Using GPS-enabled mobile phones to characterize individuals' activity patterns for epidemiology applications

E.-H. Yoo and Y.-S. Eum

University of Buffalo, SUNY, Buffalo, NY, USA
Email: {eunhye, yeum}@buffalo.edu

Abstract

We assessed the potential of global positioning system (GPS)-equipped mobile phones for health-related studies. We demonstrated the use of GPS data as a means of collecting individuals' activity patterns for personal exposure assessments and public health surveillance. The widespread use of mobile phones has enabled investigators to conduct exposure studies and to detect infectious disease at the individual level on a massive scale. However, still substantial uncertainties are present in converting raw GPS data into relevant information. To address these issues, we proposed three algorithms for pre-processing and classification of raw GPS data, and demonstrated their applications to real world data in a case study.

1. Introduction

Exposure models typically impose unrealistic assumptions such that subjects within a neighborhood are equally exposed to air pollution and/or most individuals spend their time at their residences. Similarly, a lack of understanding of human movement, which is an important component of disease transmission, has been considered as an obstacle to develop effective national communicable disease control programs. In exposure modelling, some improvements have been achieved by adopting a microenvironment (ME) approach where individuals' time spent at MEs, such as outdoors, residence, and workplace, was explicitly taken into account. However, collecting the information on individuals' time-activity patterns has been cost-, time-, and labor-intensive with limited reliability and accuracy. Comparably, aggregated data have limited efforts to reconstruct the complex and dynamic nature of real-world contact networks, which plays a critical role of contact network in an outbreak of dangerous infectious disease. The emergence of lightweight, low-cost, and accurate GPS devices has provided a promising tool for objectively assessing the geographic positions of the environmental context in which health-related behaviors take place (Schipperijn *et al.* 2014). GPS technology enabled investigators to capture daily trajectories of individuals with higher temporal resolution at increasing locational precision (Gerharz *et al.* 2013, Dias & Tchepel 2014), although the use of GPS data in health research is not without challenges. As reviewed by Krenn *et al.* (2011), the positional accuracy of GPS data collected in real world is often unacceptable in health studies, especially, over longer study periods. The data quality of GPS traces depends on the amount of data lost from signal drop-outs, loss of device battery power, and poor adherence of participants to following the specific research protocol. Despite the advancement in GPS technology, signal acquisition is still affected by the presence of tall buildings and significant uncertainties associated with the processing and classifying raw data are present.

In this study we focus on the GPS data collected from a mobile phone with and without data connection. Our primary goal is to identify major MEs associated with health-related activities using GPS data. First we developed and applied a "selective resampler" to raw GPS data for pre-processing. Using the processed data, we identify significant places and travel modes using the two automated classification schemes.

2. Study Area and Data

The GPS data are collected from Mobile phone (Moto X with Android 5.1) in a highly urbanized environment, Sydney in Australia, for 9 days in February 2016. We collected data for a week and two randomly selected weekdays from one subject (Figure 1). We used GPS Logger for Android (GPS Logger 2016), which is a lightweight, battery efficient app to log GPS coordinates at 1-minute interval to a file. This app runs in the background of the phone and upload the data to a cloud server every 30-minutes with varying positional accuracy.

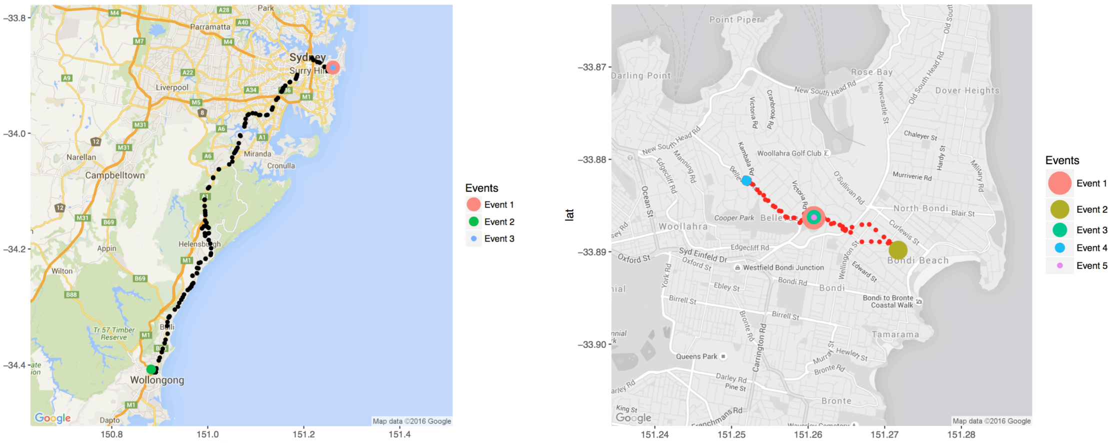

Figure 1. Activities identified by MEclassifier for weekday (left) and weekend (right)

3. Methods

3.1 Pre-processing raw GPS data

We cleaned raw GPS data by excluding data points with high positional error (e.g., above 100m) to avoid spurious results. Followed by the elimination of extreme outliers, we applied resampling to non-moving objects, i.e., GPS point associated with static activities. Our algorithm ("selective resampler") is different from the previous work (Dodge *et al.*, 2009), which enables us to avoid making unrealistic assumptions about human movement, such as it being linear, by performing resampling only on static activities. Our selective resampling algorithm requires the specification of two parameters: distance and time thresholds. These two parameters determine if any successive GPS points belong to a static activity with missing data points by comparing their separation distance and time interval to the specified thresholds. That is, resampling will be performed only if the distance and time interval between any two consecutive points is less than the distance threshold and greater than the time threshold, respectively.

3.2 Extracting significant places from GPS traces

We classified GPS traces to extract the information on static MEs using "MEclassifier", which is an extension of the density-based spatio-temporal clustering algorithm, ST-DBSCAN (Birant & Kut 2007). ST-DBSCAN detects non-linearly shaped clusters by separating clusters from noises by taking into account both spatial and temporal attributes. The unique feature of MEclassifier lies in the ability of discerning GPS-inherent error from true noise. We explicitly accounted for temporal continuity of GPS data by re-defining GPS noise, while defining moving activity as the GPS points between clusters. We developed a merging procedure to address acceptable GPS error and to simultaneously avoid situations where a single space-time cluster is divided into two MEs. We achieved this goal by defining two additional parameters to determine if any discrete cluster would be merged into a single cluster if both the minimum spatial distance between their centroids and the time interval are smaller than the specified merge parameters. Lastly, we contextualized the space-time clusters by matching with the information from a baseline survey, such as home and work address of individuals, and digital parcel data.

3.3 Detecting travel mode

We developed a travel model detection algorithm, named as "TMdetector", by combining statistical methods with machine learning algorithms. Any GPS trace connecting significant places (MEs identified from 3.2) was defined as a moving activity, which was further partitioned into multiple segments based on the mean and variance of speed. More specifically, we used a Pruned Exact Linear Time optimization method (Killick *et al.* 2012) to compute a modified Bayesian Information Criterion as a penalty function (Zhang & Siegmund 2007). The prediction of travel mode was obtained using random forest classifier based on speed, acceleration, and travel distance information.

4. Results

The GPS logger was programmed to record the location of the mobile device every minute, which yields a total of 1,440 GPS points per day. We quantified the amount of GPS data loss during the nine days of the study period, which varies day-to-day depending on the participant's activity patterns (see Table 1). The performance of MEclassifier is sensitive to missing GPS data because density-based clustering algorithms rely on the number and the location of GPS traces. Table 1 shows percentage of missing data associated with static activities, to which the resampling algorithm was applied. We selected two resampling parameter values based on the sensitivity analysis, which indicates that non-trivial amount of data was lost: all activities from 7.30% to 21.60% and static activities from 1.32% to 8.48% (sensitivity analysis results are not reported due to the limited space). After resampling, we re-examined the percentage of missing data for static activities, which shows that improvement was made on the four days— week days (day 4,8) and weekend (day 6, 7)—on which the subject participated on non- routine activities than routine activities of "home-work-home" as shown in Figure 1.

Table 1. Missing data (%) in all activities (All), static activities alone (Static), and static activities after resampling (Re-static)

Day	1	2	3	4	5	6	7	8	9
All	7.30	16.28	10.42	8.26	12.44	17.44	10.22	10.56	21.60
Static	2.62	6.11	2.64	6.81	1.32	7.44	8.48	5.77	6.25
Re-Static	2.62	6.11	2.64	**0.90**	1.32	**0.90**	**0.63**	**1.18**	6.25

We assessed the effects of resampling on classification by comparing the extracted information on MEs from raw GPS data and resampled data on day 4. Classification results were compared to the activity diary for validation, and the results are summarized in Table 2 and and Figure 1(Right). The classification of the raw GPS data failed to reproduce Event 2 while the Resampled data reproduce all the events.

Table 2. The effects of resampling on classification results (Day 4)

			Activity Diary		Before Resampling		After Resampling	
Event	ME-Type	TM-Type	Start	End	Start	End	Start	End
1	Home		00:00	10:20	00:00	10:20	00:00	10:20
		Walk	10:20	10:39	10:20	12:40	10:20	10:36

2	Other Indoor		10:39	12:00			10:36	12:02
		Walk	12:00	12:40			12:02	12:40
3	Home		12:40	14:31	12:40	14:31	12:40	14:31
		Walk	14:35	14:45			14:31	14:43
4	Other Indoor		14:45	16:36	14:43	16:36	14:43	16:36
		Walk	16:36	16:45			16:36	16:50
5	Home		16:45	23:59	16:50	23:59	16:50	23:59

5. Discussion and Conclusion

The importance of pre-processing of raw GPS data was highlighted for the applications in epidemiological studies, particularly when they are collected from mobile phones with potentially irregular data connection. The consequence of using missing GPS data may lead to biased or incorrect inference on the environments in which activities related to health outcomes occurred. Both the pre-processing and classification algorithms are an adjustment of existing methods, whereas the adjustment has critical implications to address research questions in epidemiological studies using GPS data collected from mobile phone. This paper presented a subset of our on-going project where the performance of the proposed algorithms has been tested using data collected from different regions over different periods. In near future we plan to integrate the extracted ME information with the air pollution concentrations to estimate personal exposure to air pollution. In addition, the identified time-activity patterns will be used to provide information on individuals' interactions and travel behaviours, which are the crucial information to capture dispersion process of influenza.

Acknowledgements

This work is funded from NIH/NIGMS), R01GM108731.

References

Birant, D. & Kut, A. (2007), 'ST-DBSCAN: An algorithm for clustering spatial–temporal data', Data & Knowledge Engineering 60(1), 208–221.

Dias, D. & Tchepel, O. (2014), 'Modelling of human exposure to air pollution in the urban environment: a GPS-based approach', Environmental Science and Pollution Research 21(5), 3558–3571.

Dodge, S., Weibel, R. & Forootan, E. (2009), 'Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects', Computers, Environment and Urban Systems 33(6), 419–434.

Gerharz, L. E., Klemm, O., Broich, A. V. & Pebesma, E. (2013), 'Spatio-temporal modelling of individual exposure to air pollution and its uncertainty', Atmospheric Environment 64, 56–65.

GPS Logger (2016), 'GPS Logger for Android'. URL:
https://play.google.com/store/apps/details?id=com.mendhak.gpslogger&hl=en

Killick, Rebecca, Paul Fearnhead, & IA Eckley. 2012. "Optimal Detection of Changepoints with a Linear Computational Cost." *Journal of the American Statistical Association* 107 (500). Taylor &amp; Francis: 1590–98.

Krenn, P. J., Titze, S., Oja, P., Jones, A. & Ogilvie, D. (2011), 'Use of global positioning systems to study physical activity and the environment: a systematic review', American Journal of Preventive Medicine 41(5), 508–515.

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.

Zhang, Nancy R, & David O Siegmund. 2007. "A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data." *Biometrics* 63 (1). Wiley Online Library: 22–32