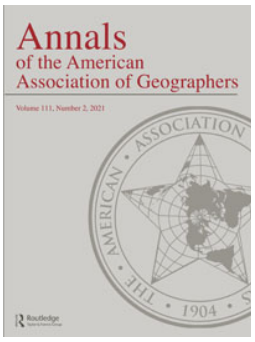


Taylor & Francis
Taylor & Francis Group

Annals of the American Association of Geographers

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/raag21

Effects of Data Preprocessing Methods on Addressing Location Uncertainty in Mobile Signaling Data

Yang Xu, Xinyu Li, Shih-Lung Shaw, Feng Lu, Ling Yin & Bi Yu Chen

To cite this article: Yang Xu, Xinyu Li, Shih-Lung Shaw, Feng Lu, Ling Yin & Bi Yu Chen (2021) Effects of Data Preprocessing Methods on Addressing Location Uncertainty in Mobile Signaling Data, Annals of the American Association of Geographers, 111:2, 515-539, DOI: 10.1080/24694452.2020.1773232

To link to this article: https://doi.org/10.1080/24694452.2020.1773232

Published online: 28 Jul 2020.

Submit your article to this journal

Article views: 629

View related articles

View Crossmark data

Effects of Data Preprocessing Methods on Addressing Location Uncertainty in Mobile Signaling Data

Yang Xu,[*,†] Xinyu Li,[*] Shih-Lung Shaw,[‡] Feng Lu,[§] Ling Yin,[#] and Bi Yu Chen[¶]

[*]*Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University*
[†]*The Hong Kong Polytechnic University Shenzhen Research Institute*
[‡]*Department of Geography, University of Tennessee*
[§]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, and the Academy of Digital China, Fuzhou University*
[#]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences*
[¶]*State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, and Collaborative Innovation Center of Geospatial Technology*

Recent years have witnessed an increasing use of big data in mobility research. Such efforts have led to many insights on the travel behavior and activity patterns of people. Despite these achievements, the data veracity issue and its impact on the processes of knowledge discovery have seldom been discussed. In this research, we investigate the veracity issue of mobile signaling data (MSD) when they are used to characterize human mobility patterns. We first discuss the location uncertainty issues in MSD that would hinder accurate estimations of human mobility patterns, followed by an examination of two existing methods for addressing these issues (clustering-based method and time window–based method). We then propose a new approach that can overcome some of the limitations of these two methods. By applying all three methods to a large-scale mobile signaling data set, we find that the choice of preprocessing methods could lead to changes in the data characteristics. Such changes, which are nontrivial, will further affect the characterization and interpretation of human mobility patterns. By computing four mobility indicators (number of origin–destination trips, number of activity locations, total stay time, and activity entropy) from the outputs of the three methods, we illustrate their varying impacts on individual mobility estimations relevant to location uncertainty issues. Our analysis results call for more attention to the veracity issue in data-driven mobility research and its implications for replicability and reproducibility of geospatial research.

Key Words: human mobility, mobile phone data, uncertainty, veracity.

Big data is no longer a buzzword but something that truly affects how academic research is performed. With the fast development of information and location-aware technologies, the types and sizes of data suitable for large-scale geographical analysis are augmented on a daily basis, bringing new questions to the field or introducing alternatives to classical problems. As we celebrate the increasing volume and velocity of big data, one crucial question that remains to be better addressed is veracity. As we bring data on board, process, and then analyze them, how much can we trust the results given the methodologies that are used?

Taking mobility research as an example, various approaches have been proposed to derive origin–destination (OD) trips from movement data sets. Although the definition of OD trips seems to be simple, extracting them from particular data sources could introduce errors and bias. For instance, OD trips can be under- or overestimated from travel surveys due to self-report errors (Stopher and Greaves 2007; Chen et al. 2010). Some studies extract OD trips from smart card transactions to examine public transport usage. Some of these data sets, however, record only the tap-in events of passengers (i.e., where they get onboard). The destinations of the trips need to be further estimated or guessed (Trépanier, Tranchant, and Chapleau 2007; Robinson et al. 2014). Social media data have also been used to derive OD matrices to support transport planning (Yang et al. 2015). The mobility traces of social media users can be sparse in time

Annals of the American Association of Geographers, 111(2) 2021, pp. 515–539 © 2020 by American Association of Geographers
Initial submission, October 2019; revised submission, February 2020; final acceptance, March 2020
Published by Taylor & Francis, LLC.

and space, though. In other words, in the contemporary big data analytics, given the peculiar characteristics of the raw data, the methodologies used will largely affect the final results, which direct the findings of the studies.

Another good example is the practice of mobile phone data. Due to the increasing adoption of mobile phones worldwide, the digital footprints documented by these devices have introduced new opportunities to human mobility research. Call detail records (CDRs)—a typical type of phone data that track individual whereabouts during phone usage activities (e.g., calls, text messages)—have been used extensively to study human travel and activity patterns (Iqbal et al. 2014; Alexander et al. 2015; Jiang et al. 2016, Jiang, Ferreira, and González 2017; Xu et al. 2018). CDRs suffer from issues of data sparsity (due to the passive data collection mechanism) and location uncertainty (e.g., cellphone signal switch), however, adding notable complexities to the estimation of travel patterns (Isaacman et al. 2012; Csáji et al. 2013; Xu et al. 2015; Zhao et al. 2016). Similar issues also exist in other types of mobile phone data (e.g., mobile sightings data, mobile signaling data) and have been discussed by previous researchers at different depths (Chen, Bian, and Ma 2014; Xu et al. 2016; F. Wang and Chen 2018).

Much of the uncertainty in mobile phone data is associated with positional inaccuracy—a key form of uncertainty in geospatial data (Goodchild 1998). In 2004, the University Consortium for Geographic Information Science identified uncertainty in spatial data as a long-term research challenge (McMaster and Usery 2004). Research attention to uncertainty issues of mobile phone data appears to have become notable only in recent years, however (W. Wu et al. 2014; Chen et al. 2016; Kwan 2016; Xu et al. 2016). A key characteristic of mobile phone data is that the locations are documented at the level of cell towers. These locations, which are usually represented as geographic coordinates of the cell towers, do not necessarily reflect the actual locations of the phone users. For instance, a cellphone's signal could oscillate between neighboring or even distant cell towers due to load balancing or signal strength variation (Kwan 2016; Xu et al. 2016). Such issues of positional inaccuracy have hindered reliable estimates of human mobility patterns that are important to many geospatial applications. Although these issues have been noticed by the research community (Kwan 2016), most efforts have been devoted to demonstrating the value of these data without questioning—or at least carefully examining—the uncertainty and veracity issues associated with the data.

The oversight of these issues is not without reason. An important fact to mention is that these mobility data sets—when they are born—are not intended for travel behavior analysis. The lack of "ground truth" makes it challenging to validate the analytical results (i.e., OD estimation). Much hope, as a result, has been put on the expectations that researchers will do it right or the errors will balance each other out when some kinds of aggregations are performed (e.g., estimating OD matrices at the level of traffic analysis zones). Although we have to acknowledge the absence of ground truth as a normality of many "big" mobility data sets, there is a need for alternatives that would look into this issue—by comparing different methodologies, their pros and cons, and the trade-offs among different practices.

In this research, we aim to investigate the veracity issue of mobile phone data when they are used to characterize human mobility patterns. In particular, we involve the usage of mobile signaling data (MSD), a typical type of phone data used in human mobility research. By applying several preprocessing methods over the data set, we examine how these methods change the data characteristics in different ways and how such changes would affect the characterization of individual human mobility patterns due to location uncertainty.

The article is organized as follows. First, we discuss uncertainty issues in MSD that would hinder accurate estimations of human mobility patterns, followed by an examination of two existing methods (clustering-based method, time window–based method) for tackling or mitigating these issues. We then propose a new approach that could overcome some of the limitations of these two methods. By processing mobile phone data using all three methods, we derive a collection of indicators to systemically compare their outputs, with the primary focus on examining their abilities to tackle oscillations (i.e., the ping-pong effect) in the data. We further derive a collection of individual mobility indicators from three sets of output—namely, the number of OD trips, number of activity locations, total stay time, and activity entropy—and evaluate the impact of preprocessing methods on mobility estimations. Finally, we discuss the implications of the results for future mobility studies and geographic knowledge discovery.

Mobile Phone Data for Human Mobility Analysis: A Brief Review

Mobile phone data have been used for human mobility analysis for more than a decade. In 2006, scholars at the Massachusetts Institute of Technology adopted cellular data to study the spatiotemporal dynamics of human activities in cities (Ratti et al. 2006). At that time, the study used Erlang data, a standard measure in the telecommunications industry that records person hours of cellphone usage (Ratti et al. 2006). Because Erlang is an aggregate measure of traffic volume in telecommunications system, the data are not suitable for studying movement patterns. Later on, call detail records, another type of phone data usually collected by cellular operators for billing purposes, began to attract academic attentions. Due to the ubiquity of mobile phones, CDRs are capable of quantifying mobility of large populations. To date, CDRs have been used to study human mobility from various perspectives, generating numerous insights into the regularities of individual movements (Gonzalez, Hidalgo, and Barabasi 2008; Song, Koren, et al. 2010; Song, Qu, et al. 2010; Pappalardo et al. 2015; Xu et al. 2018), usage of urban space (Becker et al. 2013; Silm and Ahas 2014; Xu et al. 2015; Yuan and Raubal 2016), interplay between mobility and social network structures (Cho, Myers, and Leskovec 2011; D. Wang et al. 2011; Calabrese, Smoreda, et al. 2011; Gao et al. 2013; Toole et al. 2015; Xu et al. 2017; Xu et al. 2019), and so forth (see Blondel, Decuyper, and Krings [2015] and Birenboim and Shoval [2016] for extensive reviews).

Many travel behavior studies have involved the usage of CDRs for OD estimation and mobility modeling (Alexander et al. 2015; Jiang et al. 2016; Pappalardo et al. 2016; Bwambale, Choudhury, and Hess 2017; Jiang, Ferreira, and González 2017; Xu et al. 2018). There are a few characteristics of CDRs that would complicate such tasks, however.

- CDR data are collected at the level of cell towers, of which the densities in space affect the positioning accuracy. The spacing between cell towers in a city or region could range from a few hundred meters (e.g., in densely populated urban areas) to several kilometers (e.g., in suburbs).
- The tower-to-tower balancing in the mobile network systems will produce noise for CDRs, which causes "the appearance of fake movements" (Alexander et al. 2015, 241).
- CDRs suffer from a data sparsity issue as positions of users are partially detected (e.g., during phone calls and text messages).

To overcome these issues, researchers have proposed some solutions, and the key ideas can be summarized as follows:

- Clustering-based methods are introduced to detect stay locations. A key practice is to group consecutive location observations that are close in space into clusters (Calabrese, Lorenzo, et al. 2011; Widhalm et al. 2015; Fan et al. 2018). Such clustering methods are able to filter some "fake movements" while capturing meaningful activity locations of individuals.
- Beyond this step, some studies also perform an additional step to merge the detected clusters that are close in space but might be far apart in time (Alexander et al. 2015; Xu et al. 2018). The purpose of this step is to maintain the unique identity of activity locations. For instance, two stay locations of an individual can be detected in the early morning and evening in the same day, with their representative locations (e.g., mean center or medoid of observation locations in the clusters) being different but geographically close. It is highly likely that these two stay locations refer to the same activity location of the user (e.g., home).
- OD trips of an individual can then be extracted through travels conducted between consecutive stay activities.
- To tackle the data sparsity issue, some studies filtered individuals or observation days with few records. For example, some researchers define an *active observation day* as a day where "the user has phone records in at least 8 distinct time-slots of the 48 half-hour time-slots" (Jiang, Ferreira, and González 2017, 212). This practice will partially address the data sparsity issue. The choice of the threshold, however, which is empirical and somewhat arbitrary, could have a direct impact on the analysis that follows.
- Another factor that causes fake movements is cell tower oscillation, also known as the ping-pong effect. Such effects are caused by the users' cellphone handover to nearby cell towers due to load balancing, operations by the telecommunication systems, or other factors. As a result, the documented locations of users—even when they stay still—will "bounce" back and forth between two or more base stations. Different solutions are proposed to tackle this issue. For example, W. Wu et al. (2014) proposed a few

Table 1. Summary of signaling events captured in the data set

Event type	Description
Outbound communication (OT)	Triggered by outbound phone call or text message
Inbound communication (IN)	Triggered by inbound phone call or text message
Regular update (RU)	Triggered by regular update of cellular state (active or idle)
Periodic update (PU)	Triggered by periodic tower pinging
Cellular handover (CH)	Triggered by cellphone handover from one antenna to another
Power on (ON)	Triggered when a phone is turned on and accesses the cellular network
Power off (OFF)	Triggered when a phone is turned off and disconnects from the cellular network

heuristics to identify oscillations by detecting movements at impossible speed or cell towers that appeared repeatedly. F. Wang and Chen (2018) proposed a time window–based method to detect oscillations as circular trips that occurred within a short period of time. Bayir, Demirbas, and Eagle (2010) introduced a graph-based clustering algorithm, which iteratively merges densely connected cell towers in a user's trajectory to address the oscillation effect.

Note that some studies also adopted sightings data for travel behavior analysis (Calabrese et al. 2013; Chen, Bian, and Ma 2014). Sightings data can be considered a "sibling" of CDRs. On the one hand, the data have a similar generation mechanism in that locations are passively collected during phone usage activities. On the other hand, instead of reporting the user footprint at the cell tower level, sightings data provide location estimates through triangulation technology, which further improves the spatial granularity of observations.

Despite the numerous insights into human mobility discovered from CDRs and sightings data, the observations from them are generally sparse due to the passive data collection mechanism. MSD, instead, provide a more fine-grained view of human mobility traces especially from the temporal aspect. Different from CDRs and sightings data, which are recorded during phone usage activities, MSD could capture user footprints in a more continuous manner through different types of events triggered by the telecommunications system (Janecek et al. 2015). Depending on the state of a phone—active when phone usage is detected or idle when no user activities are observed—location observations can be captured by different types of signaling events, such as cellular handover, calls, SMS, data connection, and other types of location updates. The improvement in data granularity makes MSD an appealing option for mobility studies (Z. Li et al. 2018; M. Li et al. 2019; Yan et al. 2019). Similar to CDRs and sightings data, however, issues of location uncertainty (e.g., tower-to-tower balancing, oscillation effect) still persist. There is a need to develop proper methods to handle these issues and, meanwhile, discuss their impact on travel behavior analysis.

Data

A large mobile signaling data set collected in Shanghai, China, is used. The data set captures the location traces of 7.6 million phone users during a period of one week (15–21 October 2012). The locations of phone users were tracked at the level of cell tower antennas (referred to as cells), and the location reporting was triggered through different types of signaling events. Table 1 provides a summary of the key events captured in the data set. For instance, when users engage in active phone usage, their locations will be documented by the outbound communication (OT), inbound communication (IN), or cellular handover (CH) events.[1] Even if the user has been silent for a while (i.e., no phone usage activities or movements), her location will be reported by the regular update (RU) or periodic update (PU) events.

Different from CDRs, which passively collect data during phone usage activities, the MSD set tracks user locations in a more continuous manner. This is because the RU and CH events are able to capture user movements at the cell level whether the user engages in phone usage activities or not. To elaborate, if two consecutive records (ordered by time) of a user correspond to the same cell, we can assume that the user stayed at that location during this period, because movements across cells will be recorded by the corresponding event (RU or CH).

This study takes user daily trajectory, defined as the location sequence of a user of a single day, as the basic unit for subsequent analysis. Because mobile phones can be switched on or off (Table 1),

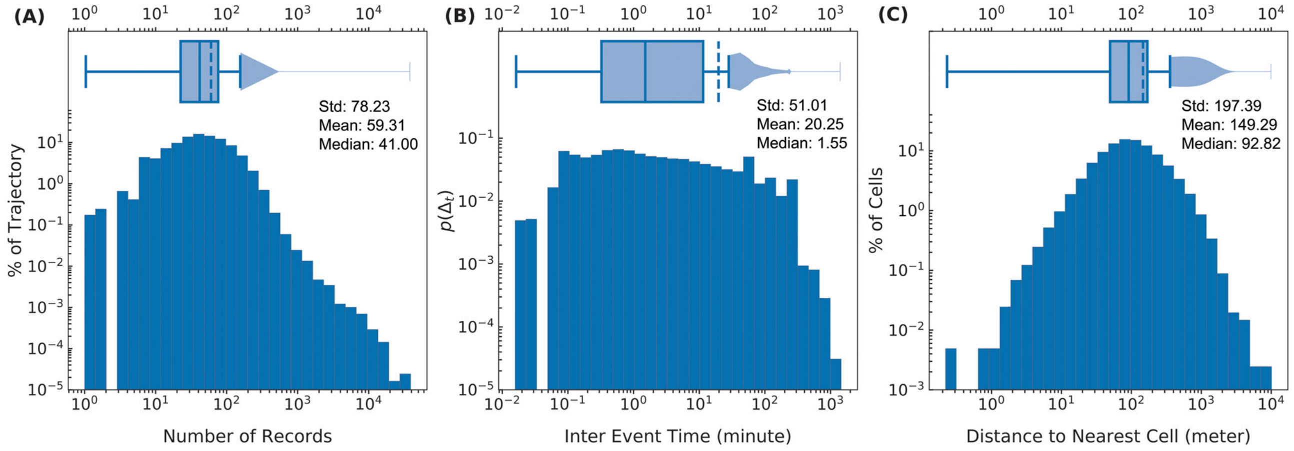


Figure 1. General statistics of the mobile signaling data: (A) number of records per trajectory; (B) distribution of interevent time; (C) spacing gap between cell tower antennas.

location of users will be unavailable during the disconnected periods. Thus, we filter the data set by removing user daily trajectories with power on (ON) or power off (OFF) events. The data set after removing such cases includes 36.7 million user daily trajectories. In other words, an average user would contribute approximately $36.7 \div 7.6 \approx 4.8$ valid trajectories to the filtered data set.

Figure 1 shows some general statistics of the data set. The number of records of a trajectory vary notably from each other, with mean and median values of 59.3 and 41.0, respectively (Figure 1A). The interevent time, measured as the duration between two consecutive records in a trajectory, ranges from a few seconds to several hours (Figure 1B). The mean and median values are 20.25 and 1.55 minutes, respectively. To better understand the density of cell tower antennas in the city, we compute, for each cell, its distance to the nearest cell. This yields a skewed distribution with mean and median values of 149.3 m and 92.8 m, respectively (Figure 1C). To obtain a better understanding of the spatial distribution of cell towers in the city, we generate a 1 km * 1 km regular grid and compute the number of towers in each grid cell. As shown in Figure 2, cell towers are unevenly distributed in Shanghai and their densities are generally higher in the core part of the city (e.g., Downtown Shanghai).

Method

In this section, we first define several key concepts that are pertinent to our analysis, followed by a summary and implementation of two existing methods for preprocessing mobile signaling data (clustering-based method, time window–based method). The analysis results will be shown in the next section to demonstrate the impact of these methods and the choice of the key parameters on data characteristics. Finally, we introduce an improved algorithm to overcome some of the limitations in these two approaches.

Definitions

A user daily trajectory (T) is defined as a sequence of tuples:

$$T = \{(l_1, t_1), (l_2, t_2), \ldots, (l_n, t_n)\}, \tag{1}$$

where l_i and t_i denote the location and the time of the ith observation.

A *displacement* (ds) is defined as the Euclidean distance between two consecutive observations (l_i, t_i) and (l_{i+1}, t_{i+1}) in a trajectory T:

$$ds = \sqrt{(\vec{l_{i+1}} - \vec{l_i})^2}. \tag{2}$$

A *type 1 oscillation pair* (O_{p1}) is defined as a subsequence of T, with a length of $N(O_p) = 3$, in which the observations "bounce" back and forth between two locations:

$$O_{p1} = \{(l_i, t_i), (l_{i+1}, t_{i+1}), (l_{i+2}, t_{i+2})\} \tag{3}$$

subject to

$$\sqrt{(\vec{l_{i+2}} - \vec{l_i})^2} = 0. \tag{4}$$

In the remainder of the article, we use *A-B-A* to describe the phenotype of such oscillation pairs.

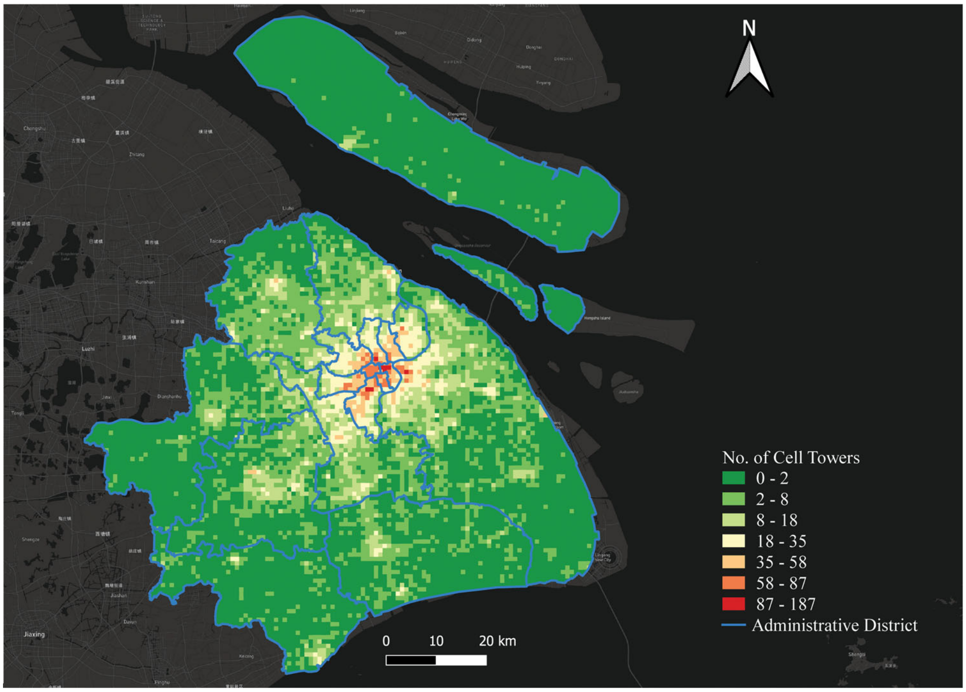


Figure 2. Number of cell towers in each grid cell (1 km * 1 km grid).

A *type 2 oscillation pair* (O_{p2}) is defined as the following subsequence:

$$O_{p2} = \{(l_i, t_i), (l_{i+1}, t_{i+1}), (l_{i+2}, t_{i+2}), (l_{i+3}, t_{i+3})\} \quad (5)$$

subject to

$$\sqrt{(\vec{l}_{i+3} - \vec{l}_i)^2} = 0 \quad (6)$$

and

$$\sqrt{(\vec{l}_{i+2} - \vec{l}_{i+1})^2} = 0. \quad (7)$$

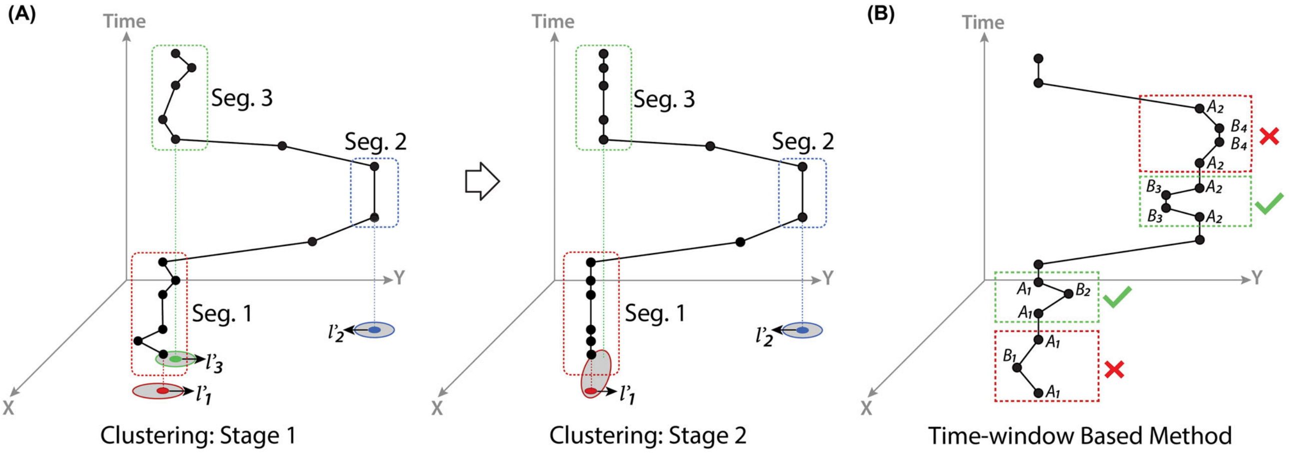
Similarly, the phenotype of type 2 oscillation pair can be represented as *A-B-B-A*.

An *oscillation sequence* (O_s) is defined as a subsequence of T that consists of continuous appearance of type 1 or type 2 oscillation pairs or a combination of both. Examples of an oscillation sequence are *A-B-A*, *A-B-A-C-A*, *A-B-B-A-B-A*, and *A-B-B-A-C-A*. Note that an oscillation sequence could consist of repeated oscillation pairs (e.g., *A-B-A-B-A*) or a combination of different ones (*A-B-A-C-A*). Note that the total number of oscillation pairs in an oscillation sequence as well as the split of two types can be easily computed. For instance, the total number of oscillation pairs, the number of O_{p1}, and the number of O_{p2} in *A-B-B-A-C-A* are two, one, and one, respectively.

Two-Stage Clustering and Time Window–Based Methods

The two-stage clustering and time window–based methods are frequently used in existing studies to tackle location uncertainty issues in mobile phone data. For instance, issues such as cellphone load balancing or signal strength variation could cause a user's documented location to switch among adjacent cell towers (Isaacman et al. 2012; Csáji et al. 2013), generating fake movements that complicate human mobility analysis. Some studies also define this issue as oscillation or a ping-pong effect, which describe that a phone's signal could switch between multiple cell towers even though the device is not moving (W. Wu et al. 2014). Given these issues, the two-stage clustering method is primarily used to generalize users' documented locations to derive their representative locations (e.g., stay locations). The time window–based method focuses explicitly on detecting and removing oscillations in the data. In this study, we apply the two-stage clustering algorithm first, followed by the time window–based method to further detect oscillations. In the next section, we introduce an alternative solution to the time window–based method and discuss their trade-offs.

The two-stage clustering algorithm used in previous studies (Alexander et al. 2015; Jiang, Ferreira,



Clustering: Stage 1

Clustering: Stage 2

Time-window Based Method

Figure 3. (A) Two-stage clustering algorithm. In Stage 1, stay segments are detected and annotated with their representative locations (l'_1, l'_2, l'_3). In Stage 2, segments that are close in space but apart in time are further grouped, with their representative locations updated. In this example, l'_1 is used as the representative location to annotate Segment 1 and Segment 3. (B) Illustration of the time window–based method: Depending on the size of the time window and the duration of the "circular events," some are identified as oscillations (green), whereas others are not (red). This reveals a potential limitation of the time window–based method for oscillation detection.

and González 2017; Xu et al. 2018) is first applied to identify stay locations from the trajectories. Note that before this clustering process, we performed a zero step to remove cellphone records with abnormal speed (i.e., $\geq$ 120 km/hr). Given $T = \{(l_1, t_1), (l_2, t_2), \ldots, (l_n, t_n)\}$, in the first stage, we compare each observation with the subsequent one and merge them into a segment if they fall within a roaming distance of Δd_1. We use the medoid of these observations—defined as the most visited cell in that segment—as its representative location. The representative location is then compared with the next observation, which will be merged into the same segment if they (representative location of the segment and the observation) fall within Δd_1. The representative location of the segment will be updated as more observations are added. The clustering process will terminate until all segments are identified. This results in a sequence $\{(l'_1, t'_1, dur_1), (l'_2, t'_2, dur_2), \ldots, (l'_n, t'_n, dur_n)\}$ where l'_i, t'_i, and dur_i denote the medoid, starting time, and the stay duration of the ith segment, respectively (Stage 1 in Figure 3A).

In the second stage, we further group the stay segments that are close in space but apart in time to further generalize a user's activity locations. In particular, we identify the stay segments with representative locations that are within a roaming distance of Δd_2. We compute the medoid of these segments, which is used as the new representative location to annotate them (Stage 2 in Figure 3A).

The clustering-based algorithm can partially tackle the issue of cellphone signal switch, but the oscillations between cells could still persist (e.g., signal switch beyond the set threshold of Δd_1 and Δd_2). Here, we apply a time window–based method proposed in F. Wang and Chen (2018) to further tackle this issue. By imposing a moving window with a fixed length (e.g., 5 minutes), the method aims to detect circular events—subsequences in a trajectory that start and end at the same cell—and identify those within the time window as oscillations (Figure 3B). The underlying assumption is that individuals are less likely to perform a circular trip within a short period of time. Although some studies perform oscillation detection over the raw data, in this study we apply the time window–based method after performing the two-stage clustering algorithm. This is because part of the oscillation issue can be addressed by the clustering step, of which the output—the stay segments annotated by the medoids—could further enhance the effectiveness of oscillation detection in the next step. In other words, we use the representative locations of the observations to perform the oscillation detection.

A New Approach for Oscillation Detection Based on Mean Absolute Deviation

In this section, we propose a new approach for oscillation detection based on the notion of mean absolute deviation (MAD). Given a user's daily trajectory, MAD measures the average deviation of

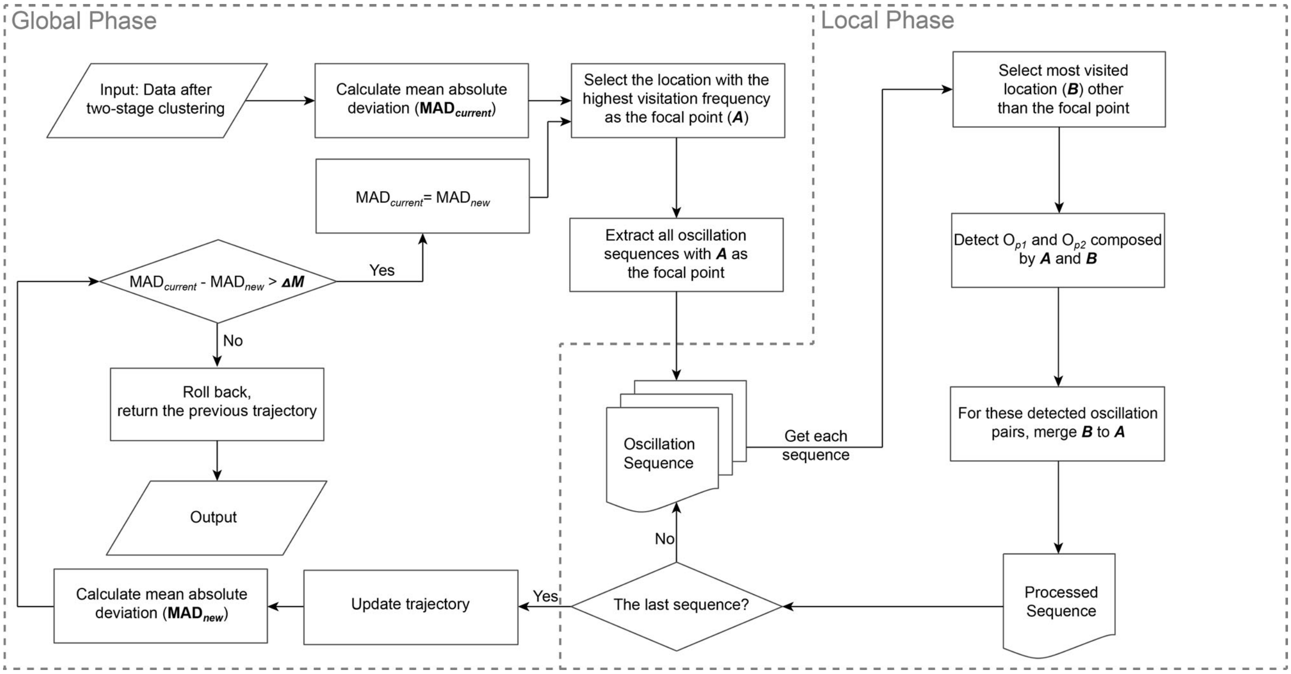

Global Phase
Local Phase
Input: Data after two-stage clustering
Calculate mean absolute deviation ($\mathbf{MAD}_{current}$)
Select the location with the highest visitation frequency as the focal point (***A***)
$\mathrm{MAD}_{current}$= MAD_{new}
$\mathrm{MAD}_{current}$ - MAD_{new} > ΔM
Yes
No
Roll back, return the previous trajectory
Output
Extract all oscillation sequences with ***A*** as the focal point
Oscillation Sequence
Get each sequence
Select most visited location (***B***) other than the focal point
Detect O_{p1} and O_{p2} composed by ***A*** and ***B***
For these detected oscillation pairs, merge ***B*** to ***A***
Processed Sequence
The last sequence?
No
Yes
Update trajectory
Calculate mean absolute deviation ($\mathbf{MAD}_{new}$)

Figure 4. Workflow of the new approach based on MAD. MAD = mean absolute deviation.

each location's visitation frequency from the median of the data set:

$$MAD = \frac{\sum_{i=1}^{n}|x_i - m(X)|}{n}. \tag{8}$$

Here, n denotes the total number of locations traversed by the user daily trajectory and x_i denotes the frequency of visits to the ith location. $m(X)$ refers to the median frequency of all locations.

The idea of the algorithm is simple. Because individuals usually pay few visits to a limited number of locations in a day, a user daily trajectory—if capturing a realistic representation of travel patterns—would have an MAD within a reasonable range. Due to oscillation effect, however, the cellphone signal would switch frequently among a collection of cells. These cells could either reflect a user's true locations or ones that were never visited by the user. The frequency of these cells will be relatively high compared to that of others (i.e., actual locations visited by user but not part of the oscillations). This would result in a suspiciously high value of MAD, which is an indication of likely oscillations. Thus, our approach aims to remove likely oscillations in an iterative manner until the value of MAD converges.

Figure 4 illustrates the workflow of the proposed approach. As mentioned previously, for each trajectory, the output of the two-stage clustering is used as the input. The algorithm contains two phases. The global phase aims to document the MAD of each iteration as well as the difference between two iterations. Meanwhile, it also identifies the focal point of each iteration, which determines where the oscillation sequences can be detected and possibly removed. The local phase aims to remove part of the oscillations in the trajectory. The new value of MAD after removing these oscillations (MAD_{new}) will be reported back to the global phase to determine whether the algorithm terminates.

In the first step, we compute the MAD of the trajectory, denoted as MAD_{current}, to document the initial state. We then identify the location with the highest visitation frequency as the focal point of this iteration. We identify this focal point because it provides an important clue about where the oscillation occurs. For simplicity, we use A to denote this focal point.

Then, the algorithm switches to the local phase by identifying all of the oscillation sequences with A as the focal point. These oscillation sequences, as mentioned previously, could consist of continuous appearance of type 1 or type 2 oscillation pairs or a combination of both. Examples of these oscillation sequences are *A-B-A-B-A*, *A-B-A-B-B-A-B-A*, and *A-B-A-C-A-C-A*, among others. For each of the detected oscillation sequences, our algorithm identifies the most frequently visited location other than the focal point *A*. Then an oscillation removal process is triggered by merging this location to the focal point. For instance, given *A-B-A-B-A*, *B* is identified as the most visited location other than the focal point *A*, and the oscillation sequence after merging the two locations becomes *A-A*. Taking *A-B-A-C-A-C-A* as another example, location *C* will be identified and the sequence after merging *C* to *A* will become *A-B-A-A*. Note that we always merge locations to the focal point *A* because it has the highest visitation frequency, which indicates that it is likely to be an actual activity location visited by the phone user. The local phase continues as all of the detected oscillation sequences are processed.

Note that for each oscillation sequence, the algorithm only merges two locations (i.e., the focal point *A* and the most visited location other than *A*) at a time. In other words, if an oscillation sequence includes more than two distinct locations, different types of oscillation pairs will be tackled at different iterations of the algorithm. For example, given *A-B-A-B-A-C-A-C-A-C-A*, *C* will be identified first and merged to *A*, and if the algorithm does not terminate when going back to the global phase, *A-B-A-B-A* will be detected and processed in the next iteration if *A* is still selected as the focal point.

Once all of the oscillation sequences are processed in the local phase, the algorithm will update the frequency of each location, from which MAD_{new} is computed. Note that when calculating the location frequency, if the same location repeats continuously over time, we only keep the first and last locations to avoid repetitive counting in next iteration (e.g., for a subsequence *D-C-C-C-C-D* in a trajectory, the frequency of *C* will be counted as two instead of four).

In this algorithm, we introduce an important parameter, ΔM, to determine whether the algorithm will terminate. In particular, if $MAD_{\text{current}} - MAD_{\text{new}} > \Delta M$, the algorithm will start a new iteration by searching the new focal point. Otherwise, the algorithm will terminate and the changes (i.e., merge of location in the local phase) made in the current iteration will be rolled back. In other words, the trajectory before conducting the local phase will be returned as the output. The choice of ΔM controls the strictness of the oscillation removal, which affects the result of the output trajectory. A small ΔM allows the algorithm to continue even when a small change in MAD is identified. A large value, however, will only remove oscillations with a high frequency.

Figure 5 shows a simple example of how the time window–based method and MAD approach could achieve different outcomes. Here we select a user trajectory (Figure 5A) and then perform the outlier removal (Figure 5B) and the two-stage clustering algorithm (Figure 5C). Then, we apply the time window–based method (window size: 5 minutes) and the MAD approach ($\Delta M = 0.5$) and compare their output. As shown in Figure 5D and Figure 5E, the time window–based method is simply a downsampling process, which only removes oscillation pairs within the 5-minute time window. It can be seen that many oscillation pairs still persist in the output. The MAD approach, in this example, tends to remove highly frequented oscillations (Figure 5F), thus achieving a more satisfactory result (Figure 5G).

We then introduce a few indicators to systematically compare the two methods. Given a particular method used, we first introduce R_{op}^{S}, to measure the total number of oscillation pairs (including both type 1 and type 2) that were removed in an oscillation sequence *S*:

$$R_{op}^{S} = \frac{\text{total number of oscillation pairs removed from } S}{\text{total number of oscillation pairs in } S}. \tag{9}$$

We also can distinguish the two types of oscillation pairs and quantify their detection ratios, respectively:

$$R_{op1}^{S} = \frac{\text{total number of } Op_1 \text{ removed from } S}{\text{total number of } Op_1 \text{ in } S} \tag{10}$$

$$R_{op2}^{S} = \frac{\text{total number of } Op_2 \text{ removed from } S}{\text{total number of } Op_2 \text{ in } S}. \tag{11}$$

The preceding three indicators measure the detection ratios from the perspective of oscillation sequence. Similarly, we can measure the detection ratios from the perspective of user daily trajectory *T*:

$$R_{op}^{T} = \frac{\text{total number of oscillation pairs removed from } T}{\text{total number of oscillation pairs in } T} \tag{12}$$

$$R_{op1}^{T} = \frac{\text{total number of } Op_1 \text{ removed from } T}{\text{total number of } Op_1 \text{ in } T} \tag{13}$$

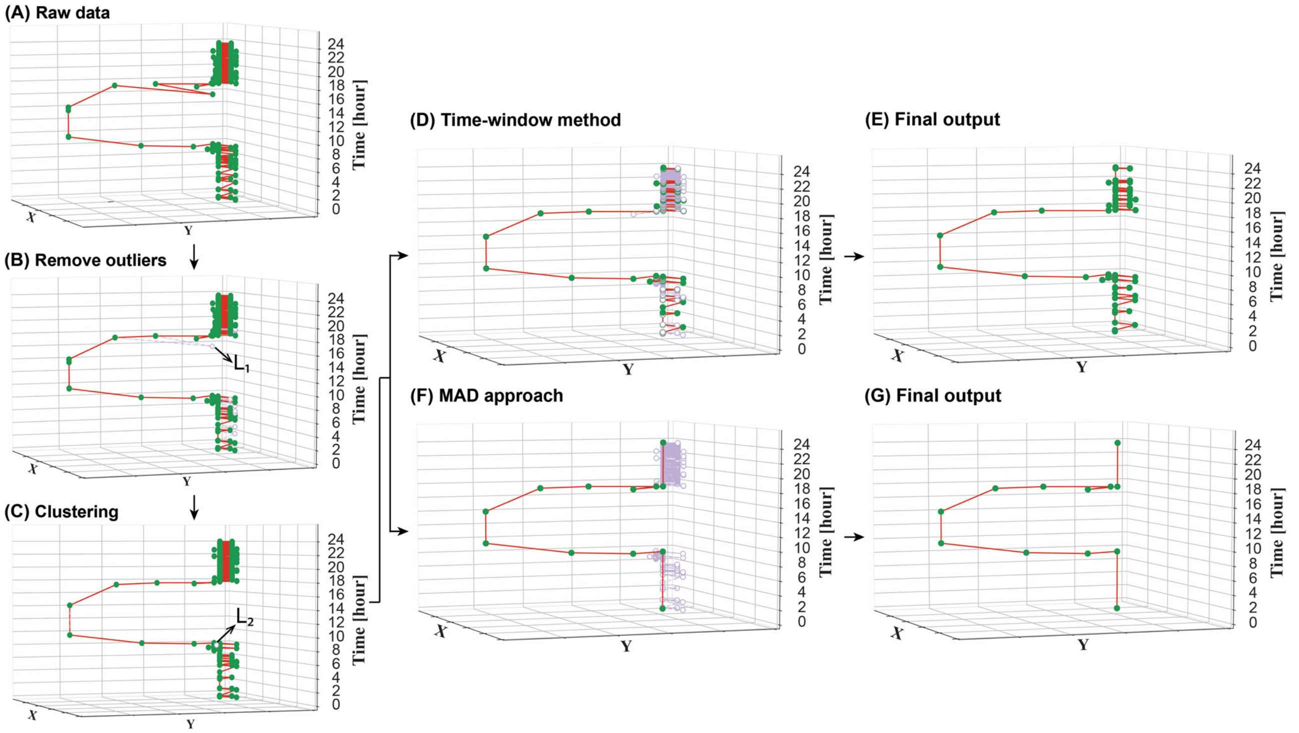


Figure 5. An example comparing the time window–based method and the new approach based on MAD: (A) raw data of a user daily trajectory; (B) L_1 is identified with abnormal speed and thus removed in the zero step; (C) the two-stage clustering merges some of the locations (e.g., L_2 is grouped with nearby cells to form a new representative location); (D–E) the time window–based method (window size: 5 minutes) only removes oscillations within the 5-minute time window; (F–G) the MAD approach tends to identify highly frequented oscillations and achieve more reasonable outcome. MAD = mean absolute deviation.

$$R^T_{op2} = \frac{\text{total number of } Op_2 \text{ removed from } T}{\text{total number of } Op_2 \text{ in } T}. \tag{14}$$

In the next section, we report the detection ratios of the two methods (time window–based method and MAD) and discuss the impact of parameter choice.

Analysis Results

Impact of Two-Stage Clustering on Data Characteristics

We first apply the two-stage clustering algorithm and evaluate its impact on data characteristics. Given that the average spacing gap between cells is roughly 150 m (Figure 1C), we set both Δd_1 and Δd_2 at 200 m and monitor the changes in oscillation sequences and oscillation pairs as the algorithm is applied. According to the results, the average number of oscillation sequence (O_s) in a trajectory is reduced from 3.68 to 2.10 (Figure 6A and Figure 6E), suggesting that the algorithm addresses part of the oscillation effect even before other dedicated methods are applied. By computing the total number of oscillation pairs in a trajectory, we find a decrease in the overall mean from 7.64 to 4.62 (Figure 6B and Figure 6F).

By further splitting the two types of oscillation pairs, we find that the algorithm has a notable impact on removing type 1 oscillation pairs (Figure 6C and Figure 6G). The average number of O_{p1} per trajectory changes from 6.95 to 3.50, but the average number of type 2 oscillation pairs (O_{p2}) increases from 0.69 to 1.13 (Figure 6D and Figure 6H), indicating that the clustering algorithm produces new instances of oscillations.

Figure 7 illustrates the number of unique locations in a trajectory and the interevent time of the two types of oscillation pairs after performing the two-stage clustering algorithm. The mean and median numbers of unique locations in a trajectory are 10.36 and 7.00, respectively (Figure 7A). The interevent time of O_{p1}, measured as the elapsed time between two A in *A-B-A*, varies notably from each other (Figure 7B). A substantial amount of type 1 oscillation pairs have an interevent time greater than 5 minutes. This reveals a notable limitation of the time window–based method that many of these oscillation pairs will be ignored given an arbitrarily selected window size (e.g.,

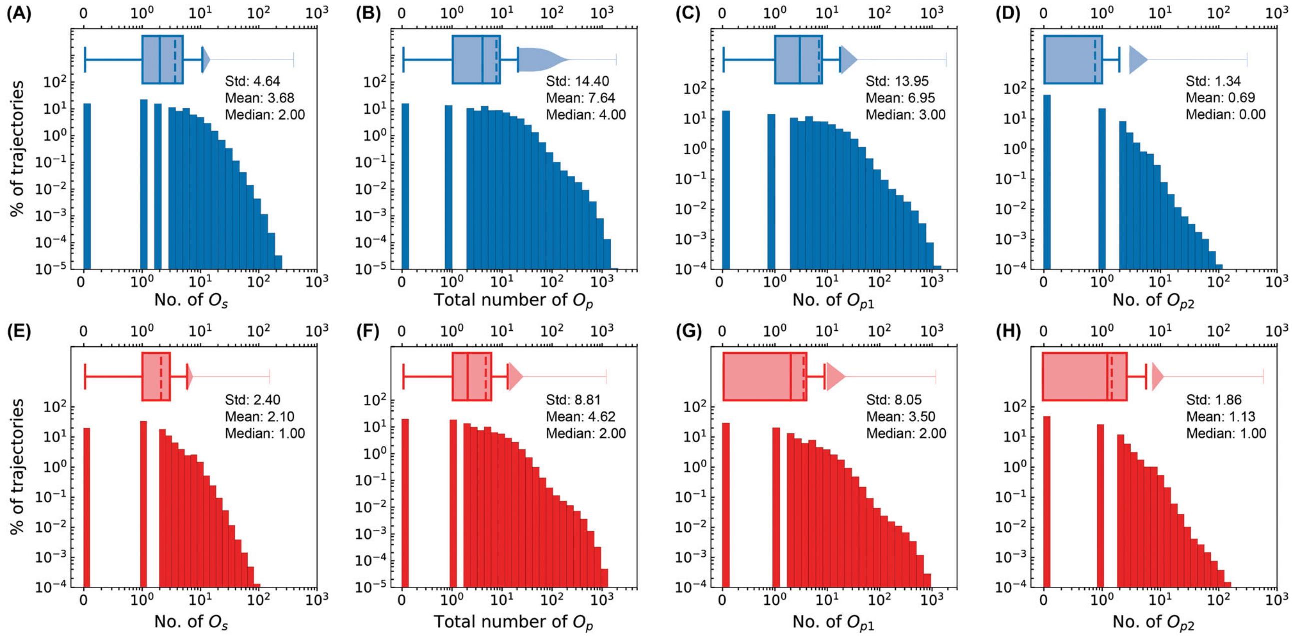


Figure 6. Distributions of the number of oscillation sequence, total number of oscillation pairs, number of type 1 and type 2 oscillation pairs in a trajectory (A–D) before and (E–H) after the two-stage clustering algorithm is performed.

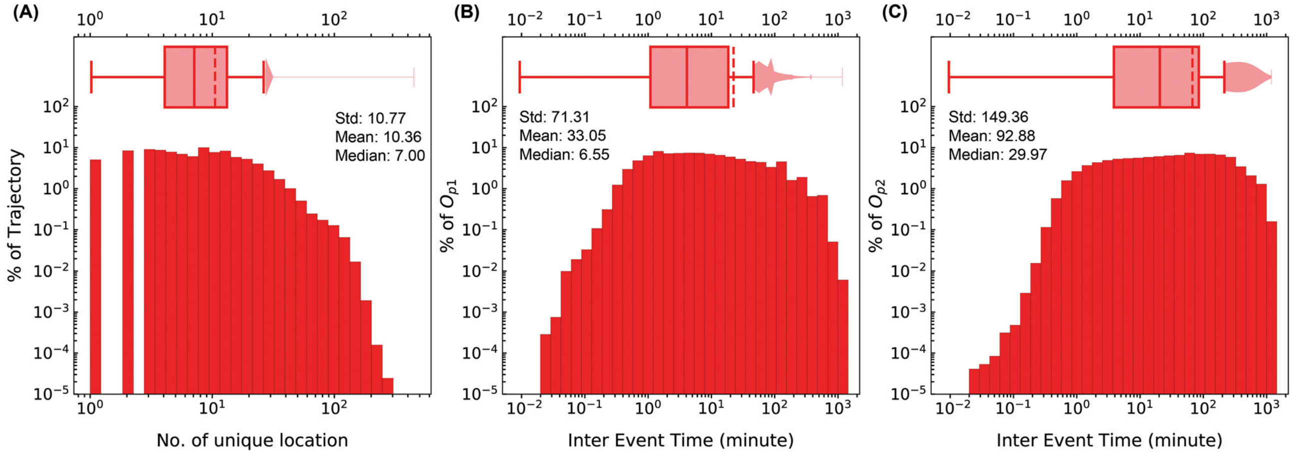


Figure 7. Distribution of (A) number of unique locations in a user trajectory, (B) interevent time of type 1 oscillation pairs, and (C) interevent time of type 2 oscillation pairs based on the output of two-stage clustering algorithm.

5 minutes). This issue persists when type 2 oscillation pairs are handled. A large variation in interevent time (the elapsed time between two *A*s in *A*-*B*-*B*-*A*) makes it extremely difficult to justify the choice of window size (Figure 7C).

Two Oscillation Detection: Time Window–Based Method versus MAD Approach

We compare the two methods using the indicators proposed earlier. For this analysis, we only consider user daily trajectories with at least one oscillation pair after performing the two-stage clustering algorithm. Figure 8 demonstrates the detection ratios of the two methods from the perspective of oscillation sequence, using $\Delta M = 0.5$ and a window size of 5 minutes as an example. By extracting all of the oscillation sequences in the trajectories, for each oscillation sequence S we compute R_{op}^{S}, R_{op1}^{S}, and R_{op2}^{S} for the two methods. This allows us to investigate not only the average detection ratio of each method ($\overline{R_{op}^{S}}, \overline{R_{op1}^{S}}$, and $\overline{R_{op2}^{S}}$) but also the difference

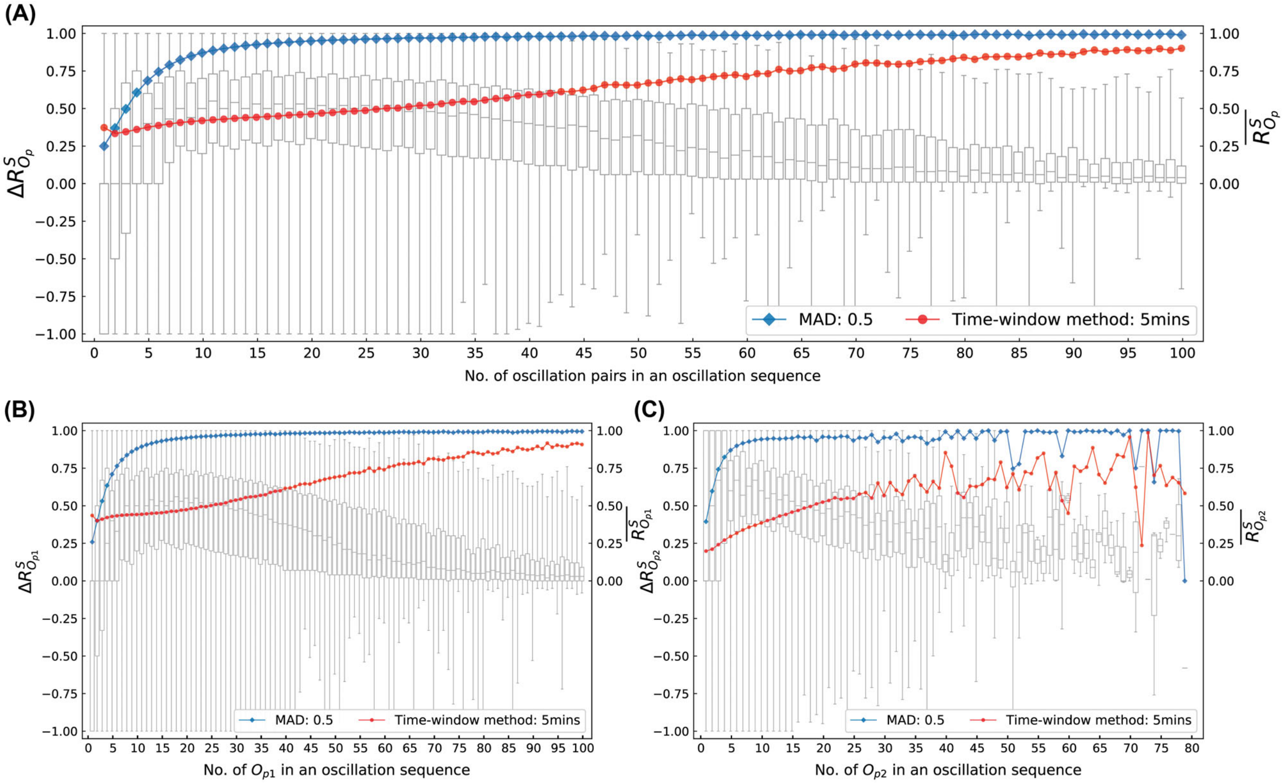


Figure 8. Detection ratio of oscillation pairs in an oscillation sequence when combining or splitting type 1 and type 2 oscillation pairs. Horizontal axes denote the number of oscillation pairs in an oscillation sequence: (A) combined; (B) type 1 oscillation pairs (O_{p1}); (C) type 2 oscillation pairs (O_{p2}). MAD = mean absolute deviation.

between the two ($\Delta R^S_{op}, \Delta R^S_{op1}$, and ΔR^S_{op2} that are measured for each S). When reporting these indicators, we also distinguish the total number of oscillation pairs in S to get a better sense of its impact.

As shown in Figure 8A, the average detection ratio of the MAD approach (diamond) tends to be higher than that of the time window–based method (circle), even when the number of oscillation pairs in S is controlled. This finding is further illustrated by the box plot showing the distribution of ΔR^S_{op}. Interestingly, we find that the detection ratio of both methods increases as the oscillation sequence becomes longer. Such an increase for the time window–based method is due to the decreased interevent time of oscillation pairs as S becomes "denser." For the MAD approach, the average detection ratio increases much faster and quickly converges to nearly 100 percent, suggesting its better performance, especially for tackling long oscillation sequences. We also find a relatively consistent difference between the two methods when splitting type 1 (Figure 8B) and type 2 (Figure 8C) oscillation pairs.

We next examine how well the two methods handle user daily trajectories. Again, the MAD approach outperforms the time window–based method by achieving a higher detection ratio, whether the two types of oscillations are combined (Figure 9A) or not (Figure 9B and Figure 9C). Note that the difference between the two methods, especially when handling O_{p1}, is smaller when a trajectory contains few or many oscillation pairs (Figure 9B). To elaborate, both methods achieve lower detection ratios when oscillations are sparse but tend to perform well when there are many oscillation pairs in a trajectory.

When computing the MAD (Equation 8), an important parameter is the number of unique locations in a trajectory (n). Here, we further evaluate the relationship between n and the performance of the two methods. As can be seen in Figure 10A, the average detection ratio of the MAD approach (diamond), $\overline{R^T_{op}}$, tends to decrease as n increases. An opposite trend is observed, however, for the time window–based method (circle). Similar patterns are observed when splitting type 1 (Figure 10B) and type 2 (Figure 10C) oscillation pairs despite the fact that the two curves (circle vs. diamond) cross each other at different values of n.

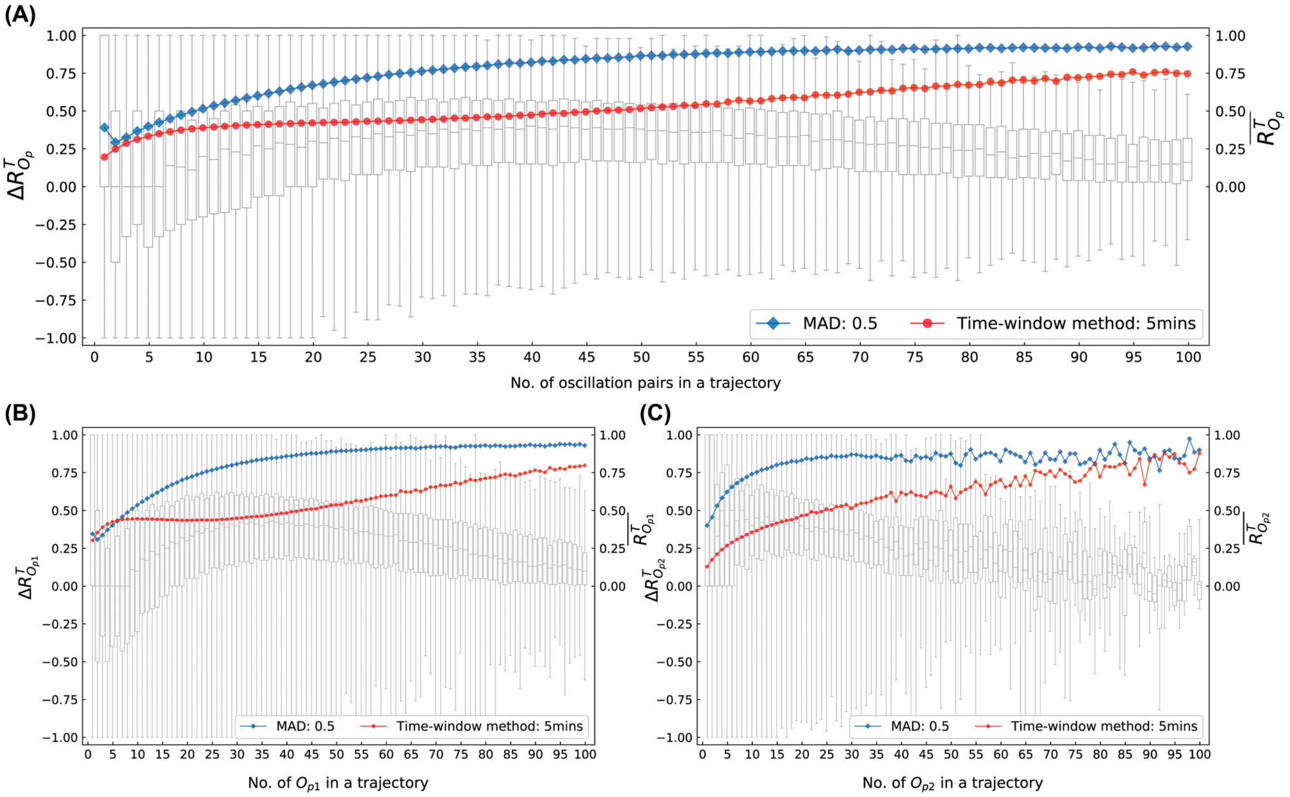


Figure 9. Detection ratio of oscillation pairs in a user daily trajectory when combining or splitting type 1 and type 2 oscillation pairs. Horizontal axes denote the number of oscillation pairs in a user daily trajectory: (A) combined; (B) type 1 oscillation pairs (O_{p1}); (C) type 2 oscillation pairs (O_{p2}). MAD = mean absolute deviation.

The low detection ratio of the MAD approach, when n is large, is partially affected by the relationship between ΔM and n. In this example, ΔM is chosen as 0.5. Because the MAD approach addresses oscillations in an iterative manner, when n is large it requires the detected oscillations (around the focal point A) to appear highly frequently to pass the current iteration; that is, to produce substantial changes to $\sum_{i=1}^{n}|x_i - m(X)|$. In other words, the MAD approach will only remove the oscillations with a high frequency when the trajectories traverse many distinct locations. This suggests that an adaptive choice of ΔM can possibly improve the MAD approach. In particular, as n becomes larger, the value of ΔM can be lowered to allow more oscillation occurrences to be removed. Evaluating this alternative is a possible direction for future research.

Next, we evaluate the impact of parameter choice on detection ratio. For the time window–based method, increasing the window size will remove more oscillation pairs in trajectories, thus increasing the average detection ratio $\overline{R^T_{op}}$. For the MAD approach, a small threshold ΔM will make the algorithm "tolerant," allowing less frequented oscillations to be removed. Here, we choose six different thresholds, $\Delta M = 0.01, \Delta M = 0.1, \Delta M = 0.25, \Delta M = 0.5, \Delta M = 1.0$, and $\Delta M = 5.0$, and compare them with the time window–based method with two parameter settings; that is, the 5-minute and 10-minute window sizes.

As shown in Figure 11, the MAD approach with the first four parameter settings ($\Delta M = 0.01, \Delta M = 0.1, \Delta M = 0.25, \Delta M = 0.5$) tends to outperform the time window–based method with the 5-minute threshold (red curve). When ΔM is set to higher values, however, such as 1.0 and 5.0, the oscillation removal process becomes more restrictive, thus achieving lower detection ratios. The result reveals both the advantage and limitation of the MAD approach. On the one hand, the time window–based method tends to achieve compatible or even higher detection ratios when the number of oscillation pairs in a trajectory is small (e.g., less than ten). This indicates that without carefully calibrating the value of ΔM, the MAD approach might produce unsatisfactory output when oscillations are sparse in a

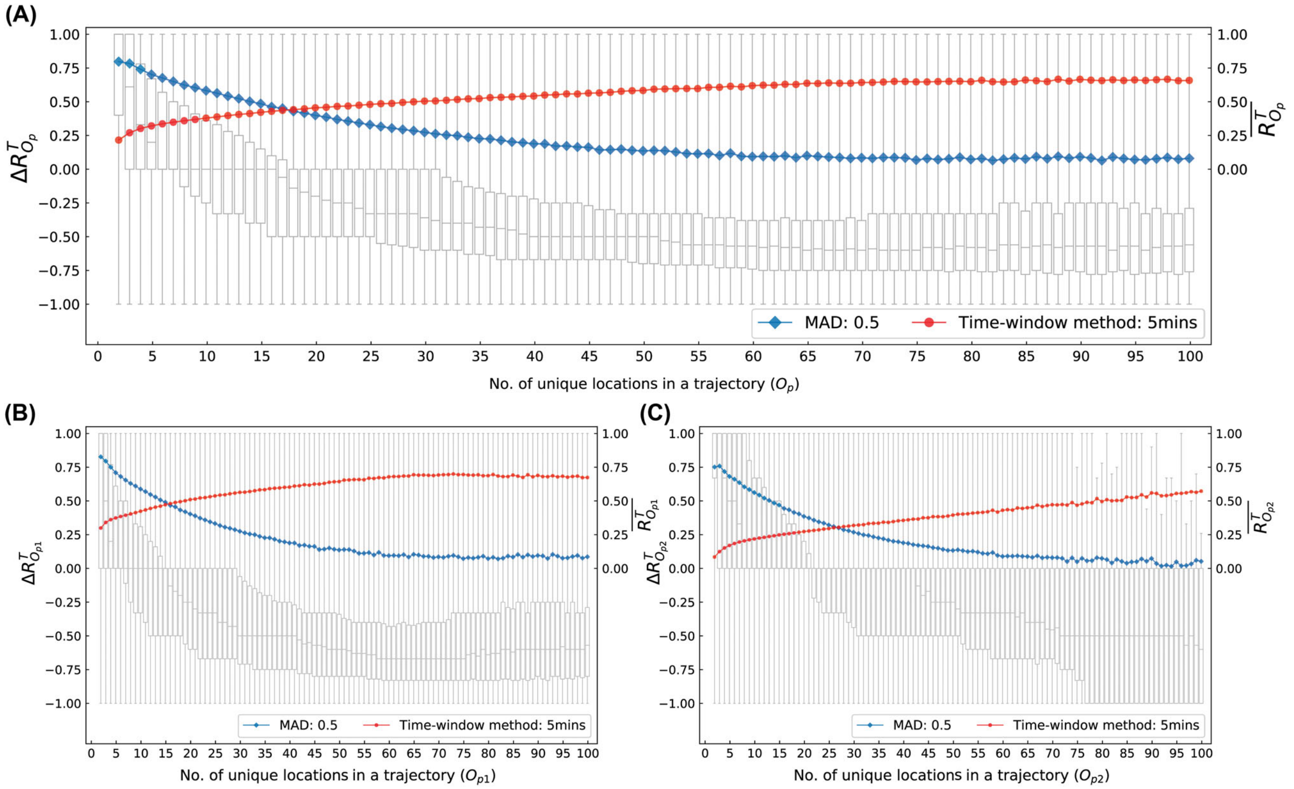


Figure 10. Detection ratio of oscillation pairs in a user daily trajectory when combining or splitting type 1 and type 2 oscillation pairs. Different from Figure 9, horizontal axes denote total number of unique locations in a trajectory. MAD = mean absolute deviation.

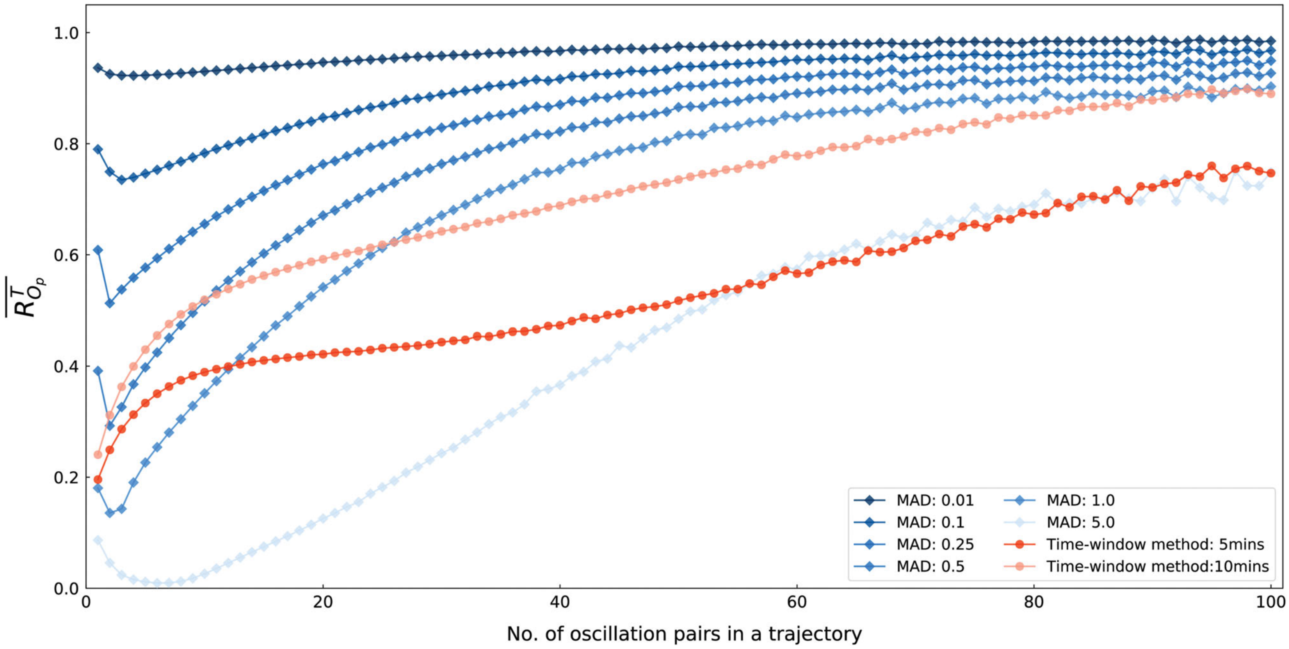


Figure 11. Average detection ratio of trajectories ($\bar{R^T_{op}}$) of the two methods under different parameter settings. Horizontal axis denotes total number of oscillation pairs in a user daily trajectory. MAD = mean absolute deviation.

trajectory. One the other hand, the MAD approach shows a clear advantage over the time window–based method when many oscillation pairs are presented in a trajectory. Among these oscillation pairs, many are not removed by the time window–based method simply because their interevent time is greater than the window size. These highly frequented oscillations are detected and removed properly by the MAD approach, however. Note that increasing the window size (e.g., to 10 minutes; see reddish line in Figure 11), as expected, will improve the overall detection ratio. The time window–based method, however, ignores the inherent structures in oscillation patterns, thus failing to remove part of the highly frequented oscillations.

Impact of Preprocessing Methods on Mobility Estimation

In this section, we further investigate the impact of the three methods on individual mobility estimations. For each user's daily trajectory, we collect the outputs generated from them; that is, two-stage clustering algorithm, time window–based method (window size = 5 minutes), and the MAD approach ($\Delta M = 0.5$). We then derive the following four mobility indicators from these outputs and evaluate their differences:

1. Number of OD trips: Given a user's daily trajectory, we first derive all of the stay activities with a duration above a threshold (e.g., 10 minutes). Then, OD trips are derived from consecutive stays. Note that if two consecutive stays for an individual correspond to the same location, we do not count the "movement" in between as a trip in this analysis. We compare the number of OD trips derived from three different methods, denoted as $OD_{\text{clustering}}$, OD_{timewin}, and OD_{mad}, respectively. We report the comparison results based on two different thresholds of stay duration: 10 minutes and 30 minutes. We use 10 minutes because the threshold was adopted in many travel behavior studies to detect travelers' meaningful stays (Alexander et al. 2015; Jiang, Ferreira, and González 2017; Xu et al. 2018). The other threshold (30 minutes) is used to evaluate whether the differences between the three methods remain consistent. The choice of the threshold can be adjusted based on specific study or application purposes.
2. Number of activity locations is simply defined as the total number of unique locations derived from a user's stay activities (above the 10-minute or 30-minute threshold). A large value indicates that the user's daily activities tend to be distributed across a variety of activity locations. We compare this number across the three methods, denoted as $A_{\text{clustering}}$, A_{timewin}, and A_{mad}, respectively.
3. Total stay time is defined as the total amount of time that a user stays across all of the activity locations; that is, $\sum duration(l_i)$. Note that when calculating this indicator, the threshold of stay time (e.g., 10 minutes or 30 minutes) is not imposed. The total stay times derived from the three methods are denoted as $S_{\text{clustering}}$, S_{timewin}, and S_{mad}, respectively.
4. Activity entropy is introduced to quantify the diversity of a user's daily activities. Given the total stay time extracted at each location l_i, we can measure the proportion of stay as $p_i = \frac{duration(l_i)}{\sum duration(l_i)}$. The activity entropy is then calculated as $H = -\sum p_i * \log(p_i)$. We use $H_{\text{clustering}}$, H_{timewin}, and H_{mad} to denote this indicator derived from the three methods.

Figure 12 reports the comparison results of OD estimation using stay duration of 10 minutes as the threshold. As shown in Figure 12A, the two-stage clustering algorithm yields a mean and median of 2.30 and 2.00, respectively. Using the time window–based method, as shown in Figure 12B, results in a slight increase in the mean value. By further measuring their difference at the level of individual trajectory (Figure 12D), we find that the two methods produce the same number of OD trips (i.e., $OD_{\text{timewin}} - OD_{\text{clustering}} = 0$) for 96.6 percent of the trajectories, whereas for the rest the time window–based method always gives a higher estimation.

Compared to these two methods, the MAD approach produces a mean of 1.97 (Figure 12C). The comparison between MAD and two-stage clustering (Figure 12E) shows that both methods give the same estimation result for 75.5 percent of the trajectories. For the remaining trajectories, however, the MAD approach gives higher estimations for 4.5 percent of the trajectories but lower estimations for 20.0 percent of the cases. A similar conclusion can be reached by comparing the MAD approach and time window–based method. The two methods give the same estimation result for 74.2 percent of the trajectories. For the remaining trajectories, the MAD approach gives higher estimations for 3.9 percent of the trajectories but lower estimations for 21.9 percent of the cases. Note that we also compare the three methods using 30 minutes as the threshold, and similar findings are observed (see Figure A.1 in the Appendix). In sum, if viewing the result of two-stage clustering as the baseline, the MAD approach tends to produce lower estimations of OD trips,

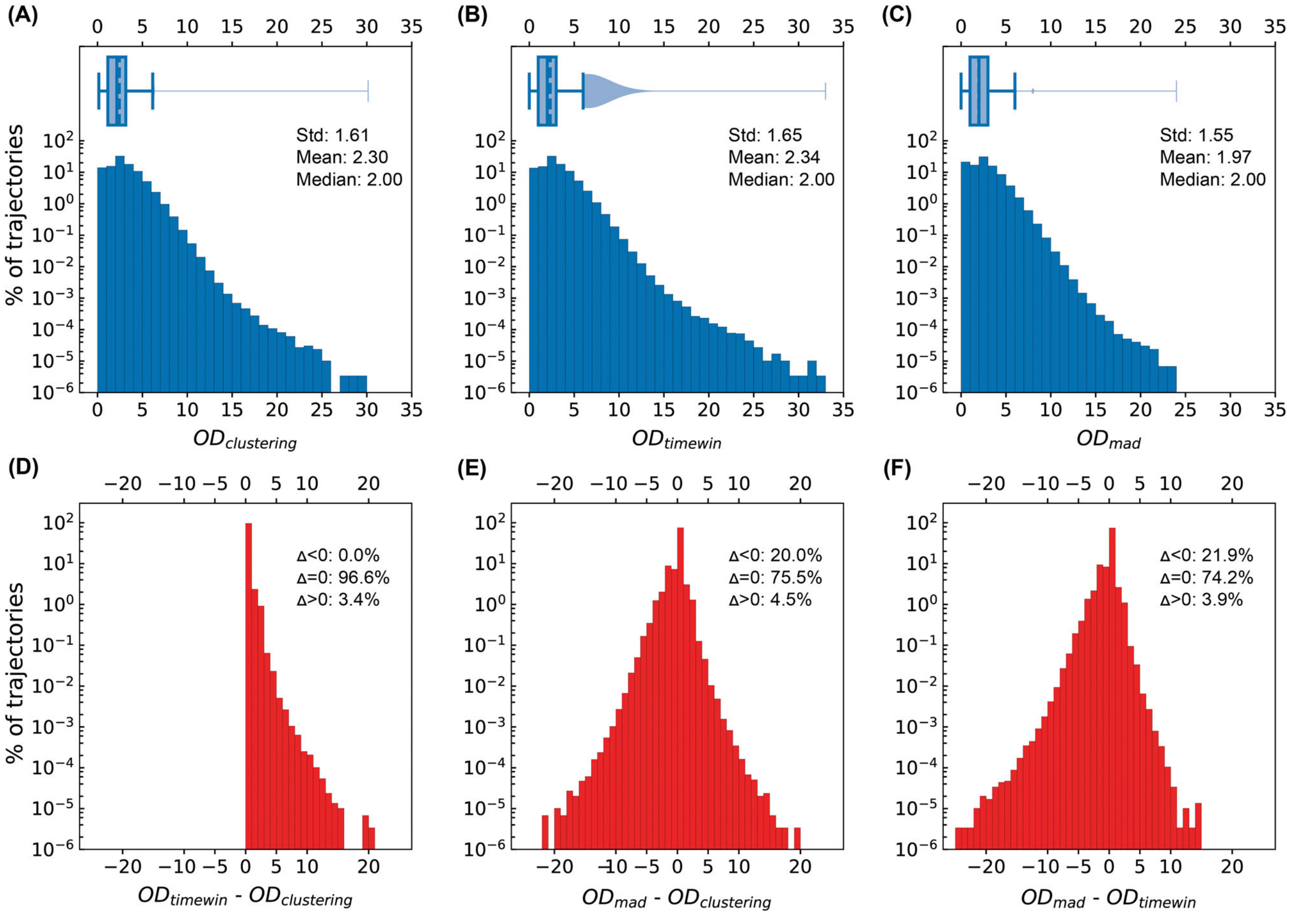


Figure 12. (A–C) Distribution of $OD_{\mathrm{clustering}}$, OD_{timewin}, and OD_{mad} and (D–F) pair-wise comparison of the three methods. Origin–destination trips are generated based on stay duration of 10 minutes.

whereas the time window–based method has a negligible impact. This is because for the MAD approach, certain cells are merged toward the focal point during oscillation removal. In other words, the movements that are part of these likely oscillations are not considered valid OD trips. The result suggests that the way the oscillations are tackled in the mobile signaling data could have a notable impact on the estimation of OD trips.

When estimating the number of activity locations (Figure 13), the three methods output means of 2.51, 2.53, and 2.26, respectively. Compared to the baseline derived from the two-stage clustering algorithm, the MAD approach produces lower estimations for 18.6 percent of the trajectories and higher estimations for only 3.0 percent of the cases (Figure 13E). The results of the time window–based method and two-stage algorithm closely resemble each other, however (Figure 13D). Again, the implication here is that performing the MAD approach will have a more obvious impact on this mobility indicator than performing the time window–based method, given that the latter is more or less a downsampling process of mobile phone trajectories. (Readers can refer to Figure A.2 in the Appendix for comparative results based on stay duration of 30 minutes.)

Regarding the total stay time, the output of two-stage clustering produces mean and median values of 831.05 and 835.65 minutes, respectively (Figure 14A). The time window–based method produces slightly larger values (Figure 14B). The MAD approach, however, gives much higher estimations (Figure 14C). The result indicates that by ignoring structural properties of oscillations in a trajectory, the estimation of total stay time can be off by several hours or even longer (Figure 14E and Figure 14F).

The estimation of stay time will also affect the characterization of activity diversity (Figure 15). Because the MAD approach is able to detect highly frequented oscillations, when these oscillations are removed or, more precisely speaking, merged toward the focal points, the observed stay time at these

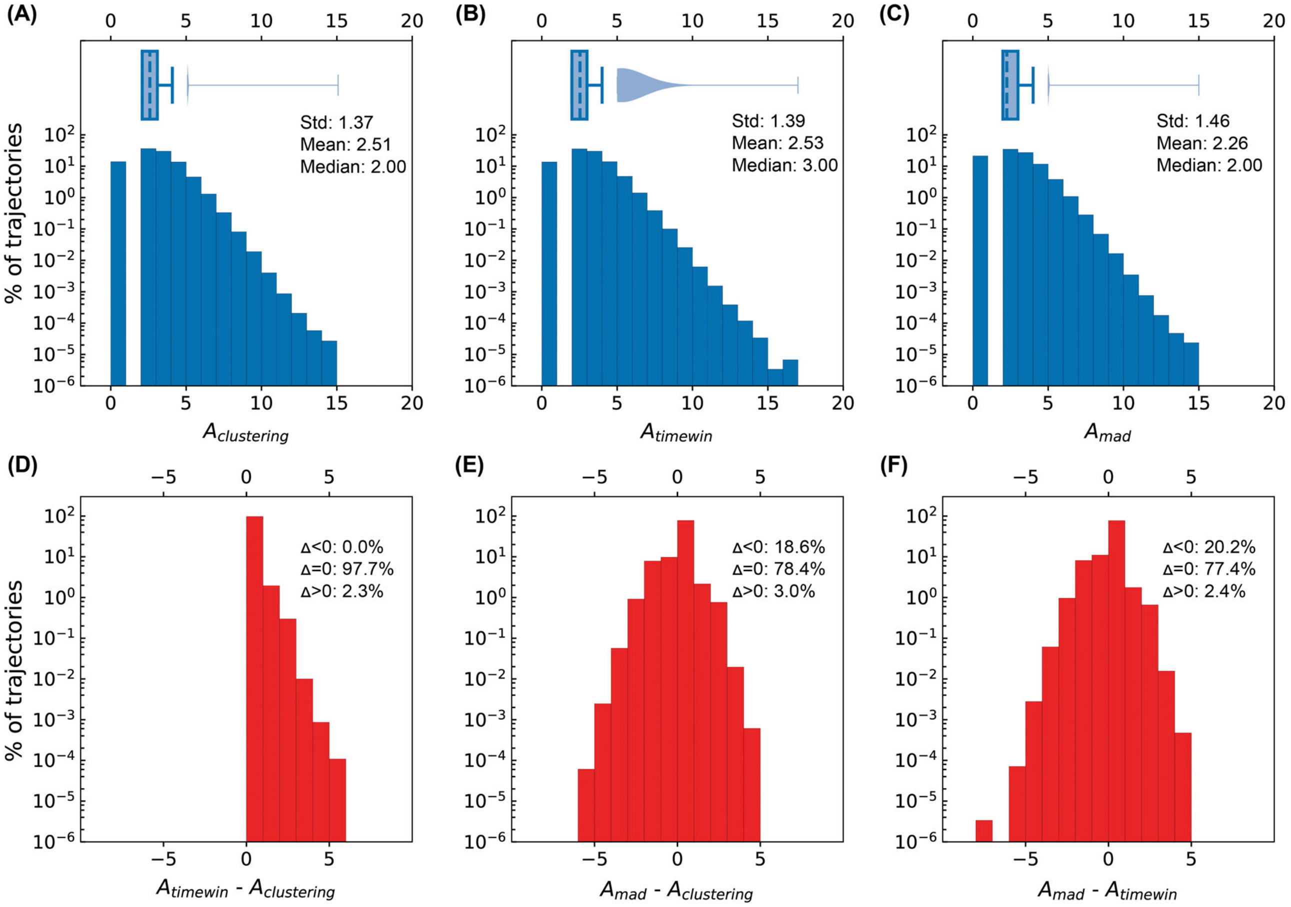


Figure 13. (A–C) Distribution of $A_{\text{clustering}}$, A_{timewin}, and A_{mad} and (D–F) pair-wise comparison of the three methods. Results are generated based on stay duration of 10 minutes.

focal points will increase. These focal points, which can be meaningful activity locations of individuals (e.g., home cell or work cell), have a notable impact on the estimation of activity entropy. As a result, as shown in Figure 15, the MAD approach produces lower entropy values, whereas the distributions of the other two methods are relatively more similar.

Discussion and Conclusion

Data veracity is an important but often neglected issue in big data analytics. This issue has been and will always be challenging to the validity of research that involves the use of big data. Without paying attention to this issue, the knowledge generated from the data, as claimed by Kwan (2016), will risk becoming an artifact of the algorithms used. In this study, we aim to reflect on this issue through the analysis of a large-scale mobile phone data set. Our results demonstrate that the choice of data preprocessing methods could lead to changes in the data characteristics. Such changes, which are nontrivial, will further affect the characterization of human mobility patterns.

By applying a two-stage clustering algorithm over the MSD, we highlight the effectiveness of this step in tackling the location uncertainty issues. Meanwhile, we find that some issues, primarily the cell tower oscillation effect, cannot be addressed completely. An example shown in Figure 5 clearly reveals this effect along with other issues (e.g., outliers with abnormal speed). The presence of these issues is likely to cause a deviation in users' documented locations from their true locations. Although we are not able to measure this deviation due to the absence of ground truth, we find that the three preprocessing methods could generate different outputs (i.e., mobile phone trajectories after preprocessing), which affect how human mobility patterns are further analyzed and interpreted.

By further applying the time window–based method, an existing practice for handling

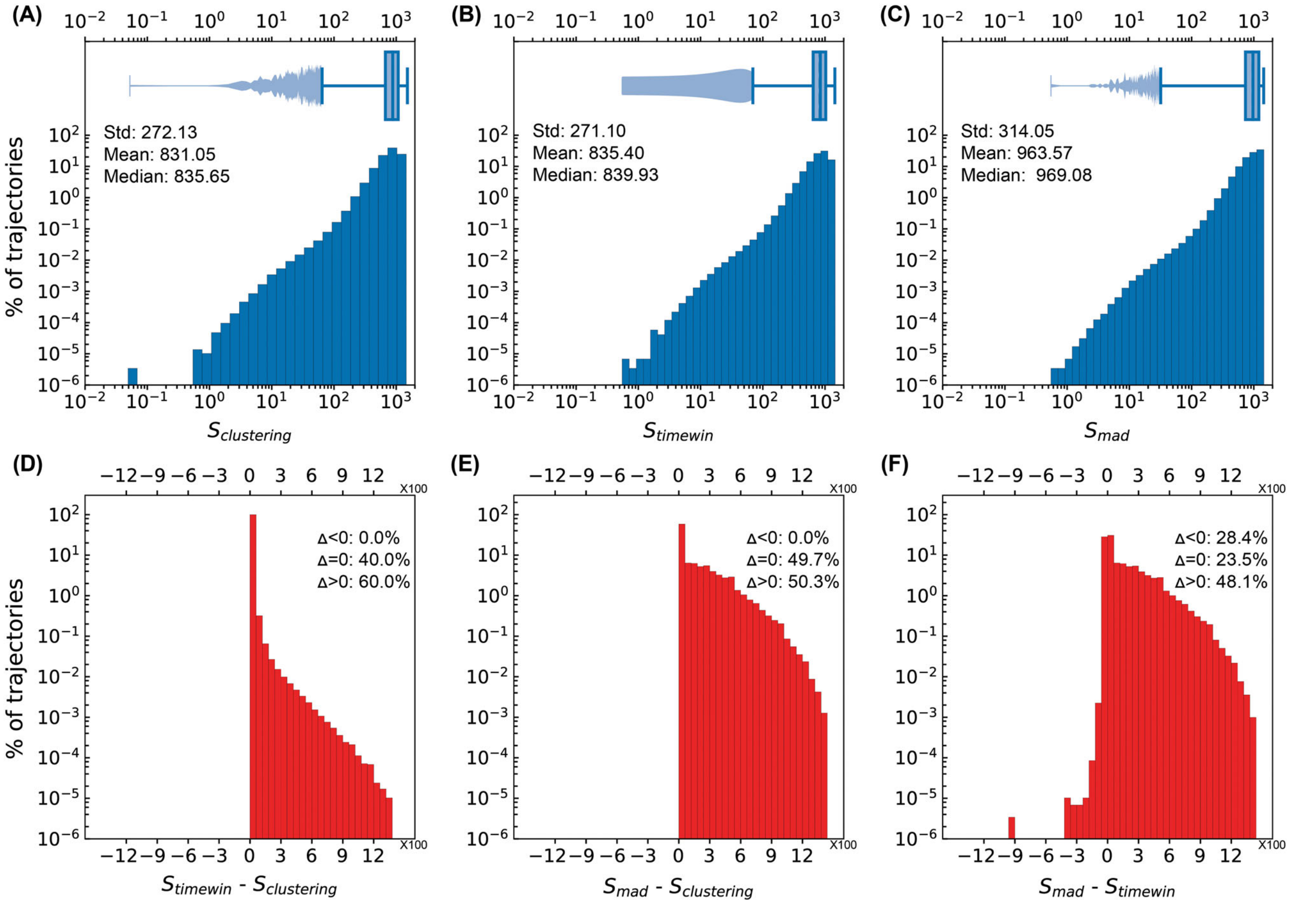


Figure 14. (A–C) Distribution of $S_{\text{clustering}}$, S_{timewin}, and S_{mad} and (D–F) pair-wise comparison of the three methods.

oscillations, we find that the oscillation issue is partially addressed. Despite the fact that some oscillation pairs are detected within the window size (e.g., 5 minutes) and then removed, there are many more that are not filtered simply because their interevent time is larger than the window size. This makes the time window–based method problematic. Because the interevent times of oscillations vary notably from each other (Figure 7), the time window–based method, depending on the choice of window size, becomes a downsampling process that removes oscillations in a somewhat random way.

We then propose an approach based on the notion of MAD to improve the oscillation detection. The MAD approach conducts the removal process by locating oscillation occurrences that appear most frequently in the trajectory and repeats this process until the value of MAD converges. The key advantage of the MAD approach is its ability to capture the frequency distribution of oscillations, from which the most suspicious ones are removed first. By comparing the MAD approach with the time window–based method through the six proposed indicators, we find that the MAD approach tends to achieve higher detection ratios, especially when there are many oscillation pairs in a trajectory. The comparison also reveals the limitation of both methods when oscillations are sparse in the data. The results shown in Figure 11 suggest that when ΔM is set to 0.5, the MAD approach tends to achieve more satisfactory results than the time window–based method. The optimal value or range of ΔM, however, should be further evaluated when ground truth data of human movements are available. We believe that the choice of ΔM is jointly affected by the spatial distribution of cell towers in the study area as well as the characteristics of MSD. Testing the proposed approach across different data sets and study areas is a meaningful task for future research.

To better understand geographic patterns of the detected oscillations, we perform an additional analysis here by counting the number of occurrences for which each cell tower is associated with oscillations (using $\Delta M = 0.5$). We summarize such information

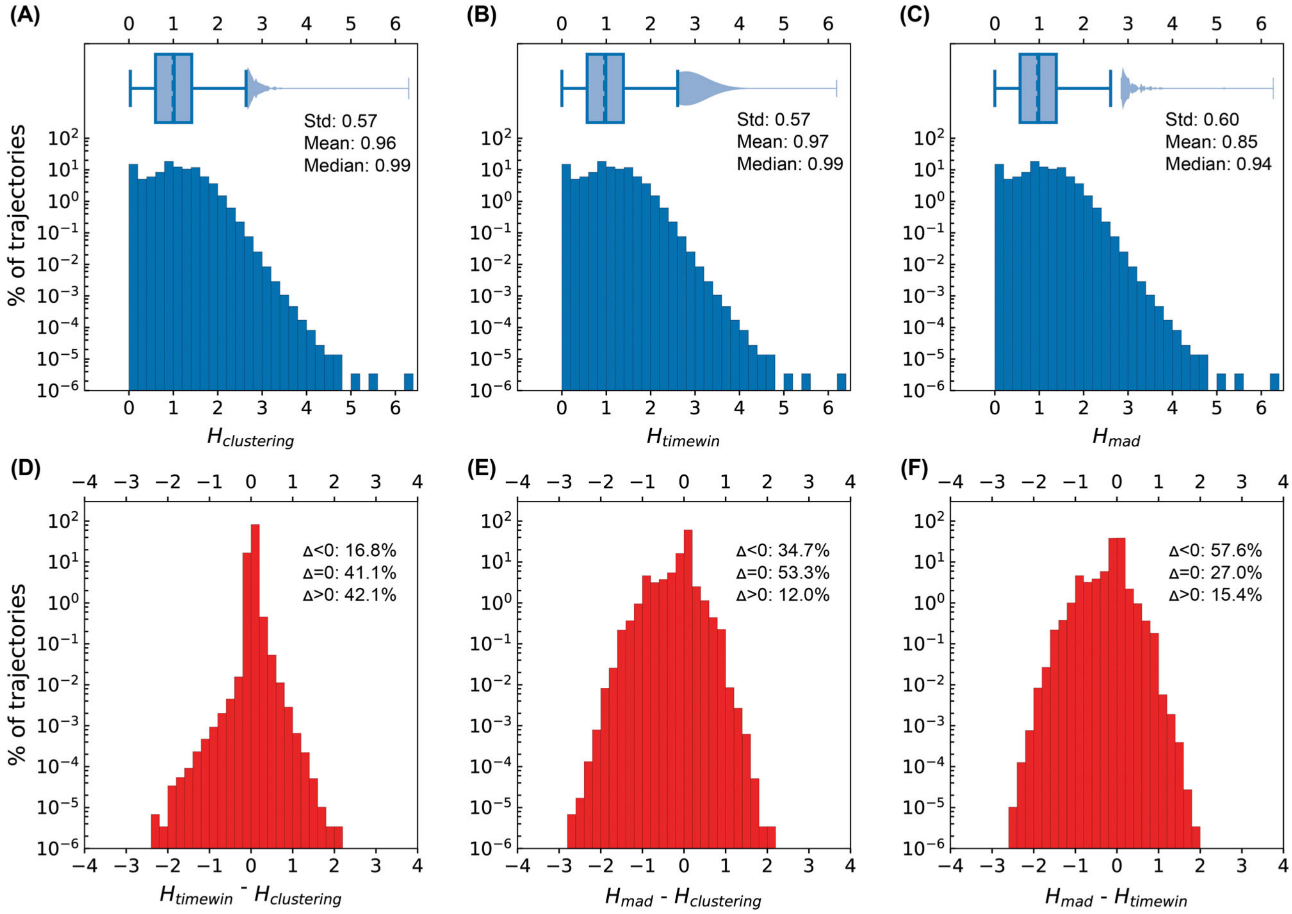


Figure 15. (A–C) Distribution of $H_{\text{clustering}}$, H_{timewin}, and H_{mad} and (D–F) pair-wise comparison of the three methods.

at the level of a 1 km * 1 km grid. As shown in Figure 16, the oscillations are observed more frequently in the core part of Shanghai, which is also the area where cell towers are densely distributed (Figure 2). These areas generally correspond to places that are frequently used by phone users (i.e., densely populated areas). Thus, the results in Figure 2 and Figure 16 suggest that cell tower oscillations tend to be more pronounced in densely populated areas. Ignoring the uncertainty issues in the mobile phone data will have a larger impact on these areas, where decision making on urban design and management is frequently needed (e.g., infrastructure investment, disease control, transport planning).

To evaluate the impact of these methods on mobility estimations, four individual mobility indicators—namely, number of OD trips, number of activity locations, total stay time, and activity entropy—are introduced. These indicators are derived from the outputs of the three different methods (two-stage clustering, time window–based method, and MAD) and then compared. Two findings are worth noting. First, the two-stage clustering algorithm and time window–based method result in similar distributions for all four mobility indicators. This suggests that although the time window–based method removes a substantial amount of oscillations in trajectories, the impact on mobility characterization is small or even trivial. This is largely due to the downsampling nature of time window–based method when it is used to detect oscillations. Second, using the MAD approach causes notable changes to the four indicators. Compared to the other two methods, the MAD approach tends to produce lower estimations of OD trips, activity locations, and activity entropy but higher values for total stay time. The comparison suggests that certain methods of handling oscillations in the mobile phone data could result in unreliable estimates of individual mobility characteristics. These uncertainties can propagate when the processed results are further used in human mobility analysis (e.g., OD estimation, dwelling time estimation, inferring individual activity purposes).

The implications are manifold. First, the varying impacts of the three methods on mobility estimations

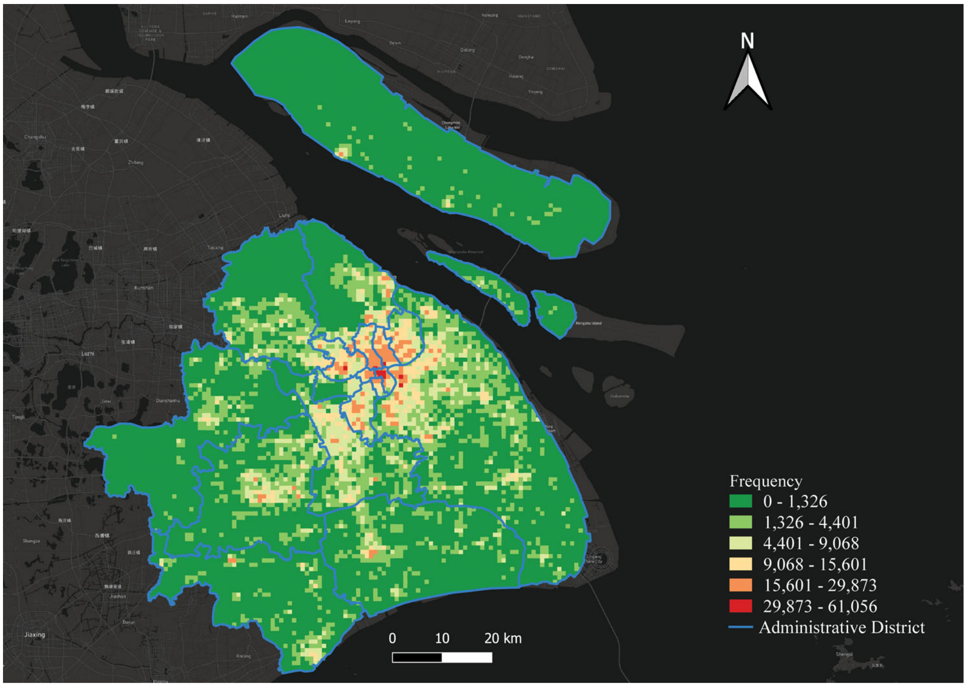


Figure 16. The number of occurrences that a cell tower is associated with oscillations (using $\Delta M = 0.5$). The numbers are aggregated and summarized at the level of a 1 km * 1 km grid.

indicate that the results generated from big data can be highly dependent on the ways they are processed. This reveals a fundamental challenge of data-driven mobility research, especially when the ground truth is difficult to acquire. A possible solution is to test the effectiveness of different methods through the integration of big data and small data. For example, experiments can be designed to collect both the mobile phone trajectories of users and their actual movement patterns through surveys. The survey-based observations can be used as a proxy for ground truth to evaluate the effectiveness of different preprocessing methods. Second, to what extent data veracity affects geographic knowledge discovery is case dependent. For instance, when estimating OD trips from mobile phone data, different methods might produce quite different estimations at an individual level. When these estimations are further aggregated, though—for example, by administrative districts or traffic analysis zones—different methods might produce similar spatial interaction patterns. In this case, even a "wrong" practice could lead to a "right" conclusion. This is not always the case, however. Recently, many studies have performed big data analytics to understand the relationship between mobility patterns and socioeconomic status of travelers (Frias-Martinez et al. 2013; Smith-Clarke, Mashhadi, and Capra 2014; Blumenstock, Cadamuro, and On 2015; Almaatouq, Prieto-Castrillo, and Pentland 2016; Pappalardo et al. 2016; Xu et al. 2018; L. Wu et al. 2019). In these studies, a collection of individual mobility (or sociality) indicators is derived and used to correlate with or to predict personal socioeconomic status. In this case, the variations in the mobility estimations from different methods could result in different conclusions (e.g., do rich or poor people conduct more trips). For studies that develop prediction models (e.g., machine learning or deep learning models), how mobility indicators are derived will affect the performance of the models and their generalization ability. This is related to an emerging discussion on the replicability and reproducibility in geospatial research (School of Geographical Sciences & Urban Planning, Arizona State University 2019). Without carefully examining the data veracity issue, the results generated from one study or geographic area might fail to be replicated in others. We believe that the method proposed in this research—once calibrated when ground truth becomes available—could help improve the estimations of human mobility patterns (e.g., OD trips, daily activity locations, and dwelling time at these locations) to support applications in transportation planning and location-based services. Because oscillation effects also occur in other types of mobility data (e.g., CDRs [Alexander et al. 2015], Wi-Fi data [Lee and Hou 2006; Bayir, Demirbas, and Eagle 2010]), the proposed method is useful for other data sets when uncertainty issues need to be tackled.

We want to point out a few limitations of this research. First, although the MAD approach tends to generate higher detection ratios of oscillations, the approach is not perfect. As demonstrated in the analysis, the choice of ΔM controls the tolerance of oscillation removal, which will affect how mobile phone trajectories are processed. It is possible, however, that some oscillations removed by the MAD approach refer to the actual movements of phone users, whereas some others that are not removed could be fake movements. A data set that documents both users' mobile phone trajectories and their actual movements (e.g., through surveys) can be useful for the calibration of ΔM and further improvement of the MAD approach (e.g., an adaptive choice of ΔM given trajectory properties). Moreover, because a significant proportion of human movements take place along roads and streets, incorporating road network–based measures (e.g., road network distance and speed) might further eliminate (or retain) some of the fake (or actual) movements. This is one direction for future research. Second, the three methods and their impact on mobility estimations have been tested and compared over one single data set. How the findings would generalize in a broader sense is worth further investigation. In this study, we have revealed the varying impacts of preprocessing methods on data characteristics. We believe that this veracity issue is not unique to the data set used in this study but exists in other mobile phone data sets. In the future, we intend to enlarge the research scope by repeating the experiments over multiple data sets and across different study areas. Nevertheless, we hope that this study provides some insights that can direct better usage of big data for future mobility studies. It also calls for more attention to the data veracity issue and its implications for geographic knowledge discovery.

Funding

This research was jointly supported by the Research Grant Council of Hong Kong (No. 25610118), the National Natural Science Foundation of China (No. 41801372), National Key Research and Development Program of China (No. 2016YFB0502104), the Alvin and Sally Beaman Professorship, Arts and Sciences Excellence Professorship, and James and Catherine Ralston Family Fund at the University of Tennessee, Knoxville.

Notes

1. When mobile phones are turned on but lose signals (e.g., traveling underground), no events or records are documented. Once phones regain signals, either emerging from underground or entering an area (e.g., subway station) where a cell tower signal is available, a CH event is triggered, which indicates a cellphone's "movement" from one cell antenna to another.

ORCID

Yang Xu http://orcid.org/0000-0003-3898-022X
Shih-Lung Shaw http://orcid.org/0000-0001-6157-5519
Bi Yu Chen http://orcid.org/0000-0003-3591-9968

References

Alexander, L., S. Jiang, M. Murga, and M. C. González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58:240–50. doi: 10.1016/j.trc.2015.02.018.

Almaatouq, A., F. Prieto-Castrillo, and A. Pentland. 2016. Mobile communication signatures of unemployment. In *International conference on social informatics*, ed. E. Spiro and Y.-Y. Ahn, 407–18. Bellevue, WA: Springer.

Bayir, M. A., M. Demirbas, and N. Eagle. 2010. Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing* 6 (4):435–54. doi: 10.1016/j.pmcj.2010.01.003.

Becker, R., R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. 2013. Human mobility characterization from cellular network data. *Communications of the* ACM 56 (1):74–82. doi: 10.1145/2398356.2398375.

Birenboim, A., and N. Shoval. 2016. Mobility research in the age of the smartphone. *Annals of the American Association of Geographers* 106 (2):283–91. doi: 10.1080/00045608.2015.1100058.

Blondel, V. D., A. Decuyper, and G. Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ Data Science* 4 (1):10. doi: 10.1140/epjds/s13688-015-0046-0.

Blumenstock, J., G. Cadamuro, and R. On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350 (6264):1073–76. doi: 10.1126/science.aac4420.

Bwambale, A., C. F. Choudhury, and S. Hess. 2017. Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography* 76:276–86.

Calabrese, F., M. Diao, G. D. Lorenzo, J. Ferreira, Jr., and C. Ratti. 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies* 26:301–13. doi: 10.1016/j.trc.2012.09.009.

Calabrese, F., G. D. Lorenzo, L. Liu, and C. Ratti. 2011. Estimating origin–destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area. *IEEE Pervasive Computing* 10 (4):36–44. doi: 10.1109/MPRV.2011.41.

Calabrese, F., Z. Smoreda, V. D. Blondel, and C. Ratti. 2011. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLoS ONE* 6 (7):e20814. doi: 10.1371/journal.pone.0020814.

Chen, C., L. Bian, and J. Ma. 2014. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies* 46:326–37. doi: 10.1016/j.trc.2014.07.001.

Chen, C., H. Gong, C. Lawson, and E. Bialostozky. 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. *Transportation Research Part A: Policy and Practice* 44 (10):830–40. doi: 10.1016/j.tra.2010.08.004.

Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies* 68:285–99. doi: 10.1016/j.trc.2016.04.005.

Cho, E., S. A. Myers, and J. Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, ed. C. Apte, 1082–90. San Diego, CA: ACM.

Csáji, B. C., A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel. 2013. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and Its Applications* 392 (6):1459–73. doi: 10.1016/j.physa.2012.11.040.

Fan, Z., T. Pei, T. Ma, Y. Du, C. Song, Z. Liu, and C. Zhou. 2018. Estimation of urban crowd flux based on mobile phone location data: A case study of Beijing, China. *Computers, Environment and Urban Systems* 69:114–23. doi: 10.1016/j.compenvurbsys.2018.01.005.

Frias-Martinez, V., C. Soguero-Ruiz, E. Frias-Martinez, and M. Josephidou. 2013. Forecasting socioeconomic trends with cell phone records. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, ed. B. Thies and A. Nanavati, 15. New York: ACM. doi: 10.1145/2442882.2442902.

Gao, S., Y. Liu, Y. Wang, and X. Ma. 2013. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS* 17 (3):463–81. doi: 10.1111/tgis.12042.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779–82. doi: 10.1038/nature06958.

Goodchild, M. F. 1998. Uncertainty: The Achilles heel of GIS. *Geo Info Systems* 8 (11):50–52.

Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies* 40:63–74. doi: 10.1016/j.trc.2014.01.002.

Isaacman, S., R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*, 239–52. ACM. doi: 10.1145/2307636.2307659.

Janecek, A., D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs. 2015. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems* 16 (5):2551–72. doi: 10.1109/TITS.2015.2413215.

Jiang, S., J. Ferreira, and M. C. González. 2017. Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data* 3 (2):208–19. doi: 10.1109/TBDATA.2016.2631141.

Jiang, S., Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González. 2016. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences of the United States of America* 113 (37):E5370–78. doi: 10.1073/pnas.1524261113.

Kwan, M.-P. 2016. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers* 106 (2):274–82.

Lee, J.-K., and J. C. Hou. 2006. Modeling steady-state and transient behaviors of user mobility: Formulation, analysis, and application. In *Proceedings of the 7th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ed. S. Palazzo, 85–96. New York: ACM.

Li, M., S. Gao, F. Lu, and H. Zhang. 2019. Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. *Computers, Environment and Urban Systems* 77:101346. doi: 10.1016/j.compenvurbsys.2019.101346.

Li, Z., L. Yu, Y. Gao, Y. Wu, G. Song, and D. Gong. 2018. Identifying temporal and spatial characteristics of residents' trips from cellular signaling data: Case study of Beijing. *Transportation Research Record: Journal of the Transportation Research Board* 2672 (42):81–90. doi: 10.1177/0361198118793495.

McMaster, R. B., and E. L. Usery. 2004. *A research agenda for geographic information science*. Boca Raton, FL: CRC.

Pappalardo, L., F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási. 2015. Returners and explorers dichotomy in human mobility. *Nature Communications* 6:8166. doi: 10.1038/ncomms9166.

Pappalardo, L., M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti. 2016. An analytical framework to nowcast well-being using mobile phone data. *International Journal of Data Science and Analytics* 2 (1–2):75–92. doi: 10.1007/s41060-016-0013-2.

Ratti, C., D. Frenchman, R. M. Pulselli, and S. Williams. 2006. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33 (5):727–48. doi: 10.1068/b32047.

Robinson, S., B. Narayanan, N. Toh, and F. Pereira. 2014. Methods for pre-processing smartcard data to improve data quality. *Transportation Research Part C: Emerging Technologies* 49:43–58. doi: 10.1016/j.trc.2014.10.006.

School of Geographical Sciences & Urban Planning, Arizona State University. 2019. Replicability and reproducibility in geospatial research: A SPARC workshop. February 11–12. Accessed June 26, 2020. https://sgsup.asu.edu/sparc/RRWorkshop.

Silm, S., and R. Ahas. 2014. Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers* 104 (3):542–59. doi: 10.1080/00045608.2014.892362.

Smith-Clarke, C., A. Mashhadi, and L. Capra. 2014. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ed. M. Jones and P. Palaque, 511–20. New York: ACM.

Song, C., T. Koren, P. Wang, and A.-L. Barabási. 2010. Modelling the scaling properties of human mobility. *Nature Physics* 6 (10):818–23. doi: 10.1038/nphys1760.

Song, C., Z. Qu, N. Blumm, and A.-L. Barabási. 2010. Limits of predictability in human mobility. *Science* 327 (5968):1018–21. doi: 10.1126/science.1177170.

Stopher, P. R., and S. P. Greaves. 2007. Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice* 41 (5):367–81. doi: 10.1016/j.tra.2006.09.005.

Toole, J. L., C. Herrera-Yaqüe, C. M. Schneider, and M. C. González. 2015. Coupling human mobility and social ties. *Journal of the Royal Society Interface* 12 (105). doi: 10.1098/rsif.2014.1128.

Trépanier, M., N. Tranchant, and R. Chapleau. 2007. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems* 11 (1):1–14. doi: 10.1080/15472450601122256.

Wang, D., D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ed. C. Apte, 1100–1108. New York: ACM. doi: 10.1145/2020408.2020581.

Wang, F., and C. Chen. 2018. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies* 87:58–74. doi: 10.1016/j.trc.2017.12.003.

Widhalm, P., Y. Yang, M. Ulm, S. Athavale, and M. C. González. 2015. Discovering urban activity patterns in cell phone data. *Transportation* 42 (4):597–623. doi: 10.1007/s11116-015-9598-x.

Wu, L., L. Yang, Z. Huang, Y. Wang, Y. Chai, X. Peng, and Y. Liu. 2019. Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems* 77. doi: 10.1016/j.compenvurbsys.2019.101368.

Wu, W., Y. Wang, J. B. Gomes, D. T. Anh, S. Antonatos, M. Xue, P. Yang, et al. 2014. Oscillation resolution for mobile phone cellular tower data to enable mobility modelling. In *2014 IEEE 15th International Conference on Mobile Data Management*, Vol. 1, ed. IEEE Computer Society, 321–28. Washington, DC: IEEE.

Xu, Y., A. Belyi, I. Bojic, and C. Ratti. 2017. How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data. *Transactions in GIS* 21 (3):468–87. doi: 10.1111/tgis.12285.

Xu, Y., A. Belyi, I. Bojic, and C. Ratti. 2018. Human mobility and socioeconomic status: Analysis of Singapore and Boston. *Computers, Environment and Urban Systems* 72:51–67. doi: 10.1016/j.compenvurbsys.2018.04.001.

Xu, Y., A. Belyi, P. Santi, and C. Ratti. 2019. Quantifying segregation in an integrated urban physical-social space. *Journal of the Royal Society, Interface* 16 (160). doi: 10.1098/rsif.2019.0536.

Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li. 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation* 42 (4):625–46. doi: 10.1007/s11116-015-9597-y.

Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, F. Lu, J. Chen, Z. Fang, and Q. Li. 2016. Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers* 106 (2):489–502.

Yan, L., D. Wang, S. Zhang, and D. Xie. 2019. Evaluating the multi-scale patterns of jobs–residence balance and commuting time–cost using cellular signaling data: A case study in Shanghai. *Transportation* 46 (3):777–816. doi: 10.1007/s11116-018-9894-3.

Yang, F., P. J. Jin, Y. Cheng, J. Zhang, and B. Ran. 2015. Origin-destination estimation for non-commuting trips using location-based social networking data. *International Journal of Sustainable Transportation* 9 (8):551–64. doi: 10.1080/15568318.2013.826312.

Yuan, Y., and M. Raubal. 2016. Analyzing the distribution of human activity space from mobile phone usage: An individual and urban-oriented study. *International Journal of Geographical Information Science* 30 (8):1594–621. doi: 10.1080/13658816.2016.1143555.

Zhao, Z., S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin. 2016. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science* 30 (9):1738–62. doi: 10.1080/13658816.2015.1137298.

YANG XU is an Assistant Professor in the Department of Land Surveying and Geo-Informatics at the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: yang.ls.xu@polyu.edu.hk. His research interests include GIScience, human mobility, and urban informatics.

XINYU LI is a PhD Student in the Department of Land Surveying and Geo-Informatics at the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: joeylee.li@connect.polyu.hk. His research interests include spatiotemporal data mining, deep learning, and geospatial artificial intelligence.

SHIH-LUNG SHAW is Alvin and Sally Beaman Professor and Arts and Sciences Excellence Professor in the Department of Geography at the University of Tennessee, Knoxville, TN 37996. E-mail: sshaw@utk.edu. His research interests include transportation geography, human dynamics, GIScience, space–time geographic information systems (GIS), and GIS for transportation.

FENG LU is a Professor of the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China. E-mail: luf@lreis.ac.cn. His research interests involve spatial data modeling, trajectory data mining, complex network analysis, knowledge graphs, and GIS for transportation.

LING YIN is an Associate Professor in the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, Guangdong Province 518055, China. E-mail: yinling@siat.ac.cn. Her research interests include human dynamics, spatial epidemic models, space–time GIS, and GIS for transportation.

BI YU CHEN is a Professor in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China. E-mail: chen.biyu@whu.edu.cn. His research interests include GIS for transportation, transport geography, and spatiotemporal big data analytics.

Appendix

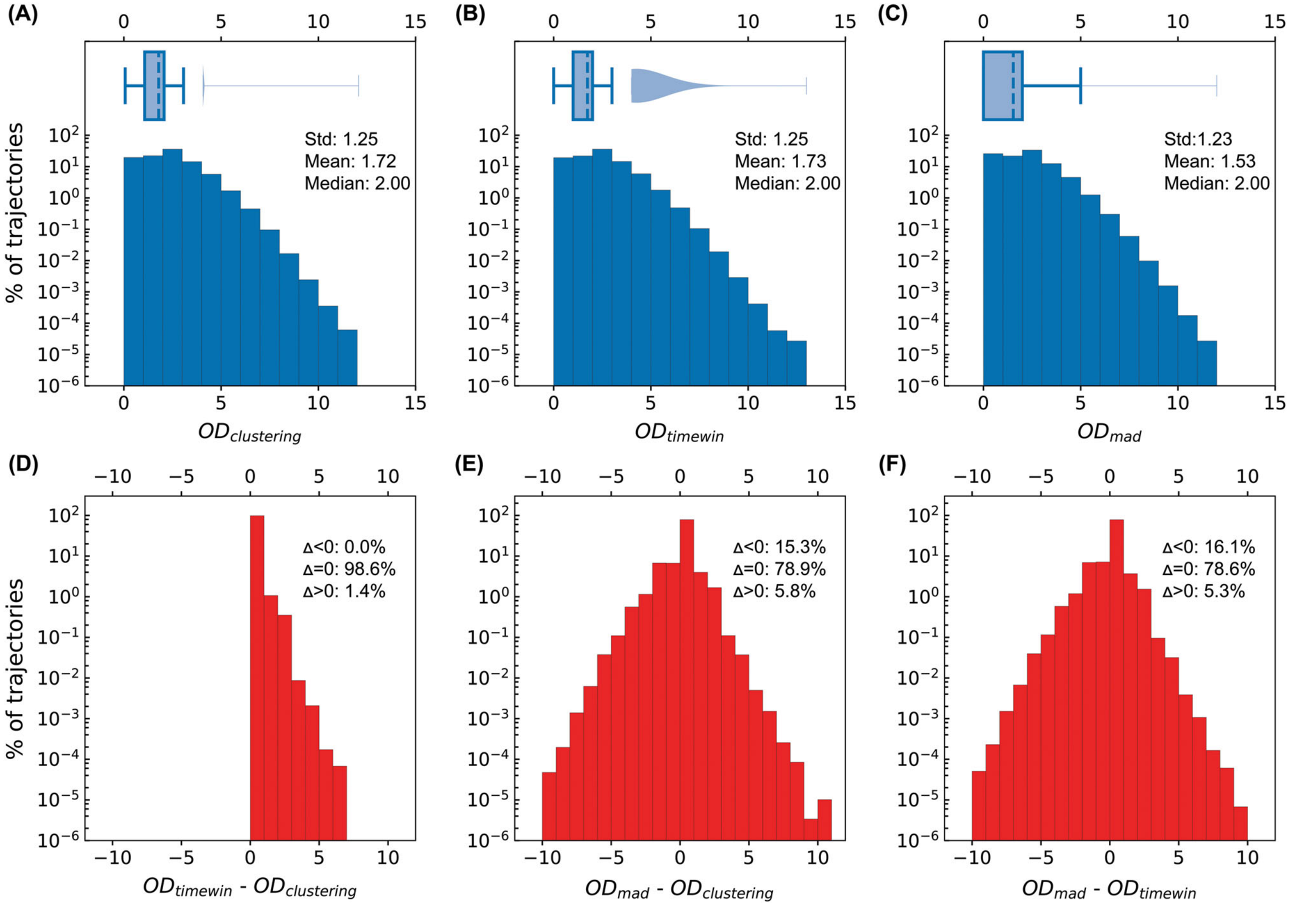


Figure A.1. (A–C) Distribution of $OD_{\text{clustering}}$, OD_{timewin}, and OD_{mad} and (D–F) pair-wise comparison of the three methods. Origin–destination (OD) trips are generated based on stay duration of 30 minutes.

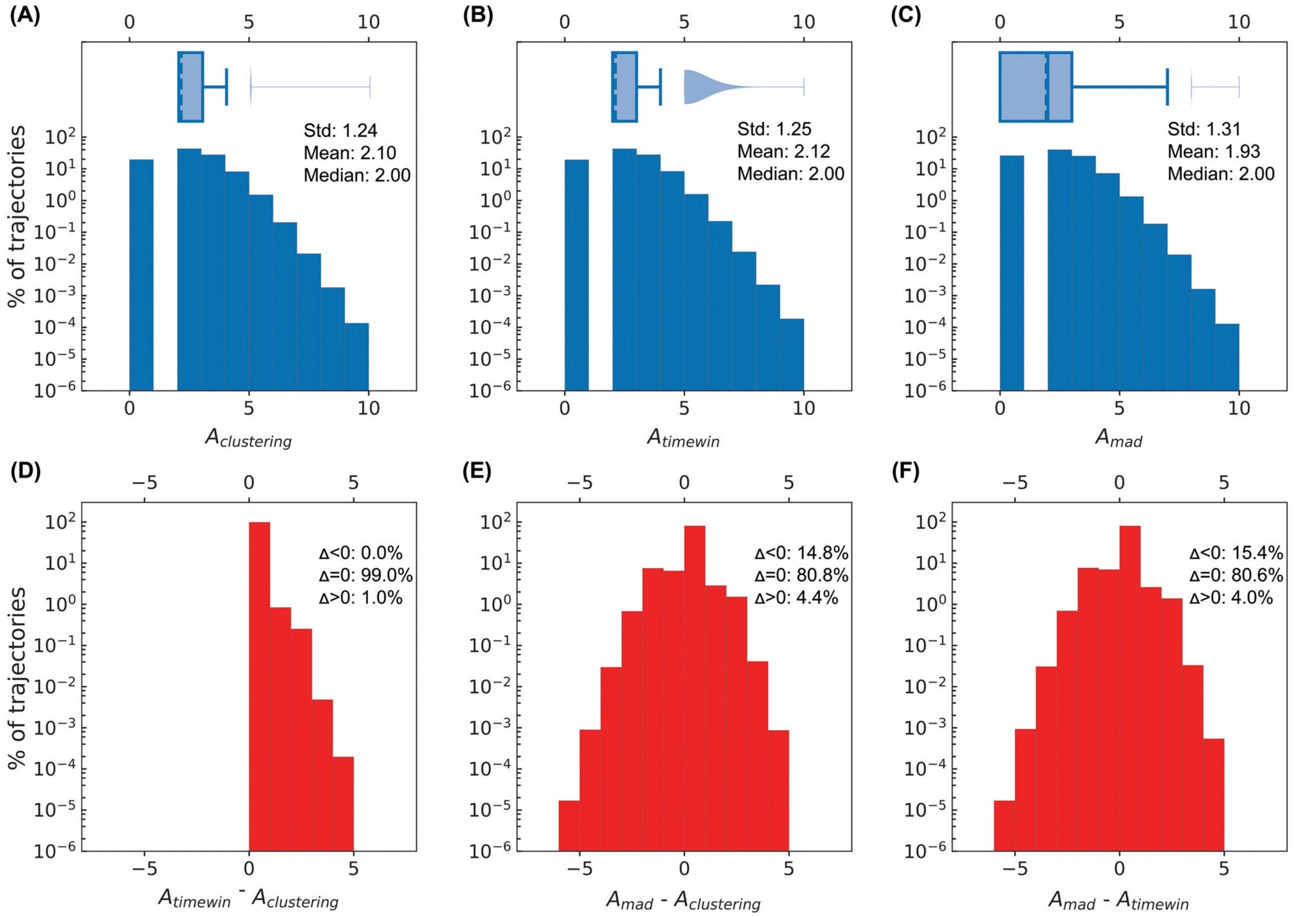


Figure A.2. (A–C) Distribution of $A_{\text{clustering}}$, A_{timewin}, and A_{mad} and (D–F) pair-wise comparison of the three methods. Results are generated based on stay duration of 30 minutes.