9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Comparing Methodologies for Constructing Epidemic Networks from GPS Mobility Data

Keywords: GPS data, normalization, imputation, epidemic modeling, human mobility

Extended Abstract

The increased availability of high-resolution GPS human mobility data has had an enormous impact on our understanding of human behavior, with applications ranging from urbanism and transportation to epidemic modeling. In particular, the Covid-19 pandemic and subsequent data-for-good initiatives from several data companies have resulted in a sharp increase in the availability of this type of data for use in epidemiological research. However, this data is often composed of a small sample of individuals and is known to feature sparse and irregular signals at the individual level. To effectively integrate this data into epidemic models, appropriate methods must be employed to normalize the data. Despite this, the most suitable method for constructing epidemic networks from this data is still not clearly established.

In order to analyze the effects of normalization in GPS human-mobility data for constructing origin-destination networks in the context of epidemics, we compare four different methodologies [3, 4, 1, 2]. Our focus is on within-city networks, as these networks can be used to inform local policy design and reveal the heterogeneity of human contact patterns. The varying methodologies are distinguished by their models and assumptions used to predict unobserved activity in the typically sparse mobility data. Specifically, they vary on three main aspects: (1) whether individuals in a certain geographic area are aggregated and assumed to have homogeneous behavior; (2) how origin-destination flows are scaled to obtain unbiased population estimates; and (3) whether and how information about individuals staying at home is incorporated. We replicate all methodologies in the same dataset.

The significance of our research lies in how it demonstrates that properties of epidemic networks and their impacts on models can be drastically altered by the underlying methodologies used to normalize the data. This prompts a reexamination of the validity of certain results in the literature which did not sufficiently take into account sparsity and robustness issues. Furthermore, our research highlights the limitations of this type of data, and brings further attention to the need for caution when using GPS human-mobility data for epidemic-modeling research.

Data. We use GPS human-mobility data from Spectus, a leading location intelligence company, which provides data to academic and humanitarian initiatives through its Data for Good program. These data are first-party and collected from anonymous users who have opted in to share their location data. The privacy of the data is enhanced by removing pings associated with sensitive locations and by upleveling the location of users' "personal areas", such as home locations, to the centroid of Census block groups. The data consists of time-stamped GPS coordinates (pings) tied to unique anonymous device identifiers. We model flows to and from census block groups (CBGs). For normalization purposes, we use the 2020 5-year American Community Survey. Our analysis is restricted to GPS pings from users estimated to reside in the city of Philadelphia in the first 2 weeks of April 2020; corresponding to a sample of 27,551 users.

Methods. Constructing O-D flows from individual visitation data requires trips to be aggregated from individuals in O visiting D. This is challenging because the sample of devices in

each CBG is only a fraction of the population, and this rate can vary significantly. Furthermore, the amount of data for each individual varies and is usually sparse. Thus, it is unclear how to weight each visit in the aggregate. This sparsity shown in Figure 1, where more than half of users have activity in less than 15% of the hours. Filtering out users with few data points could reduce the sample size, leading to increased bias, as seen in Figure 2, where panel (a) shows the ratio of devices to population of each CBG in the original sample, and panel (b) shows the same ratios when restricting to devices with at least 10 hours with activity. We can see a sharp reduction in sample size, leaving virtually no devices in a large number of CBGs.
We first cluster individual-level pings to obtain hourly visitation data of users to destinations by first applying a time-augmented version of the DBScan clustering algorithm which extracts user stops. Then, we map each stop to a CBG and split the stop into multiple visits, one for each hour the stop straddles (see Panel (a) of Figure 1). Additionally, Spectus provides estimates of devices' CBG of residence based on frequent nighttime locations over the preceding months.
The four normalization methodologies for O-D networks we compare are **Method 1:** Hourly O-D flows estimated using the hourly number of visitors in each destination—upscaled by the population-wide sampling rate—and a daily estimate of individuals staying home in each CBG, which uses a baseline monthly O-D flow for regularization [3]; **Method 2:** Hourly O-D flows estimated using hourly number of visitors in each destination and the weekly distribution of visitors' origin, using a Bayesian hierarchical model to smooth the CBG-level sampling rates and the highly-sparse hourly visits [1]; **Method 3:** Daily O-D flows obtained by a straightforward upscaling the counts of visitors from O to D by the origin CBG's sampling rate [4]; and **Method 4:** An imputation of the location of each individual user each hour, either at destinations or home, using neural collaborative filtering, and then aggregation by CBG of origin to form O-D networks [2].

Results. We show that depending on the methodology used, the resulting networks exhibit significantly different attributes in terms of their degree distribution, their modularity, most central nodes, and other features. Additionally, we show that an epidemic simulated with an SEIR model overlaid on these networks produces very different results; namely the identification of hotspots that depend on the normalization technique as well as different epidemic sizes and asymptotic behavior. Additionally, we compare each methodology in the ability to predict the destinations of a held-out set of users and in their ability to predict mobility in the future.

References

[1] Francisco Barreras, Mikhail Hayhoe, Hamed Hassani, and Victor M Preciado. Autoekf: Scalable system identification for covid-19 forecasting from large-scale gps data. *arXiv preprint arXiv:2106.14357*, 2021.

[2] Francisco Barreras, Bethany Hsiao, Hamed Hassani, Victor M Preciado, and Duncan J Watts. *29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, forthcoming.

[3] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of covid-19 explain inequities and inform reopening. *Nature*, 589(7840):82–87, 2021.

[4] Yuhao Kang, Song Gao, Yunlei Liang, Mingxiao Li, Jinmeng Rao, and Jake Kruse. Multiscale dynamic human mobility flow dataset in the us during the covid-19 epidemic. *Scientific data*, 7(1):1–13, 2020.

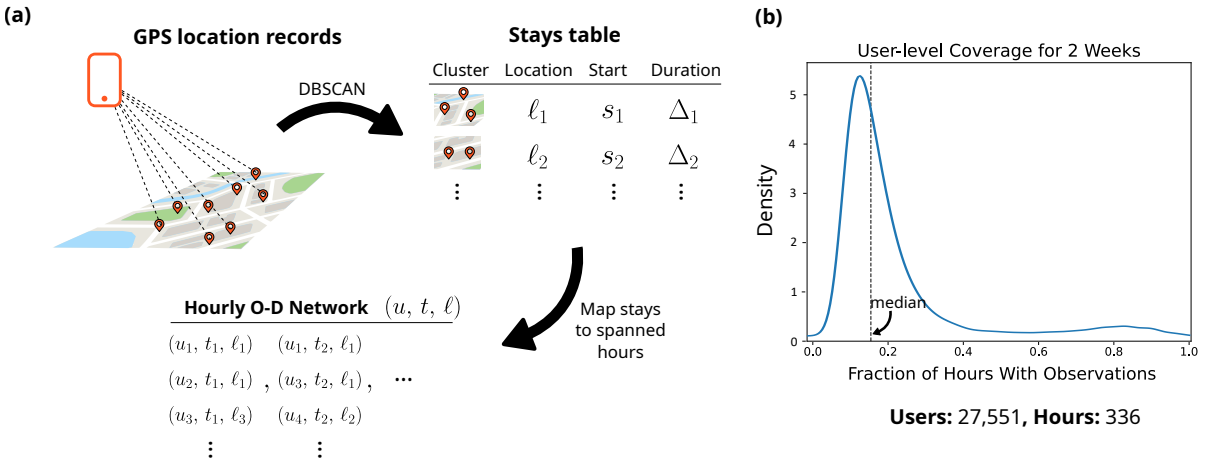

Figure 1: **(a)** Methodology to process location data and construct a bipartite network between individuals and destinations visited. GPS pings for each user are clustered using a time-augmented version of DBScan to extract stops, which are joined to the destination closest to their centroids; these stays are split by hours to form an hourly network between users and destinations. **(b)** Kernel density estimation of individual-level coverage for users estimated to reside in Philadelphia over the first two weeks of April, 2022; defined as the fraction of hours for which a user has any data points. This figure reveals the sparsity in this type of data and the need for normalization.

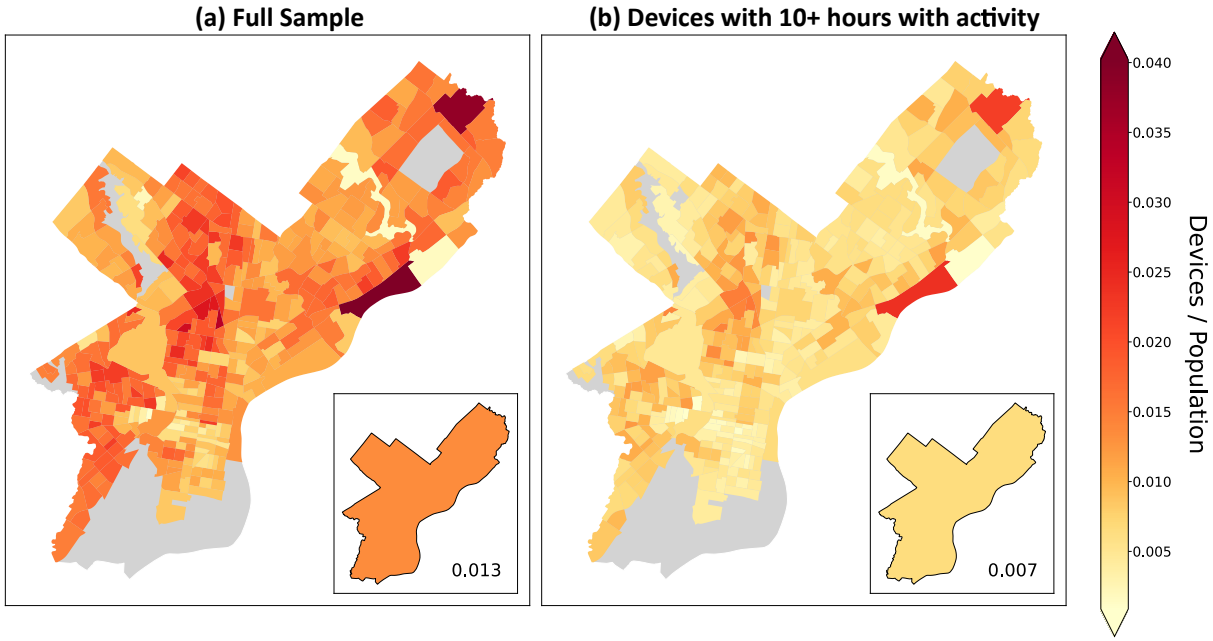

Figure 2: Distribution of the number of devices estimated to reside in each CBG expressed as a fraction of the population of each CBG. **(a)** shows the uneven distribution of devices across CBGs and the very high sparsity pattern that makes normalization challenging. **(b)** shows the same distribution for the sample of high-coverage devices with 10+ hours of activity for which the distribution is even more sparse and uneven. The differences between the two heatmaps show that a restriction to high-coverage users reduces the sample steeply and would make normalization more difficult.