Handling missing data in smartphone location logs

Boaz Sobrado

16 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 compare the performance of the model compared to currently used methods, such as linear interpolation.

Introduction

How active people are and how they interact with their environment affects a wide range of measures including health , income and social capital [Goodchild and 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 et al., 2003] and the underestimation of the duration of commutes [Delclòs-Alió et al., 2017].

Objective data on human mobility has become available through the Global Positioning System (GPS) which uses the distance between a device and a number of satellites to determine location. 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 et al., 2009], the movement correlates of personality and academic performance [Harari et al., 2016, Wang et al., 2015] and detecting bipolar disorder [Palmius et al., 2017].

In most of these studies participants are given a specialised devices, 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 may be biased because of the introduction of a new device to the participant's life. Because of this, specialised logs usually span a short amount of time. Barnett and Onnela [2016] advocate installing 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] (*secondary logs*). These logs can be accessed and shared by users. 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.

There is currently no golden standard in how to deal with missing data in custom or secondary logs [Barnett and Onnela, 2016]. Jankowska et al. [2015] have pointed out that there is often little transparency regarding decisions of how to deal with missing data. Methods frequently 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. Traditional missing data methods, such as multiple imputation, cannot be used easily due to the geospatial nature of the data. On the other hand, noisy data can lead to inaccurate conclusions if it is not accounted for, such as underestimating the movement of individuals. In this paper we will compare methods used to deal with measurement error and missing data in location information. Specifically, we are interested in establishing accurate mobility patterns from smartphone GPS logs.

Problem description & literature review

Given that there is next to no literature on missing data in custom or secondary logs it is worth illustrating the typical characteristics of this data using an example data set. Moreover, although we could find no published papers on dealing with missing data in secondary logs, there are methods which deal with similar problems. In this section we will describe the problems with the data and why they arise as well as the models which could be applied to it.

Missing data and noise in location logs

This example data set comes from Google Location History and spans 3 years from January 2013 to January 2017 with multiple different Android devices. It contains 814 941 measurements, with approximately 742 measurements per day (*SD*=868.15). The data set contains a wide range of variables. For the purposes of this paper we will focus only on latitude, longitude, accuracy and time. Accuracy is defined as the size in meters of the radius of a circle such that 67% of the measurements fall within it.

Accuracy in location logs

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 [Schipperijn et al., 2014, Duncan et al., 2013]. Unfortunately for social scientists, this happens to be where most people in the developed world tend to spend most of their time.

Given that Android phones collect location information from WiFi access points, cellphone triangulation, and GPS measurements due to computational and battery constraints [LaMarca et al., 2005, Chen et al., 2006], the accuracy is substantially lower than in professional grade GPS trackers. Based on this data set, less than 40% of measures fall with a 67% confidence radius of 20 meters. Palmius et al. [2017] note that in their custom logs inaccurate location values are interspersed between more accurate location values at higher sample rates per hour. I observe similar patterns in secondary logs.

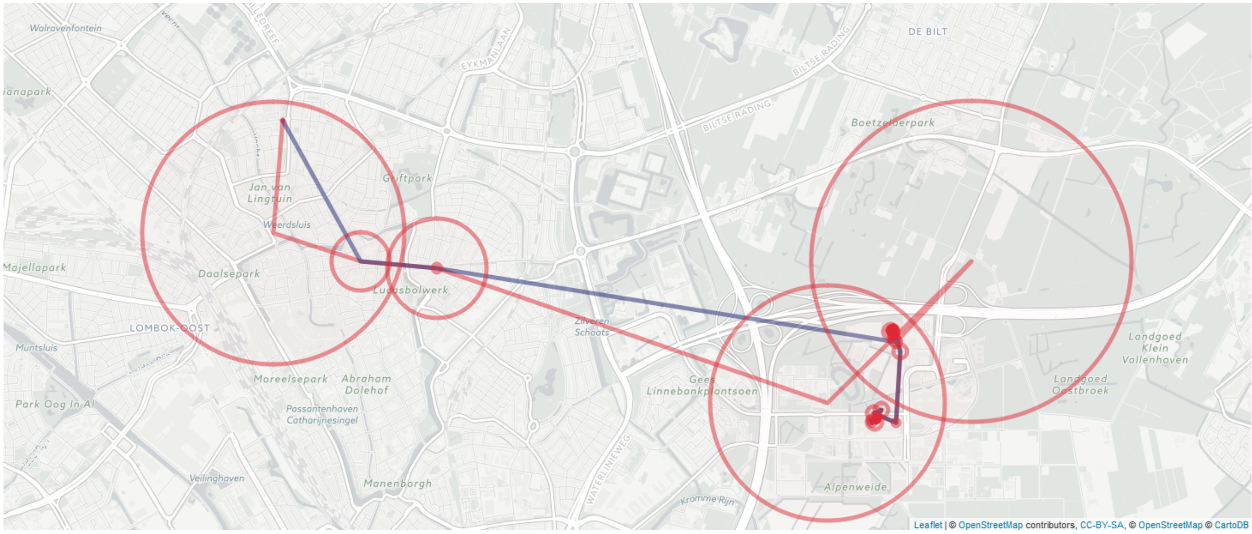

Figure 1: Measurement accuracy of each logged measurement. The x-axis denotes the accuracy in meters such that the user's actual location is expected to be within that radius with 67 percent confidence. Measures were classified into high accuracy (under 800m), medium accuracy (under 5000m) and low accuracy (over 5000m). Measures with accuracy of over 20000m were omitted from the graph. The maximum accuracy measure was over 40000m. Low accuracy measures are often near in time to more high accuracy measures.

Missingness

Missing data is a pervasive issue as it can arise due to multiple factors, both technical and behavioural. 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% [Saeb et al., 2015, Grünerbl et al., 2015, Palmius et al., 2017]. Similar trends are consistently reported in other fields [e.g. Harari et al., 2016, Jankowska et al., 2015].

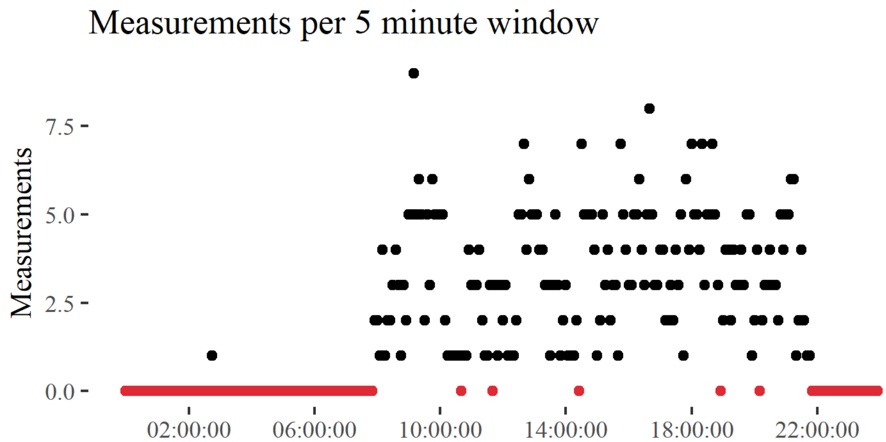

Figure 2: Missing data over time for the author. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a day. The fill of the points shows the amount of hours of captured data each day.

State of the art in spatiotemporal models

Research with respect to the analysis of GPS data is wideranging, highly interdisciplinary and serves vastly different purposes. The following section briefly illustrates some methods used to deal with measurement inaccuracy and missing data problems in spatiotemporal data, as well as their applicability to our question. Moreover, we discuss in detail two approaches which deal explicitly with missing data in smartphone measured human location.

Let $d \in \mathbb{R}^3$ be the recorded data containing lattitude, longitude, accuracy and $l \in \mathbb{R}^3$ be actual location in space. d_k and l_k denote the measured location and the true location at time k.

Spatiotemporal Imputation Methods

Given fixed measurement stations there are several imputation methods for spatiotemporal measurements. 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.

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.

State Space Models

There is a vast literature of using state space models (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. The extended Kalman filter is the de facto standard for GPS navigation [Chen and Brown, 2013].

The state equation of the Kalman filter models the current location:

$$l_k = Al_{k-1} + Bu_{k-1} + w_{k-1}$$

Here w_k is known as the process noise and is conceptually equivalent to the error term in a linear regression. A is a square matrix of order n which relates our current state to the previous state. B is an $n \times x$ matrix which relates the optional control input $u \in \mathbb{R}^x$ to the state l.

The measurement equation models the measurements at any given time:

$$d_k = Hl_k + v_k$$

v_k is known as the measurement noise and represents the errors in measurement and H is an $m \times n$ matrix relating the state to the measurement.

The noise in both equations assumed to be:

$$p(w) \sim N(0, Q)$$

$$p(v) \sim N(0, R)$$

Where Q is the process noise covariance matrix and R is the measurement noise covariance.

Then, let $\widehat{l}_k^- \in \mathbb{R}^n$ be the a priori state estimate given our knowledge of the process and $\widehat{l}_k \in \mathbb{R}^n$ be the a posteriori state estimate at step k given measurement d_k. These two state estimates have error covariances denoted P_k^- and P_k respectively. Algebraically the Kalman filter can then be written as:

$$\widehat{x}_k = \widehat{x}_k^- + K(z_k - H\widehat{x}_k^-)$$

Where K (the weight or Kalman gain) is a square matrix order $n \times m$ matrix which minimizes the posteriori error covariance.

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 our purposes the main limitation is that these models are based on the Markov property. Thus, the estimated location at timepoint k is often based only upon measurements at k and the previous timepoint $k - 1$. This may be suitable for ships at sea, but it ignores the highly periodic nature of human movement. Hierarchical structuring and conditioning on a larger context have been suggested as ways to improve their performance, but these are often computationally intractable or unfeasible [Sadilek and Krumm, 2016].

Alternative models

Alternatives to state space models include long range-persistence models, such as cascading walks models and the FarOut model which rely on self-similarity and autoregressive characteristics [Han et al., 2015, Sadilek and Krumm, 2016]. The latter uses Fourier analysis and PCA to extract cyclical patterns in an individual's behaviour and reduce the dimensionality of the extracted features and yields interpretable predictions for an individuals location months in advance.

Filtering & Mean imputation

Palmius et al. [2017] deal with measuremement inaccuracy of D by removing from the data set all unique low-accuracy 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:

> 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's model

Barnett and Onnela [2016] deal with custom logs where location is measured for 2 minutes and subsequently not measured for 10 minutes. They 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.

This paper is to the best of our knowledge the only attempt at establishing a principled approach to this problem

Methods

Datasets & Analyses

The data used was collected between 2014 and 2017 on different Android devices from X individuals across X continents. The table below provides more details:

Log duration	Observations	Missing days	Mean Accuracy	SD Accuracy
From 2014-01-22 to 2017-01-23	814941	80	121.83	414.86

In addition to the secondary logs, participants also volunteered to carry with them a specialised GPS tracker for week. Therefore we also used a week of parallel secondary and specialised logs.

Analyses were performed using R and a multitude of other packages [Wickham, 2009, Wickham and Francois, 2016, **?**, Arnold, 2013, R Core Team, 2017, Pebesma and Bivand, 2005, Bivand et al., 2013].

Data pre-processing

The data was discretised into 5 minute windows following Palmius et al. [2017] (as described above, for more detail please consult the original paper).

Imputation methods

Four imputation methods were selected in order to cover techniques applied in the literature. Briefly, the filtered mean imputation method described by Palmius et al. [2017], the model developed by Barnett and Onnela [2016] as well as an adapted FarOut model [Sadilek and Krumm, 2016]. Simple Kalman filtered linear interpolation was used as a benchmark models.

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 on two computed measures: amount of trips made and distance traveled.

Measures of performance used were root mean square error (RMSE) and mean absolute error (MAE).

Results

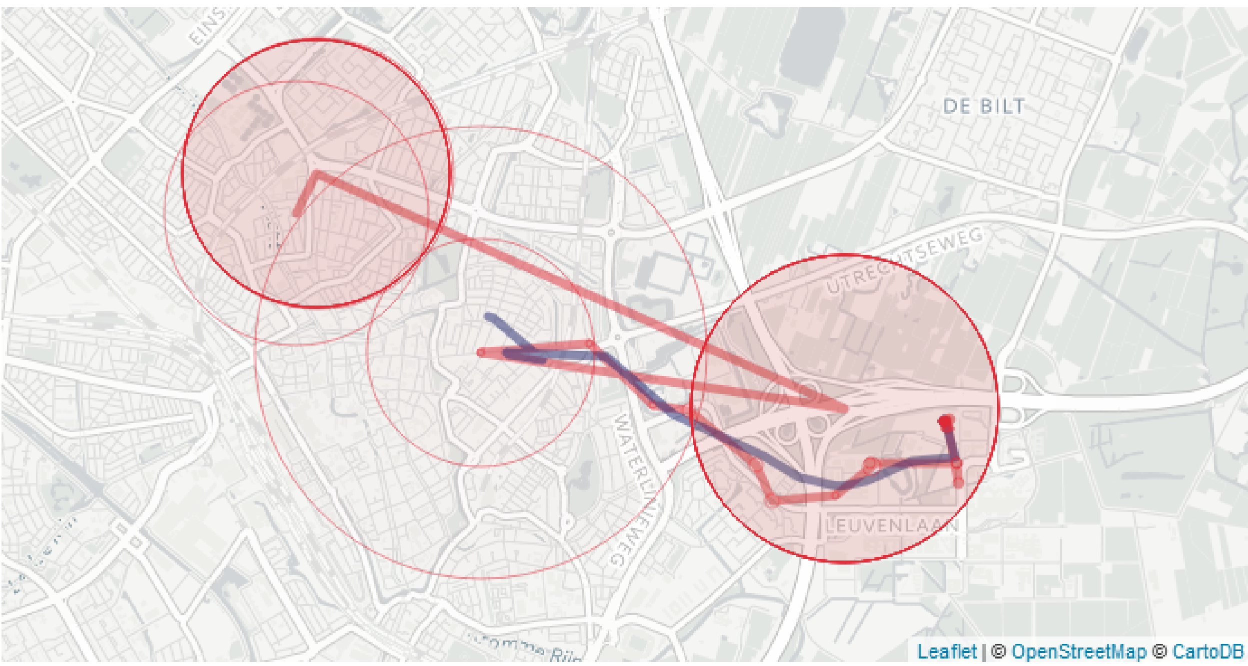

Figure 3: Raw GPS measurements of a journey from de Uithof to Tuinwijk on February 17th 2017. The measurements are in red, the filtered path is in blue. The circles denote 67% confidence intervals of the given GPS measurement. Measurements and fitted points which follow each other in time are connected by lines. The inacurate measurements lead to estimates of irregular movements. The filtered movement estimate seems more accurate in terms of the path, but lags behind the unfiltered movement.

Discussion

References

References

Jeffrey B. Arnold. *ggthemes: Extra themes, scales and geoms for ggplot*, 2013. URL https://CRAN.R-project.org/package=ggthemes. R package version 1.5.1.

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.

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.

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.

Xiao-Pu Han, Xiang-Wen Wang, Xiao-Yong Yan, and Bing-Hong Wang. Cascading Walks Model for Human Mobility Patterns. *PLOS ONE*, 10(4):e0124800, October 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0124800. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0124800.

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. 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/.

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.