# Implications of incomplete GPS-enabled mobile phone data on human mobility studies

EunHye Yoo
Department of Geography, University at Buffalo SUNY
Buffalo, NY, USA
eunhye@buffalo.edu

Youngseob Eum
Department of Geography, University at Buffalo SUNY
Buffalo, NY, USA
yeum@buffalo.edu

John E. Roberts
Department of Psychology, University at Buffalo SUNY
Buffalo, NY, USA
robertsj@buffalo.edu

### Abstract

Little is known about the quality of GPS-enabled mobile phone data in which the location of mobile phone is determined with special queries with pre-determined sampling intervals. This particular type of mobile phone traces, referred to as active mobile phone data, are of interest of this paper. We focus on the systematic gaps in active mobile phone tracking data and investigate their effects on human mobility pattern analyses. To address this under-sampling, we developed a strategy to impute missing data and assessed its impact on a mobility pattern analysis. We identified each participant's frequently visited places from both raw and imputed mobile phone data using a density-based clustering algorithm and compared the results with external data sources (i.e., participants' online survey responses). The results indicated that retrieval rates from imputed data were superior to raw data. Our study highlights the importance of understanding and addressing limitations of mobile phone derived positioning data prior to their use in human mobility studies.

*Keywords*: active mobile phone data, imputation, human mobility studies, under-sampling problem

## 1 Introduction

The increasing availability of tracking data collected from the Global Positioning System (GPS)-equipped devices has the potential to lead to paradigmatic shifts in scientific discovery across disciplines. For example, mobile phone data have been increasingly viewed as an alternative to traditional travel survey based activity-travel data in transportation research (Chung and Shalaby, 2005). Studies on physical activity and sedentary behaviour have also actively engaged in GPS technologies in recent years, typically combined with other movement sensors (e.g., accelerometers or pedometers).

For tracking individual movements, mobile phones are often preferred over GPS-standalone devices due to their portability, lower cost, and readily available connections and integration with Geographic Information System (GIS) databases. Although they have been particularly useful in resource-poor environments, the vast number of mobile phone users provides an opportunity for tracking human movement patterns within diverse physical environments throughout both the developed and developing world (Wesolowski et al., 2012; Chen et al., 2018). Despite the popularity of mobile phone data, studies on the accuracy or the coverage of location data are relatively sparse. Here it is worth noting that the present paper focuses on active mobile phone data in which the location of the mobile phone is determined in response to special queries using a radio wave (Ahas et al., 2008). They are typically collected using a mobile phone app that is specifically developed for that purpose with the permission from the phone holder. These 'GPS trace' data are distinct from `call detail records (CDR)' or 'signaling' data, which have been widely used in transportation and urban studies (Song et al., 2010; Jiang et al., 2013; Wang and Chen, 2018, Huang et al., 2019). The latter, ``passively generated mobile phone data'', are based on phone uses, e.g., calls, texting, or internet browsing, or when it communicates to the cellular network, and consequently these data tend to be sparse and less accurate compared to active mobile phone data that are based on GPS traces.

There are a number of studies that have evaluated either the accuracy or feasibility of standalone GPS devices in applications that require consistent and regularly sampled location information (Wiehe et al.; 2008, Xiao et al., 2012). Fewer studies have reported the properties and issues associated with passively generated mobile phone data. For example, Gonzalez et al. (2008), Song et al. (2010), and Calabrese et al. (2013) opted to select highly active users to address the highly irregular temporal gaps of CDR data for pre-processing. In contrast, even less is known about the quality of GPS trace data from GPS-enabled mobile phones despite the growing popularity and potential of this approach (Chan et al., 2018; Yoo, 2019). Unlike the passive mobile phone data, which estimate locations by cellular towers, GPS-enabled mobile phone tracking data (`active mobile phone data' hereafter) contain both a series of positions that approximate the movements of a user and network-driven positioning data. Thus, the quality of active mobile phone data is determined by the frequency of positioning operations (the sampling rate) and by the quality of positioning technology used (Renso et al., 2008). However, at present we do not have a clear view of how they are related to different aspects of the quality of GPS-enabled mobile phone derived tracking data. Nor is much known about the impact that such data will have on the results of mobility analyses.

Spatial data quality assessment and its impact on subsequent analyses and modeling have been one of primary research themes of Geographical Information Science (GIScience) (Goodchild, 1992; Haining, 2003). Given that mobile phone-based location data are rapidly becoming an important part of geospatial database, a careful assessment of its quality is needed. Among various facets of spatial data quality, the incompleteness of active mobile tracking data is directly associated with spatial data *consistency*. In the present paper, we aimed to estimate the degree of consistency in spatial data derived from GPS-enabled mobile phones using the analytical framework established in GIScience. We developed an imputation strategy that utilizes on local environment information, such as parcel boundaries, and the time intervals of consecutive recordings and evaluated its performance in the case studies.

## 2 Materials and Methods

### 2.1 Study Area and Location Data Collection

We used time-stamped location data obtained from 1,464 residents of the Buffalo-Niagara region within Erie and Niagara counties of western New York. Participants used their own GPS-equipped iPhone to collect time-location data over a six-month period (24 October 2016 to 4 June 2017). They also provided information about home and work addresses.

The information on each participant's movements was collected using a phone app developed by our project team. The app took into consideration of feasibility issues (e.g., battery drain), issues of privacy and surveillance, and the need for sufficient temporal resolution and stability by adopting the significant-change location' service [https://developer.apple.com/documentation/corelocation/getting_the_user_s_location] for phone users' positioning data collection. The location service ran the application in the background to save the battery power, and positioning data were collected only when participants were in motion and no records were available if a user remained static. Specifically, a phone user's location was updated only when the positional change was 500 m or greater with at least a 5 minutes time differences.

Participants also listed their five most frequently visited locations each week using an online survey. Based on the name and address of survey responses, participant's frequently visited locations were geocoded and used as reference data in the case study.

GIS parcel data were used to provide contextual information. The parcel data of the counties of Erie and Niagara were obtained from the New York State GIS clearinghouse (https://gis.ny.gov) and contained the information about the size and the property type of each parcel.

### 2.2 Imputation strategy

Let sequence $T_i = \{p_1, \dots p_k, \dots p_{n_i}\}$ be the mobile phone-based location data for the participant $i$ whose $k$-th position $p_k = (\boldsymbol{s}_k, t_k, A_k)$ consists of the spatial location (longitude, latitude), the time at which the position was recorded, and a list of additional information about the $k$-th position, such as horizontal and vertical accuracy, respectively.

The significant-change location service of iPhone provides an overall information on each participant's whereabouts, but creates systematic gaps in space and time and requires pre-processing. For instance, the sampling rate at night is likely to be lower than other time of day due to the inactivity of participants, which results in under-sampling at night. In turn, this sampling strategy is likely to affect the results of mobility patterns inferred from the mobile phone data.

To address this systematic sampling bias, we propose an imputation strategy based on both the time-stamp and spatial location of recording. We also combined parcel information of the study area to provide the environmental contextual information at each individual's trajectories. Specifically, we assessed the gap between two consecutive points in space and time and applied the following four steps of imputation rules.

Step 1. For each participant's trace $T_i,\ i = 1, \dots, n$, we searched for any pair of consecutive positions $(p_k, p_{k+1})$ whose spatial gap was greater than 500 m and the time interval between two observations was greater than or equal to 30 minutes and less than 24 hours, respectively. Some participants had a temporal gap greater than 24 hours in their traces, potentially due to turning off the phone. We decided not to apply the proposed imputation strategy in such cases without additional information.

Step 2. When the spatial and temporal gaps between two consecutive observations exceeded the thresholds described in Step 1, imputation was performed by generating missing time-location points at the original location $s_k$ at 30-minute time intervals. The imputation interval of 30 minutes was sufficient for the subsequent analyses of the weekly frequently visited places.

Step 3. We also searched for any consecutive points that fell in parcels in the study area whose size was greater than 1x1 km$^2$. If their time gap of the consecutive observations was greater than 30 minutes and less than 24 hours, we imputed points at the location of original point at the 30 minutes intervals.

Step 4. At known locations of each participants, such as home or work addresses, we searched for points that were located within 200 meters with the reported horizontal accuracy. If the selected points meet the condition described in Step 1, we imputed points at the original location $s_k$.

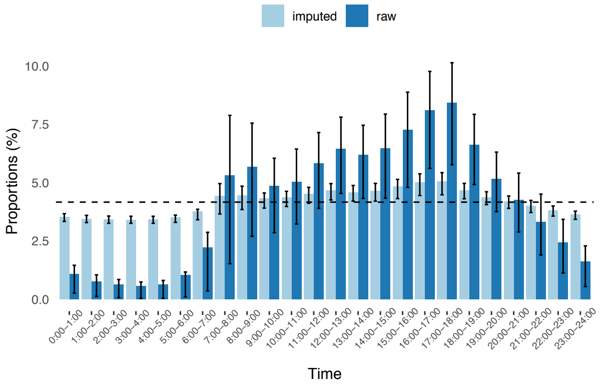
The performance of the imputation strategy was evaluated by comparing each participant's frequently visited places inferred from active mobile phone data to their responses from the weekly online survey. Specifically, each participant's frequently visited places were inferred by applying the algorithm of density-based spatial clustering of applications with noise (DBSCAN) to each participant's weekly mobile phone traces. We applied DBSCAN to both raw and imputed data and compared their results, i.e., locations of spatial clusters, to the survey-based frequently visited locations. To quantify the comparison results, we calculated a *retrieval rate* each week per participant. For example, if the DBSCAN results identified two out of four survey-based frequently visited places on a given week, the retrieval rate would be 50%. To facilitate the comparison between raw and imputed data, we calculated an average retrieval rate across entire study period and all participants for raw and imputed data, respectively.

We also conducted a sensitivity analysis of the retrieval rates with respect to DBSCAN parameters---the distance threshold for searching spatial neighbors (Eps) and the number of minimum data points to determine a cluster (MinPts). As the number of observations varies from week to week and a specification for a constant value of MinPts can be problematic, we chose the percentage of a total number of observations obtained in each week per participant. For example, 3% implies the case when the MinPts is set to the 3% of the total number of observations in that week.

## 3 Results

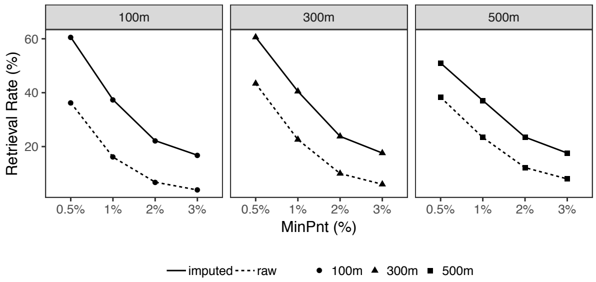
Imputed tracking data consisted of 14,734,667 spatial and temporal observations, whereas the raw data consisted of 5,076,774 records. The frequency of time-varying observations in raw and imputed data is summarized in pairs of barplots of Figure 1, and show substantial differences between the two types of data. Each pair of bars, which are shaded in light and dark colors, denotes imputed and raw data, respectively. Height represents the average proportions of tracking data collected within that hour. The horizontal dashed line represents an expected proportion of positional data under regular sampling. That is, if the same number of traces were observed each hour over 24-hour period, the proportion at each hour would be 1/24. The hourly proportions obtained from the imputed data closely approximated the horizontal line, whereas those of raw data substantially deviated from the dashed line over the time of day; there were relatively lower sampling rates (under-sampling) at night time (23:00–7:00) and higher rates (over-sampling) during day time (7:00–18:00). This is likely due to the significant-change location service of iPhones used in our data collection, as well as from GPS signal lapses commonly occurring inside buildings, such as being at home in early mornings. Movements during morning and afternoon commutes are likely to trigger more frequent tracking data collection according to the significant-change location service of iPhone, leading to the higher proportion of tracking data during those time periods.

Figure 1: Frequency of time-varying observations in raw and imputed mobile phone data



The performance of the proposed imputation strategy was evaluated by comparing the average retrieval rates calculated from raw and imputed data. The retrieval rate based on imputed data is as high as 60.6 %, whereas the corresponding rate from raw data is 36.3 %. To account for the sensitivity of the DBSCAN algorithm to the parameter specification, we calculated the average retrieval rates for a combination of different MinPts and Eps values and summarized the results in Figure 2. Imputed data yielded higher retrieval rates than raw data regardless of a parameter specification of DBSCAN. The results also suggest that retrieval rates are more sensitive to MinPts than Eps, as the retrieval rate of raw data increases from 3.9 to 36.3% by the change of MinPts in comparison to 3.9 to 8.0% improvement due to the change of Eps.

Figure 2 Retrieval rates of survey-based frequently visited places from raw and imputed mobile phone data



## 4 Discussion and Conclusions

Despite the popularity of passively generated mobile phone data in human mobility and transportation research, there are issues associated with their locational inaccuracy, as well as their irregular spacing and sparse resolution. Similarly, GPS tracking data from dedicated GPS loggers can provide frequent and accurate positioning data, but they tend to be used in research involving relatively small sample sizes and that cover relatively brief study periods. In contrast, active mobile phone data from GPS-enabled mobile phones has greater accuracy and resolution than passive phone data and can be more widely employed than dedicated GPS loggers. However, critical data quality issues need to be addressed including data consistency.

We proposed a simple but effective imputation strategy that utilized spatial and temporal contextual information to fill the gaps in space and time of active mobile phone data. This imputation strategy was validated by comparing mobility indicators derived from imputed versus raw data to weekly online survey responses collected from each participant. Of course, our focus on testing validity in terms of the frequently visited places are not exhaustive and there are numerous other aspects of human mobility that could be investigated using widely accepted indicators, such as the radius of gyration and the trip distance distribution (Schneider et al., 2013). However, our analyses are sufficient to demonstrate the need for data pre-processing.

Given that the survey data itself has not been validated and the quality of online survey data may vary across participants, caution is warranted in interpreting the results of our validation study. We are also confident that our imputation approach can be further improved to account for the accuracy of positioning data because the accuracy of positioning data is time- and space-dependent. Much work still needs to be done in both understanding and analyzing active mobile phone data in human mobility studies.

## Acknowledgements

This work is funded in part by US National Institutional Health (NIH) Grant R01GM108731. I am grateful for helpful comments from Dr. Bian. The opinions expressed herein are those of the author and do not necessarily reflect the views of NIH.

## References

Ahas, R., Aasa, A., Roose, A., Mark, U., and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management*, 29(3):469–486.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies* 26:301–313.

Chan, Y.-F. Y., Bot, B. M., Zweig, M., Tignor, N., Ma, W., Suver, C., Cedeno, R., Scott, E. R., Hershman, S. G., Schadt, E. E., et al. (2018). The asthma mobile health study, smartphone data collected using Researchkit. *Scientific Data*, 5:180096

Chen, B. Y., Wang, Y., Wang, D., Li, Q., Lam, W. H., and Shaw, S.-L. (2018). Understanding the impacts of human mobility on accessibility using massive mobile phone tracking data. *Annals of the American Association of Geographers*, pages 1–19.

Chung, E.-H. and Shalaby, A. (2005). A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology*, 28(5):381–401.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Goodchild, M. F. (1992). Geographical information science. *International Journal of Geographical Information Systems*, 6(1):31–45.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. pages 61– 74, Cambridge University Press, New York.

Huang, H., Cheng, Y., and Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*. https://doi.org/10.1016/j.trc.2019.02.008

Renso, C., Puntoni, S., Frentzos, E., Mazzoni, A., Moelans, B., Pelekis, N., and Pini, F. (2008). Wireless network data sources: tracking and synthesizing trajectories. In *Mobility, Data Mining and Privacy*, pages 73– 100. Springer.

Schneider, C. M., Belik, V., Couronn´e, T., Smoreda, Z., and Gonz´alez, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246.

Song, C., Qu, Z., Blumm, N., and Barab´asi, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Wang, F. and Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87:58–74.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270.

Wiehe, S. E., Hoch, S. C., Liu, G. C., Carroll, A. E., Wilson, J. S., and Fortenberry, J. D. (2008). Adolescent travel patterns: pilot data indicating distance from home varies by time of day and day of week. *Journal of Adolescent Health*, 42(4):418–420.

Xiao, Y., Low, D., Bandara, T., Pathak, P., Lim, H. B., Goyal, D., Santos, J., Cottrill, C., Pereira, F., Zegras, C., et al. (2012). Transportation activity analysis using smartphones. In 2012 *IEEE Consumer Communications and Networking Conference* (CCNC), pages 60–61. IEEE.

Yoo, E.-H. (2019). How short is long enough?: Modeling temporal aspects of human mobility behavior using mobile phone data. *Annals of the Association of American Geographers*, http://10.1080/24694452.2019.1586516