Handling missing data in smartphone location logs

Boaz Sobrado

11 November 2017

Abstract

There is a great interest in using objective location data to infer the mobility measures of individuals. 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, Oliveira, and Thompson 2003) and the underestimation of the duration of commutes (Delclòs-Alió, Marquet, and Miralles-Guasch 2017).

Objective data on human mobility has become available throgh 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, Schulz, and Odoms-Young 2009), the movement correlates of personality and academic performance (G. M. 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 GPS devices or a smartphone with a custom made app. 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 and usually span a short amount of time. A solution is to take advantage of the already exisiting smartphone location logs, such as Google Location History, which store location information of millions of users spanning several years(Location History 2017). These logs can be accessed and shared by users. Yet, because GPS sensors consume a significant amounts of battery, these 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 this context (Barnett and Onnela 2016). Jankowska, Schipperijn, and Kerr (2015) have pointed out that there is often little transparency regarding decisions of how to deal with it. 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. On the other hand, noisy data can lead to inaccurate conclusions if it is not accounted for. 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 smartphone location 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 this narrow domain, there are however methods which deal with similar problems. In this section we will describe the problems with the data and models which could be applied to it.

Missing data and noise in location logs

This example data set spans 3 years from January 2013 to January 2017. It contains 814 941 measurements, with approximately 742 measurements per day (SD=868.15). Moreover, it also contains estimates of velocity, altitude, heading and activity (e.g. in a vehicle, still). For the purposes of this paper we will focus only on lattitude, longitude and time.

Accuracy

Even in professional grade GPS trackers less than 80% of measurements fall within 10 meters of the true location. Moreover, 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.

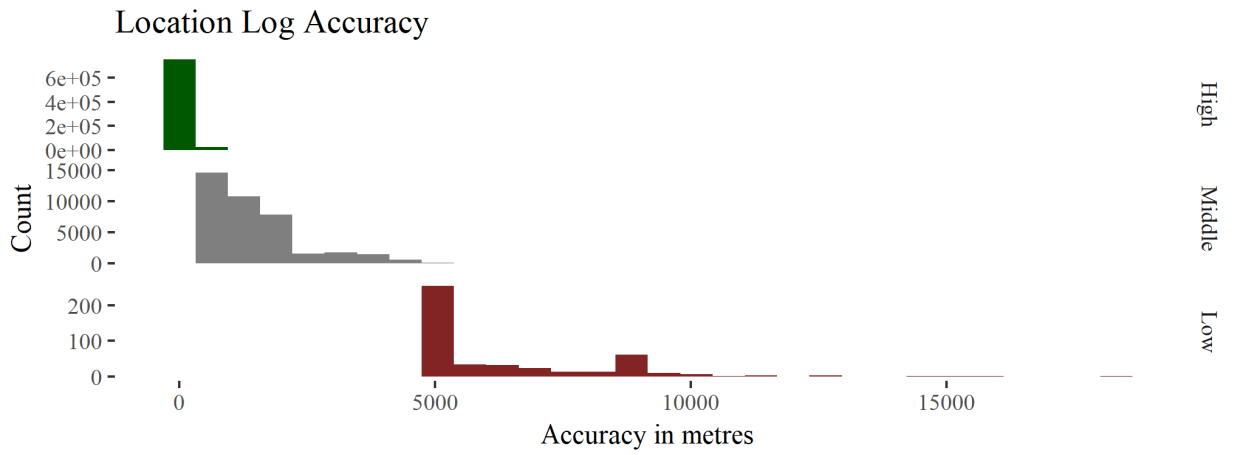

Figure 1: Measurement accuracy of each logged measurement. The x-axis denotes the accuracy in meters such that the user's actual location is expected to be within that radius with 67 percent confidence. Measures were classified into high accuracy (under 800m), medium accuracy (under 5000m) and low accuracy (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.

Given that Android phones collect location information from WiFi access points, cellphone triangluation, and GPS measurements due to computational and battery constraints (LaMarca et al. 2005; M. Y. 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.

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. G. M. Harari et al. 2016; Jankowska, Schipperijn, and Kerr 2015).

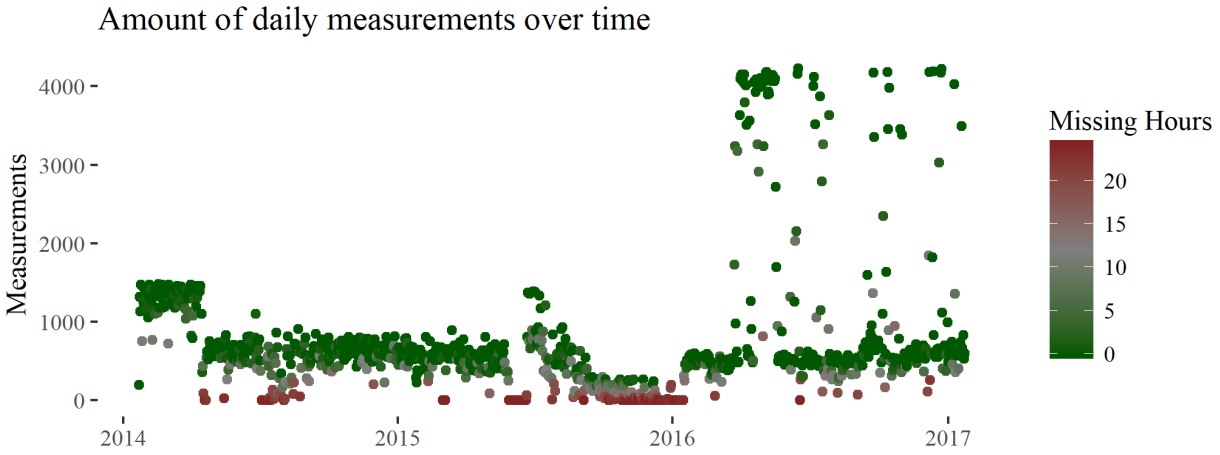

Figure 2: Missing data over time for the author. The x-axis denotes time, the y-axis shows how many measurements are made and each point is a day. The fill of the points shows the amount of hours of captured data each day.

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.

Spatiotemporal Imputation Methods

Given fixed measurement stations there are sevaral 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, Z. 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, ostentibly 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 (non-linear) Kalman filter is the de facto standard for GPS navigation (Z. 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. For our pourposes 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 infeasible (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.

Linear interpolation

This simple model is used by

Barnett's model

Methods

An asside as to my conversation with Peter on 07.11

We can get some trackers on people to measure movement constantly and be a "golden standard" to evaluate the models.

Data & 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:

```
datadetails <- readRDS("datadetails.rds")
kable(datadetails,digits = 2,col.names = c("Log duration",
                                           "Observations",
                                           "Missing days",
                                           "Mean Accuracy",
                                           "SD Accuracy"))
```

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

Analyses were performed using R and a multitude of other packages (Wickham 2009; Wickham and Francois 2016; **???**; Arnold 2013; R Core Team 2017; E. J. Pebesma and Bivand 2005; Bivand, Pebesma, and Gomez-Rubio 2013).

Model evaluation

Models were evaluated using a test set and a train set. During a period of 7 days participants wore a professional GPS device (TRX-100) which reported location every minute. In addition, they continued using their phones as normal. Missing data was imputed for the week for which more accurate data was available.

Measures for which the accuracy was estimated include *barnett's path paper*.

Results

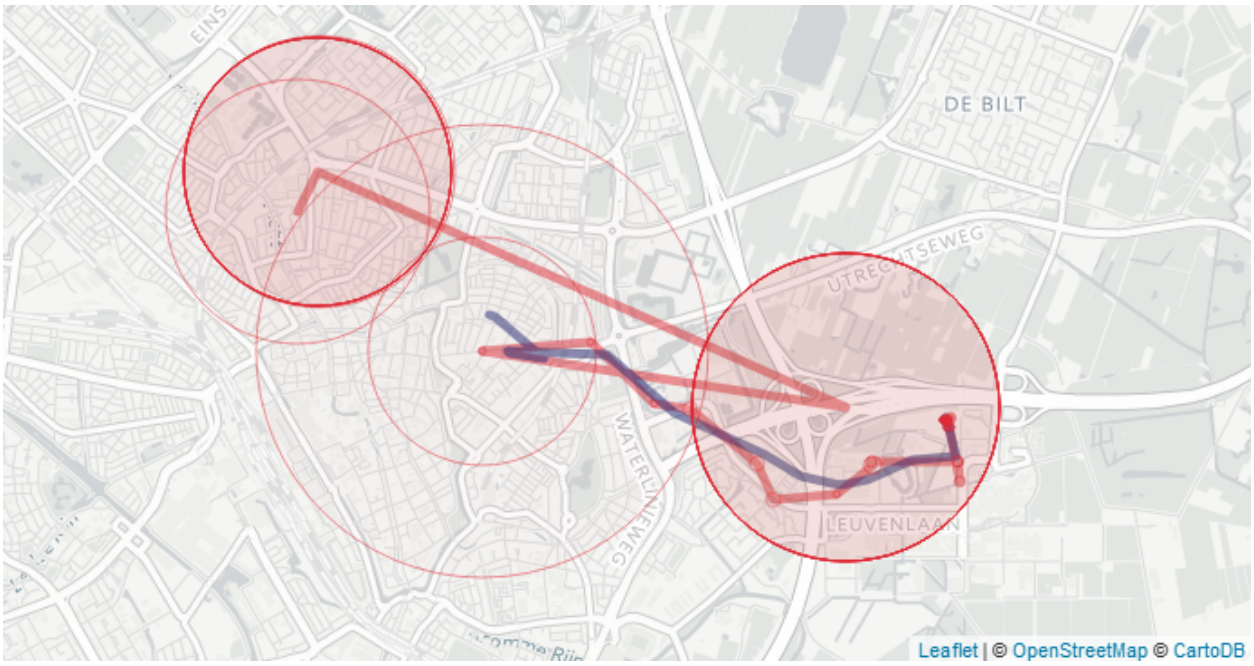

Figure 3: Raw GPS measurements of a journey from de Uithof to Tuinwijk on February 17th 2017. The measurements are in red, the filtered path is in blue. The circles denote 67% confidence intervals of the given GPS measurement. Measurements and fitted points which follow each other in time are connected 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

Arnold, Jeffrey B. 2013. *Ggthemes: Extra Themes, Scales and Geoms for Ggplot*. https://CRAN.R-project.org/package=ggthemes.

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

Bivand, Roger S., Edzer Pebesma, and Virgilio Gomez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition*. Springer, NY. http://www.asdar-book.org/.

Chen, Mike Y., Timothy Sohn, Dmitri Chmelev, Dirk Haehnel, Jeffrey Hightower, Jeff Hughes, Anthony LaMarca, Fred Potter, Ian Smith, and Alex Varshavsky. 2006. "Practical Metropolitan-Scale Positioning for GSM Phones." In *UbiComp 2006: Ubiquitous Computing*, 225–42. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. doi:10.1007/11853565_14.

Chen, Zhe, and Emery N. Brown. 2013. "State Space Model." *Scholarpedia* 8 (3): 30868. doi:10.4249/scholarpedia.30868.

Delclòs-Alió, Xavier, Oriol Marquet, and Carme Miralles-Guasch. 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.

Duncan, Scott, Tom I. Stewart, Melody Oliver, Suzanne Mavoa, Deborah MacRae, Hannah M. Badland, and Mitch J. Duncan. 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, Lingbing, Gen Nowak, T.J. O'Neill, and A Welsh. 2014. "CUTOFF: A Spatio-Temporal Imputation Method." *Journal of Hydrology* 519 (November): 3591–3605. doi:10.1016/j.jhydrol.2014.11.012.

Goodchild, Michael F., and Donald G. Janelle. 2010. "Toward Critical Spatial Thinking in the Social Sciences and Humanities." *GeoJournal* 75 (1): 3–13. doi:10.1007/s10708-010-9340-3.

Grünerbl, Agnes, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Ohler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2015. "Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients." *IEEE Journal of Biomedical and Health Informatics* 19 (1): 140–48. doi:10.1109/JBHI.2014.2343154.

Han, Xiao-Pu, Xiang-Wen Wang, Xiao-Yong Yan, and Bing-Hong Wang. 2015. "Cascading Walks Model for Human Mobility Patterns." *PLOS ONE* 10 (4): e0124800. doi:10.1371/journal.pone.0124800.

Harari, Gabriella M., Nicholas D. Lane, Rui Wang, Benjamin S. Crosier, Andrew T. Campbell, and Samuel D. Gosling. 2016. "Using Smartphones to Collect Behavioral Data in Psychological Science: Opportunities, Practical Considerations, and Challenges." *Perspectives on Psychological Science* 11 (6): 838–54. doi:10.1177/1745691616650285.

Jankowska, Marta M., Jasper Schipperijn, and Jacqueline Kerr. 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, Anthony, Yatin Chawathe, Sunny Consolvo, Jeffrey Hightower, Ian Smith, James Scott, Timothy Sohn, et al. 2005. "Place Lab: Device Positioning Using Radio Beacons in the Wild." In *Pervasive Computing*, 116–33. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. doi:10.1007/11428572_8.

Location History, Google. 2017. "Timeline." Accessed November 9. https://www.google.com/maps/timeline?pb.

Palmius, N., A. Tsanas, K. E. A. Saunders, A. C. Bilderbeck, J. R. Geddes, G. M. Goodwin, and M. De Vos. 2017. "Detecting Bipolar Depression from Geographic Location Data." *IEEE Transactions on Biomedical*

Engineering 64 (8): 1761–71. doi:10.1109/TBME.2016.2611862.

Patterson, Toby A., Len Thomas, Chris Wilcox, Otso Ovaskainen, and Jason Matthiopoulos. 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, Edzer J., and Roger S. Bivand. 2005. "Classes and Methods for Spatial Data in R." *R News* 5 (2): 9–13. https://CRAN.R-project.org/doc/Rnews/.

Preisler, Haiganoush K., Alan A. Ager, Bruce K. Johnson, and John G. Kie. 2004. "Modeling Animal Movements Using Stochastic Differential Equations." *Environmetrics 15: P. 643-657*. 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. https://www.R-project.org/.

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

Saeb, Sohrab, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 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, Jasper, Jacqueline Kerr, Scott Duncan, Thomas Madsen, Charlotte Demant Klinker, and Jens Troelsen. 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, Rui, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. 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*, 295–306. UbiComp '15. New York, NY, USA: ACM. doi:10.1145/2750858.2804251.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

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

Wolf, Jean, Marcelo Oliveira, and Miriam Thompson. 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 (January): 189–98. doi:10.3141/1854-21.

Zenk, Shannon N., Amy J. Schulz, and Angela Odoms-Young. 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, Zhongrong, Xuan Yang, Hao Li, Weide Li, Haowen Yan, and Fei Shi. 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): 384–97. doi:10.1016/j.jhydrol.2017.07.053.