Handling missing data in smartphone location logs

Boaz Sobrado[1]

[1] Utrecht University

Author Note

Department of Methodology & Statistics

Submitted as a research report.

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

Abstract

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

Keywords: GPS, Human Mobility

Word count: X

Handling missing data in smartphone location logs

How active people are and how they interact with their environment affects a wide range of outcomes including health, income and social capital (Goodchild & Janelle, 2010). A better understanding of both within-person and between-person variability in geospatial patterns could be conducive to better social, health and urban-planning policies. Yet a large part of studies on human mobility are largely based on pen-and-paper travel diaries. These surveys have known methodological flaws, such as the short period of data collection (due to costs and burden to respondents), the underreporting of short trips (Wolf, Oliveira, & Thompson, 2003) and the underestimation of the duration of commutes (Delclòs-Alió, Marquet, & Miralles-Guasch, 2017).

These obstacles can be overcome by using objective data on human mobility. The Global Positioning System (GPS), which uses the distance between a device and a number of satellites to determine location, provides such data. Within behavioural science, this type of data has been used to investigate topics such as the effects of the food environment on eating patterns (Zenk, Schulz, & Odoms-Young, 2009), the movement correlates of personality (G. M. Harari et al., 2016), academic performance (Wang, Harari, Hao, Zhou, & Campbell, 2015) and bipolar disorder (Palmius et al., 2017).

In most of these studies participants are given a specialised device, resulting in accurate mobility GPS data (*specialised logs*). However, Barnett and Onnela (2016) point out that these studies are not scalable due to cost and burden to participants. Moreover, they may be biased because of the introduction of a new device to the participant's life. Because of these drawbacks, specialised logs usually span a short amount of time. Barnett and Onnela (2016) advocate installing a custom-made tracking app on user's phones (*custom logs*). Another solution is to take advantage of existing smartphone location logs, such as Google Location History, which store location information of millions of users spanning several years (Location History, 2017) (*secondary logs*). By law, logs can be accessed and shared by users for free (Commission, 2017). Yet, because they were created for non-academic purposes under engineering constraints, the sensors do not monitor continuously and 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 multiple reasons. Technical reasons include signal loss, battery failure and device failure. Behavioural reasons include leaving the phone at home, switching the phone off, switching location measurements off, and so on. As a result, applied researchers are often left with wide temporal gaps with no measurements. For instance, different groups studying the effect of bipolar disorder on human movement have reported missing data rates between 30% to 50% (Grünerbl et al., 2015; Palmius et al., 2017; Saeb et al., 2015). Similar trends are consistently reported in other fields (e.g. G. M. Harari et al., 2016; Jankowska, Schipperijn, & Kerr, 2015).

There is currently no golden standard in how to deal with missing data in custom or secondary logs (Barnett & Onnela, 2016). Traditional missing data methods, such as mean imputation, cannot be used easily in spatiotemporal data because the measurements are correlated in time and space. For example, assume the simplistic case that an individual spends almost half her time at work, half her time at home and a small amount of time commuting between the two along a angled path. Using mean imputation would result in imputed values of her being at the geometric midpoint between the home and work for all missing values, even though she has never been there and never will be. Worringly, Jankowska et al. (2015) have pointed out that there is often little transparency regarding decisions of how to deal with missing data.

The accuracy of GPS measurements in smartphones is substantially lower than in professional grade GPS trackers. For example, Android phones collect location information through a variety of methods, such as from WiFi access points, cellphone triangulation, and GPS measurements. They use different methods due to computational and battery constraints (M. Y. Chen et al., 2006; LaMarca et al., 2005).In professional grade GPS trackers less than 80% of measurements fall within 10 meters of the true location. GPS measures are reported to be most inaccurate in high density urban locations and indoors (S. Duncan et al., 2013; Schipperijn et al., 2014). Unfortunately for social scientists, this happens to be where most people in the developed world tend to spend most of their time.

On the other hand, noisy data can lead to inaccurate conclusions if it is not accounted for, such as overestimating the movement of individuals. For instance, suppose that a naive researcher calculates the distance travelled by an individual by drawing a line between each measured point and calculating the sum of the length of all of these lines. If there is noise, the measurements will vary even though the individual is not moving. If the measurements are frequent, then the researcher will end up with a lot of movement, even though the individual did not move at all. The problem is further complicated by the fact that missing data and noisy measurements are related to each other. For instance, 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 compare methods used to deal with measurement error and missing data in mobility patterns from secondary GPS logs.

A concrete example

Given that there is little literature on dealing with missing data in custom or secondary logs it is worth illustrating the typical characteristics of this data using an example data set. The example dataset comes from the Google Location History of a single individual and spans from January 2013 to January 2017. It was recorded with multiple different Android devices and contains 814 941 measurements, with approximately 742 measurements per day ($\widehat{\sigma}$=868.15). The dataset contains a wide range of variables including inferred activity and velocity. For the purposes of this paper we will focus only on latitude, longitude, accuracy (defined below) and a timestamp.

Location logs and notation

GPS measurements report our location on a three dimensional planet, yet we are interested in placing these measurements on a two dimensional map. Projecting three dimensional measurements onto a two dimensional plane results in errors, in order to minimise these errors we borrow an error minimising projection method from Barnett and Onnela (2016).

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

Accuracy in location logs. In the example data set the distribution of a 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 1 shows how accuracy tends to vary as a function of user behaviour, time and location. There are several recurring low-accuracy points, most likely cell-phone triangulation towers.

Missingness. Over 54% of the data is missing for the entire duration of the log. However,this is misleading as there are several long periods with no measurements whatsoever. The structure of missingness of a day with measurements is shown below.

Current methods for imputing missing spatiotemporal data

As we have mentioned before, Jankowska et al. (2015) point out that there is little transparency regarding decisions of how to deal with missing data in GPS mobility logs. We suspect this is because the highly interdisciplinary nature of the problem means researchers are unaware of potential solutions. For this reason, we consider it important to briefly discuss a few methods one could consider in order to deal with measurement inaccuracy and missing data problems in spatiotemporal data. In doing so, we will argue that state space models (SSMs) and spatiotemporal imputation methods, which are extensively used to model missing and noisy spatiotemporal data, are not well suited to deal with human mobility patterns. Moreover, we discuss in detail two approaches by Palmius et al. (2017) and Barnett and Onnela (2016) which deal explicitly with missing data in mobility patterns from smartphone GPS logs.

There is a vast literature of using SSMs to improve measurements accuracy and deal with missing data. Behavioural ecologists for instance, have used SSMs extensively to explain how animals interact with their environment (Patterson, Thomas, Wilcox, Ovaskainen, & Matthiopoulos, 2008). These models can be quite complex, for example 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. In fact, 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.

However, the main limitation of SSMs is that they ignore the fact that humans have regular movement routines, such as going to work or shopping for groceries on weekends. This limitation is due to the fact that SSMs are based on the Markov property. Thus, 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 improve the performance of Markovian models, but 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, although they could be of use in filtering noise.

In addition to SSMs there are spatiotemporal imputation methods often used in climate or geological research for estimating missing data. For instance, Feng, Nowak, O'Neill, and Welsh (2014) illustrate their CUTOFF method, which relies on estimating missing values using the nearest observed neighbours in time, using rainfall data from dozens of gauging stations across Australia. Similarly, Z. Zhang et al. (2017) use a variety of machine learning methods to present their model based on underground water data in China.

The difficulty of applying this class of spatiotemporal imputation models to human mobility tracks lies in the fact that these models generally assume fixed measurement stations (such as rainfall gauging stations). While Feng et al. (2014) claim their model could be used to establish mobility patterns, ostensibly by dividing the sample space into rasters analogous to measurement stations indicating a probability of the individual being there, this seems to be computationally unfeasible. To our knowledge such models have not been implemented for mobility traces.

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

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

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

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

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

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

Methods

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 (**???**; Bivand, Pebesma, & Gomez-Rubio, 2013; E. J. Pebesma & Bivand, 2005; R Core Team, 2017; Wickham, 2009; Wickham & Francois, 2016).

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

1. Mean imputation as described by Palmius et al. (2017).
2. The model developed by Barnett and Onnela (2016) using spatial, temporal and circadian 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.

References

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

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

Doust, P. (2013). *Smoothing - Smooth GPS data - Stack Overflow*. Retrieved from https://stackoverflow.com/questions/1134579/smooth-gps-data

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

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

doi:10.1016/j.jhydrol.2014.11.012

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

Grünerbl, A., Muaremi, A., Osmani, V., Bahle, G., Ohler, S., Tröster, G., … Lukowicz, P. (2015). Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE Journal of Biomedical and Health Informatics*, *19*(1), 140–148. doi:10.1109/JBHI.2014.2343154

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. doi:10.1177/1745691616650285

Jankowska, M. M., Schipperijn, J., & Kerr, J. (2015). A Framework For Using GPS Data In Physical Activity And Sedentary Behavior Studies. *Exercise and Sport Sciences Reviews*, *43*(1), 48–56. doi:10.1249/JES.0000000000000035

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., … Schilit, B. (2005). Place Lab: Device Positioning Using Radio Beacons in the Wild. In *Pervasive Computing* (pp. 116–133). Springer, Berlin, Heidelberg. doi:10.1007/11428572_8

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

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

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., & Francois, R. (2016). *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 Research Board*, *1854*, 189–198. doi:10.3141/1854-21

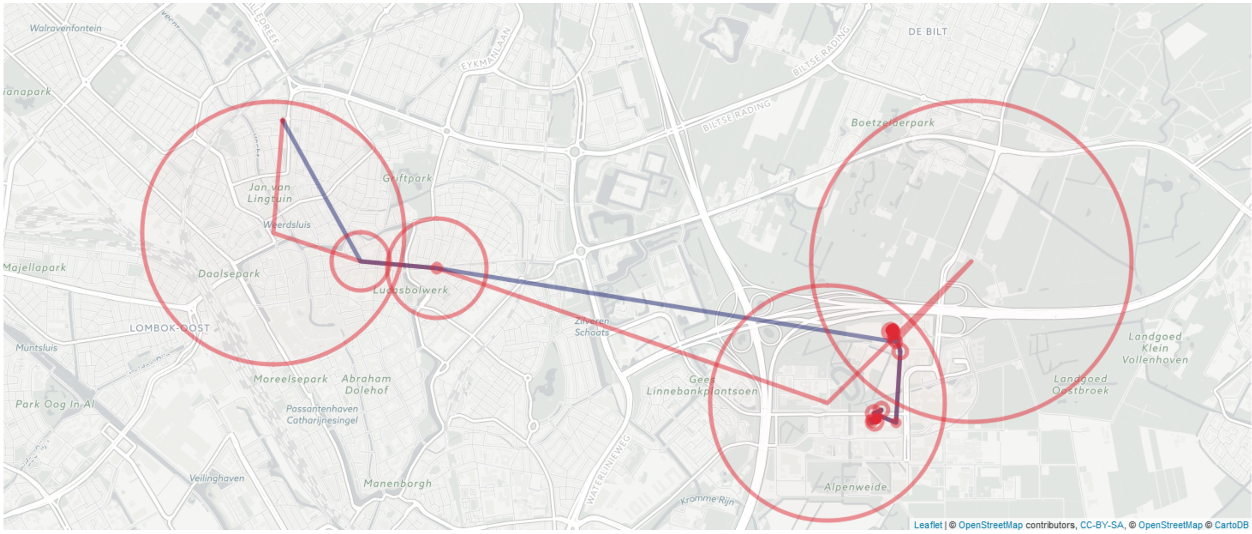
Zenk, S. N., Schulz, A. J., & Odoms-Young, A. (2009). How Neighborhood Environments Contribute to Obesity. *The American Journal of Nursing*, *109*(7), 61–64. doi:10.1097/01.NAJ.0000357175.86507.c8

Zhang, Z., Yang, X., Li, H., Li, W., Yan, H., & Shi, F. (2017). Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level. *Journal of Hydrology*, *553*(Supplement C), 384–397. doi:10.1016/j.jhydrol.2017.07.053

Table 1

Table with descriptives about the data sets used to build the imputation methods.

Log duration	Logged days	Observations	Missing days	Missing data	Mean Accu
From 2013-02-06 to 2017-03-29	1512	646376	635	0.22	12
From 2016-07-14 to 2017-05-10	300	158382	3	0.41	139
From 2014-01-22 to 2017-01-23	1097	814941	80	0.25	12



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

Figure 1. Measurement accuracy of each logged measurement in a morning journey. 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.

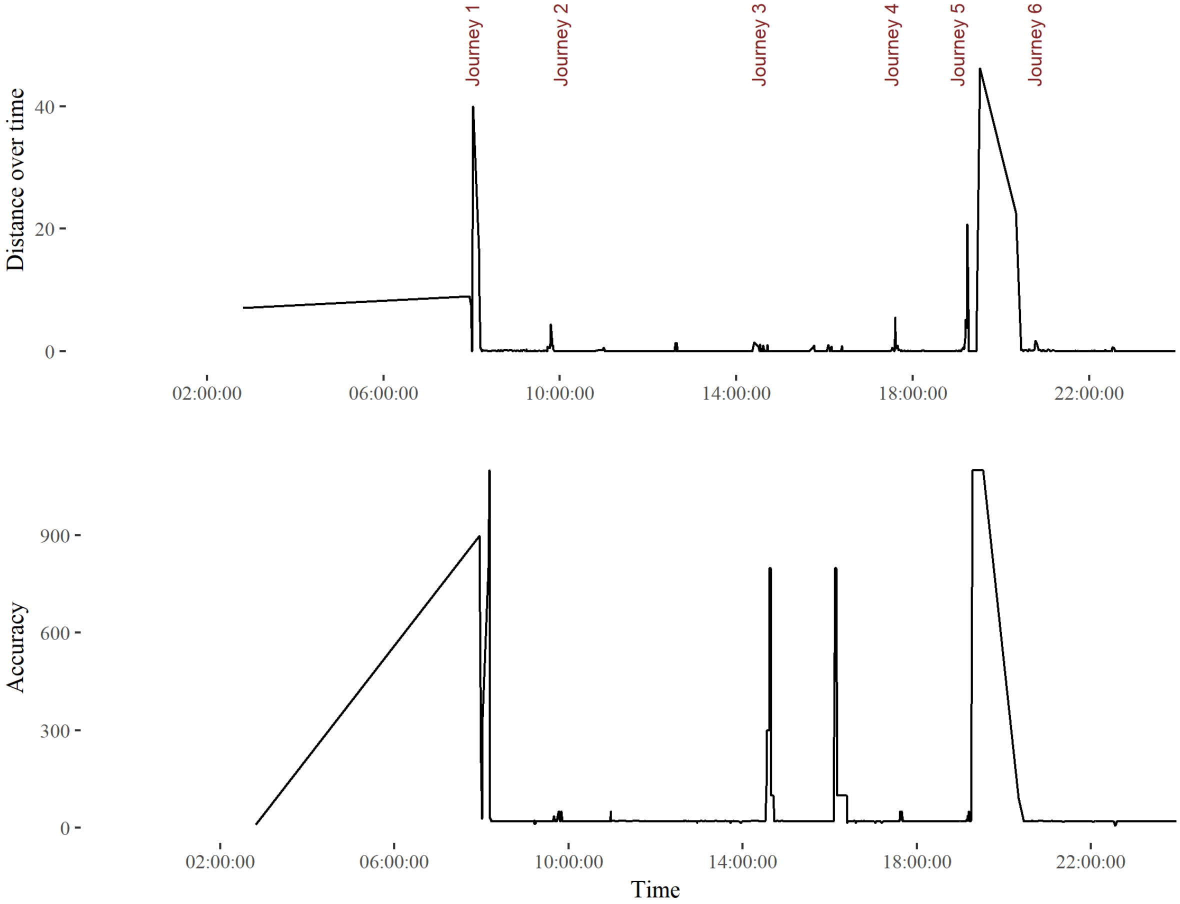


Figure 2. The accuracy of raw measures with time on the x-axis and accuracy on the y-axis. The colour scale shows the distance between the measurement and the previous point.

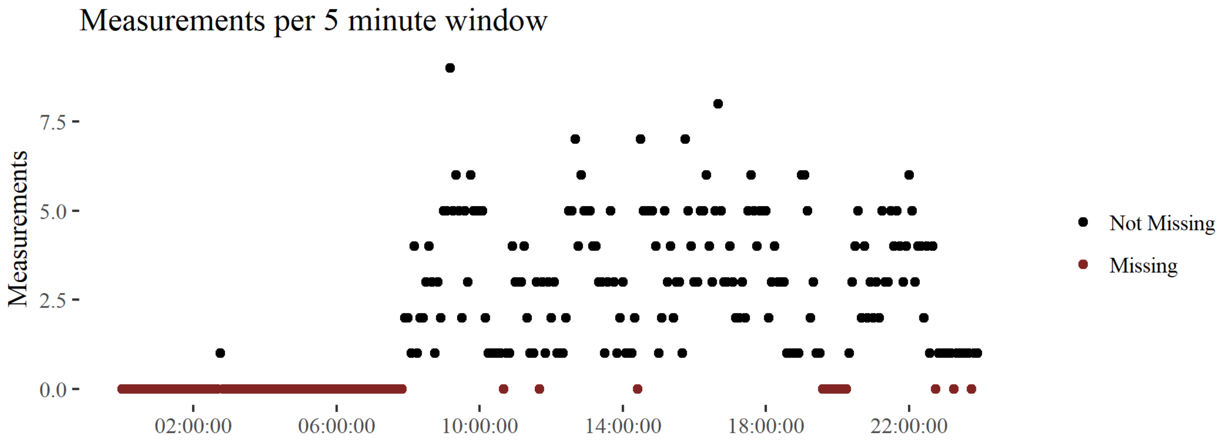


Figure 3. Example of missing data over a day. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points are filled with red and lie on the x-axis.