1        Handling missing data in smartphone location logs

2                          Boaz Sobrado[1]

3                          [1] Utrecht University

4                              Author Note

5        Department of Methodology & Statistics

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

7        Correspondence concerning this article should be addressed to Boaz Sobrado, . E-mail:

8        boaz@boazsobrado.com

9                                    Abstract

10  Using objective location data to infer the mobility measures of individuals is highly desirable,

11  but methodologically difficult. Using commercially gathered location logs from smartphones

12  holds great promise, as they have already been gathered, often span years and can be

13  associated to individuals. However, due to technical constraints this data is more sparse and

14  inaccurate than that produced by specialised equipment. In this paper we present a model

15  which leverages the periodicity of human mobility in order to impute missing data values.

16  Moreover, we will assess the performance of the model relative to currently used methods,

17  such as linear interpolation.

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

19      Word count: 2297

<sub>20</sub>                          Handling missing data in smartphone location logs

<sub>21</sub>          How people move about in their environment affects a wide range of outcomes, such as

<sub>22</sub>  health, income and social capital (Goodchild & Janelle, 2010). A better understanding of

<sub>23</sub>  mobility could lead to better health and urban-planning policies. Yet a large part of studies

<sub>24</sub>  on human mobility are conducted with pen-and-paper travel diaries, despite well known

<sub>25</sub>  methodological flaws. The high cost and burden to respondents limits the span of data

<sub>26</sub>  collection. Short trips are frequently under-reported (Wolf, Oliveira, & Thompson, 2003) .

<sub>27</sub>  The self-reported duration of commutes is often underestimated (Delclòs-Alió, Marquet, &

<sub>28</sub>  Miralles-Guasch, 2017).

<sub>29</sub>          These obstacles can be overcome by using objective data on human mobility. Such data

<sub>30</sub>  can now be obtained using the Global Positioning System (GPS). GPS uses the distance

<sub>31</sub>  between a device and several satellites to determine location. GPS measurements can be

<sub>32</sub>  used to infer a vast range of socioeconomic and behavioural measures, including where the

<sub>33</sub>  individual lives, how much time he or she spends at home and where (and how) they travel.

<sub>34</sub>          Within behavioural sciences, researchers have used GPS data to investigate

<sub>35</sub>  wideranging topics. For example, Zenk, Schulz, and Odoms-Young (2009) investigate the

<sub>36</sub>  effects of the food environment on eating patterns. G. M. Harari et al. (2016) look at the

<sub>37</sub>  movement correlates of personality, finding that extrovers individuals spend less time at

<sub>38</sub>  home than introverts. Wang, Harari, Hao, Zhou, and Campbell (2015) looks at how

<sub>39</sub>  academic performance is affected by movement patterns. Palmius et al. (2017) use mobility

<sub>40</sub>  patterns to predict bipolar depression.

<sub>41</sub>          In most studies participants receive a specialised GPS devices to track their movement.

<sub>42</sub>  We call resulting logs *specialised logs*. Barnett and Onnela (2016) point out several

<sub>43</sub>  methodological issues with these studies. Like studies using pen-and-paper travel diaries,

<sub>44</sub>  collecting specialised logs is costly and places a high burden on participants. Besides,

<sub>45</sub>  introducing a new device to the participant's life may bias their behaviour. Due to these

<sub>46</sub>  drawbacks specialised logs usually span a short amount of time. Barnett and Onnela (2016)

advocate installing a custom-made tracking app on user's phones (*custom logs*). Another

solution is to take advantage of existing smartphone location logs (*secondary logs*) . For

instance, Google Location History contains information on millions of users (Location

History, 2017). Often, secondary logs span several years. By law, secondary logs are

accessible to users for free (Commission, 2017). Yet, secondary logs also present

methodological challenges. They were created for non-academic purposes under engineering

constraints (detailed subsequently). These constraints mean that sensors do not track users

continuously, meaning that the resulting logs can be sparse and inaccurate. Hence, two

important challenges are dealing with measurement noise and missing data.

Missing data is a pervasive issue as it can arise due to several reasons. Technical

reasons include signal loss, battery failure and device failure. Behavioural reasons include

leaving the phone at home or switching the device off. As a result, secondary logs often

contain wide temporal gaps with no measurements. For instance, several research groups

studying mental health report missing data rates between 30% to 50% (Grünerbl et al., 2015;

Palmius et al., 2017; Saeb et al., 2015). Other researchers report similar trends in different

fields (e.g. G. M. Harari et al., 2016; Jankowska, Schipperijn, & Kerr, 2015).

There is no golden standard for dealing with missing data in GPS logs (Barnett &

Onnela, 2016). Importantly, spatiotemporal measurements are often correlated in time and

space. This means that common methods, such as mean imputation, are unsuitable. For

example, imagine an individual who splits almost all her time between work and home.

Suppose she spends a small amount of time commuting between the two along a circular path.

Using mean imputation to estimate her missing coordinates, we impute her to be at the

midpoint between home and work. She has never and will never be there! Worryingly, there

is little transparency on how researchers deal with missing data (Jankowska et al., 2015).

The accuracy of smartphone location measurements is substantially lower than that of

professional GPS trackers. Android phones collect location information through a variety of

methods. Other than GPS measurements, Androids use less-accurate heuristics such as WiFi

74 access points and cellphone triangulation. Different methods are used because of

75 computational and battery constraints. GPS is the most energy consuming sensor on most

76 smartphones (M. Y. Chen et al., 2006; LaMarca et al., 2005). In professional GPS trackers

77 less than 80% of measurements fall within 10 meters of the true location. GPS measures are

78 most inaccurate in dense urban locations and indoors (S. Duncan et al., 2013; Schipperijn et

79 al., 2014). Unfortunately for researchers, this is where people in the developed world spend

80 most of their time.

81        Noisy data can lead to inaccurate conclusions if it is not accounted for. Suppose we

82 wish to calculate an individual's movement in a day. A simple approach would be to

83 calculate the sum of the distance between each measurement. But if there is noise, the

84 coordinates will vary even though the individual is not moving. If the measurements are

85 frequent and noisy, we will calculate a lot of movement, even if the individual did not move

86 at all! This issue is also visualised in Figure 2. The problem is further complicated because

87 missing data and noisy measurements are related. Methods used by researchers to reduce

88 noise, such as throwing out inaccurate measurements (e.g. Palmius et al., 2017), can

89 exacerbate the severity of the missing data problem.

90        In this paper we will explore in detail the problem of missing data and measurement

91 error in secondary location logs. Moreover, we will compare methods used to deal with these

92 problems.

## A concrete example

94        There is little literature on dealing with missing data in custom or secondary logs.

95 Thus it is worth illustrating the typical characteristics using an example data set. The

96 example data set comes from the Google Location History of a single individual. It spans

97 from January 2013 to January 2017 and contains 814 941 measurements. The data set

98 contains a multitude of variables, including inferred activity and velocity. We will focus on

99 measurements of latitude, longitude, accuracy (defined below). All measurements are paired

with a timestamp.

**Aggregate measures.** Social scientists are most interested in aggregating spatiotemporal data to more socially relevant information. For instance, Palmius et al. (2017) calculate several measures to predict bipolar depression, such as the total distance travelled, the amount of time spent at home, travelling and the regularity of diurnal movement. As an example we calculate two of these measures: time spent at home and distance travelled for the month of February. Figure 1 shows how the aggregate measures are influenced by the noise in the measurements and the missingness. In the case of distance travelled, simply filtering out the most inaccurate measures leads to significantly lower estimates. This phenomenon is mapped out on Figure 2 . On the other hand, the time spent at home would be greatly influenced at times by the imputation method, as several days with long periods of missing data.

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

**Missingness.** Over 54% of the data is missing for the entire duration of the log. This may be misleading as there are several long periods with no measurements whatsoever (see Figure 4). For days which were not entirely missing, approximately 22% of all five minute segments were missing. The structure of missingness of a day with measurements is shown in

127 Figure 5. As you can see, there are several long periods over the course of the log for which

128 there are no measurements. Moreover, even during a single day there are continuous periods

129 where there is missing data, mostly during the late hours of the night in this case.

## Methods

### Two-dimensional projections and notation

132     GPS measurements provide us with coordinates on the surface of the earth. Because

133 most mobility metrics are computed for data in $\mathbb{R}^2$, we are interested in mapping these $\mathbb{R}^3$

134 measurements on a 2D Euclidean plane. Projecting three dimensional measurements onto a

135 two dimensional plane results in distortion. To minimise errors we borrow an error

136 minimising projection method from Barnett and Onnela (2016).

137     Having thus converted lattitude and longitude onto coordinates unique to each

138 individual, let a person's true location on this two-dimensional plane be $G(t) = [G_x(t)G_y(t)]$

139 where $G_x(t)$ and $G_y(t)$ denote the location of the individual at time $t$ on the x-axis and

140 y-axis respectively. Moreover, let $D \in \mathbb{R}^2$ be the recorded data containing lattitude and

141 longitude. In addition, let $a$ denote the estimated accuracy of the recorded data. accuracy.

142 $G(t)$, $D$ and $a$ are indexed by time labled by the countable set $t = t_1 < ... < t_{n+1}$. For

143 simplicity, let each entry in the discrete index set $t$ represent a 5 minute window. The

144 measure of accuracy $a_t$ is given in meters such that it represents the radius of a 67%

145 confidence circle. If $D_t = \emptyset$ it is considered *missing* and it is not missing otherwise.

146     When several data sets are available from individuals living in overlapping areas we

147 can construct a $t \times i$ matrix $M$ where the entry $M(t, i)$ contains $G(t)$ for the individual $i$.

### Selecting candidate models

149     There is no golden standard or established practice in how to deal with missing data in

150 GPS logs. Researchers are generally vague about what practices they follow (Jankowska et

151 al., 2015). Ostensibly this is because they are unaware of possible solutions. In an attempt

to elucidate the topic, we explore potential solutions. We will argue that extensively used

spatiotemporal methods, such as state space models (SSMs), are not well suited to deal with

human mobility patterns. We also discuss in detail two approaches which deal explicitly with

mobility patterns from custom or secondary logs (Barnett & Onnela, 2016; Palmius et al.,

2017 ).

There is a vast literature on using SSMs in spatiotemporal statistics. For example,

ecologists have used SSMs to explain how animals interact with their environment

(Patterson, Thomas, Wilcox, Ovaskainen, & Matthiopoulos, 2008). These models can be

quite complex. 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 us, the main limitation of SSMs is that they ignore regular movement routines. For

instance, humans tend to go to work on weekdays and sleep at night. Because SSMs are

based on the Markov property, they cannot incorporate this information. The estimated

location $G(t)$ at timepoint $t$ is often based only upon measurements $D_t$, $D_{t-1}$ and ignores all

$D_{t-i}|i \geq 2$. Hierarchical structuring and conditioning on a larger context have been suggested

as ways to add periodicity to Markovian models. These solutions are often computationally

intractable or unfeasible (Sadilek & Krumm, 2016). For this reason we do not consider SSMs

to be useful for imputing missing data. Nonetheless, they could be of use in filtering noise.

In climate or geological research spatiotemporal imputation methods are often used.

For instance, the CUTOFF method estimates missing values using the nearest observed

neighbours (Feng, Nowak, O'Neill, & Welsh, 2014 ). The authors illustrate their example

using rainfall data from gauging stations across Australia. Similarly, Z. Zhang et al. (2017)

179 use a variety of machine learning methods to impute missing values. The example provided

180 relates to underground water data. Generally these models assume fixed measurement

181 stations (such as rainfall gauging stations).

182 For this reason they cannot be easily applied to missing mobility tracks. Feng et al.

183 (2014) claim their model could be used to establish mobility patterns. This may be possible

184 by dividing the sample space into rasters. Each raster would be analogous to a measurement

185 station. These artificial stations could "measure" the probability of the individual being

186 there. To our knowledge such models have not been implemented for mobility traces and

187 seems computationally inefficient.

188 On the other hand, a few researchers have explicitly attempted to impute missing data

189 from human mobility patterns. Palmius et al. (2017) deal with the measuremement

190 inaccuracy of $D$ in custom logs by removing from the data set all unique low-accuracy $a$ data

191 points that had $\frac{d}{dt}D > 100\frac{km}{h}$. Subsequently the researchers down sample the data to a

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

193 "If the standard deviation of $[D]$ in both latitude and longitude within a 1 h

194 epoch was less than 0.01 km, then all samples within the hour were set to the

195 mean value of the recorded data, otherwise a 5 min median filter window was

196 applied to the recorded latitude and longitude in the epoch".

197 Missing data was imputed using the mean of measurements close in time if the

198 participant was recorded within 500m of either end of a missing section and the missing

199 section had a length of $\leq 2h$ or $\leq 12h$ after 9pm.

200 Barnett and Onnela (2016) follow a different approach which is, to the best of our

201 knowledge, the only pricipled approach to dealing with missing data in human mobility data.

202 Barnett and Onnela (2016) work with custom logs where location is measured for 2 minutes

203 and subsequently not measured for 10 minutes. In the words of the authors, Barnett and

204 Onnela (2016) handle missing data by:

"simulat[ing] flights and pauses over the period of missingness where the direction,

duration, and spatial length of each flight, the fraction of flights versus the

fraction of pauses, and the duration of pauses are sampled from observed data."

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

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

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

Analyses were performed using R and a multitude of other statistical packages (Arnold, 2017; Auguie, 2017; Aust, 2016; Bivand, Pebesma, & Gomez-Rubio, 2013; Cheng, Karambelkar, & Xie, 2017; Francois, 2017; E. J. Pebesma & Bivand, 2005; R Core Team, 2017; RStudio Team, 2015; Thoen, 2017; Vaughan & Dancho, 2017; Wickham, 2009, 2017; Wickham & Henry, 2017; Wickham, Francois, Henry, & Muller, 2017).

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

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

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

**Imputation methods.** Four imputation methods were selected in order to cover a wide range of techniques applied in the literature:

1. Mean imputation as described by Palmius et al. (2017).

2. The model developed by Barnett and Onnela (2016) using both spatial and temporal proximity.

3. Simple linear interpolation was used as a benchmark model.

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

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

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

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

# References

Arnold, J. B. (2017). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'.* Retrieved

from https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2017). *GridExtra: Miscellaneous functions for "grid" graphics.* Retrieved from

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

Aust, F. (2016). *Citr: 'RStudio' add-in to insert markdown citations.* Retrieved from

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

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

Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R,

second edition.* Springer, NY. Retrieved from http://www.asdar-book.org/

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. (2017). *Leaflet: Create interactive web maps with the

javascript 'leaflet' library.* Retrieved from

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

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

Retrieved from

http://ec.europa.eu/justice/data-protection/individuals/rights/index_en.htm

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.

276    *Travel Behaviour and Society, 9*(Supplement C), 1–9. doi:10.1016/j.tbs.2017.07.001

277  Doust, P. (2013). *Smoothing - Smooth GPS data - Stack Overflow.* Retrieved from

278    https://stackoverflow.com/questions/1134579/smooth-gps-data

279  Duncan, S., Stewart, T. I., Oliver, M., Mavoa, S., MacRae, D., Badland, H. M., & Duncan,

280    M. J. (2013). Portable global positioning system receivers: Static validity and

281    environmental conditions. *American Journal of Preventive Medicine, 44*(2), e19–29.

282    doi:10.1016/j.amepre.2012.10.013

283  Feng, L., Nowak, G., O'Neill, T., & Welsh, A. (2014). CUTOFF: A spatio-temporal

284    imputation method. *Journal of Hydrology, 519*, 3591–3605.

285    doi:10.1016/j.jhydrol.2014.11.012

286  Francois, R. (2017). *Bibtex: Bibtex parser.* Retrieved from

287    https://CRAN.R-project.org/package=bibtex

288  Goodchild, M. F., & Janelle, D. G. (2010). Toward critical spatial thinking in the social

289    sciences and humanities. *GeoJournal, 75*(1), 3–13. doi:10.1007/s10708-010-9340-3

290  Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Tröster, G., . . . Lukowicz, P.

291    (2015). Smartphone-based recognition of states and state changes in bipolar disorder

292    patients. *IEEE Journal of Biomedical and Health Informatics, 19*(1), 140–148.

293    doi:10.1109/JBHI.2014.2343154

294  Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D.

295    (2016). Using Smartphones to Collect Behavioral Data in Psychological Science:

296    Opportunities, Practical Considerations, and Challenges. *Perspectives on*

297    *Psychological Science, 11*(6), 838–854. doi:10.1177/1745691616650285

298  Jankowska, M. M., Schipperijn, J., & Kerr, J. (2015). A Framework For Using GPS Data In

299    Physical Activity And Sedentary Behavior Studies. *Exercise and Sport Sciences*

300    *Reviews, 43*(1), 48–56. doi:10.1249/JES.0000000000000035

301  LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., . . . Schilit, B.

302    (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

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

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/

Rhee, I., Shin, M., Hong, S., Lee, K., & Chong, S. (2007). Human Mobility Patterns and
Their Impact on Routing in Human-Driven Mobile Networks. *ACM HotNets 2007.*
Retrieved from http://koasas.kaist.ac.kr/handle/10203/160927

RStudio Team. (2015). *RStudio: Integrated development environment for r.* Boston, MA:
RStudio, Inc. Retrieved from http://www.rstudio.com/

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

Thoen, E. (2017). *Padr: Quickly get datetime data ready for analysis*. Retrieved from
https://CRAN.R-project.org/package=padr

Vaughan, D., & Dancho, M. (2017). *Tibbletime: Time aware tibbles*. Retrieved from
https://CRAN.R-project.org/package=tibbletime

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. (2017). *Tidyr: Easily tidy data with 'spread()' and 'gather()'
functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

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

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*

357        *Research Board, 1854*, 189–198. doi:10.3141/1854-21

358   Zenk, S. N., Schulz, A. J., & Odoms-Young, A. (2009). How Neighborhood Environments

359        Contribute to Obesity. *The American Journal of Nursing, 109*(7), 61–64.

360        doi:10.1097/01.NAJ.0000357175.86507.c8

361   Zhang, Z., Yang, X., Li, H., Li, W., Yan, H., & Shi, F. (2017). Application of a novel hybrid

362        method for spatiotemporal data imputation: A case study of the Minqin County

363        groundwater level. *Journal of Hydrology, 553*(Supplement C), 384–397.

364        doi:10.1016/j.jhydrol.2017.07.053

Table 1

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

| Log duration | Logged days | Observations | Missing days | Missing data | Mean Accuracy |
|---|---|---|---|---|---|
| 2013-02-06 to 2017-03-29 | 1,512.00 | 646,376.00 | 635.00 | 0.22 | 127.78 |
| 2016-07-14 to 2017-05-10 | 300.00 | 158,382.00 | 3.00 | 0.41 | 1,394.60 |
| 2014-01-22 to 2017-01-23 | 1,097.00 | 814,941.00 | 80.00 | 0.25 | 121.83 |
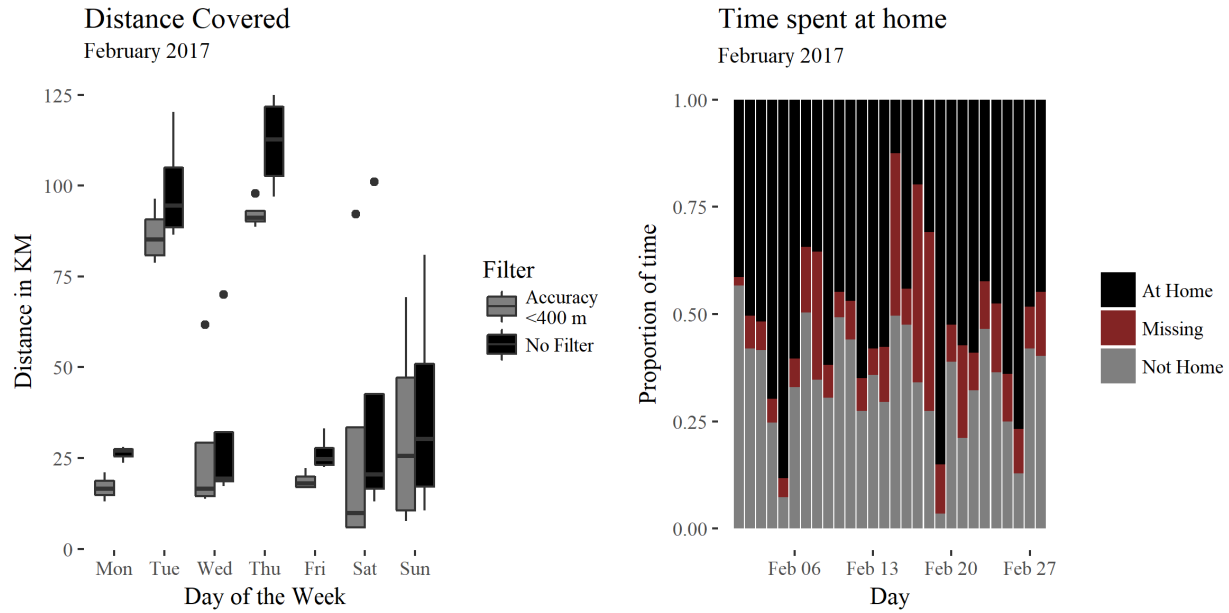
*Figure 1*. Proportion of time spent at home and distance travelled in February 2017. We estimate time spent at home by calculating the mean lattitude and longitude for every 5 minute time period in the month. The user is coded as at home if within 250 meters of the home coordiantes. One can see that several days have long periods with missing data. More accurate estimates can be reached by imputing the missing measurements. As for distance travelled, it is evident that although the behavioural trends remain the same, the absolute estimate of distance travelled varies depending on the filter.
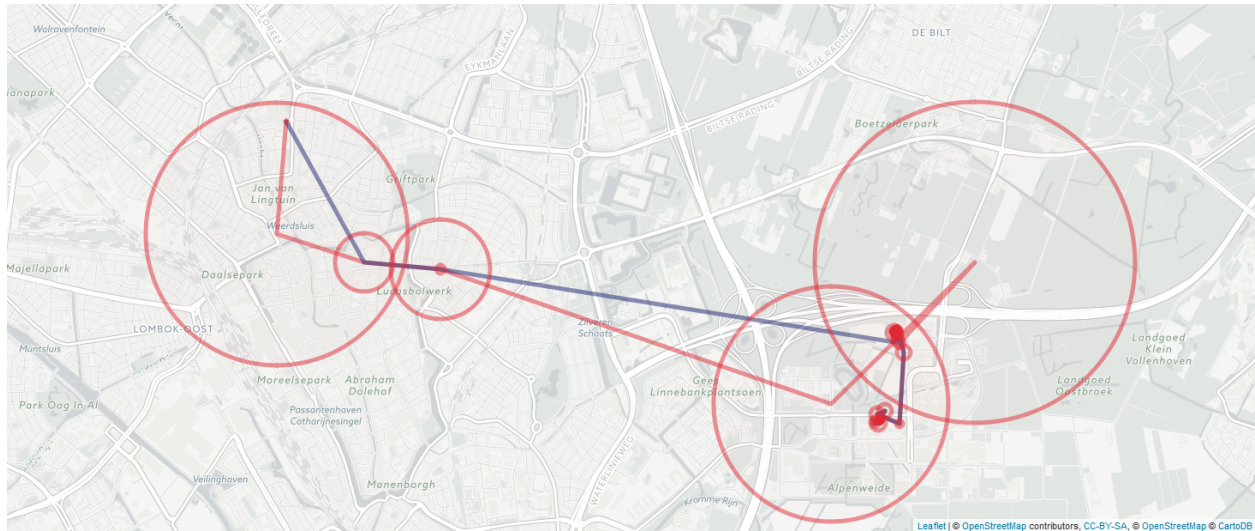
*Figure 2.* Measurement accuracy of each logged measurement of a morning journey on February 15th 2017. This includes all measurements from midnight to midday. The red circles denote the accuracy of all logged measurement points (the raw data). The points connected in time are connected by a line. The blue line shows the path without the most inaccurate (accuracy > 400 meters) points filtered out. The red line shows the path with all measurements included.
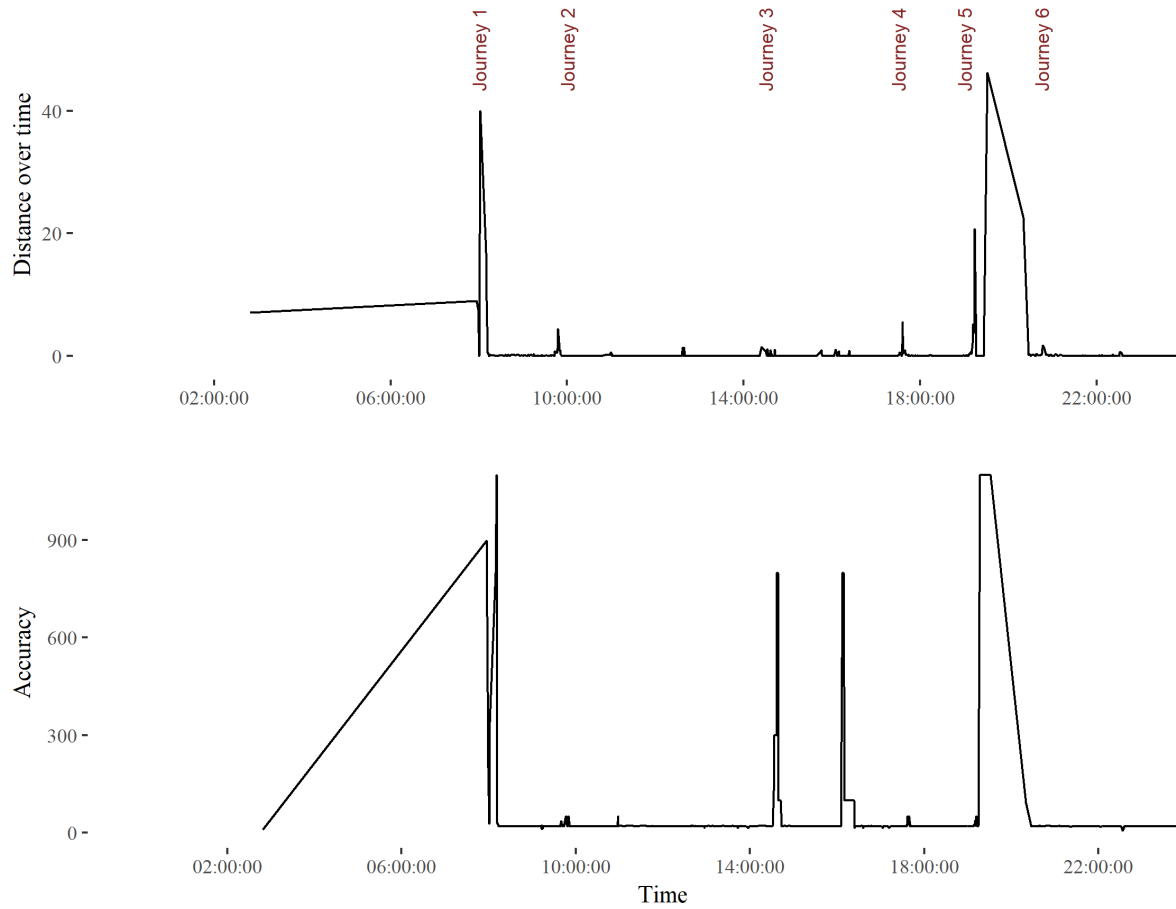
*Figure 3*. Measures of user activity and measurement accuracy on February 15th 2017.The upper chart shows the distance from the next measured point in meters over the course of the day. All journeys are labled at the top of the chart. The first peak corresponds to the first journey from the user's home 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. All other journeys are not shown 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.
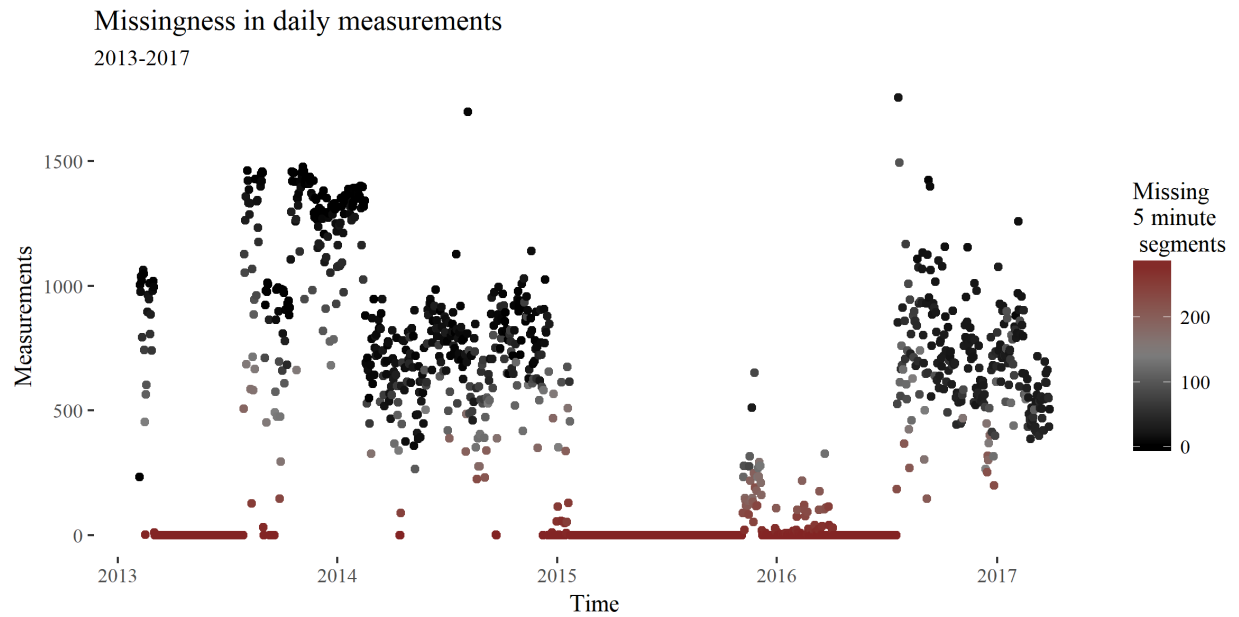
*Figure 4.* Example of missing data over the entire duration of the log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points are filled with red and lie on the x-axis.

Measurements per 5 minute window
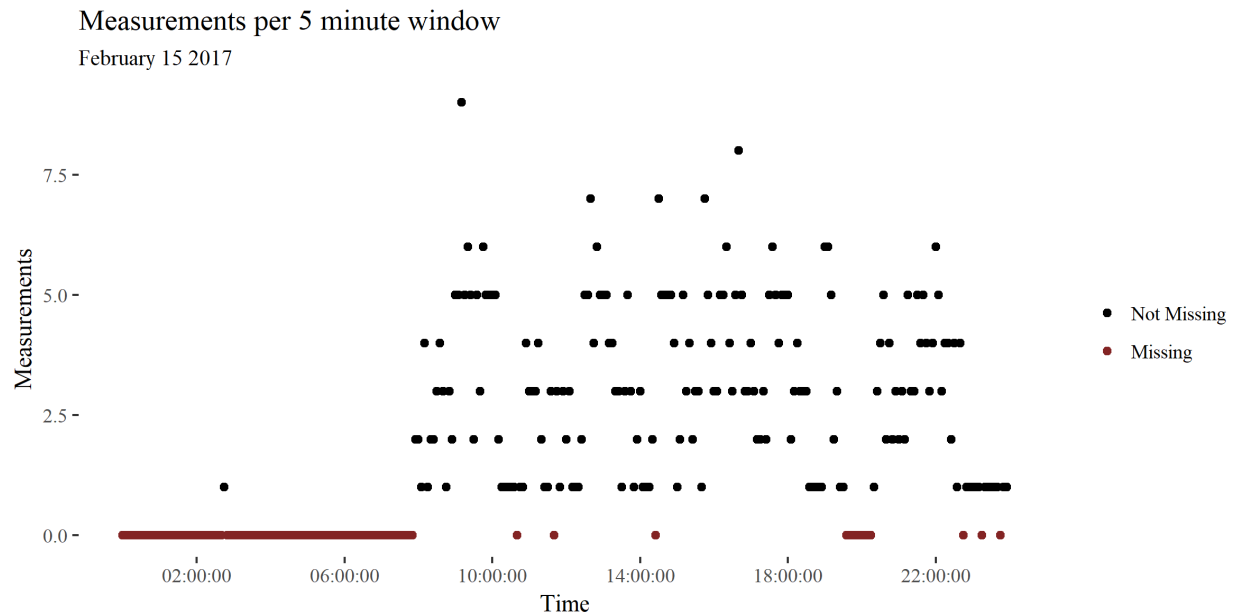
February 15 2017



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