

Research Master's programme Methodology & Statistics
Utrecht University, the Netherlands

MSc Thesis *Boaz Sobrado 5891272*

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

May 2018

Supervisors:

Prof. Dr. *Peter Lugtig*

Second grader:

(Prof.) Dr. *Herbert Hoijtink*

Preferred journal of publication: Psychological Methods

Word count: 8319

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

3 Boaz Sobrado¹

4 ¹ Utrecht University

5 Author Note

6 Department of Methodology & Statistics

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

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

11

Abstract

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

31

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

32

Word count: 8319

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

35 **Introduction**

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

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

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

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

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

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

96 **Background**

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

109 Researchers often develop custom-made tracking applications which participants are
110 instructed to download on their phone. Alternatively participants are given a phone to use
111 for a given period of time with the custom-made tracking app pre-installed. We call location
112 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 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

167 method to impute missing periods.

168 **Related Work**

169 How have researchers dealt with missing data in human mobility logs thus far?

170 Unfortunately there is no golden standard in how to deal with this type of missing data.

171 Researchers are generally vague about what practices they follow (Jankowska et al., 2015).

172 This vagueness is worrisome as it invites solutions which contain significant researcher

173 degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). Most researchers simply

174 down-sample temporally and remove missing observations or use some sort of rule-based

175 common sense imputations (e.g. Palmius et al., 2017). The only principled approach (that

176 we know of) aiming to solve the issue of missing data in this context (i.e. location logs as

177 they relate to human mobility) is that of Barnett and Onnela (2016). We will explore the

178 methods of Barnett and Onnela (2016) and Palmius et al. (2017) in detail subsequently,

179 after introducing other spatiotemporal methods.

180 A lack in methods for missing data imputation for human mobility patterns does not

181 imply there is not a vast literature on modelling movement. The most widespread models

182 are SSMs, therefore we shall detail a few examples and subsequently argue that they are

183 nonetheless unsuited for long term human mobility logs. Ecologists have used SSMs to

184 explain how animals interact with their environment (T. A. Patterson, Thomas, Wilcox,

185 Ovaskainen, & Matthiopoulos, 2008). Preisler, Ager, Johnson, and Kie (2004) uses

186 Markovian movement processes to characterise the effect of roads, food patches and streams

187 on cyclical elk movements. The most well studied SSM is the Kalman filter, which is the

188 optimal algorithm for inferring linear Gaussian systems. The extended Kalman filter is the

189 de facto standard for GPS navigation (Z. Chen & Brown, 2013). The advantage of state

190 space models is that they are flexible, deal with measurement inaccuracy, include

191 information from different sources and can be used in real time.

192 For secondary logs the main limitation of SSM implementations is that they ignore

193 movement routines. For instance, humans tend to go to work on weekdays and sleep at night.
194 Because SSMs are based on the Markov property, they cannot incorporate this information.
195 In other words, the estimated location at a time-point t is often based only upon
196 measurements at timepoints t and $t - 1$. Hierarchical structuring and conditioning on a
197 larger context have been suggested as ways to add periodicity to Markovian models. These
198 solutions are often computationally intractable or unfeasible (Sadilek & Krumm, 2016).
199 Moreover, these models often assume time and space invariance (i.e. location is not a direct
200 function of time or space). These mathematical assumptions are violated in the case of
201 human movement patterns. For this reason we do not consider existing SSMs to be useful for
202 imputing missing data in smartphone mobility logs.

203 In the wider realm of spatiotemporal statistics there are numerous missing data
204 imputation methods. These often come from climate or geological research and rely on
205 spatiotemporal auto-correlations. For instance, Feng, Nowak, O'Neill, and Welsh (2014)
206 estimate missing values by incorporating similar observed temporal information from the
207 value's nearest spatial neighbors. The authors illustrate their example using rainfall data
208 from gauging stations across Australia. Similarly, Z. Zhang et al. (2017) use a variety of
209 machine learning methods to impute missing values. The example provided relates to
210 underground water data. Generally these models assume fixed measurement stations (such
211 as rainfall gauging stations). For this reason they cannot be easily applied to missing
212 mobility tracks without significant pre-processing.

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

220 of 12 per hour using a median filter. Moreover, Palmius et al. (2017) explain:

221 “If the standard deviation of in both latitude and longitude within a 1 h epoch
222 was less than 0.01 km, then all samples within the hour were set to the mean
223 value of the recorded data, otherwise a 5 min median filter window was applied
224 to the recorded latitude and longitude in the epoch”.

225 Missing data was imputed using the mean of measurements close in time if the
226 participant was recorded within 500m of either end of a missing section and the missing
227 section had a length of $\leq 2h$ or $\leq 12h$ if this occurred after 9pm. In cases where the
228 previous conditions are not met no values are imputed.

229 Barnett and Onnela (2016) follow a different approach which is, to the best of our
230 knowledge, the only principled approach to dealing with missing data in human mobility
231 data. Barnett and Onnela (2016) work with custom logs where location is measured for 2
232 minutes and subsequently not measured for 10 minutes. Barnett and Onnela (2016) handle
233 missing data by first converting data to mobility traces, which are defined as a sequence of
234 flights and pauses. Flights are segments of linear movements and pauses corresponding to
235 periods of time where a person does not move. Subsequently, the authors impute missing
236 data by:

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

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

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

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

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

266 Methodology

267 Notation

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

272 on an ellipsoid using the WGS84 ellipsoid parameters.

273 Subsequently let us simplify by assuming that a persons location is on a
274 two-dimensional Euclidean plane. Let a person's true location on this two-dimensional plane
275 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
276 on the x-axis and y-axis respectively. For simplicity, we can assume that the x-axis is the
277 longitude and the y-axis is the latitude. Moreover, let $D \in \mathbb{R}^2$ be the recorded data
278 containing the latitude and longitude. In addition, let a denote the estimated accuracy of
279 the recorded data. Furthermore, $G(t)$, D and a are indexed by time labeled by the countable
280 set $t = t_1 < \dots < t_{n+1}$. If $D_t = \emptyset$ it is considered *missing*. The measure of accuracy a_t is
281 given in meters such that it represents the radius of a 67% confidence circle. The reader
282 should keep in mind that a high accuracy value a_t for a measurement D_t means that the
283 measurement is inaccurate (this can be rather confusing).

284 **Personalised Map Matched Imputation**

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

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

$$t_a = t_i < \dots < t_{i+n}$$

292 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
293 is the first pause after the end of t_a :

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

²⁹⁴ Then it follows that between the two pauses there must be a flight indexed by t_x of length
²⁹⁵ $j - i + n$.

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

²⁹⁶ We define pause locations $G(t_a), G(t_b) | t_a, t_b \in P$ as locations where an individual spends an
²⁹⁷ extended amount of time in the same space (e.g. school, home, work, train station, barber
²⁹⁸ shop, bar, gym). Importantly, our model assumes period and cyclic human movement such
²⁹⁹ that there are many pauses $t_{a1}, t_{a2}, \dots, t_{an}$ such that $G(t_{a1}) = G(t_{a2}) = \dots = G(t_{an})$.

³⁰⁰ Moreover, it is possible for two consecutive pauses t_a and t_b to share a location $G(t_a) = G(t_b)$
³⁰¹ such that $t_a \neq t_b$. For example, if the individual leaves home for a run and returns home
³⁰² without stopping anywhere else.

³⁰³ Let us define as $Flight_{ab}^x$ the set of all points belonging to a flight between $G(t_a)$ and
³⁰⁴ $G(t_b)$ at time-point t_x .

$$Flight_{ab}^x = G(t_x) | t_x \in F = \{G(t_{i+n}), \dots, G(t_j)\}$$

³⁰⁵ Again, there are many flights $t_{x1}, t_{x2}, \dots, t_{xn}$ such that $Flight_{ab}^{x1} = Flight_{ab}^{x2} = \dots = Flight_{ab}^{xn}$.
³⁰⁶ Then, we can define as $Path_{ab}$ the set of all flights between $G(t_a)$ and $G(t_b)$ at all
³⁰⁷ time-points. For simplicity, we assume that $Path_{ab} = Path_{ba}$.

³⁰⁸ In addition, we consider all measurements $D(t)$ to be imperfect measurements of $G(t)$:

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

³⁰⁹ Personalised Map Matched Imputation algorithm

³¹⁰ Our algorithm performs the following steps:

- ³¹¹ 1. *Map building*: Extract from measurements D all pause locations and path locations to
³¹² create a personalised map consisting of discrete bins.
- ³¹³ 2. *Binning*: Assign each measurement D to a unique discrete location.
- ³¹⁴ 3. *Imputing*: Use a classification method to predict missing measurements based on all
³¹⁵ 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

$X = D_{t,t+1,\dots,t+n}$. 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:

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

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

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

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

369 Thus, all measurements D are then assigned to a discrete bin on the personal map.

370 This includes previous measurements which were discarded from the map building exercise
 371 due to an accuracy a value which exceeded $a_{P\ lim}$ or $a_{F\ lim}$. In this implementation we used a
 372 simple assigning function, whereby the measurements where assigned to the nearest bin.

373 **Classification.** At this point the objective of PMMI is to take all the information

374 available about the mobility history of an individual and impute the missing value. In this
 375 implementation we trained an artificial neural network (ANN) to do so. The ANN has a
 376 simple architecture: it consists of an input layer, a hidden layer and an output layer. The
 377 input layer and the hidden layer employ a rectified linear unit activation function and the
 378 output layer uses a softmax activation function.

379 The input variables to the ANN are:

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

385 For the encoding of the time of day we took the cosine and the sine transforms of the

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

390 For a missing time-point at $D_t \in \emptyset$, the output of the model is a set of probability

391 estimates associated with every location bin. That is, for each missing time-point the model
 392 returns a vector of probability estimates (with one estimate per bin) associated with where
 393 the individual is.

³⁹⁴ **Datasets & Analyses**

³⁹⁵ The secondary location log used to train the imputation methods was collected
³⁹⁶ between 2013 and 2017 on different Android devices from a single individual. About 54% of
³⁹⁷ the data is missing for the entire duration of the log. There are several long periods with no
³⁹⁸ measurements whatsoever, as well as periods with far less missing data. For days which were
³⁹⁹ not entirely missing, approximately 22% of all five minute segments were missing. The
⁴⁰⁰ structure of missingness of a non-missing day is shown in Figure 4. As you can see, there are
⁴⁰¹ several long periods over the course of the log for which there are no measurements. The
⁴⁰² median sampling frequency per day for non-missing days is around 0.006 Hz.

⁴⁰³ For simplicity, we subsequently used a time period when the individual was living in
⁴⁰⁴ the Netherlands. This subset contains 156,000 measurements over a period of less than six
⁴⁰⁵ months.

⁴⁰⁶ Palmius et al. (2017)'s algorithm was implemented in R based on the original
⁴⁰⁷ MATLAB code and pseudocode kindly provided by the authors. Barnett and Onnela
⁴⁰⁸ (2016)'s method was slightly adapted in R to fit the data structure from the original R code
⁴⁰⁹ provided by the authors. When executing their models we used the same parameters as the
⁴¹⁰ authors did. To represent Barnett and Onnela (2016)'s model we used the variant where
⁴¹¹ movements were sampled based on spatial proximity.

⁴¹² All analyses were performed using R (Version 3.4.3; R Core Team, 2017)¹.

¹We, 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 *tidyverse* (Version 0.8.0;

413 All the code is available on a public repository².

414 **Results & Evaluation Metrics**

415 The results will consist of multiple steps:

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

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

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

Wickham & Henry, 2018).

²[Github](#) (Sobrado, 2018)

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

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

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

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

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

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

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

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

483 Results

484 Map building & binning evaluation

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

489 The model was tuned based on a visual inspection of the resulting personal map

490 meeting the following criteria:

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

498 When selecting parameters there is a trade off between bias and precision. This is

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

504 Map building results in a personalised map with pause and path bins. An excerpt can

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

512 For the entire period examined the we find a mean deviance δ_{dev} of 40 meters

513 and a median deviance of 15 meters. Around 69% of the deviance values are within their
514 corresponding accuracy value, which is close to the expected theoretical 67%. Approximately

515 9% of values were not taken into account when creating the bins, given that their the
 516 accuracy a exceeded $a_{\text{Path Lim}} = 300$.

517 In comparison, Palmius et al. (2017)'s method has a median deviance δ_{dev} of 3 meters,
 518 with a mean of 115 due to high deviance outliers. On the other hand, the Barnett and
 519 Onnela (2016) method has mean deviance δ_{dev} of 343 meters and a median deviance of 8
 520 meters. Barnett and Onnela (2016)'s deviance is necessarily higher than Palmius et al.
 521 (2017), as they down-sample temporally (like Palmius et al., 2017) and subsequently
 522 aggregate into pauses and linear flights.

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

532 **Imputation algorithm performance evaluation in cross-validation.** The first
 533 factor we must take into account is the coverage of each imputation model. The results show
 534 us that while PPMI and the baseline models impute all missing values, the other two models
 535 have difficulties doing so (Table 1). In particular, the models of Palmius et al. (2017) and
 536 Barnett and Onnela (2016)'s fail to impute an increasing amount of missing values with an
 537 increase in the duration of the missing period. Palmius et al. (2017)'s method in particular
 538 failed to impute a single missing time period when the duration of the removed periods was
 539 one day. The longer a log, the higher the probability that long time periods will be missing.
 540 Secondary logs are typically long. Hence missing data imputation methods designed to work
 541 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.

559 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

569 reported by the public transportation travel log in the five minutes prior to the measurement
570 16% of the time. Needless to say the coverage was 100% as all missing values were imputed.

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

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

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

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

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

596 imputation methods, all three examined models suggest that the user spent approximately
597 60% of their time at home.

598 **Discussion & Conclusion**

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

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

612 PPMI's increased coverage and flexibility comes from its ability to make complex
613 non-linear predictions. For instance, in a given missing period it might make sense to predict
614 that the individual is either at home, or at the office, or at a shop with equal probability.
615 While PPMI can make such an imputation, none of the alternative methods can do this.
616 Moreover, the ability to take the prediction probability values from the neural network also
617 helps in dealing with uncertainty. A known-drawback of single imputation is that it takes an
618 imputed value and treats it as observed. Simple rule-based methods such as Palmius et al.
619 (2017)'s are essentially algorithmic single imputation methods. With PPMI it is possible to
620 model uncertainty using the predicted probabilities of each estimate. For instance, in the
621 previous example, we could choose to only take estimates with a high degree of confidence,

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

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

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

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

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

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

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

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

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

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

References

- 698
- 699 Allaire, J., & Chollet, F. (2018). *Keras: R interface to 'keras'*. Retrieved from
700 <https://CRAN.R-project.org/package=keras>
- 701 Arnold, J. B. (2017). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'*. Retrieved
702 from <https://CRAN.R-project.org/package=ggthemes>
- 703 Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*.
704 Retrieved from <https://github.com/crsh/papaja>
- 705 Barnett, I., & Onnela, J.-P. (2016). Inferring mobility measures from GPS traces with
706 missing data. *arXiv:1606.06328 [Stat]*. Retrieved from
707 <http://arxiv.org/abs/1606.06328>
- 708 Birant, D., & Kut, A. (2007). ST-dbscan: An algorithm for clustering spatial–temporal data.
709 *Data & Knowledge Engineering*, 60(1), 208–221.
- 710 Bivand, R., Keitt, T., & Rowlingson, B. (2017). *Rgdal: Bindings for the 'geospatial' data
711 abstraction library*. Retrieved from <https://CRAN.R-project.org/package=rgdal>
- 712 Brunsdon, C. (2007). Path estimation from GPS tracks. *Proceedings of the 9th International
713 Conference on GeoComputation*. Retrieved from
714 <http://eprints.maynoothuniversity.ie/6148/>
- 715 Chen, M. Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., ... Varshavsky,
716 A. (2006). Practical metropolitan-scale positioning for GSM phones. In *UbiComp
717 2006: Ubiquitous computing* (pp. 225–242). Springer, Berlin, Heidelberg.
718 doi:10.1007/11853565_14
- 719 Chen, Z., & Brown, E. N. (2013). State space model. *Scholarpedia*, 8(3), 30868.
720 doi:10.4249/scholarpedia.30868
- 721 Cheng, J., Karambelkar, B., & Xie, Y. (n.d.). *Leaflet: Create interactive web maps with the
722 javascript 'leaflet' library*. Retrieved from <http://rstudio.github.io/leaflet/>
- 723 Delclòs-Alió, X., Marquet, O., & Miralles-Guasch, C. (2017). Keeping track of time: A
724 smartphone-based analysis of travel time perception in a suburban environment.

- 725 *Travel Behaviour and Society*, 9(Supplement C), 1–9. doi:[10.1016/j.tbs.2017.07.001](https://doi.org/10.1016/j.tbs.2017.07.001)
- 726 Duncan, S., Stewart, T. I., Oliver, M., Mavoa, S., MacRae, D., Badland, H. M., & Duncan,
727 M. J. (2013). Portable global positioning system receivers: Static validity and
728 environmental conditions. *American Journal of Preventive Medicine*, 44(2), e19–29.
729 doi:[10.1016/j.amepre.2012.10.013](https://doi.org/10.1016/j.amepre.2012.10.013)
- 730 European Commission. (2017). *Protecting your data: Your rights - european commission.*
- 731 Retrieved from
- 732 http://ec.europa.eu/justice/data-protection/individuals/rights/index_en.htm
- 733 Feng, L., Nowak, G., O'Neill, T., & Welsh, A. (2014). CUTOFF: A spatio-temporal
734 imputation method. *Journal of Hydrology*, 519, 3591–3605.
- 735 doi:[10.1016/j.jhydrol.2014.11.012](https://doi.org/10.1016/j.jhydrol.2014.11.012)
- 736 Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., &
737 Gough, H. G. (2006). The international personality item pool and the future of
738 public-domain personality measures. *Journal of Research in Personality*, 40(1),
739 84–96.
- 740 Goodchild, M. F., & Janelle, D. G. (2010). Toward critical spatial thinking in the social
741 sciences and humanities. *GeoJournal*, 75(1), 3–13. doi:[10.1007/s10708-010-9340-3](https://doi.org/10.1007/s10708-010-9340-3)
- 742 Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Tröster, G., ... Lukowicz, P.
743 (2015). Smartphone-based recognition of states and state changes in bipolar disorder
744 patients. *IEEE Journal of Biomedical and Health Informatics*, 19(1), 140–148.
745 doi:[10.1109/JBHI.2014.2343154](https://doi.org/10.1109/JBHI.2014.2343154)
- 746 Harari, G., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016).
747 Using smartphones to collect behavioral data in psychological science: Opportunities,
748 practical considerations, and challenges. *Perspectives on Psychological Science*, 11(6),
749 838–854. doi:[10.1177/1745691616650285](https://doi.org/10.1177/1745691616650285)
- 750 Hijmans, R. J. (2017a). *Geosphere: Spherical trigonometry*. Retrieved from

- 751 <https://CRAN.R-project.org/package=geosphere>
- 752 Hijmans, R. J. (2017b). *Raster: Geographic data analysis and modeling*. Retrieved from
753 <https://CRAN.R-project.org/package=raster>
- 754 Jankowska, M. M., Schipperijn, J., & Kerr, J. (2015). A framework for using GPS data in
755 physical activity and sedentary behavior studies. *Exercise and Sport Sciences
Reviews*, 43(1), 48–56. doi:[10.1249/JES.00000000000000035](https://doi.org/10.1249/JES.00000000000000035)
- 756 LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., ... Schilit, B.
757 (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](https://doi.org/10.1007/11428572_8)
- 760 Liao, L., Patterson, D. J., Fox, D., & Kautz, H. (2007). Learning and inferring transportation
761 routines. *Artificial Intelligence*, 171(5), 311–331. doi:[10.1016/j.artint.2007.01.006](https://doi.org/10.1016/j.artint.2007.01.006)
- 762 London, I. (2016). *Encoding cyclical continuous features - 24-hour time. Ian london's blog*.
763 Retrieved April 27, 2018, from
764 [//ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/](https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/)
- 765 Müller, K. (2017). *Bindrcpp: An 'rcpp' interface to active bindings*. Retrieved from
766 <https://CRAN.R-project.org/package=bindrcpp>
- 767 Neuwirth, E. (2014). *RColorBrewer: ColorBrewer palettes*. Retrieved from
768 <https://CRAN.R-project.org/package=RColorBrewer>
- 769 Palmius, N., Tsanas, A., Saunders, K. E. A., Bilderbeck, A. C., Geddes, J. R., Goodwin, G.
770 M., & Vos, M. D. (2017). Detecting bipolar depression from geographic location data.
771 *IEEE Transactions on Biomedical Engineering*, 64(8), 1761–1771.
772 doi:[10.1109/TBME.2016.2611862](https://doi.org/10.1109/TBME.2016.2611862)
- 773 Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., & Matthiopoulos, J. (2008).
774 State-space models of individual animal movement. *Trends in Ecology & Evolution*,
775 23(2), 87–94. doi:[10.1016/j.tree.2007.10.009](https://doi.org/10.1016/j.tree.2007.10.009)
- 776 Pebesma, E. J., & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R News*,

- 777 5(2), 9–13. Retrieved from <https://CRAN.R-project.org/doc/Rnews/>
- 778 Preisler, H. K., Ager, A. A., Johnson, B. K., & Kie, J. G. (2004). Modeling animal
779 movements using stochastic differential equations. *Environmetrics* 15: P. 643-657.
- 780 Retrieved from <https://www.fs.usda.gov/treesearch/pubs/33038>
- 781 R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna,
782 Austria: R Foundation for Statistical Computing. Retrieved from
783 <https://www.R-project.org/>
- 784 Sadilek, A., & Krumm, J. (2016). Far out: Predicting long-term human mobility. *Microsoft
785 Research*. Retrieved from <https://www.microsoft.com/en-us/research/publication/far-predicting-long-term-human-mobility/>
- 787 Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D.
788 C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life
789 behavior: An exploratory study. *Journal of Medical Internet Research*, 17(7), e175.
790 doi:10.2196/jmir.4273
- 791 Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C. D., & Troelsen, J. (2014).
792 Dynamic accuracy of GPS receivers for use in health research: A novel method to
793 assess GPS accuracy in real-world settings. *Frontiers in Public Health*, 2, 21.
794 doi:10.3389/fpubh.2014.00021
- 795 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology:
796 Undisclosed flexibility in data collection and analysis allows presenting anything as
797 significant. *Psychological Science*, 22(11), 1359–1366. doi:10.1177/0956797611417632
- 798 Sobrado, B. (2018). *My thesis on missing data in GPS measurements*. Retrieved from
799 <https://github.com/sobradob/thesis>
- 800 Tatsiramos, K. (2009). Geographic labour mobility and unemployment insurance in europe.
801 *Journal of Population Economics*, 22(2), 267–283. doi:10.1007/s00148-008-0194-7
- 802 Thoen, E. (2017). *Padr: Quickly get datetime data ready for analysis*. Retrieved from

- 803 <https://CRAN.R-project.org/package=padr>
- 804 *Timeline*. (2017). Retrieved from <https://www.google.com/maps/timeline?pb>
- 805 Vaughan, D., & Dancho, M. (2018). *Tibbletime: Time aware tibbles*. Retrieved from
806 <https://CRAN.R-project.org/package=tibbletime>
- 807 Vaughan, D., & Dancho, M. (2018). *Tibbletime: Time aware tibbles*. Retrieved from
808 <https://CRAN.R-project.org/package=tibbletime>
- 809 Villanueva, C., & Aggarwal, B. (2013). The association between neighborhood
810 socioeconomic status and clinical outcomes among patients 1 year after
811 hospitalization for cardiovascular disease. *Journal of Community Health*, 38(4),
812 690–697. doi:[10.1007/s10900-013-9666-0](https://doi.org/10.1007/s10900-013-9666-0)
- 813 Wang, R., Harari, G., Hao, P., Zhou, X., & Campbell, A. T. (2015). SmartGPA: How
814 smartphones can assess and predict academic performance of college students. In
815 *Proceedings of the 2015 ACM international joint conference on pervasive and
816 ubiquitous computing* (pp. 295–306). New York, NY, USA: ACM.
817 doi:[10.1145/2750858.2804251](https://doi.org/10.1145/2750858.2804251)
- 818 Wickham, H. (2009). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.
819 Retrieved from <http://ggplot2.org>
- 820 Wickham, H. (2017). *Scales: Scale functions for visualization*. Retrieved from
821 <https://CRAN.R-project.org/package=scales>
- 822 Wickham, H., & Henry, L. (2018). *Tidyr: Easily tidy data with 'spread()' and 'gather()'*
823 functions. Retrieved from <https://CRAN.R-project.org/package=tidyr>
- 824 Wickham, H., & Ruiz, E. (2018). *Dplyr: A 'dplyr' back end for databases*. Retrieved from
825 <https://CRAN.R-project.org/package=dbplyr>
- 826 Wickham, H., Francois, R., Henry, L., & Müller, K. (2017). *Dplyr: A grammar of data
827 manipulation*. Retrieved from <https://CRAN.R-project.org/package=dplyr>
- 828 Wickham, H., Hester, J., & Francois, R. (2017). *Readr: Read rectangular text data*.
829 Retrieved from <https://CRAN.R-project.org/package=readr>
- 829 Wolf, J., Oliveira, M., & Thompson, M. (2003). Impact of underreporting on mileage and
830 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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%
Palnius	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%

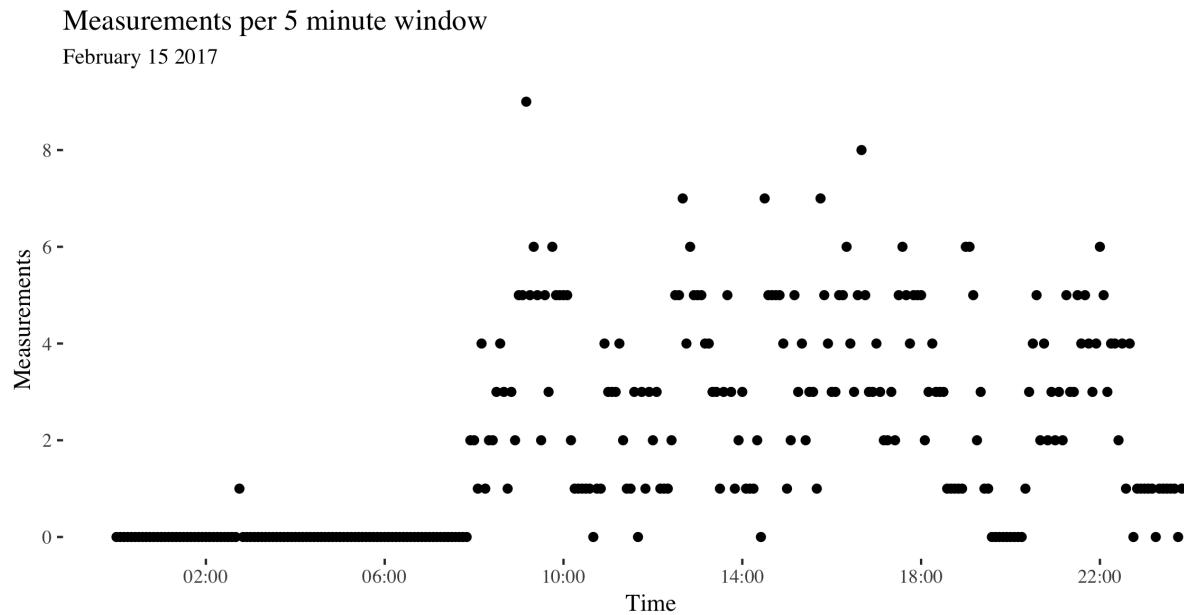


Figure 1. Example of missing data over the entire duration of a secondary log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points lie on the x-axis.

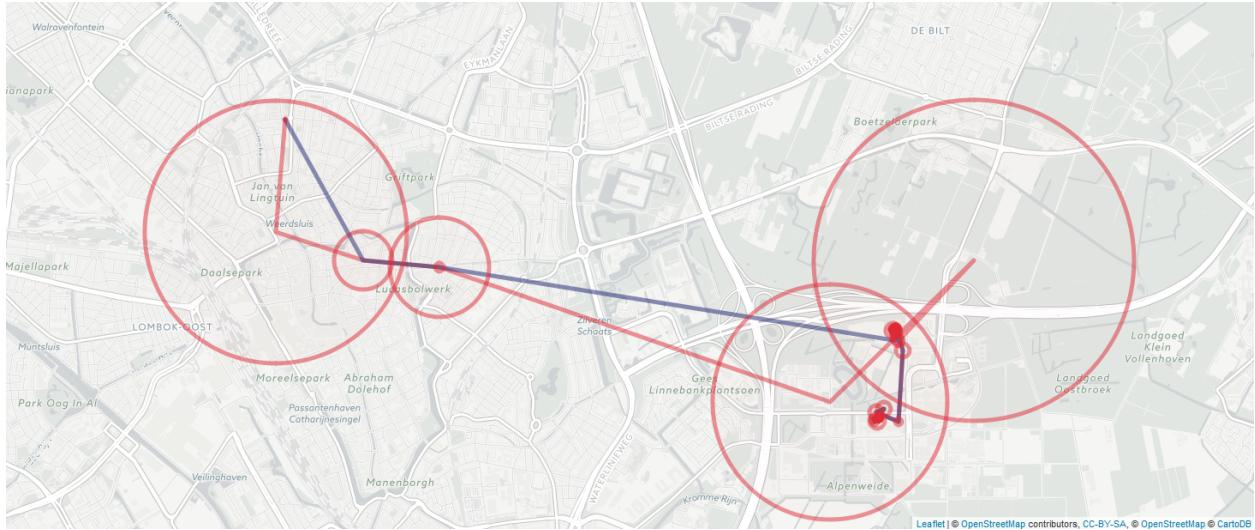


Figure 2. Measurement accuracy of each logged measurement of a morning journey on February 15th 2017. This includes all measurements from midnight to midday. The red circles denote the accuracy of all logged measurement points (the raw data). The points connected in time are connected by a line. The 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.

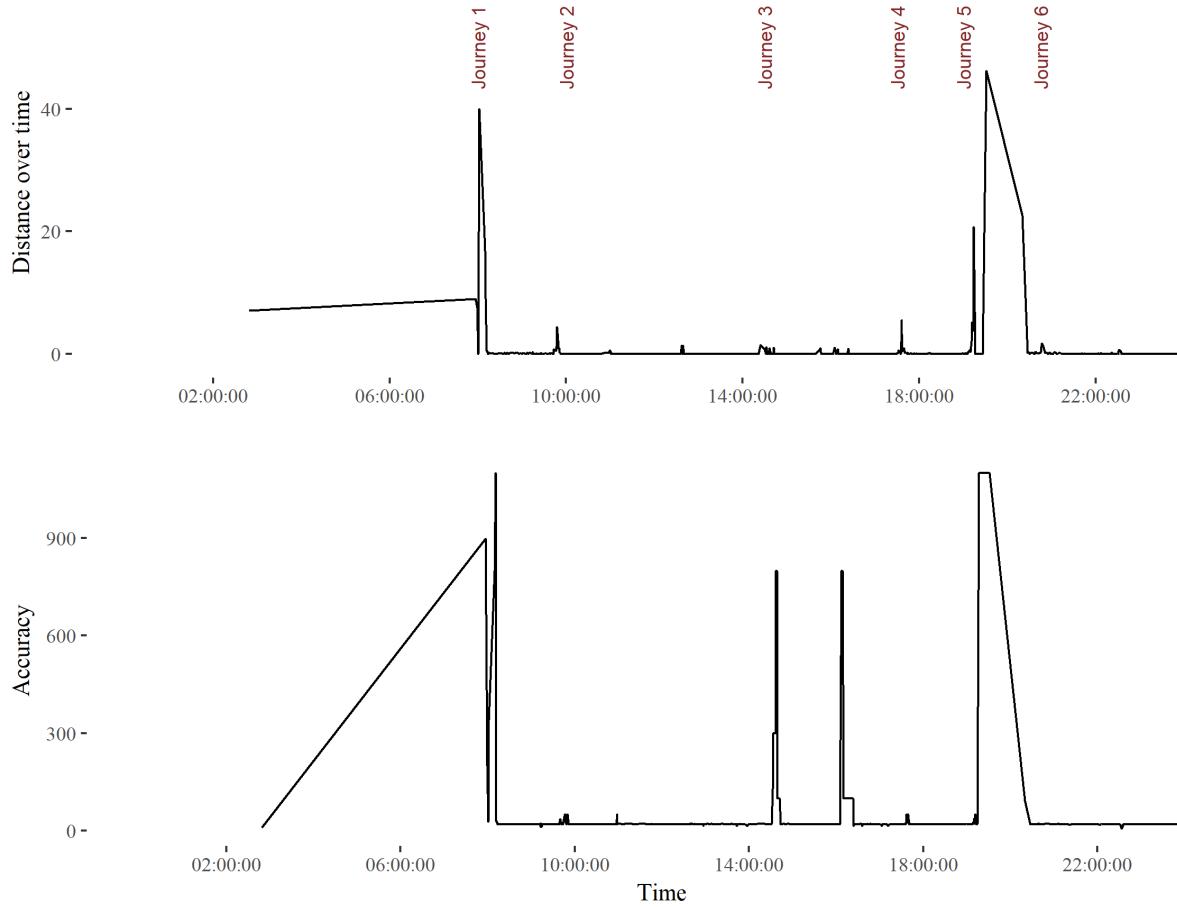


Figure 3. Measures of user activity and measurement accuracy on February 15th 2017. The upper chart shows the distance from the next measured point in meters over the course of the day. The first peak corresponds to the first journey from the user's home to a gym around 8am. The second, smaller peak 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.

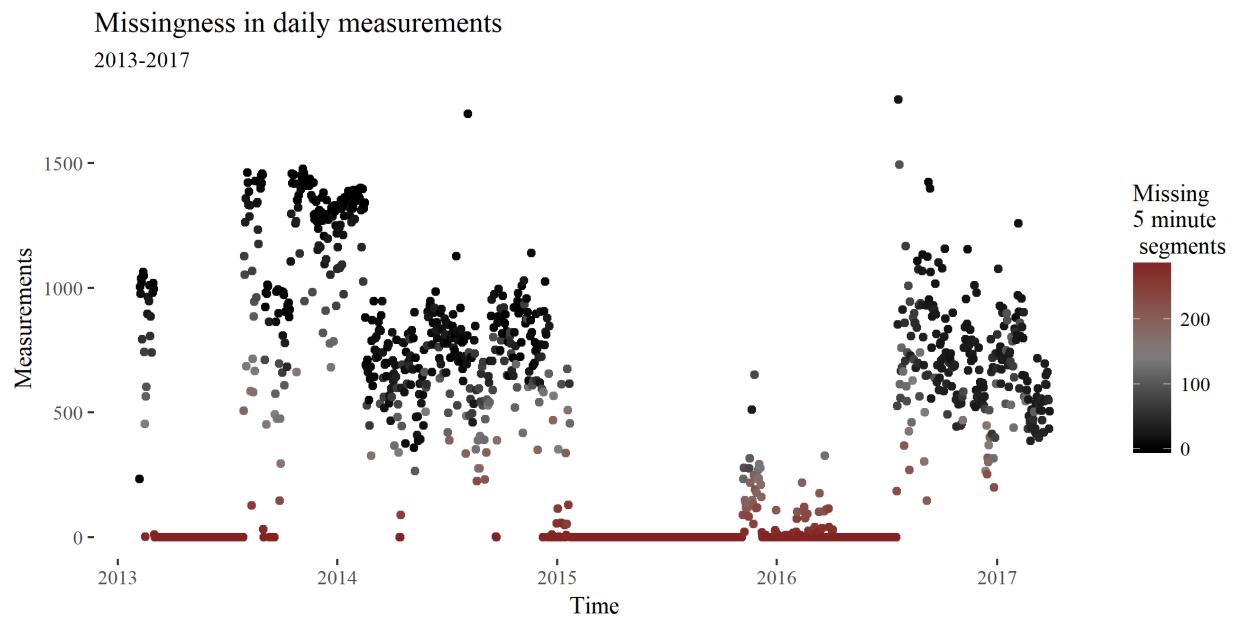


Figure 4. Missing data for the entire duration of the log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. The entire log contains several long periods with no information. These points are filled with red and lie on the x-axis.

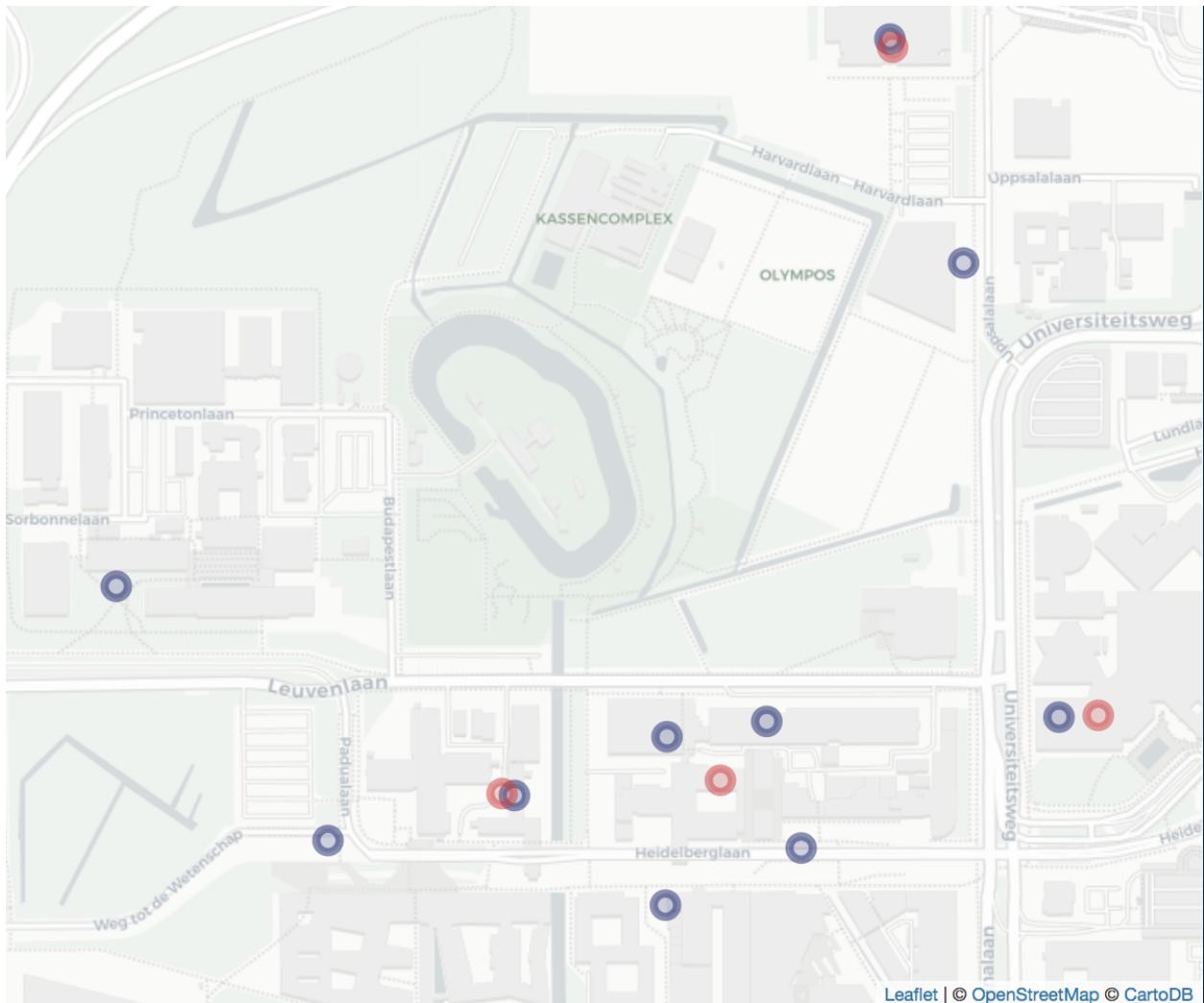


Figure 5. Example of pause locations in De Uithof university campus using 150 meters (blue) and 400 meters (red) as clustering parameters.

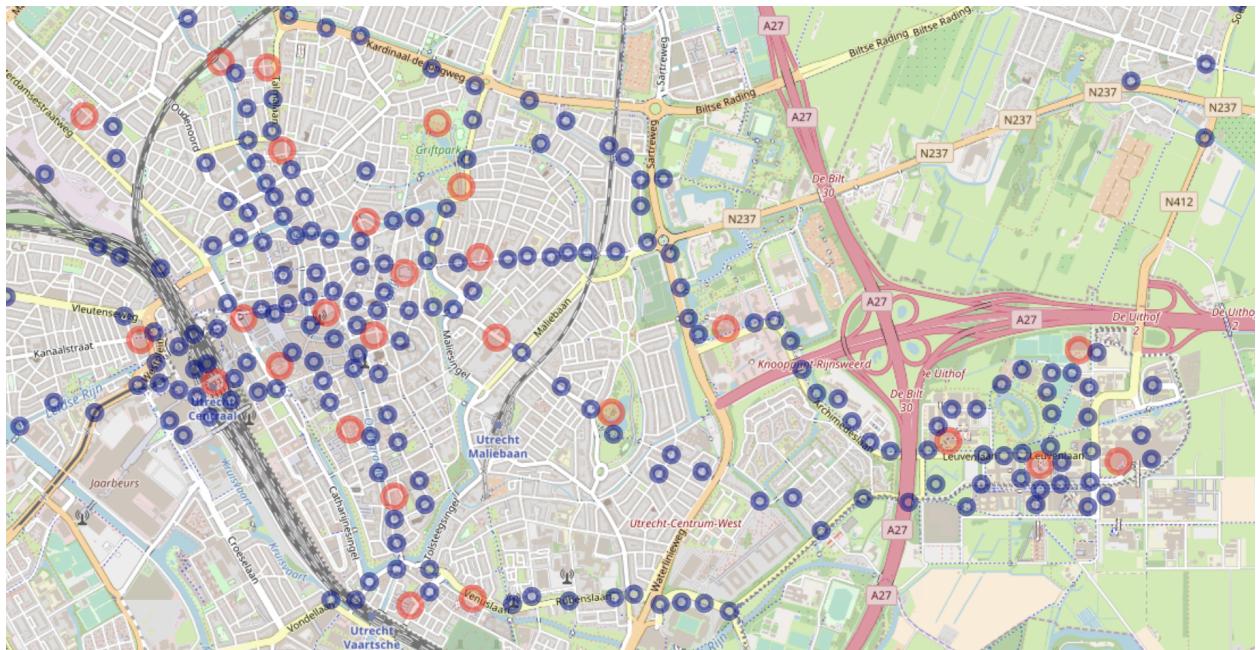


Figure 6. Excerpt of the personalised binned map of an individual. Red points are pause locations, blue points are path locations.

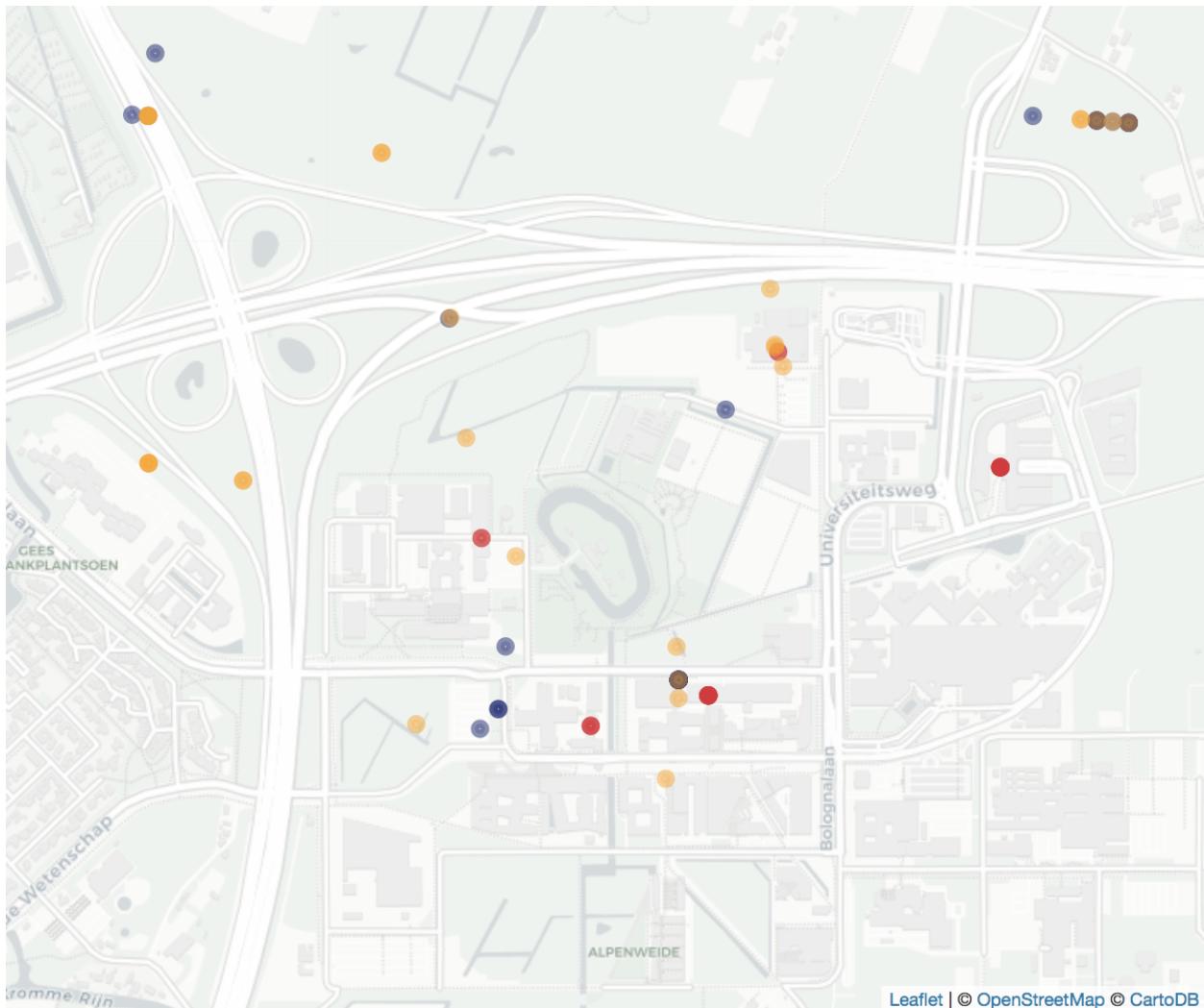


Figure 7. The difference between temporal and spatial downsampling. The blue circles are raw measurements, the yellow circles are temporally downsampled locations. Spatially downsampled locations are in red. Due to measurement sparsity and inaccuracy many of the temporally downsampled locations are in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

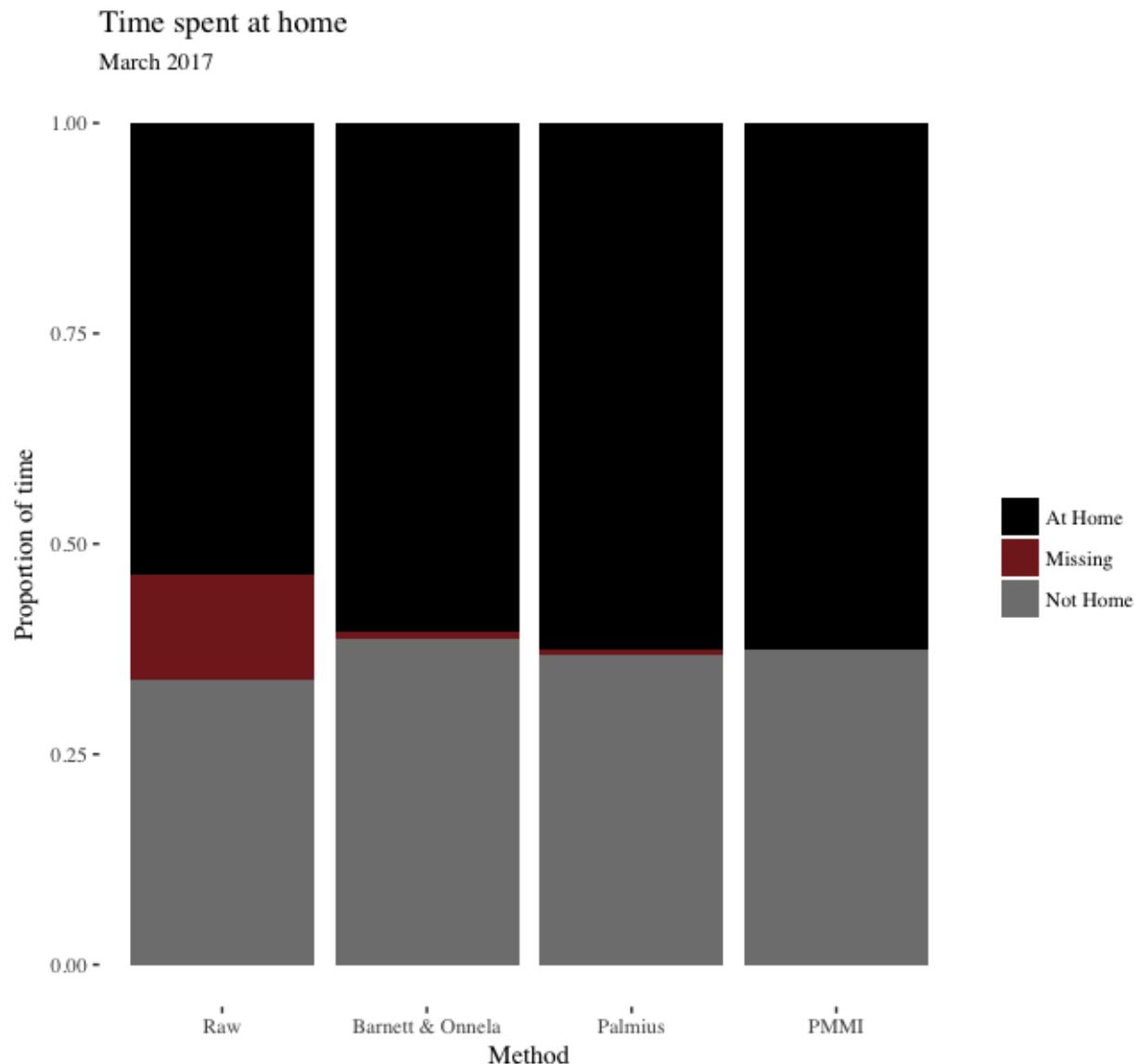


Figure 8. Proportion of time spent at home in March 2017. The raw values are estimated by downsampling temporally the latitude and longitude for every 5 minute time period in the month. We used each method's own binning method and classified as at home if the downsampled measurement was within 250 meters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