Research Report: Longitudinal models of human mobility

Boaz Sobrado

03 November 2017

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).

Longitudinal data on human temporospatial behaviour has become available through smartphones and other sensors. This data has been used to investigate topics such as the effects of the food environment on eating patterns *Zenk*, 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). While there have been advances in processing this data (Zheng et al. 2009) extracting meaningful information remains a difficult task. Two important challenges are dealing with measurement noise and missing data.

With regards to measurement accuracy, studies of professional grade GPS trackers suggest that 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 researchers, this happens to be where most people in the developed world tend to spend most of their time. In addition, location data collected by more ubiquitous (but less specialised) smartphones are an amalgamation of sensor data. For instance, Android phones collect location information from WiFi access points, cellphone triangluation, and GPS measurements due to computational and battery constraints (LaMarca et al. 2005; Chen et al. 2006). This adds another layer of complexity as each of these measures has its own characteristics in terms of measurement accuracy and bias.

As for missing data it 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 tracking off, etc. 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).

There is currently no golden standard in how to deal with missing temporospatial data (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 increase 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. Specifically, we are interested in location tracks as measured by commerically available smartphones.

State of the art

Research with respect to the analysis of GPS data is wideranging, highly interdisciplinary and often serves different purposes. The following section briefly illustrates the different methods used to solve the measurement inaccuracy and missing data problem.

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 to explain how animals interact with their environment (Patterson et al. 2008). These models can be quite complex, for example *Priesler* 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. *Gurram Kalman Chen Brown.* The extended Kalman filter is the de facto standard for GPS measurements.

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 assumption ignores the periodic nature of human movement, whereby people generally spend the nights at home and the day at work. 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).

Spatiotemporal Imputation Methods

For highly correlated in time.

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.

Methods

Data & Analyses

The data used was collected between 2014 and 2017 on different Android devices. A total of X individuals contributed to the analysis, yielding a total of Y data points. Moreover we used diary information from **SOMEWHERE**.

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)).

Plots: data points over time

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.

Filtering

Carlson et al 2015 describe a personal activity measurement system (PALMS), which can filter data points (among others). Their system is validated using cameras.

Like others (e.g. Palmius et al. (2017)) they detect invalid points using extreme speed.

Google offers an out of the box activity inferral mechanism.

Discussion on missing data

Our missing data patterns fit the broader pattern of missing data reported by others (Palmius et al. 2017; Saeb et al. 2015; Grünerbl et al. 2015; G. M. Harari et al. 2016).

Missing data can Importantly, one cannot say that the data is missing at random as there are a myriad of reasons why missingness can be related to location, such as turning off the phone on the plane before a long flight.

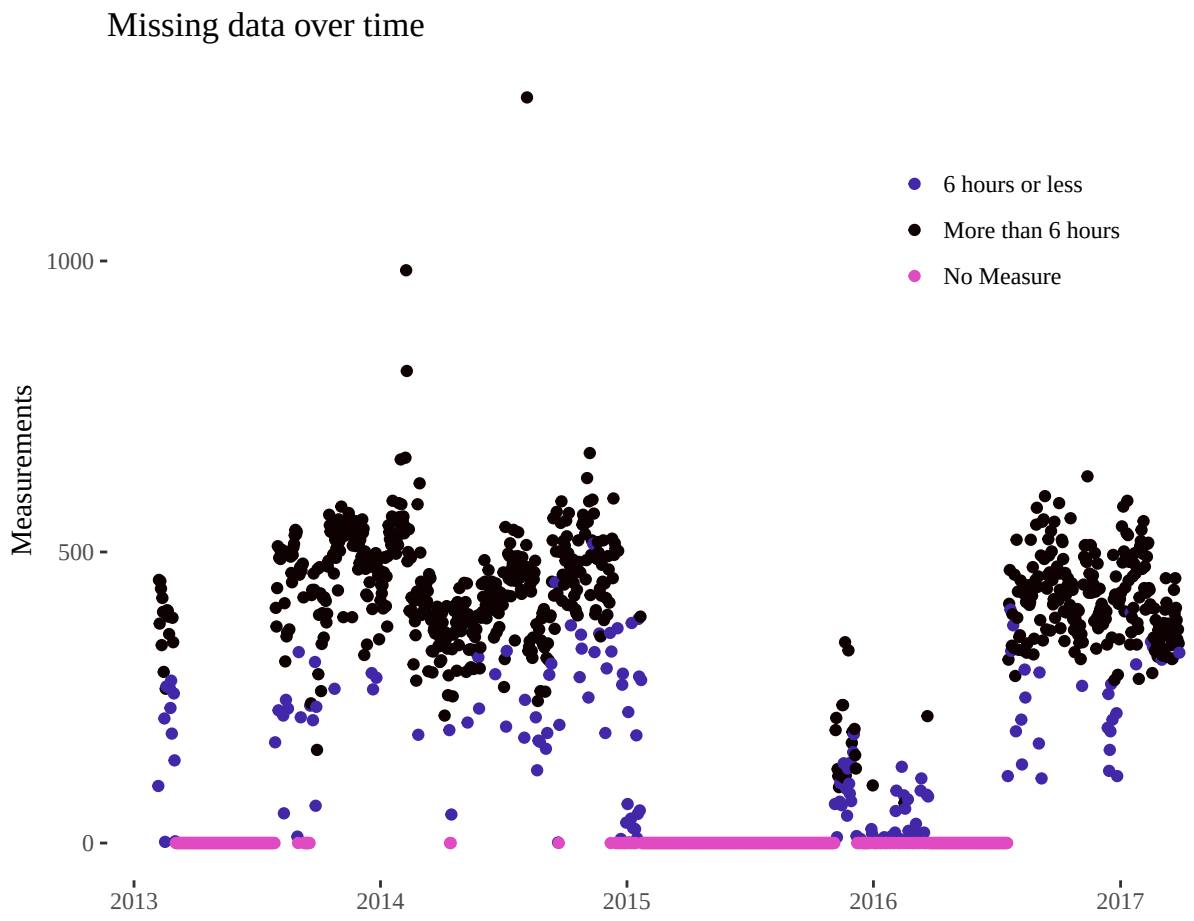

Figure 1: Missing data over time for the author. The x-axis denotes time, the y-axis shows how many measurements are made on each day. The fill of the points shows the amount of hours of captured data each day.

Models

Spatiotemporal imputation methods

Results

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.

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.

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.

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/.

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.

Zheng, Yu, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. 2009. “Mining Interesting Locations and Travel Sequences from GPS Trajectories.” In *Proceedings of the 18th International Conference on World Wide Web*, 791–800. WWW ’09. New York, NY, USA: ACM. doi:10.1145/1526709.1526816.