$$t_a = t_i < ... < t_{i+n}$$

Let t_b be a pause of length m such that there is no temporal overlap between t_a and t_b and t_b is the first pause after the end of t_a:

$$t_b = t_j < ... < t_{j+m} | t_{i+n} < t_j$$

Then it follows that between the two pauses there must be a flight indexed by t_x of length $j - i + n$.

$$t_x = t_{i+n} < ... < t_j | t_x \in F$$

We define pause locations $G(t_a), G(t_b) | t_a, t_b \in P$ as locations where an individual spends an extended amount of time in the same space (e.g. school, home, work, train station, barber shop, bar, gym). Importantly, our model assumes period and cyclic human movement such that there are many pauses $t_{a1}, t_{a2}, ..., t_{an}$ such that $G(t_{a1}) = G(t_{a2}) = ... = G(t_{an})$. Moreover,it is possible for two consecutive pauses t_a and t_b to share a location $G(t_a) = G(t_b)$ such that $t_a \neq t_b$. For example, if the individual leaves home for a run and returns home without stopping anywhere else.

Let us define as $Flight^x_{ab}$ the set of all points belonging to a flight between $G(t_a)$ and $G(t_b)$ at time-point t_x.

$$Flight^x_{ab} = G(t_x) | t_x \in F = \{G(t_{i+n}), ..., G(t_j)\}$$

Again, there are many flights $t_{x1}, t_{x2}, ..., t_{xn}$ such that $Flight^{x1}_{ab} = Flight^{x2}_{ab} = ... = Flight^{xn}_{ab}$. Then, we can define as $Path_{ab}$ the set of all flights between $G(t_a)$ and $G(t_b)$ at all time-points. For simplicity, we assume that $Path_{ab} = Path_{ba}$.

In addition, we consider all measurements $D(t)$ to be imperfect measurements of $G(t)$:

$$G(t) = D(t) + \text{Measurement Error}$$

Personalised Map Matched Imputation algorithm

Our algorithm performs the following steps:

1. *Map building*: Extract from measurements D all pause locations and path locations to create a personalised map consisting of discrete bins.
2. *Binning*: Assign each measurement D to a unique discrete location.
3. *Imputing*: Use a classification method to predict missing measurements based on all the available information.

Map building. Following Wu et al. (2017)'s spatial-temporal-semantic (STS) feature extraction algorithm our aim is to transform pause and path locations into machine learning friendly discrete locations, which we will call location bins. There are multiple ways of extracting such measurement clusters in the literature, such as Spatio-Temporal Density-Based Spatial Clustering of Applications with Noise (ST-DBSCAN) and sequence oriented clustering (SOC) (Birant & Kut, 2007). We will focus on methods which explicitly deal with mobility patterns extracted from smartphone logs (Barnett & Onnela, 2016; Palmius et al., 2017). Both of these methods pre-process the data and subsequently use two steps to extract pause locations: first they extract pauses and their corresponding locations, then they cluster pause locations based on spatial proximity. This implementation of PMMI uses a stricter version of Barnett and Onnela (2016)'s approach to extract pauses.

First the measurements D are filtered such that only measurements with an accuracy value lower than $a_{\text{P lim}}$ remain within the sample. Then, a measurement D_t belongs to a pause if and only if:

1. The next measurement D_{t+1} is within $t_{\text{Pause lim}}$ amount of seconds (so it is not missing)
2. The next measurement D_{t+1} is within $d_{\text{Pause lim}}$ meters.
3. The duration of the pause is more than $\delta_{\text{Pause lim}}$ seconds.
4. Let the measurements of a possible pause which fit the aforementioned criteria be $X = D_{t,t+1,\ldots,t+n}$. Let $\bar{x}$ be the mean coordinates of X, let $x_{\text{dist}} \in X$ be the furthest away points from $\bar{x}$ and $\bar{a}$ be the the mean accuracy of X. Then X is a pause if and only if the distances between $\bar{x}$ and x_{dist} are $\leq 2 \times \bar{a}$.

The resulting set of points were then hierarchically clustered using a distance matrix, such that all points within d meters of each other were clustered into a pause location. Each pause location is a bin.

For all remaining measurements we assume that they belong to paths. In this implementation of PMMI we use the following algorithm to estimate paths:

1. Take all measurements which are not pauses, filter them based on an accuracy threshold $a_{\text{Path Lim}}$.
2. Create a distance matrix for all remaining measurements $D_t \in F$ and hierarchically cluster it accordingly, such that all points within $d_{\text{Path Lim}}$ meters of each other are clustered into a single flight bin.

At this point all empirically observed path bins and pause bins are extracted. However, there may be some overlap between pause bins and path bins. Thus, the bins are clustered again, such that the pause bins retain priority. This means that if a pause bin and a path bin are within less than d meters of each other, the path bin is removed. The reasoning for this is that the threshold for not being in a pause bin should be higher, as individuals spend the majority of time at a pause bin. The end result is a discretised map which contains pause and flight bins based on the entire log history of the individual.

Binning. Wu et al. (2017)'s spatial-temporal-semantic (STS) feature extraction algorithm uses map matching as a ground truth to assign noisy measurements into discrete bins along roads. In other words, in addition to the measured data they also use a geographic database that contains information about the area in which the individual is (e.g. where precisely the roads are), and sort measurements into bins based on both the measurement and the geographic data base. For example, if an individual is measured as moving closely in parallel to a road A in an area where there is no other parallel road, Wu et al. (2017)'s method will assume that the individual is on the road A.

PMMI uses a similar logic, but without using any external geographic database. The key modification in PMMI is that whilst Wu et al. (2017) uses an external map, we use the total location history of the individual to create a personalised map. This map is subsequently used to bin measurements. This is feasible for two reasons: humans tend to have repetitive movement habits and secondary logs tend to be long. To put it in simpler terms, we consider each measurement at D_x as a sample of $Path_{ab}$, and by aggregating many measurements we can use them to map out $Path_{ab}$.

Thus, all measurements D are then assigned to a discrete bin on the personal map. This includes previous measurements which were discarded from the map building exercise due to an accuracy a value which exceeded $a_{\mathrm{P\ lim}}$ or $a_{\mathrm{F\ lim}}$. In this implementation we used a simple assigning function, whereby the measurements where assigned to the nearest bin.

Classification. At this point the objective of PMMI is to take all the information available about the mobility history of an individual and impute the missing value. In this implementation we trained an artificial neural network (ANN) to do so. The ANN has a simple architecture: it consists of an input layer, a hidden layer and an output layer. The input layer and the hidden layer employ a rectified linear unit activation function and the output layer uses a softmax activation function.

The input variables to the ANN are:

1. The previous and subsequent observed bin as a binary class matrix.
2. The distance in time to the next & previous bin.
3. The time of the day encoded as a cyclical two-dimensional feature.
4. The day of the week as a binary class matrix.
5. The month of the year as a binary class matrix.

For the encoding of the time of day we took the cosine and the sine transforms of the amount of seconds that have elapsed after midnight (London, 2016). This is necessary for the model to be able to understand that one second past midnight and one second before midnight are in fact two seconds away from each other. Moreover we scaled the non-binary values to occupy a range between 0 and 1 in order to ensure convergence.

For a missing time-point at $D_t \in \emptyset$, the output of the model is a set of probability estimates associated with every location bin. That is, for each missing time-point the model returns a vector of probability estimates (with one estimate per bin) associated with where the individual is.

Datasets & Analyses

The secondary location log used to train the imputation methods was collected between 2013 and 2017 on different Android devices from a single individual. About 54% of the data is missing for the entire duration of the log. There are several long periods with no measurements whatsoever, as well as periods with far less missing data. For days which were not entirely missing, approximately 22% of all five minute segments were missing. The structure of missingness of a non-missing day is shown in Figure 4. As you can see, there are several long periods over the course of the log for which there are no measurements. The median sampling frequency per day for non-missing days is around 0.006 Hz.

For simplicity, we subsequently used a time period when the individual was living in the Netherlands. This subset contains 156,000 measurements over a period of less than six months.

Palmius et al. (2017)'s algorithm was implemented in R based on the original MATLAB code and pseudocode kindly provided by the authors. Barnett and Onnela (2016)'s method was slightly adapted in R to fit the data structure from the original R code provided by the authors. When executing their models we used the same parameters as the authors did. To represent Barnett and Onnela (2016)'s model we used the variant where movements were sampled based on spatial proximity.

All analyses were performed using R (Version 3.4.3; R Core Team, 2017)[1].

[1]We, furthermore, used the R-packages *bindrcpp* (Version 0.2; Müller, 2017), *dbplyr* (Version 1.2.0; Wickham & Ruiz, 2018), *dplyr* (Version 0.7.4; Wickham, Francois, Henry, & Müller, 2017), *geosphere* (Version 1.5.7; Hijmans, 2017a), *ggplot2* (Version 2.2.1; Wickham, 2009), *ggthemes* (Version 3.4.0; Arnold, 2017), *kableExtra* (Version 0.7.0; Zhu, 2018), *keras* (Version 2.1.4; Allaire & Chollet, 2018), *knitr* (Version 1.20; Xie, 2015), *leaflet* (Version 2.0.0; Cheng, Karambelkar, & Xie, n.d.), *padr* (Version 0.4.0; Thoen, 2017), *papaja* (Version 0.1.0.9709; Aust & Barth, 2018), *raster* (Version 2.6.7; Hijmans, 2017b), *RColorBrewer* (Version 1.1.2; Neuwirth, 2014), *readr* (Version 1.1.1; Wickham, Hester, & Francois, 2017), *rgdal* (Version 1.2.16; R. Bivand, Keitt, & Rowlingson, 2017), *scales* (Version 0.5.0; Wickham, 2017), *sp* (Version 1.2.7; Hijmans, 2017a; Pebesma & Bivand, 2005), *tibbletime* (Version 0.1.0; Vaughan & Dancho, 2018), and *tidyr* (Version 0.8.0;

All the code is available on a public repository[2] .

Results & Evaluation Metrics

The results will consist of multiple steps:

1. Evaluating the performance of the map building and assigning functions.
2. Comparing the performance of PMMI compared to alternative methods (Barnett & Onnela, 2016; Palmius et al., 2017) and a baseline model using cross-validation.
3. Comparing the performance of PPMI compared to the aformentioned methods using objective ground-truth data (public transportation time-stamped locations).

Map building & binning evaluation. Before we can evaluate the accuracy of the imputations, it is essential to evaluate how well noise in the data has been smoothed. All methods smoothen noise differently in an attempt to extract true location $G(t)$ from measurements D_t. In the subsequent cross-validation step we will remove random smoothened 5 minute periods and compare the imputed value to the smoothened value. For this reason we want to ensure that the smoothened value actually does reflect an approximation to the true location $G(t)$. Otherwise we run the risk of over-estimating the accuracy of a model in the sense that we will compute the extent to which an imputation method can correctly impute a smoothened value which has little to do with the true location $G(t)$.

In order to evaluate the map building and binning we will first visually evaluate the paths and pause locations. A visual evaluation of paths superimposed on a map is an established way to heuristically check their accuracy (e.g. Brunsdon, 2007). Then, let the average distance between the actual measured point and the binned point be the *deviation distance* δ_{dev}. With respect to the deviation distance δ_{dev}, we expect:

Wickham & Henry, 2018).

[2] Github (Sobrado, 2018)

1. A positive relationship between the deviation distance and the accuracy a (in meters) of each measurement.
2. Roughly 67% of the deviation distances δ_{dev} are within accuracy a of each measurement.

Imputation algorithm performance in cross-validation. We will compare the performance of PMMI, Palmius et al. (2017) and Barnett and Onnela (2016). In addition, we will also compute two baseline models. The first one is a naive model, which simply imputes the previous observed value. The second one we will call the "home" model, which imputes that the individual is at home at all missing timepoints. Using these two models as baselines makes sense given the high degree of autocorrelation in the data and that individuals tend to spend most of their time at home. To compare the performance of the aforementioned methods we will remove 25% of measured time intervals at random within a four week period. We will make our comparisons for intervals of 5 minutes, 1 hour and 1 day. In other words, we will remove 25% of time intervals at random, while varying the duration of the time intervals removed.

For the Barnett and Onnela (2016) and PMMI models we will use all the available data to train the models with the exception of the time periods being investigated. Palmius et al. (2017)'s model does not require training.

To compare all methods with each other we will compute a distance measure (how far was the removed location from the predicted location) and the coverage. The coverage refers to the percentage of missing cases which were imputed. For PMMI's imputations we will use a weighted mean for the distance measures whereby each 5 minute period is weighted equally. This is necessary because the other two models downsample temporally to 5 minutes and because the frequency of measurements in a 5 minute period is correlated with imprecise measurements and visiting infrequently travelled locations. Hence weights must be used to avoid inflating error rates for PPMI in comparison to the other models.

In addition, other measures of interest for PMMI are *accuracy* (in what percentage of

the cases was the appropriate bin predicted), the *confidence* and the *distance expectation*. The confidence is the probability with which the model predicts the most likely bin. For instance, if the model predicts the missing bin to be bin A with a probability of 0.9 then we can say this is a high confidence prediction. Similarly, the distance expectation is the cross-product of the estimated probabilities that the individual is at any of all given bins and the distances between the bins to the true bin. For example, if the true location of an individual is bin A (bin A is 1000 meters away from bin B) and the model assigns a probability of 0.9 at bin A and 0.1 at bin B, then the distance expectation would be 100 meters.

Imputation algorithm performance using external benchmarks. We will compare the performance of PMMI, Palmius et al. (2017), Barnett and Onnela (2016) and the naive baseline model with regards to imputing real missing measurments. To do so we will use information from the Dutch public transportation card. The Dutch public transportation service provides users with time-stamped location data of when and where they board, change lines or leave public transportation. In this paper we used 97 such events in a single month.

To be able to make a comparison between models we will remove all measurements from within the 5 minute period that a time-stamped measurement is available. Then we will use each model to impute the location of the individual within that period and compare it to the true location.

Results

Map building & binning evaluation

We used the following parameters to extract pauses: an accuracy limit $a_{\text{Pause lim}}$ of 250 meters, a time limit $t_{\text{Pause lim}}$ of 300 seconds, a distance limit of $d_{\text{Pause lim}}$ 50 meters and a minimum pause duration limit $\delta_{\text{Pause lim}}$ of 100 seconds. Moreover, to extract path bins we used the parameters: $a_{\text{Path Lim}} = 150\text{meters}$ and $d_{\text{Path Lim}} = 300\text{meters}$.

The model was tuned based on a visual inspection of the resulting personal map meeting the following criteria:

1. Plausible extraction of pause locations. For example, a pause in a field with no path leading to it is less plausible than a pause at a train station, airport, or office building.
2. A suitable extraction of paths. Ensuring that individual paths are logically consistent (e.g. do not go over rivers or other unplausable behaviour).
3. Less than a fifth of extracted bins are assigned only one measurement. Unique bins are unavoidable as people sometimes visit a location briefly and only once, nonetheless they pose problems for the classification algorithms.

When selecting parameters there is a trade off is between bias and precision. This is because an increase in precision (in the form of a higher resolution of bin locations) comes at the expense of bias (assigning measurements to bins becomes more difficult). For instance, by increasing the clustering distance parameter $d_{\text{Pause lim}}$ we can extract more valid pause locations at the risk of falsely categorising certain measurements to the wrong bin. This is illustrated in Figure 5.

Map building results in a personalised map with pause and path bins. An excerpt can be seen in Figure 6. It is important to remind the reader that PMMI is map agnostic and uses no external information. Therefore, the close overlap with features on the map, such as pause bins at relevant buildings and transportation locations, as well as the flight bins following roads and railway lines indicate a high degree of precision in personal map building. As expected, PMMI's path extraction yields greater accuracy for frequently occurring paths. For example, the frequently travelled Amsterdam-Utrecht railway line has been extracted almost perfectly, while the less frequently travelled Utrecht-Enschede line is far sparser.

For the entire period examined period the we find a mean deviance δ_{dev} of 40 meters and a median deviance of 15 meters. Around 69% of the deviance values are within their corresponding accuracy value, which is close to the expected theoretical 67%. Approximately

9% of values were not taken into account when creating the bins, given that their the accuracy a exceeded $a_{\text{Path Lim}} = 300$.

In comparison, Palmius et al. (2017)'s method has a median deviance δ_{dev} of 3 meters, with a mean of 115 due to high deviance outliers. On the other hand, the Barnett and Onnela (2016) method has mean deviance δ_{dev} of 343 meters and a median deviance of 8 meters. Barnett and Onnela (2016)'s deviance is necessarily higher than Palmius et al. (2017), as they down-sample temporally (like Palmius et al., 2017) and subsequently aggregate into pauses and linear flights.

The key difference between temporal and spatial down-sampling is shown in Figure 7. Temporal down-sampling is much more sensitive to noise in sparsely measured periods because it averages out values within five minute periods. Often there are only a few noisy measurements in those periods (see Figure 1 which shows that a large proportion of 5 minute periods have less than 3 measurements), which leads to a noise in the down-sampled values. Unsurprisingly, there is a positive relationship between deviance and the amount of measurements in each down-sampled interval. The fact that over 90% of deviance values are within accuracy (substantially higher than the expected theoretical 67%) confirms that temporal down-sampling is not sufficiently filtering out the noise.

Imputation algorithm performance evaluation in cross-validation. The first factor we must take into account is the coverage of each imputation model. The results show us that while PPMI and the baseline models impute all missing values, the other two models have difficulties doing so (Table 1). In particular, the models of Palmius et al. (2017) and Barnett and Onnela (2016)'s fail to impute an increasing amount of missing values with an increase in the duration of the missing period. Palmius et al. (2017)'s method in particular failed to impute a single missing time period when the duration of the removed periods was one day. The longer a log, the higher the probability that long time periods will be missing. Secondary logs are typically long. Hence missing data imputation methods designed to work with secondary logs must be capable of imputing longer time periods.

With respect to the distance metrics, Table 1 shows the results of the distance metrics for each method. We can see that PPMI predicted the location perfectly for at least about half of the removed values in all three time categories. To be precise PPMI's prediction accuracy was 88%, 73% and 47% for the 5 minute, 1 hour and 1 day periods respectively. As a comparison, the baseline naive model's prediction accuracy was 87%, 68% and 22%, and the home baseline model's were 46%, 52% and 40%. Although the home baseline model and PPMI achieved similar distance metrics in the 1 day condition, this does not mean that PPMI defaulted to simply predicting the individual is home at all times (only 40% of predictions were home). PPMI outperformed the alternative models when taking into account the median scores, but not the mean. The explanation is that the alternative models failed to impute more difficult time periods. This increases the magnitude of errors of the PPMI method.

On the other hand, the confidence score for each time period were heavily skewed left, with a median of 1. Accordingly, the distance expectation (the crossproduct of the model probabilities of an estimate with the distances of the true value to the estimates) values were quite similar to the distance scores, albeit approximately 5% higher. Based on this we can say that this implementation of PPMI was overconfident in its predictions.

Imputation algorithm performance evaluation using external benchmarks. Once we removed the measurements within the 5 minute period of each time-stamp in the public transportation log, we employed each of the aforementioned methods to impute the location of the individual at the time-stamped period. Then we calculated the distance metrics between the true location (based on the public transport log) and the imputed location. We did not compute the home baseline model as it would have always been wrong for this subset of data (travelling and being at home are mutually exclusive).

First, the median distance between the imputed value and the true value for the baseline naive model was 555 meters (with a mean of 3303). For this method, the prediction accuracy was 16%. This means that the binning method places the individual at the location

reported by the public transportation travel log in the five minutes prior to the measurement 16% of the time. Needless to say the coverage was 100% as all missing values were imputed.

For the PPMI the median distance metric was 1037 meters (mean 5637). The prediction accuracy was higher than that of the baseline naive model at 24%. Although PPMI imputed more points correctly than the naive baseline model, those it imputed incorrectly were significantly more incorrect. All missing values were imputed, hence resulting in a coverage of 100%.

The same could not be said about Palmius et al. (2017)'s method, as it failed to impute almost 40% of the time periods. This results in a coverage of 60%. For the time periods the method did impute, there was a median distance of 1617 meters between the true location and the imputed value (with a mean of 1517 meters).

Barnett and Onnela (2016)'s method performed better in terms of coverage at 81%. The median distance metric was 1342 meters and the mean value was 5506 meters.

Example: effect on aggregate measures. Social scientists are most interested in aggregating spatiotemporal data to more socially relevant metrics, such as the amount of time spent at home, frequency of travels to new locations, the amount of trips made or distance covered. As an example we calculated the time spent at home of the user without imputing any missing data and with all three of the investigated methods. The results can be seen in Figure 8. Without any form of missing imputation the individual's time spent at home is unknown for 12% of the time, which adds up to almost four days over the course of a 30 day month.

The amount of time spent at home has been found to be a reliable predictor of extroversion (Harari et al., 2016) and the onset of depressive episodes in bipolar patients (Palmius et al., 2017). The 12% of time that remains unaccounted for could mean the difference between an introvert and an extrovert, or a healthy individual and an individual suffering of depression. Filling in these gaps is the exactly the sort of problem that can be solved using missing data imputation. Interestingly, despite their radically different

imputation methods, all three examined models suggest that the user spent approximately 60% of their time at home.

Discussion & Conclusion

Overall the PMMI performed better than the alternative models, particularly during longer missing periods and with objective data gathered from the Dutch public transport service. However, PPMI did not perform substantially better than the baseline models for all time periods and with the objective Dutch public transport data, suggesting that the classification method can be improved. In addition, the comparison to the performance of the Barnett and Onnela (2016) and Palmius et al. (2017) models is somewhat unfair, as they were created for custom logs, not secondary logs. Nonetheless, the comparison remains valid as they are the closest we found to a missing data imputation methods in smartphone location logs.

In addition to higher accuracy under the conditions typical of secondary logs, the advantages of PPMI are increased coverage and flexibility for missing data imputation, robustness to irregular sampling, the ability to model complex non-linear interactions in its imputations, and the ability to use historical records to smooth movement noise.

PPMI's increased coverage and flexibility comes from its ability to make complex non-linear predictions. For instance, in a given missing period it might make sense to predict that the individual is either at home, or at the office, or at a shop with equal probability. While PPMI can make such an imputation, none of the alternative methods can do this. Moreover, the ability to take the prediction probability values from the neural network also helps in dealing with uncertainty. A known-drawback of single imputation is that it takes an imputed value and treats it as observed. Simple rule-based methods such as Palmius et al. (2017)'s are essentially algorithmic single imputation methods. With PPMI it is possible to model uncertainty using the predicted probabilities of each estimate. For instance, in the previous example, we could choose to only take estimates with a high degree of confidence,

thus creating confidence intervals by adding and subtracting the amount of cases where the location of the individual is ambiguous.

With respect to irregular sampling, alternative methods use temporally based down-sampling in order to reduce noise. This leads to deterioration in resolution not only over space, but also over time. A combination of irregular sampling with the fluctuating accuracy values can lead to nonsensical results. For instance, consider a case where there are two inaccurate measurements in movement at 12:00:01 and 12:04:59. Down-sampling over 5 minute periods will lead to a value that will be the mean of the two inaccurate samples, which is likely to be a location the individual is certainly not. PPMI instead down-samples spatially, which ensures that the binned location is one which is composed of the mean of hundreds of observations, not just the few that happen within a single period.

While both Barnett and Onnela (2016) and Palmius et al. (2017) use historical data to smoothen pause locations by clustering pause locations with a close degree of spatial proximity, neither of them do the same for non-pause locations. This may be feasible with high frequency, regularly sampled short duration logs but creates noise with secondary logs. Moreover, with secondary logs it is feasible to spatially "average out" multiple samples of the same path in order to recreate it in its entirety. For instance, although the mean sampling frequency during train travels on the Amsterdam-Utrecht line is low (about 0.01 Hz) the personalised map manages to recreate the train line almost perfectly, despite being completely map agnostic. Map agnosticism makes the model more flexible as it can be used for areas where no good geographical databases are available.

There are multiple methodological limitations in this paper. Most importantly, the evaluation methods are imperfect. The golden standard would be to use at least one highly accurate professional grade GPS device with high sampling frequency to validate results. Until that is available, the use of public transport data and cross-validation is just a substitute.

Furthermore, PPMI can be further developed. The map building function, the

assignment function and the classification model remain simplistic and could be improved.

In map building, the probability of a pause at a given location is certainly related to other factors, such as the time of the day as well as the prior history of pauses at that location. These factors are not taken into account in the pause extraction function. Improved methods would do well to do so. As for paths, a drawback of the current method is that the density of the bins is a function of the clustering parameter d, the distance between the observed points and their sampling density. It does not take into account the length of the path as well as the average sampling frequency of the path. This is an issue because it can lead to bins to which data points are seldom assigned. For example, while the Amsterdam-Utrecht line has been mapped out almost perfectly, many of the clusters along the route have only been assigned few measurements. This leads to difficulties in the classification part of the model, as infrequently observed clusters are hard for the model to predict.

The current assignment function simply assigns each measurement to the nearest bin. It does not take into account any contextual information that can be gleaned from the entire movement history of the information, such as what path they are on. For instance, assume that it is known that an individual is travelling from point A to point B along path AB, and there is an inaccurate measurement closest to a bin which belongs to path AC. By only taking distance into account, the measurement can get assigned to the wrong bin on path AC. An improvement would be use a Bayesian method, whereby assignment is a function of both the measurement and a model of the individuals movement history. In terms of state-space models the state equation would represent a probabilistic representation of where the individual could be at that given time based on the individuals entire movement history. The space side of the model would be a measurement equation representing the measurement and the uncertainty surrounding it in the form of a. This would be a similar implementation to that used in Liao et al. (2007), but rather taking the personalised map instead of information about road networks.

As for the simplicity of the classification method, the neural network which was used to generate predictions used no information on sequence patterns longer than the previous and next bin. It also has no contextual information about how far away the bins are from each other. A model with more input variables and a more sophisticated design, such as a recurrent neural network (RNN), or a long short-term memory recurrent neural network (LSTM) would likely perform significantly better. Moreover, more inputs can be added from alternative sensors in smartphones such as the accelerometer, barometer, measures of phone activity and so on. These can be used to supplement or improve the raw location measurements, or indeed inform the predictive model.

In conclusion, with this paper we aim to start filling the noticeable gap in the social science methodology literature in using smartphone location logs to study human movement. Clearly there is room for improvement, but this is to be expected with new methods. We are in the early days of using smartphone location measurements in social science. Nonetheless, the methodological advantages are clear: millions of individuals have years long location logs containing objective measurements. In addition, these measurements can be easily obtained. This is an unprecedented opportunity. However, with great opportunity comes a great responsibility, as this data raises difficult questions about privacy that researchers must be prepared to answer. Privacy concerns should be addressed, but should not dissuade researchers from developing new methods. Objective location logs are vastly superior to alternatives, such as questionnaires and travel diares, which rely on accurate self-reporting. Social science researchers must take advantage of regulatory changes with regard to data portability and put the vast wealth of data collected by commercial entities to scientific use.

References

Allaire, J., & Chollet, F. (2018). *Keras: R interface to 'keras'*. Retrieved from https://CRAN.R-project.org/package=keras

Arnold, J. B. (2017). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggthemes

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Barnett, I., & Onnela, J.-P. (2016). Inferring mobility measures from GPS traces with missing data. *arXiv:1606.06328 [Stat]*. Retrieved from http://arxiv.org/abs/1606.06328

Birant, D., & Kut, A. (2007). ST-dbscan: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, *60*(1), 208–221.

Bivand, R., Keitt, T., & Rowlingson, B. (2017). *Rgdal: Bindings for the 'geospatial' data abstraction library*. Retrieved from https://CRAN.R-project.org/package=rgdal

Brunsdon, C. (2007). Path estimation from GPS tracks. *Proceedings of the 9th International Conference on GeoComputation*. Retrieved from http://eprints.maynoothuniversity.ie/6148/

Chen, M. Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., ... Varshavsky, A. (2006). Practical metropolitan-scale positioning for GSM phones. In *UbiComp 2006: Ubiquitous computing* (pp. 225–242). Springer, Berlin, Heidelberg. doi:10.1007/11853565_14

Chen, Z., & Brown, E. N. (2013). State space model. *Scholarpedia*, *8*(3), 30868. doi:10.4249/scholarpedia.30868

Cheng, J., Karambelkar, B., & Xie, Y. (n.d.). *Leaflet: Create interactive web maps with the javascript 'leaflet' library*. Retrieved from http://rstudio.github.io/leaflet/

Delclòs-Alió, X., Marquet, O., & 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. doi:10.1016/j.tbs.2017.07.001

Duncan, S., Stewart, T. I., Oliver, M., Mavoa, S., MacRae, D., Badland, H. M., & Duncan, M. J. (2013). Portable global positioning system receivers: Static validity and environmental conditions. *American Journal of Preventive Medicine*, *44*(2), e19–29. doi:10.1016/j.amepre.2012.10.013

European Commission. (2017). *Protecting your data: Your rights - european commission*. Retrieved from http://ec.europa.eu/justice/data-protection/individuals/rights/index_en.htm

Feng, L., Nowak, G., O'Neill, T., & Welsh, A. (2014). CUTOFF: A spatio-temporal imputation method. *Journal of Hydrology*, *519*, 3591–3605. doi:10.1016/j.jhydrol.2014.11.012

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96.

Goodchild, M. F., & Janelle, D. G. (2010). Toward critical spatial thinking in the social sciences and humanities. *GeoJournal*, *75*(1), 3–13. doi:10.1007/s10708-010-9340-3

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Tröster, G., … 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. doi:10.1109/JBHI.2014.2343154

Harari, G., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & 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. doi:10.1177/1745691616650285

Hijmans, R. J. (2017a). *Geosphere: Spherical trigonometry*. Retrieved from

https://CRAN.R-project.org/package=geosphere

Hijmans, R. J. (2017b). *Raster: Geographic data analysis and modeling*. Retrieved from https://CRAN.R-project.org/package=raster

Jankowska, M. M., Schipperijn, J., & 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. doi:10.1249/JES.0000000000000035

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., ... Schilit, B. (2005). Place lab: Device positioning using radio beacons in the wild. In *Pervasive computing* (pp. 116–133). Springer, Berlin, Heidelberg. doi:10.1007/11428572_8

Liao, L., Patterson, D. J., Fox, D., & Kautz, H. (2007). Learning and inferring transportation routines. *Artificial Intelligence*, *171*(5), 311–331. doi:10.1016/j.artint.2007.01.006

London, I. (2016). *Encoding cyclical continuous features - 24-hour time. Ian london's blog*. Retrieved April 27, 2018, from //ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/

Müller, K. (2017). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from https://CRAN.R-project.org/package=bindrcpp

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. Retrieved from https://CRAN.R-project.org/package=RColorBrewer

Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G. M., & Vos, M. D. (2017). Detecting bipolar depression from geographic location data. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1761–1771. doi:10.1109/TBME.2016.2611862

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in Ecology & Evolution*, *23*(2), 87–94. doi:10.1016/j.tree.2007.10.009

Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*,

5(2), 9–13. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Preisler, H. K., Ager, A. A., Johnson, B. K., & Kie, J. G. (2004). Modeling animal movements using stochastic differential equations. *Environmetrics 15: P. 643-657.* Retrieved from https://www.fs.usda.gov/treesearch/pubs/33038

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Sadilek, A., & Krumm, J. (2016). Far out: Predicting long-term human mobility. *Microsoft Research*. Retrieved from https://www.microsoft.com/en-us/research/publication/far-predicting-long-term-human-mobility/

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & 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. doi:10.2196/jmir.4273

Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C. D., & 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. doi:10.3389/fpubh.2014.00021

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Sobrado, B. (2018). *My thesis on missing data in GPS measurements*. Retrieved from https://github.com/sobradob/thesis

Tatsiramos, K. (2009). Geographic labour mobility and unemployment insurance in europe. *Journal of Population Economics*, *22*(2), 267–283. doi:10.1007/s00148-008-0194-7

Thoen, E. (2017). *Padr: Quickly get datetime data ready for analysis*. Retrieved from

https://CRAN.R-project.org/package=padr

Timeline. (2017). Retrieved from https://www.google.com/maps/timeline?pb

Vaughan, D., & Dancho, M. (2018). *Tibbletime: Time aware tibbles*. Retrieved from https://CRAN.R-project.org/package=tibbletime

Villanueva, C., & Aggarwal, B. (2013). The association between neighborhood socioeconomic status and clinical outcomes among patients 1 year after hospitalization for cardiovascular disease. *Journal of Community Health*, *38*(4), 690–697. doi:10.1007/s10900-013-9666-0

Wang, R., Harari, G., Hao, P., Zhou, X., & 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* (pp. 295–306). New York, NY, USA: ACM. doi:10.1145/2750858.2804251

Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017). *Scales: Scale functions for visualization*. Retrieved from https://CRAN.R-project.org/package=scales

Wickham, H., & Henry, L. (2018). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., & Ruiz, E. (2018). *Dbplyr: A 'dplyr' back end for databases*. Retrieved from https://CRAN.R-project.org/package=dbplyr

Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Francois, R. (2017). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Wolf, J., Oliveira, M., & 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. doi:10.3141/1854-21

Wu, F., Fu, K., Wang, Y., Xiao, Z., & Fu, X. (2017). A spatial-temporal-semantic neural network algorithm for location prediction on moving objects. *Algorithms*, *10*(2), 37. doi:10.3390/a10020037

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/

Zenk, S. N., Schulz, A. J., & Odoms-Young, A. (2009). How neighborhood environments contribute to obesity. *The American Journal of Nursing*, *109*(7), 61–64. doi:10.1097/01.NAJ.0000357175.86507.c8

Zhang, Z., Yang, X., Li, H., Li, W., Yan, H., & 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. doi:10.1016/j.jhydrol.2017.07.053

Zhu, H. (2018). *KableExtra: Construct complex table with 'kable' and pipe syntax*. Retrieved from https://CRAN.R-project.org/package=kableExtra

Table 1

Distance in meters between the removed time period and the imputed value.

	Five minutes			One Hour			One Day		
	Mean	Median	Coverage	Mean	Median	Coverage	Mean	Median	Coverage
Barnett & Onella	240	5	97%	33	6	88%	62	6	85%
Palmius	43	9	97%	497	4	89%	NA	NA	0%
PPMI	269	0	100%	908	0	100%	5,757	0	100%
Naive Baseline	426	0	100%	1,502	0	100%	14,266	1,288	100%
Home Baseline	5,599	0	100%	5,667	0	100%	5,757	0	100%

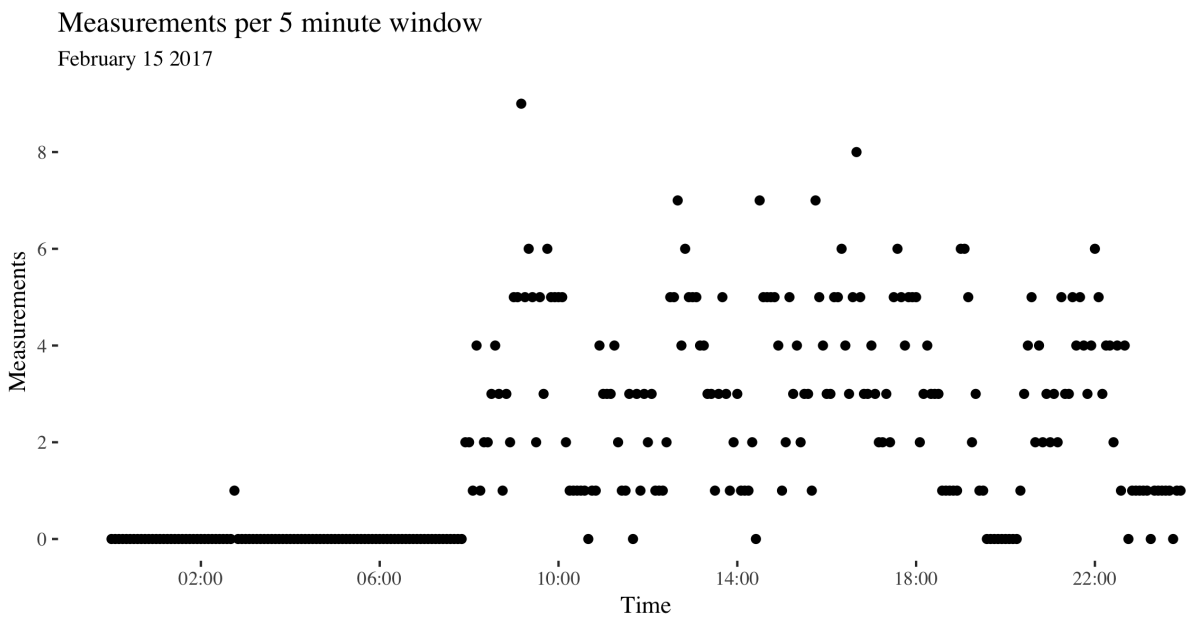

Figure 1. Example of missing data over the entire duration of a secondary log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points lie on the x-axis.

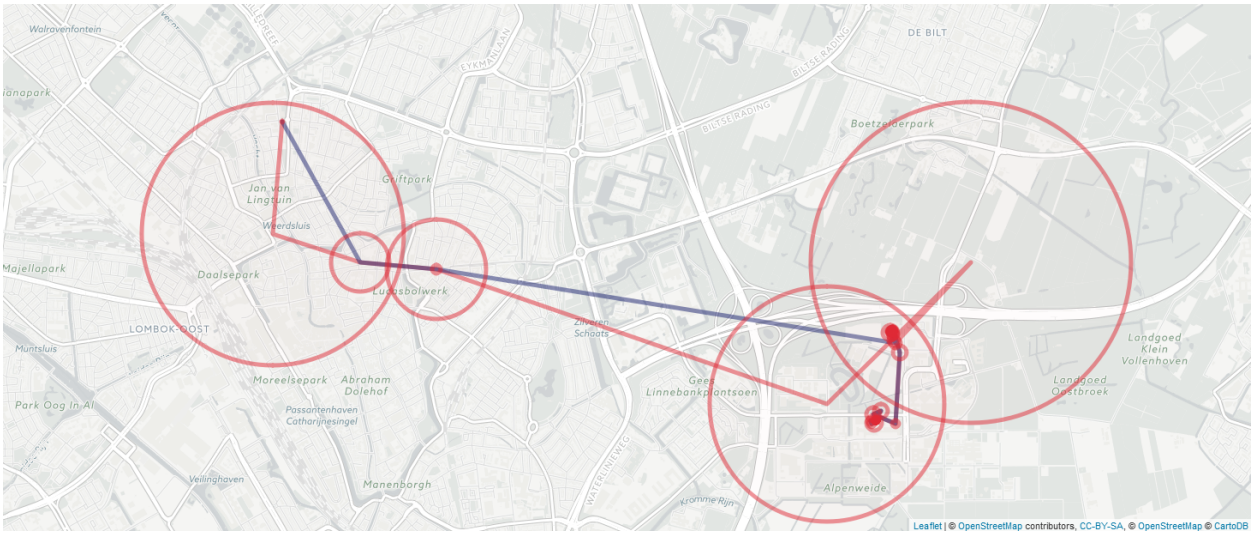

Figure 2. Measurement accuracy of each logged measurement of a morning journey on February 15th 2017. This includes all measurements from midnight to midday. The red circles denote the accuracy of all logged measurement points (the raw data). The points connected in time are connected by a line. The 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. In smartphone logs inaccurate location values are interspersed between more accurate location values at higher sample rates per hour. Inaccurate measures are often followed by more accurate measures. There are several recurring low-accuracy points, such as the one in the northwest corner, possibly the result of cellphone tower triangulation.

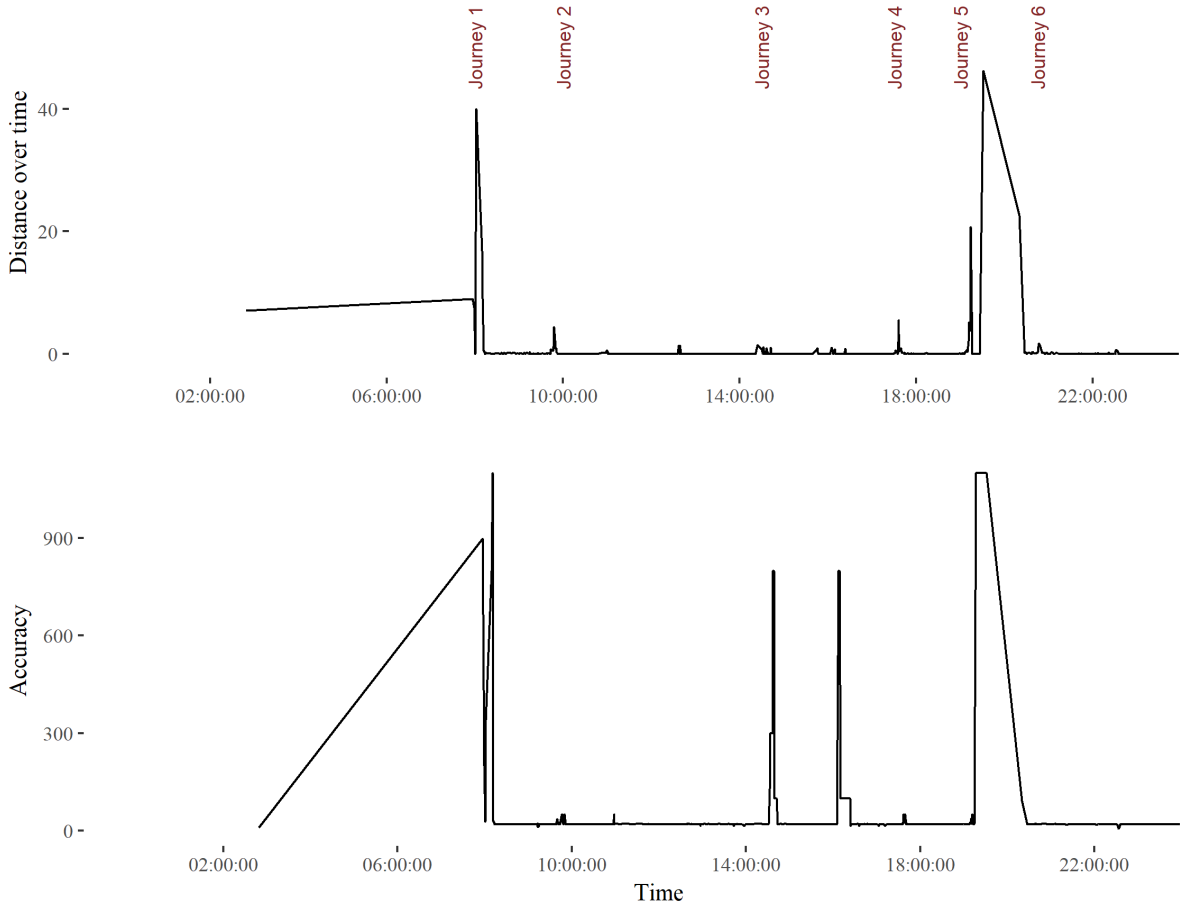

Figure 3. Measures of user activity and measurement accuracy on February 15th 2017.The upper chart shows the distance from the next measured point in meters over the course of the day. The first peak corresponds to the first journey from the user's home to a gym around 8am. The second, smaller peak before 10 reflects a journey from the gym to the nearby lecture theatre. Both journeys can be seen in Figure 2. 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. Stationary accuracy varies depending on phone battery level, wifi connection and user phone use.

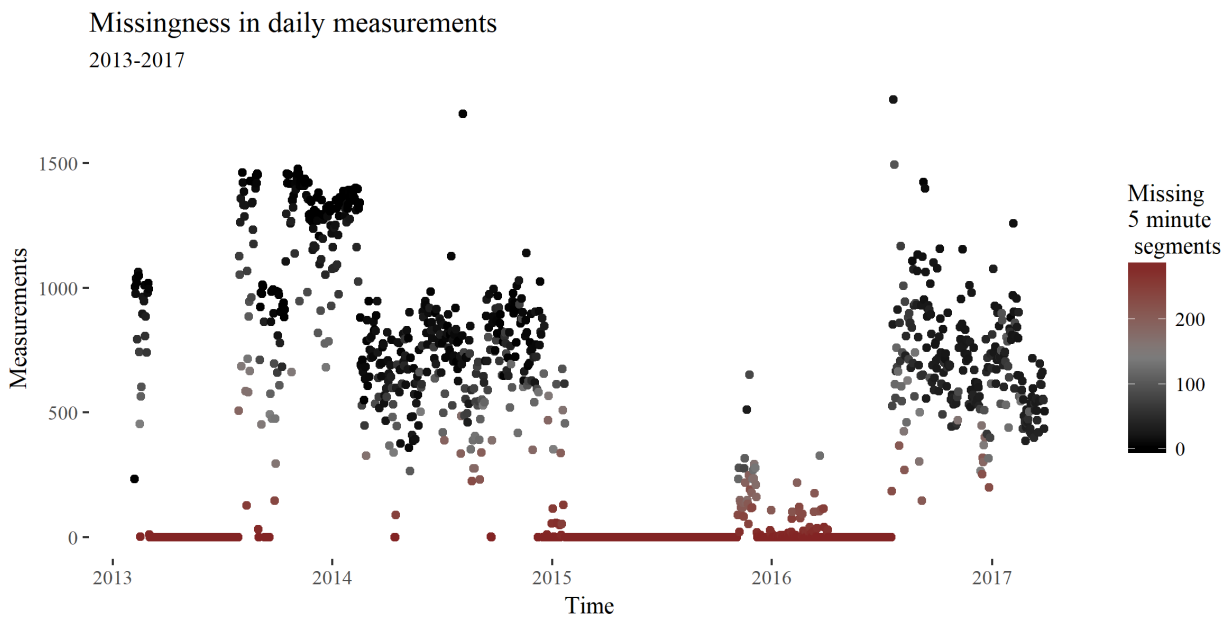

Figure 4. Missing data for the entire duration of the log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. The entire log contains several long periods with no information. These points are filled with red and lie on the x-axis.

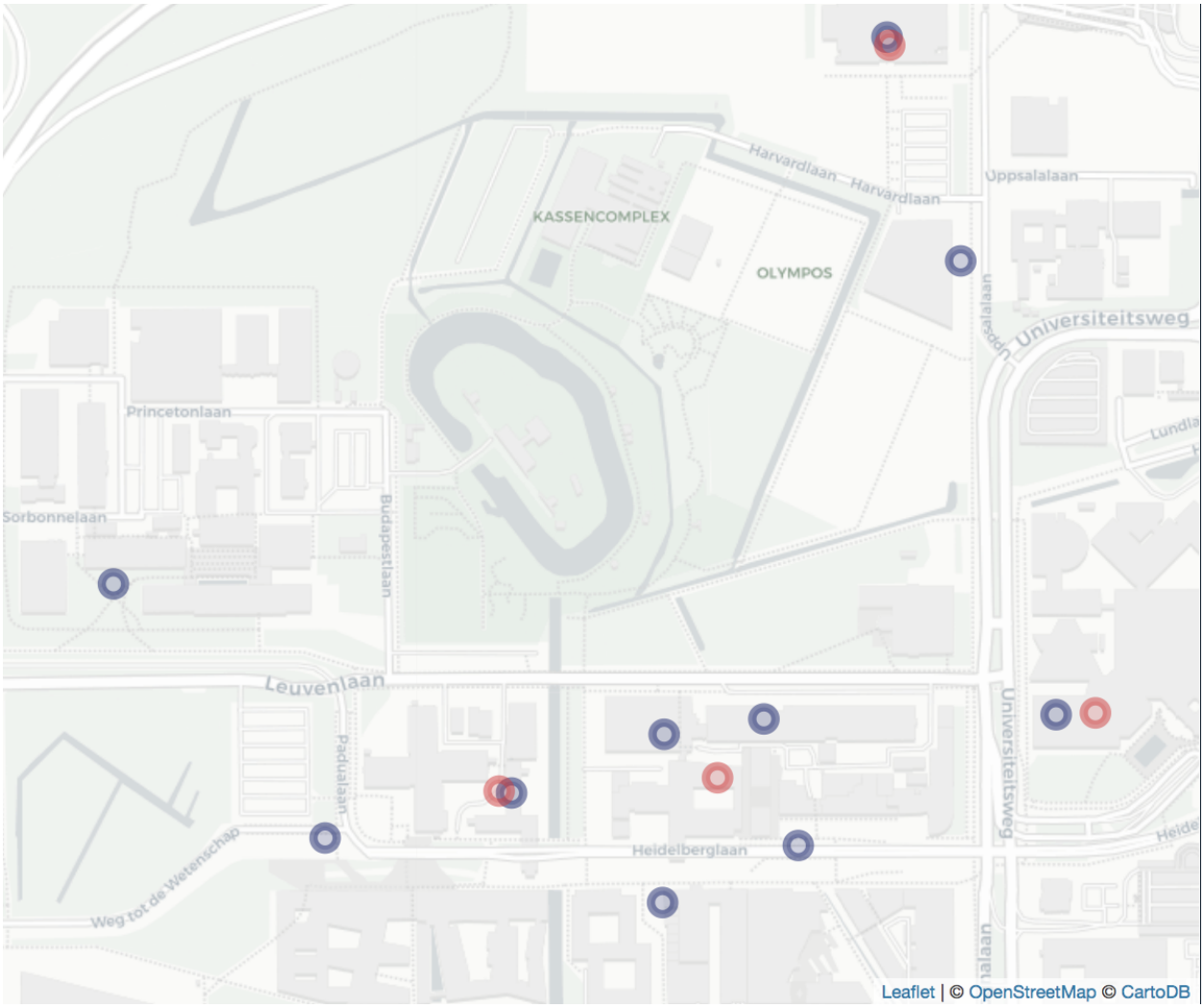

Figure 5. Example of pause locations in De Uithof university campus using 150 meters (blue) and 400 meters (red) as clustering parameters.

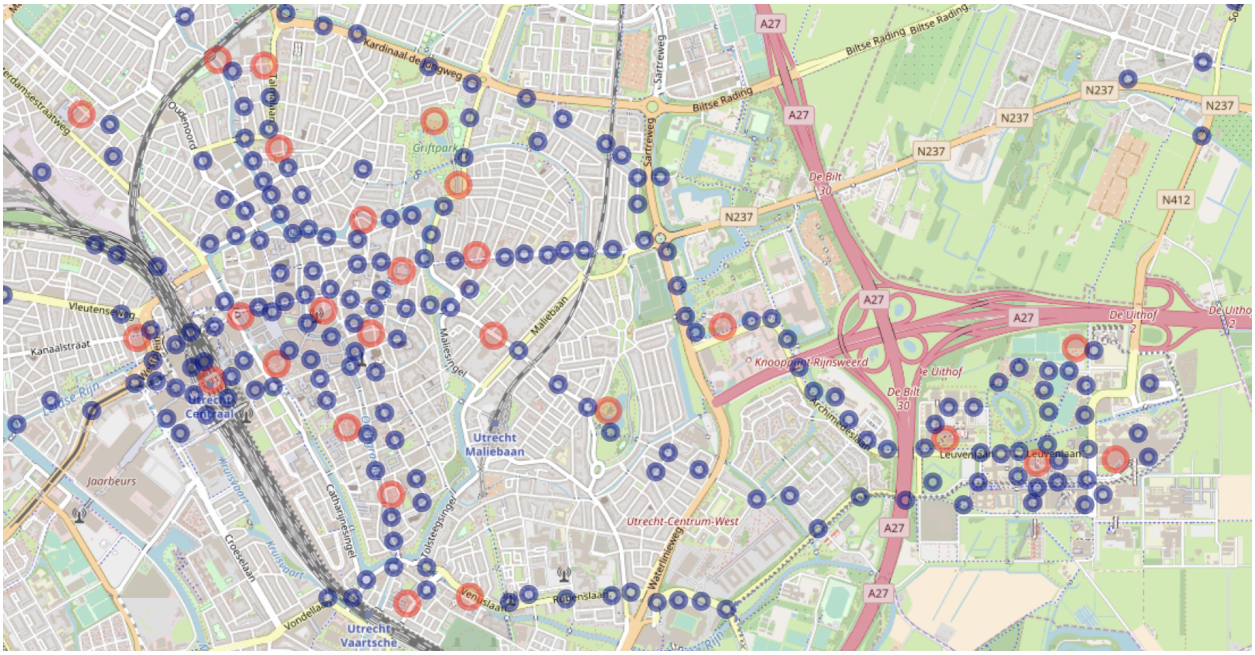

Figure 6. Excerpt of the personalised binned map of an individual. Red points are pause locations, blue points are path locations.

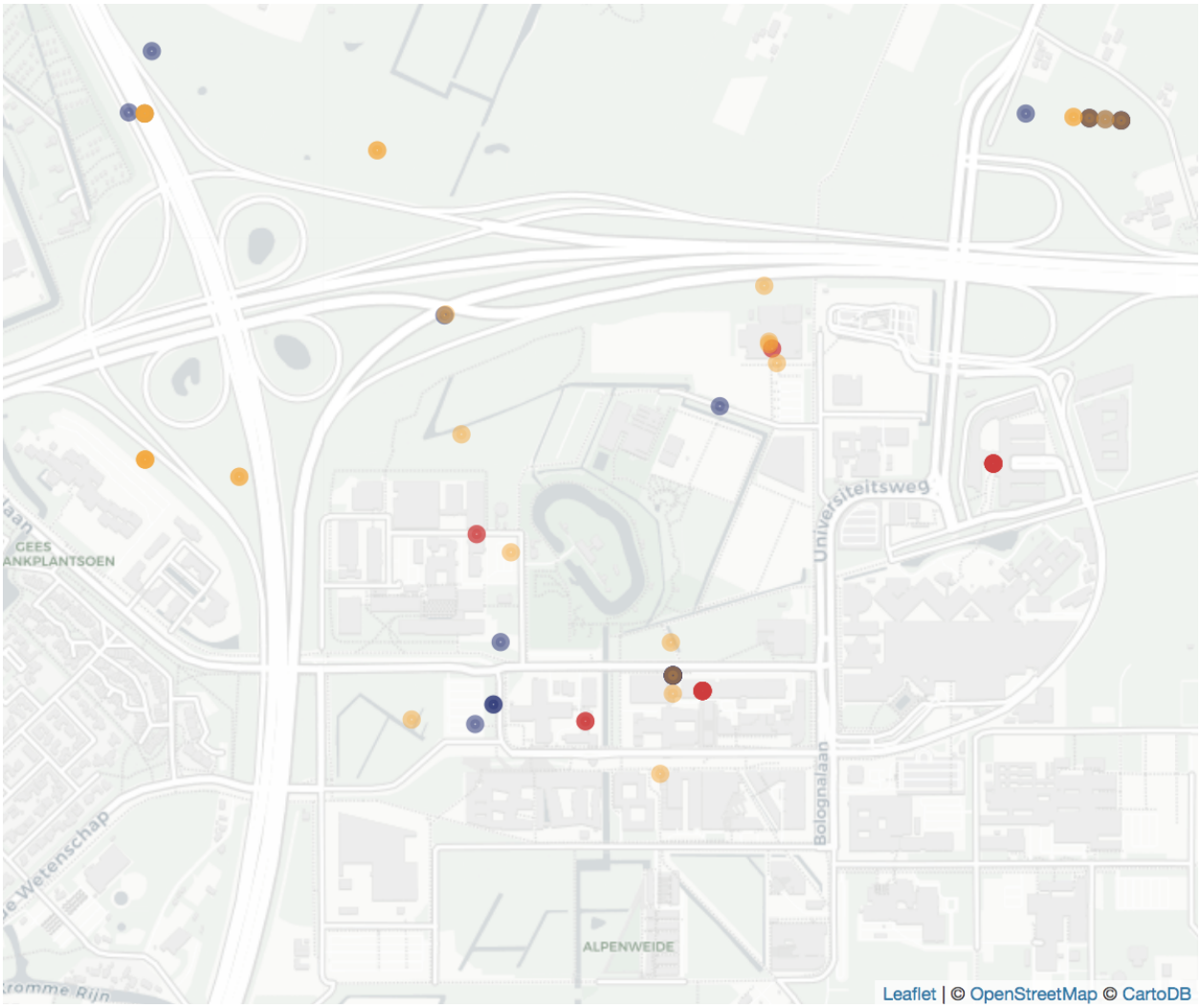

Figure 7. The difference between temporal and spatial downsampling. The blue circles are raw measurments, the yellow circles are temporally downsampled locations. Spatially downsampled locations are in red. Due to measurement sparcity and inaccuracy many of the temporally downsampled locations are in unfeasable locations.

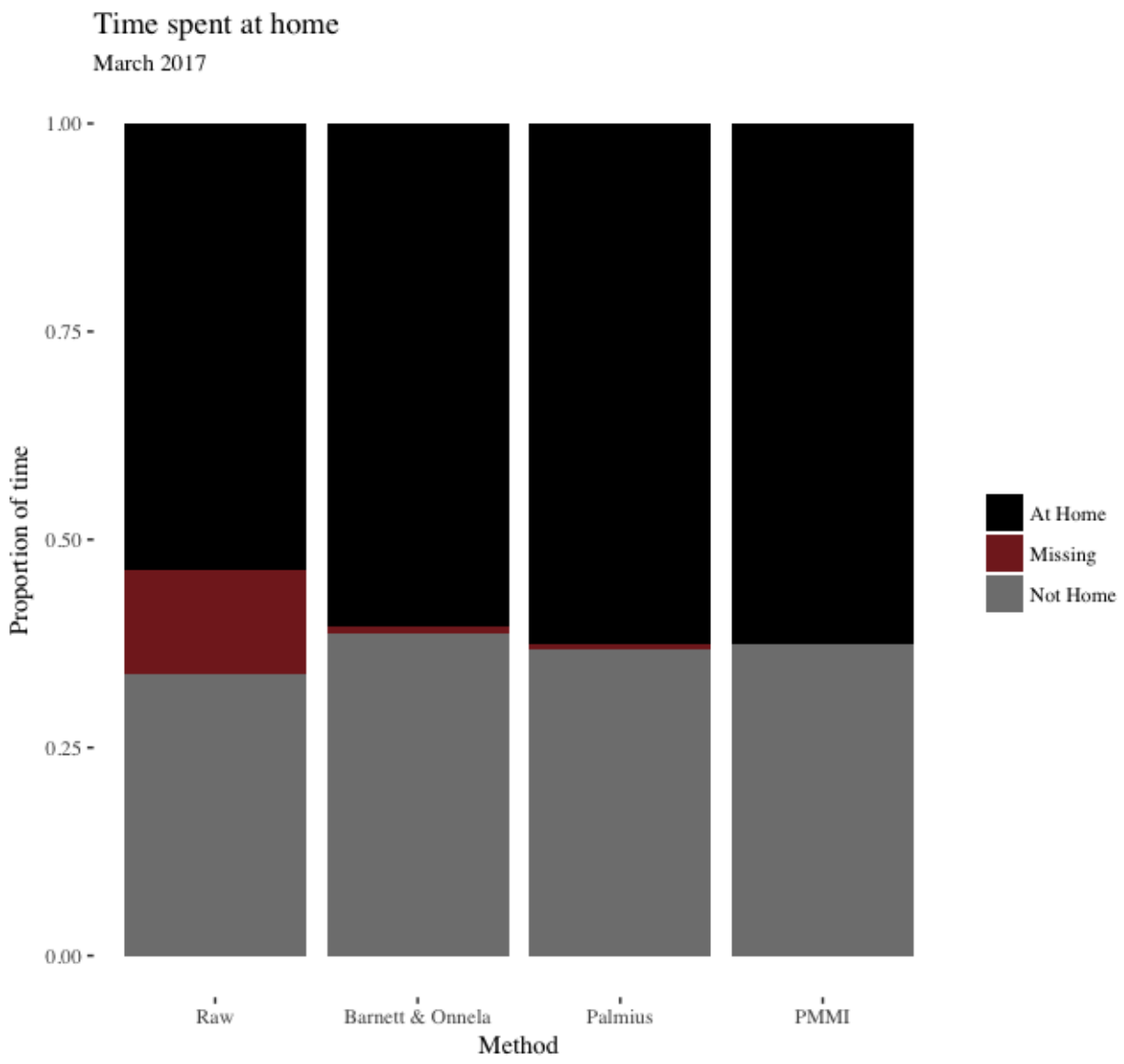

Figure 8. Proportion of time spent at home in March 2017. The raw values are estimated by downsampling temporally the lattitude and longitude for every 5 minute time period in the month. We used each method's own binning method and classified as at home if the downsampled measurement was within 250 meters from home.