Handling missing data in smartphone location logs

Boaz Sobrado

26 November 2017

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.

Introduction

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

These obstacles can be overcome by using objective data on human mobility. Such data can now be obtained using the Global Positioning System (GPS). GPS uses the distance between a device and several satellites to determine location. Within behavioural sciences, researchers have used GPS data to investigate wideranging topics. For example, Zenk et al. [2009] investigate the effects of the food environment on eating patterns. Harari et al. [2016] look at the movement correlates of personality. Wang et al. [2015] does the same for academic performance. Palmius et al. [2017] use mobility patterns to predict bipolar depression.

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

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

There is no golden standard for dealing with missing data in GPS logs [Barnett and Onnela, 2016]. Traditional methods, such as mean imputation, are not suitable for spatiotemporal data. This is because measurements

are often correlated in time and space. For example, imagine an individual who splits almost all her time between work or at 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].

Smartphone location measurements are substantially less accurate than those of professional GPS trackers. Android phones collect location information through a variety of less accurate methods. Other than GPS measurements, Androids use WiFi access points and cellphone triangulation. Different methods are used because of computational and battery constraints [LaMarca et al., 2005, Chen et al., 2006]. In professional GPS trackers less than 80% of measurements fall within 10 meters of the true location. GPS measures are most inaccurate in dense urban locations and indoors [Schipperijn et al., 2014, Duncan et al., 2013]. Unfortunately for researchers, this is where people in the developed world spend most of their time.

Noisy data can lead to inaccurate conclusions if it is not accounted for. Suppose we wish to calculate an individual's movement in a day. A simple approach would be to calculate the sum of the distance between each measurement. But if there is noise, the coordinates will vary even though the individual is not moving. If the measurements are frequent and noisy, we will calculate a lot of movement. Even if the individual did not move at all! 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 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 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 2). 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.

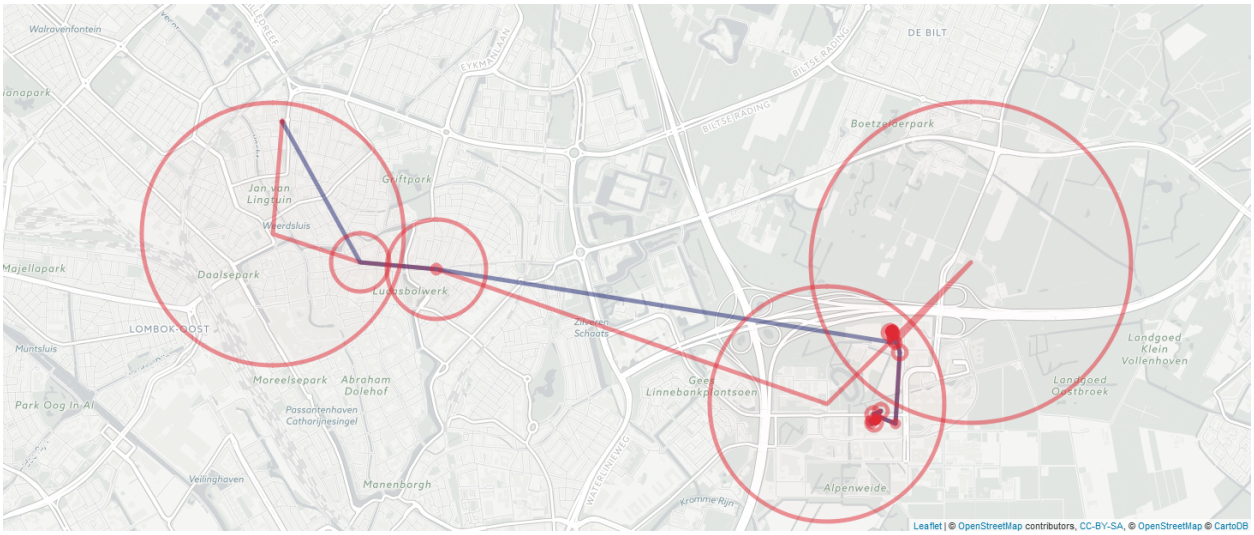

Figure 1: Measurement accuracy of each logged measurement on a morning journey. The red circles denote the accuracy of all logged measurement points (the raw data). The points connected in time are connected by a line. The blue line shows the path without the most inaccurate (accuracy > 400 meters) points filtered out. The red line shows the path with all measurements included.

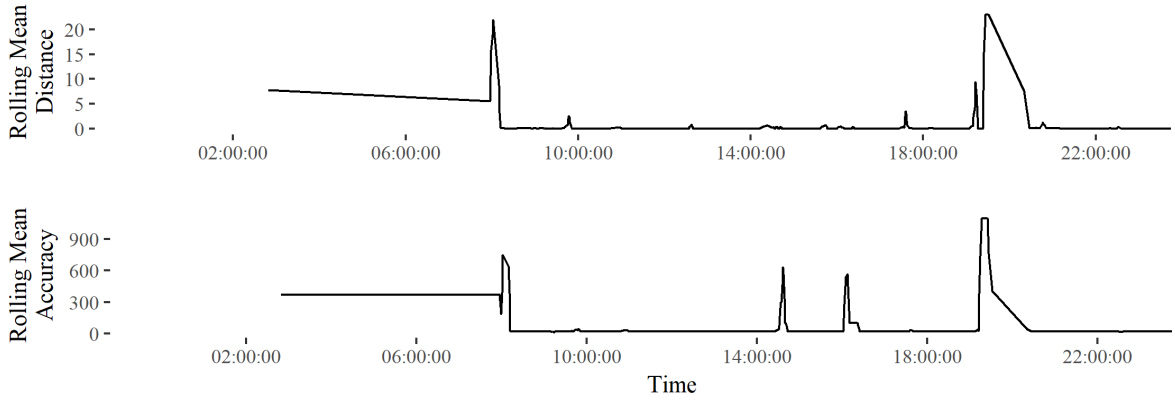

Figure 2: The upper chart shows the rolling mean distance from the next point in meters over the course of the day. The lower chart shows the rolling mean accuracy over the course of the day. In both cases, the rolling window is 3.

Missingness

Over 54% of the data is missing for the entire duration of the log (as shown in Figure X a). 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 in Figure X b. As you can see, there are several long periods over the course of the log for which there are no measurements. Moreover, even during a single day there are continuous periods where there is missing data, mostly during the late hours of the night in this case.

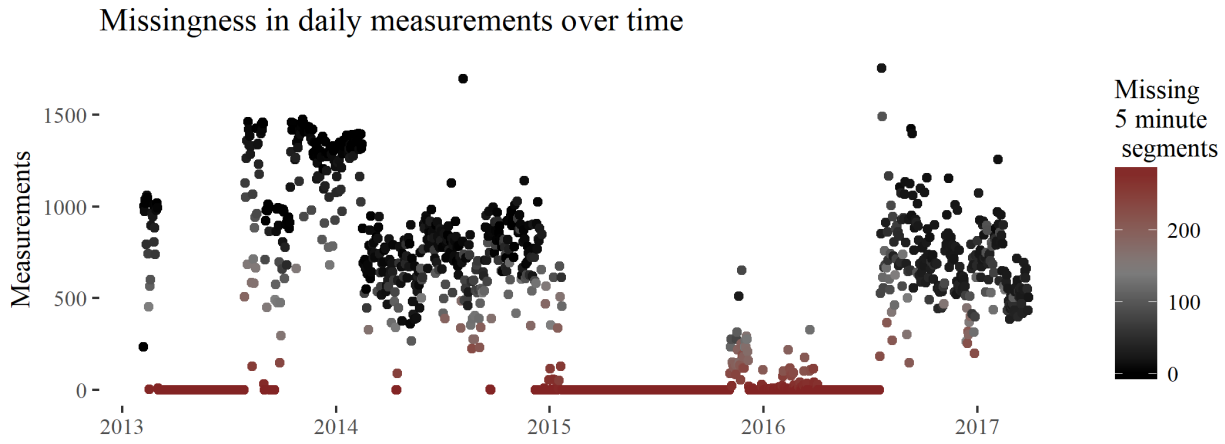

Figure 3: Example of missing data over the entire duration of the log. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points are filled with red and lie on the x-axis.

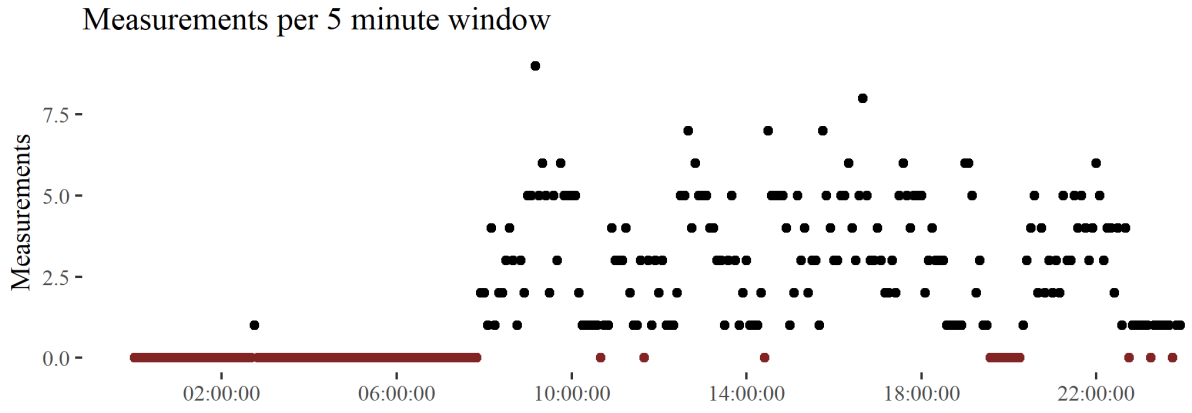

Figure 4: Example of missing data over a day. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a five minute window. For this day there were several periods with no information. These points are filled with red and lie on the x-axis.

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 et al., 2008]. These models can be quite complex, for example Preisler et al. [2004] uses Markovian movement processes to characterise the effect of roads, food patches and streams on cyclical elk movements. The most well studied SSM is the Kalman filter, which is the optimal algorithm for inferring linear Gaussian systems. In fact, the extended Kalman filter is the de facto standard for GPS navigation [Chen and Brown, 2013]. The advantage of state space models is that they are flexible, deal with measurement inaccuracy, include information from different sources and can be used in real time.

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 and 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 et al. [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, 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 measuremement 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).

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

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 [Wickham, 2009, Wickham and Francois, 2016, **?**, R Core Team, 2017, Pebesma and Bivand, 2005, Bivand et al., 2013].

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 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 et al. [2007] for each imputed dataset. Like Barnett and Onnela [2016] we calculate bias by substracting the estimated measure under each approach for the same measure calculated on the full data. For simulation-based imputation approaches a mean value over 100 samples was taken.

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

References

References

Ian Barnett and Jukka-Pekka Onnela. Inferring Mobility Measures from GPS Traces with Missing Data. *arXiv:1606.06328 [stat]*, June 2016. URL http://arxiv.org/abs/1606.06328.

Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013. URL http://www.asdar-book.org/.

Mike Y. Chen, Timothy Sohn, Dmitri Chmelev, Dirk Haehnel, Jeffrey Hightower, Jeff Hughes, Anthony LaMarca, Fred Potter, Ian Smith, and Alex Varshavsky. Practical Metropolitan-Scale Positioning for GSM Phones. In *UbiComp 2006: Ubiquitous Computing*, Lecture Notes in Computer Science, pages 225–242. Springer, Berlin, Heidelberg, September 2006. ISBN 978-3-540-39634-5 978-3-540-39635-2. doi: 10.1007/11853565_14. URL https://link.springer.com/chapter/10.1007/11853565_14.

Zhe Chen and Emery N. Brown. State space model. *Scholarpedia*, 8(3):30868, March 2013. ISSN 1941-6016. doi: 10.4249/scholarpedia.30868. URL http://www.scholarpedia.org/article/State_space_model.

European Commission. Protecting your data: your rights - European Commission, 2017. URL http://ec.europa.eu/justice/data-protection/individuals/rights/index_en.htm.

Xavier Delclòs-Alió, Oriol Marquet, and Carme Miralles-Guasch. 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, October 2017. ISSN 2214-367X. doi: 10.1016/j.tbs.2017.07.001. URL http://www.sciencedirect.com/science/article/pii/S2214367X16301466.

Paul Doust. *smoothing - Smooth GPS data - Stack Overflow*. 2013. URL https://stackoverflow.com/questions/1134579/smooth-gps-data.

Scott Duncan, Tom I. Stewart, Melody Oliver, Suzanne Mavoa, Deborah MacRae, Hannah M. Badland, and Mitch J. Duncan. Portable global positioning system receivers: static validity and environmental conditions. *American Journal of Preventive Medicine*, 44(2):e19–29, February 2013. ISSN 1873-2607. doi: 10.1016/j.amepre.2012.10.013.

Lingbing Feng, Gen Nowak, T.J. O'Neill, and A Welsh. CUTOFF: A spatio-temporal imputation method. *Journal of Hydrology*, 519:3591–3605, November 2014. doi: 10.1016/j.jhydrol.2014.11.012.

Michael F. Goodchild and Donald G. Janelle. Toward critical spatial thinking in the social sciences and humanities. *GeoJournal*, 75(1):3–13, February 2010. ISSN 0343-2521. doi: 10.1007/s10708-010-9340-3. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2863328/.

Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Ohler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE journal of biomedical and health informatics*, 19(1):140–148, January 2015. ISSN 2168-2208. doi: 10.1109/JBHI.2014.2343154.

Gabriella M. Harari, Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges. *Perspectives on Psychological Science*, 11(6):838–854, November 2016. ISSN 1745-6916. doi: 10.1177/1745691616650285. URL https://doi.org/10.1177/1745691616650285.

Marta M. Jankowska, Jasper Schipperijn, and Jacqueline Kerr. A Framework For Using GPS Data In Physical Activity And Sedentary Behavior Studies. *Exercise and sport sciences reviews*, 43(1):48–56, January 2015. ISSN 0091-6331. doi: 10.1249/JES.0000000000000035. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4272622/.

Anthony LaMarca, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, James Howard, Jeff Hughes, Fred Potter, Jason Tabert, Pauline Powledge, Gaetano Borriello, and Bill Schilit. Place Lab: Device Positioning Using Radio Beacons in the Wild. In *Pervasive Computing*, Lecture Notes in Computer Science, pages 116–133. Springer, Berlin, Heidelberg, May 2005. ISBN 978-3-540-26008-0 978-3-540-32034-0. doi: 10.1007/11428572_8. URL https://link.springer.com/chapter/10.1007/11428572_8.

Google Location History. Timeline, 2017. URL https://www.google.com/maps/timeline?pb.

N. Palmius, A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. De Vos. Detecting Bipolar Depression From Geographic Location Data. *IEEE Transactions on Biomedical Engineering*, 64(8):1761–1771, August 2017. ISSN 0018-9294. doi: 10.1109/TBME.2016.2611862.

Toby A. Patterson, Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos. State–space models of individual animal movement. *Trends in Ecology & Evolution*, 23(2):87–94, February 2008. ISSN 0169-5347. doi: 10.1016/j.tree.2007.10.009. URL http://www.sciencedirect.com/science/article/pii/S0169534707003588.

Edzer J. Pebesma and Roger S. Bivand. Classes and methods for spatial data in R. *R News*, 5(2):9–13, November 2005. URL https://CRAN.R-project.org/doc/Rnews/.

Haiganoush K. Preisler, Alan A. Ager, Bruce K. Johnson, and John G. Kie. Modeling animal movements using stochastic differential equations. *Environmetrics 15: p. 643-657*, 2004. URL https://www.fs.usda.gov/treesearch/pubs/33038.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, and Song Chong. Human Mobility Patterns and Their Impact on Routing in Human-Driven Mobile Networks. *ACM HotNets 2007*, November 2007. URL http://koasas.kaist.ac.kr/handle/10203/160927.

Adam Sadilek and John Krumm. Far Out: Predicting Long-Term Human Mobility. *Microsoft Research*, December 2016. URL https://www.microsoft.com/en-us/research/publication/far-predicting-long-term-human-mobility/.

Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, 17(7):e175, 2015. doi: 10.2196/jmir.4273. URL https://www.jmir.org/2015/7/e175/.

Jasper Schipperijn, Jacqueline Kerr, Scott Duncan, Thomas Madsen, Charlotte Demant Klinker, and Jens Troelsen. 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, 2014. ISSN 2296-2565. doi: 10.3389/fpubh.2014.00021.

Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. SmartGPA: How Smartphones Can Assess and Predict Academic Performance of College Students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 295–306, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4. doi: 10.1145/2750858.2804251. URL http://doi.acm.org/10.1145/2750858.2804251.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL http://ggplot2.org.

Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2016. URL https://CRAN.R-project.org/package=dplyr. R package version 0.5.0.

Jean Wolf, Marcelo Oliveira, and Miriam Thompson. 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, January 2003. ISSN 0361-1981. doi: 10.3141/1854-21. URL http://trrjournalonline.trb.org/doi/abs/10.3141/1854-21.

Shannon N. Zenk, Amy J. Schulz, and Angela Odoms-Young. How Neighborhood Environments Contribute to Obesity. *The American journal of nursing*, 109(7):61–64, July 2009. ISSN 0002-936X. doi: 10.1097/01.NAJ.0000357175.86507.c8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2789291/.

Zhongrong Zhang, Xuan Yang, Hao Li, Weide Li, Haowen Yan, and Fei Shi. 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, October 2017. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2017.07.053. URL http://www.sciencedirect.com/science/article/pii/S0022169417305188.