Effects of data preprocessing methods on addressing location uncertainty in mobile signaling data

Yang Xu*[1,2], Xinyu Li[1], Shih-Lung Shaw[3], Feng Lu[4,5,6], Ling Yin[7], and Biyu Chen[8,9]

[1]Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University
[2]The Hong Kong Polytechnic University Shenzhen Research Institute
[3]Department of Geography, University of Tennessee, Knoxville
[4]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
[5]University of Chinese Academy of Sciences, Beijing, China
[6]The Academy of Digital China, Fuzhou University
[7]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[8]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University
[9]Collaborative Innovation Center of Geospatial Technology, Wuhan, China

Abstract

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 has 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 dataset, we find that the choice of preprocessing methods could lead to changes of the data characteristics. Such changes, which are non-trivial, will further affect the characterization and interpretation of human mobility patterns. By computing four mobility indicators (number of OD 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 to replicability and reproducibility of geospatial research.

Key words: human mobility; mobile phone data; uncertainty; veracity

*Electronic address: yang.ls.xu@polyu.edu.hk

1 Introduction

"Big Data" is no longer a buzzword, but something that truly impacts 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 datasets. 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 over-estimated from travel surveys due to self-report errors (Chen et al. 2010; Stopher and Greaves 2007). Some studies extract OD trips from smart card transactions to examine public transport usage. However, some of these datasets 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 et al. 2007; Robinson et al. 2014). Social media data have also been used to derived OD matrices to support transport planning (Yang et al. 2015). However, the mobility traces of social media users can be sparse in time and space. 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., call, text message) — have been used extensively to study human travel and activity patterns (Iqbal et al. 2014; Alexander et al. 2015; Jiang et al. 2016; 2017; Xu et al. 2018). However, CDRs suffer from issues of data sparsity (due to the passive data collec-

tion mechanism) and location uncertainty (e.g., cellphone signal switch), which add 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 they have been discussed by previous researchers at different depths (Chen et al. 2014; Xu et al. 2016; 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 (UCGIS) identified uncertainty in spatial data as a long-term research challenge (McMaster and Usery 2004). However, research attentions on uncertainty issues of mobile phone data appears to become notable only in the recent years (Wu et al. 2014; Xu et al. 2016; Kwan 2016; Chen et al. 2016). A key characteristic of mobile phone data is that the locations are documented at the level of cellphone towers. These locations, which are usually represented as geographic coordinates of the cellphone 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 cellphone towers due to load balancing or signal strength variation (Xu et al. 2016; Kwan 2016). Such issues of positional inaccuracy have hindered reliable estimates of human mobility patterns that are important to many geospatial applications. Although the above-mentioned issues have been noticed by the research community (Kwan 2016), it seems that the major effort has 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 overlook of these issues is not without reasons. An important fact to mention is that these mobility datasets — when they are born — are not intended for travel behaviour 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 normality of many "big" mobility datasets, there is a need for alternatives that would look into this

issue — by comparing different methodologies, their pros and cons, and the tradeoffs 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 dataset, we examine how these methods change the data characteristics in different ways, and how such changes would impact 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.

2 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 Massachusetts Institute of Technology (MIT) 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 telecommunications industry that records person hours of cellphone usage (Ratti et al. 2006). Since Erlang is an aggre-

gate measure of traffic volume in telecommunications system, the data are not suitable for studying movement patterns. Later on, *Call Detail Records (CDRs)*, another type of phone data usually collected by cellular operators for billing purpose, 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 et al. 2008; Song et al. 2010a;b; 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 et al. 2011; Calabrese et al. 2011b; Wang et al. 2011; Gao et al. 2013; Toole et al. 2015; Xu et al. 2017; 2019), and so forth (See (Blondel et al. 2015; Birenboim and Shoval 2016) for extensive reviews).

Many travel behaviour studies have involved the usage of CDRs for OD estimation and mobility modelling (Alexander et al. 2015; Jiang et al. 2016; Pappalardo et al. 2016; Jiang et al. 2017; Bwambale et al. 2017; Xu et al. 2018). However, there are a few characteristics of CDRs that would complicate such tasks:

- CDR data are collected at the level of cellphone towers, of which the densities in space affect the positioning accuracy. The spacing gap between cellphone 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 noises for CDRs, which cause "the appearance of fake movements" (Alexander et al. 2015).

- CDRs suffer from 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 (Cal-

abrese et al. 2011a; Fan et al. 2018; Widhalm et al. 2015). Such clustering methods are able to filter some "fake movements" while capturing meaningful activity locations of individuals.

- Beyond the above step, some studies also perform an additional step to merge the detected clusters that are close in space but may 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 et al. 2017). This practice will partially address the data sparsity issue. But the choice of the threshold, which is empirical and somewhat arbitrary, could have a direct impact on the analysis that follows.

- Another factor that causes fake movements is *celltower oscillation*, also known as the ping-pong effect. Such effects are caused by the users' cellphone handover to nearby cellphone 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, (Wu et al. 2014) proposed a few heuristics to identify oscillations by detecting movements at impossible speed or cellphone towers that appeared repeatedly. (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 et al. 2010) introduced a graph-based clustering algorithm, which iteratively merge densly connected cellphone towers in a user's trajectory, to address the oscillation effect.

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

Despite the numerous insights of human mobility discovered from CDRs and sightings data, the observations from them are generally sparse due to the passive data collection mechanism. *Mobile signaling data (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 usages are 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 update. The improvement in data granularity makes MSD an appealing option for mobility studies (Yan et al. 2018; Li et al. 2018; 2019). However, similar to CDRs and sightings data, 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 behaviour analysis.

3 Dataset

A large mobile signaling dataset collected in Shanghai, China is used. The dataset captures the location traces of 7.6 million phone users during a period of one week (October 15 - 21, 2012). The locations of phone users were tracked at the level of cellphone 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 dataset. 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) event[1]. Even 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) event.

Table 1: Summary of signaling events captured in the dataset

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

Different from CDRs that passively collect data during phone usage activities, the mobile signaling dataset 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 no matter 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. Since mobile phones can be switched on or off (Table 1), location of users will be unavailable during the disconnected periods.

[1]When mobile phones are turned on but lose signals (e.g., travelling underground), there will be no events/records documented during such periods. Once phones regain signals, either emerging from underground or entering an area (e.g, subway station) where cell tower signal is available, it will trigger a Cellular Handover (CH) event, which indicates a cellphone's "movement" from one cell antenna to another.

Thus, we filter the dataset by removing user daily trajectories with Power On (ON) or Power Off (OFF) events. The dataset after removing such cases include 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 dataset.

Figure 1 shows some general statistics of the dataset. The number of records of a trajectory varies notably from each other, with the mean and median value as 59.3 and 41.0, respectively (Figure 1A). The inter-event 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 value are 20.25 and 1.55 minutes, respectively. To better understand the density of cellphone tower antennas in the city, we compute, for each cell, its distance to the nearest cell. This yields a skewed distribution with the mean and median value as 149.3m and 92.8m, respectively (Figure 1C). To obtain a better understanding of the spatial distribution of cellphone towers in the city, we generate a 1km*1km regular grid and compute the number of towers in each grid cell. As shown in Figure 2, cellphone towers are unevenly distributed in Shanghai and their densities are generally higher in the core part of the city (e.g., Downtown Shanghai).

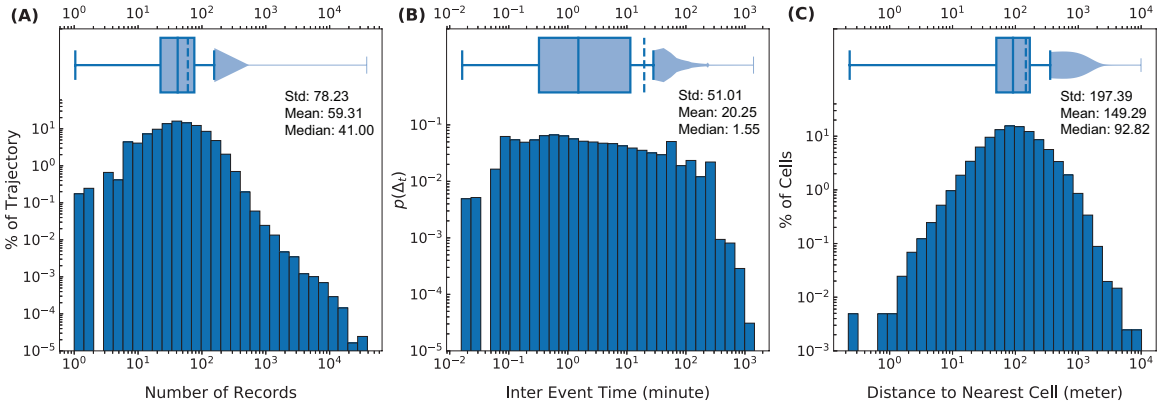

Figure 1: General statistics of the mobile signaling data: (A) Number of records per trajectory; (B) Distribution of inter-event time; (C) Spacing gap between cellphone tower antennas.

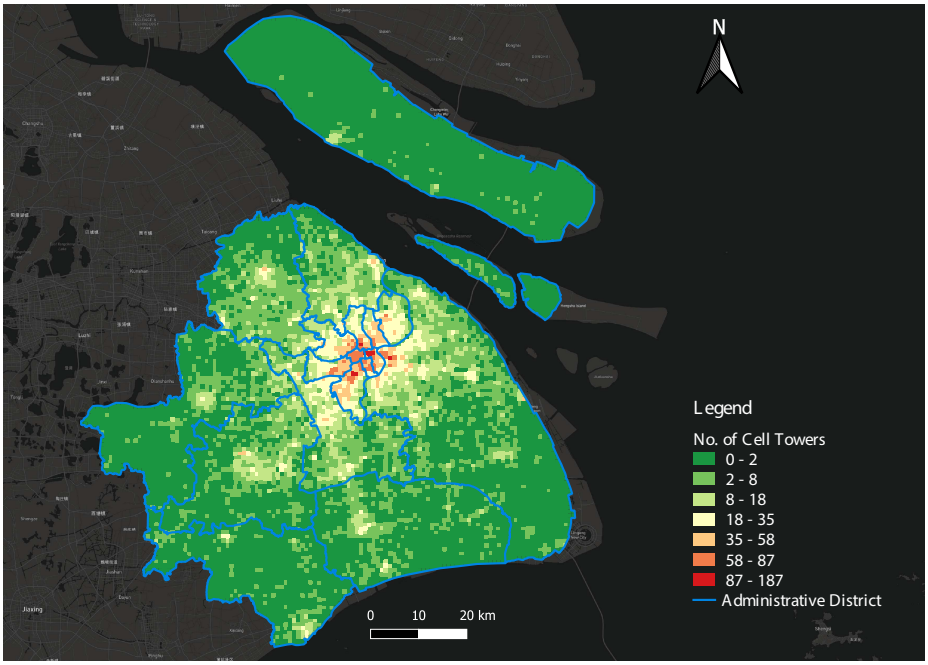

Figure 2: Number of cellphone towers in each grid cell (1km*1km grid).

4 Methodology

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.

4.1 Definitions

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

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

where l_i and t_i denote the location and the time of the i^{th} 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:

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

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:

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

In the remaining of the paper, we will use "*A*-*B*-*A*" to describe the "phenotype" of such oscillation pairs.

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

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

subject to:

$$\sqrt{(\overrightarrow{l_{i+3}} - \overrightarrow{l_i})^2} = 0 \tag{6}$$

and

$$\sqrt{(\overrightarrow{l_{i+2}} - \overrightarrow{l_{i+1}})^2} = 0 \tag{7}$$

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}, the number of O_{p2}

in "*A-B-B-A-C-A*" are two, one, and one, respectively.

4.2 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 cellphone towers (Csáji et al. 2013; Isaacman et al. 2012), generating "fake movements" that complicate human mobility analysis. Some studies also define this issue as *oscillation* or *ping-pong effect*, which describe that a phone's signal could switch between multiple cellular towers even though the device is not moving (Wu et al. 2014). Given these issues, the two-stage clustering method is primarily used to generalize a user's 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 tradeoff.

The two-stage clustering algorithm used in previous studies (Alexander et al. 2015; Jiang et al. 2017; Xu et al. 2018) is first applied to identify stay locations from the trajectories. Note that before this clustering process, we have performed a *zero step* to remove cellphone records with abnormal speed (i.e., $\geq 120km/h$). Given $T = \{(l_1, t_1), (l_2, t_2), ..., (l_n, t_n)\}$, 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 — which is 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 (representation 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 the segments are identified. This results in a sequence $\{(l'_1, t'_1, dur_1), (l'_2, t'_2, dur_2), ..., (l'_n, t'_n, dur_n)\}$ where l'_i, t'_i, and dur_i denote the medoid, starting time, and the stay duration of the i^{th} 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 an user's activity locations. In particular, we identify the stay segments whose representative locations 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).

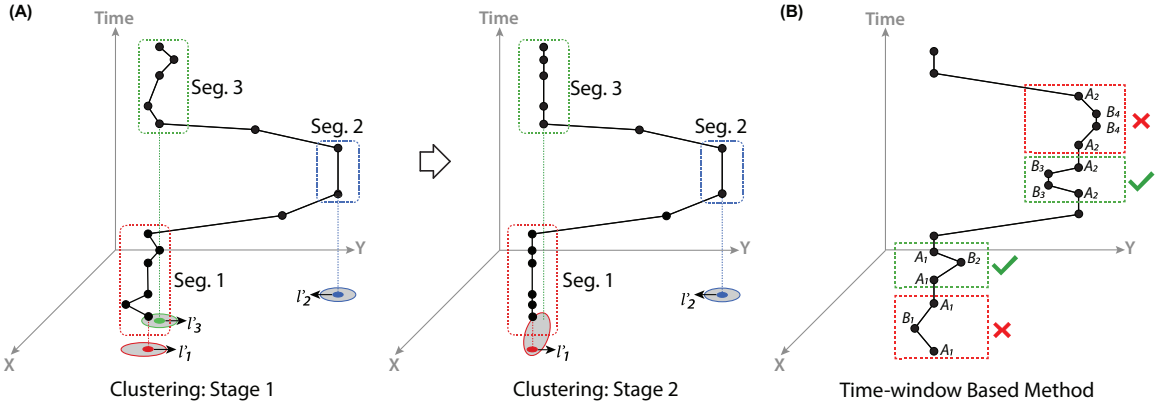

Figure 3: (A) Two-stage clustering algorithm: In stage 1, stay segments are detected and annotated with their representative locations (l'_1, l'_2, l'_3). In stage 2, segments that are close in space but apart in time are further grouped, with their representative locations updated. In this example, 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) while others are not (red). This reveals a potential limitation of time-window based method for oscillation detection.

The clustering based algorithm can partially tackle the issue of cellphone signal switch. However, 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 (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 — subquences in a trajectory which 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.

4.3 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 daily trajectory, MAD measures the average deviation of each location's visitation frequency from the median of the dataset:

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

Here, n denotes the total number of locations traversed by the user daily trajectory and x_i denotes the frequency of visit to the i^{th} location. $m(X)$ refers to the median frequency of all locations.

The idea of the algorithm is simple. Since normal 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 a MAD within a reasonable range. However, due to oscillation effect, the cellphone signal would switch frequently among a collection of cells. These cells could either reflect an 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 are 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.

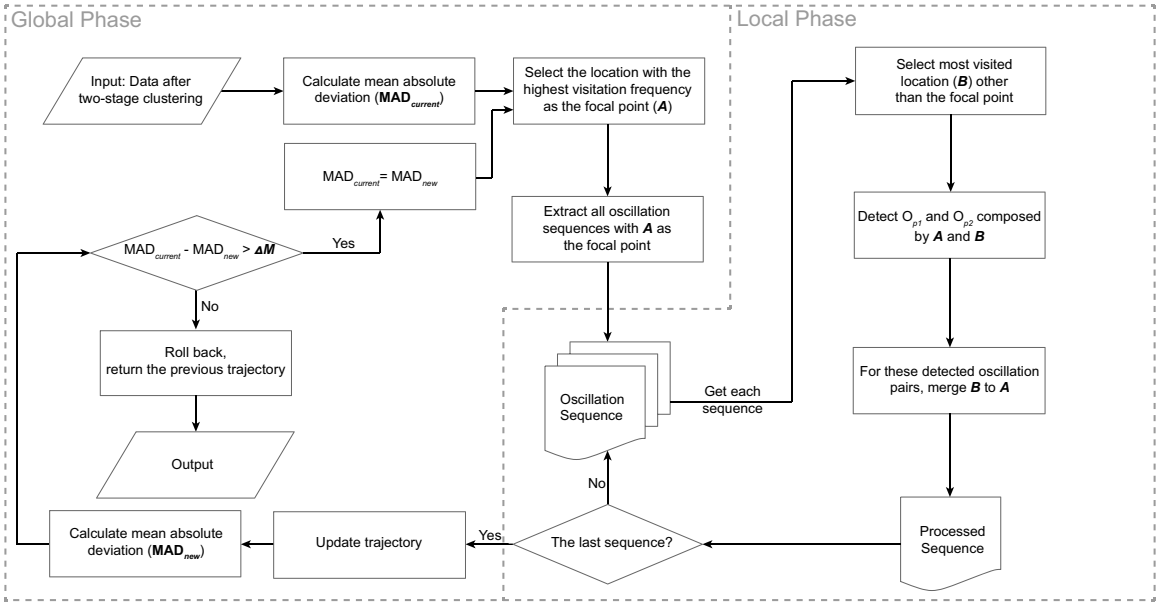

Figure 4: Workflow of the new approach based on mean absolute deviation (MAD).

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 much 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 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", "A-B-A-C-A-C-A", among others. For each of the detected oscillation sequence, 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 become "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 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 the oscillation sequences are processed in the local phase, the algorithm will update the frequency of each location, from which the 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 the last location 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 $MAD_{current} - MAD_{new} > \Delta M$, 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 of MAD is identified. A large value, however, will only remove oscillations with a high frequency.

Figure 5 shows a simple example of how 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-min 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 more satisfactory result (Figure 5G).

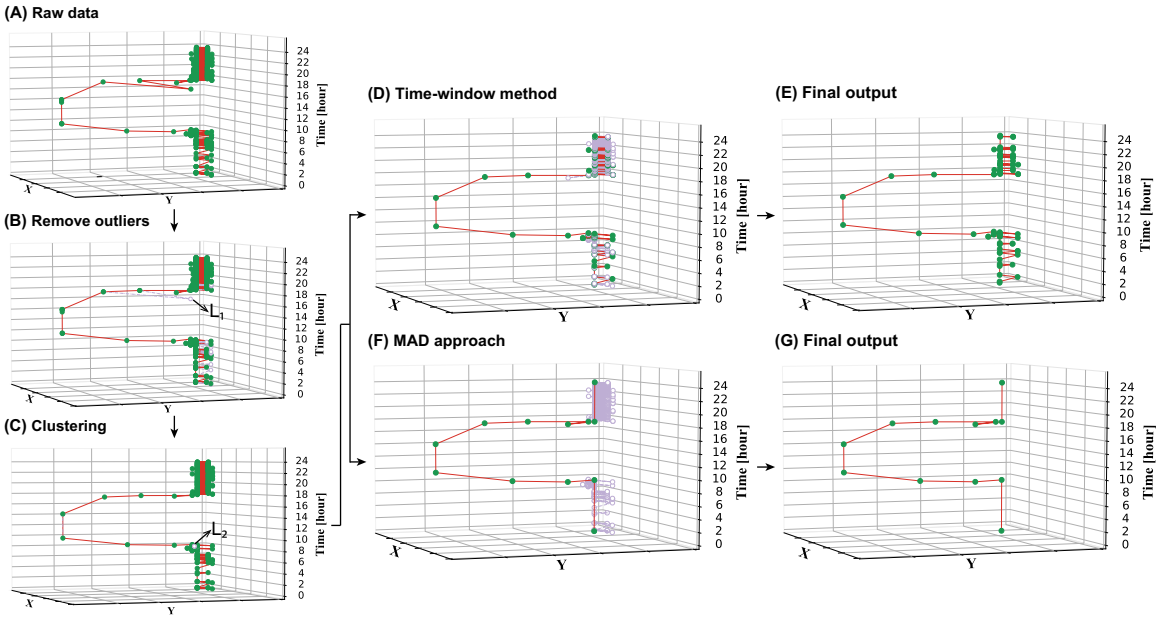

Figure 5: An example comparing the time-window based method and the new approach based on mean absolute deviation (MAD): (A) raw data of a user daily trajectory; (B) L_1 is identified with abnormal speed and thus removed in the zero step; (C) The two-stage clustering merges some of the locations (e.g., 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-min time window; (F-G) the MAD approach tends to identify highly frequented oscillations and achieve more reasonable outcome.

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 ratio 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 above 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:

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

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

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

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

5 Analysis Results

5.1 Impact of two-stage clustering on data characteristics

We first apply the two-stage clustering algorithm and evaluate its impact on data characteristics. Given the average spacing gap between cells as roughly 150m (Figure 1C), we set both Δd_1 and Δd_2 at 200m, and monitor the changes of 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 of 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. However, 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.

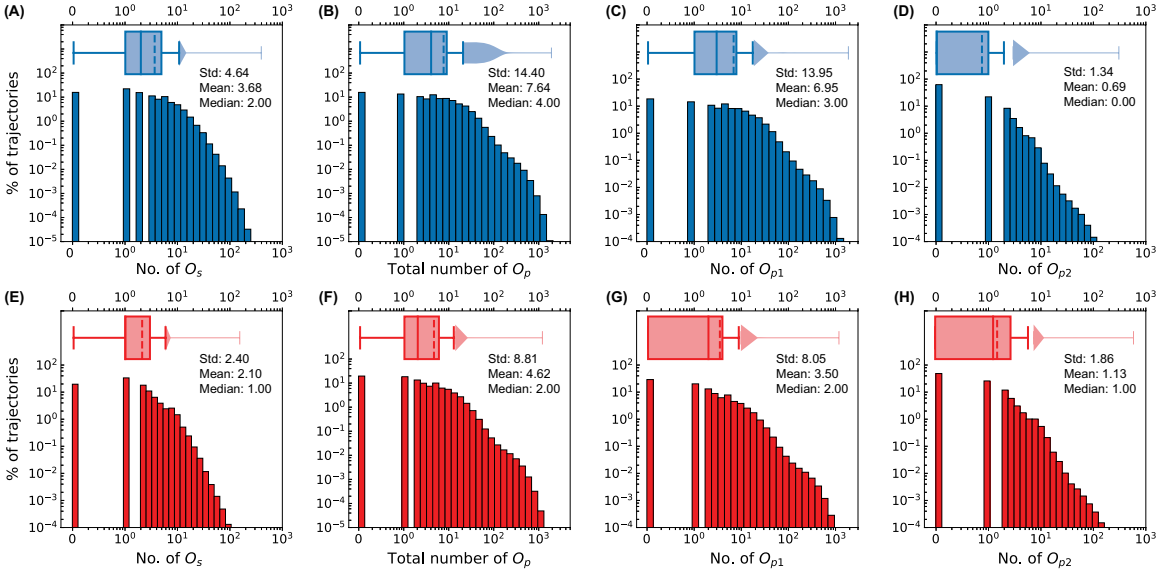

Figure 6: Distributions of the number of oscillation sequence, total number of oscillation pairs, number of type 1 and type 2 oscillation pairs in a trajectory before (A-D) and after (E-H) the two-stage clustering algorithm is performed.

Figure 7 illustrates the number of unique locations in a trajectory and the inter-event time of the two types of oscillation pairs after performing the two-stage clustering algorithm. The mean and median number of unique locations in a trajectory are 10.36 and 7.00, respectively (Figure 7A). The inter-event 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 inter-event time greater than 5 minutes. This reveals a notable limitation of time-window based method that many of these oscillation pairs will be ignored given an arbitrarily selected window size (e.g., 5 minutes). This issue will persist when type 2 oscillation pairs are handled. A large variation of inter-event time (the elapsed time between two A in "A-B-B-A") makes it extremely difficult to justify the choice of window size (Figure 7C).

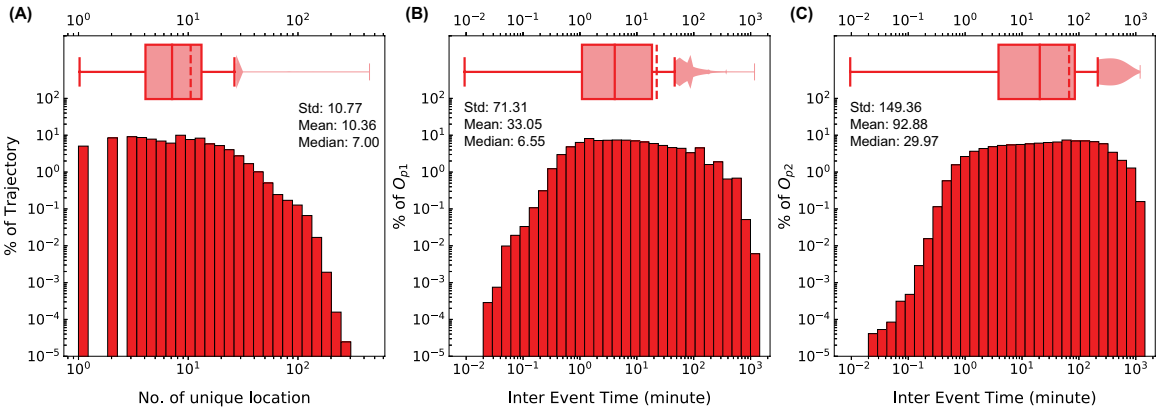

Figure 7: Distribution of: (A) number of unique locations in a user trajectory, (B) inter-event time of type 1 oscillation pairs, and (C) inter-event time of type 2 oscillation pairs based on the output of two-stage clustering algorithm.

5.2 Oscillation detection: time-window based method vs. MAD approach

We compare the two methods using the indicators proposed in section 4.3. 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 window size of 5 minutes as an example. By extracting all the oscillation sequences in the trajectories, for each oscillation sequence S, we compute R^S_{op}, R^S_{op1} and R^S_{op2} for the two methods. This allow us to not only investigate the average detection ratio of each method ($\overline{R^S_{op}}$, $\overline{R^S_{op1}}$ and $\overline{R^S_{op2}}$), but also the difference between the two (ΔR^S_{op}, ΔR^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 inter-event 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%, suggesting its better performance especially

for tackling long oscillation sequences. We also find relatively consistent difference between the two methods when splitting type 1 (Figure 8B) and type 2 oscillation pairs (Figure 8C).

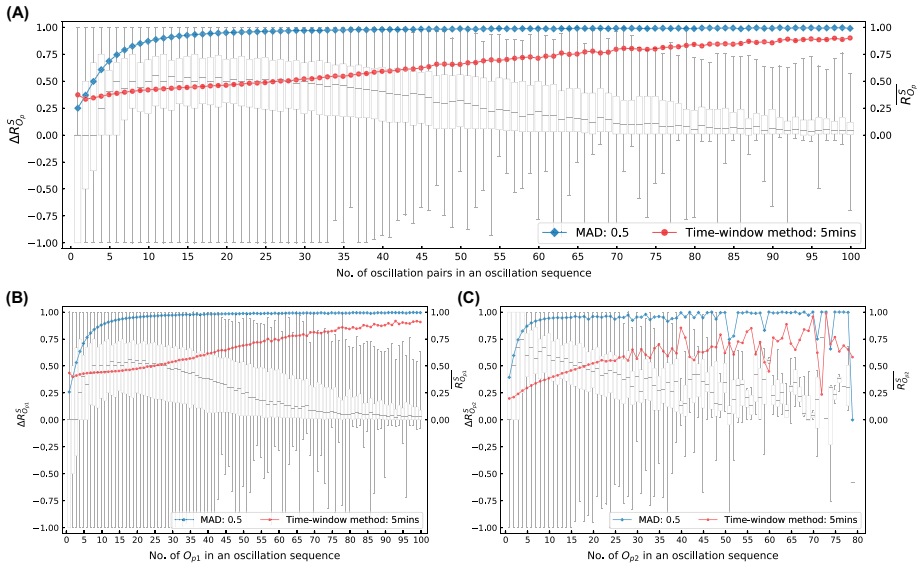

Figure 8: Detection ratio of oscillation pairs in an oscillation sequence when combining or splitting type 1 and type 2 oscillation pairs. Horizontal axes denote the number of oscillation pairs in an oscillation sequence: (A) combined; (B) type 1 oscillation pairs (O_{p1}); (C) type 2 oscillation pairs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

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, no matter 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}, are 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 mean absolute deviation (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 MAD approach (diamond), $\overline{R^T_{op}}$, tends to decrease as n increases. However, an opposite trend is observed for the time-window based method (circle). Sim-

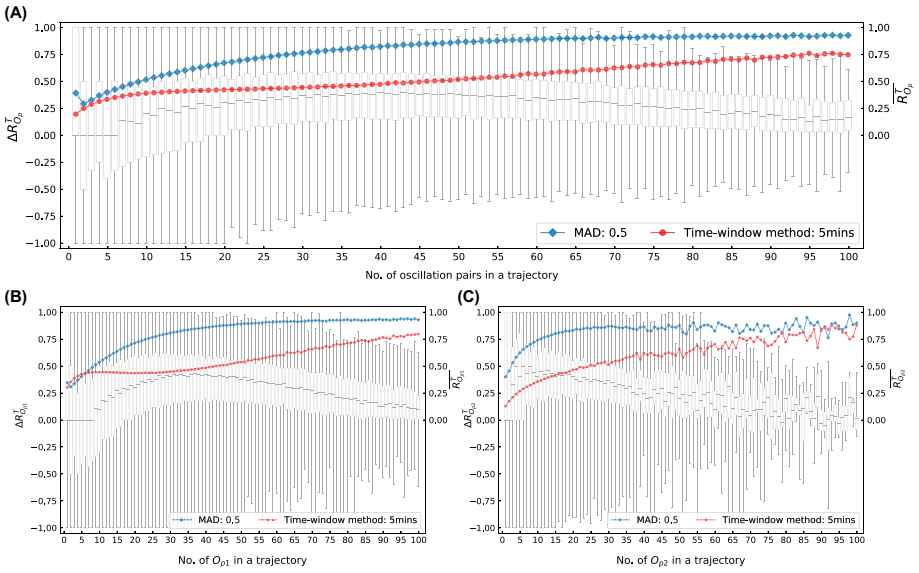

Figure 9: Detection ratio of oscillation pairs in a user daily trajectory when combining or splitting type 1 and type 2 oscillation pairs. Horizontal axes denote the number of oscillation pairs in a user daily trajectory: (A) combined; (B) type 1 oscillation pairs (O_{p1}); (C) type 2 oscillation pairs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

ilar patterns are observed when splitting type 1 (Figure 10B) and type 2 oscillation pairs (Figure 10C) despite that the two curves (circle vs. diamond) cross each other at different values of n.

The low detection ratio of 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. Since 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 in order to pass the current iteration, i.e., 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.

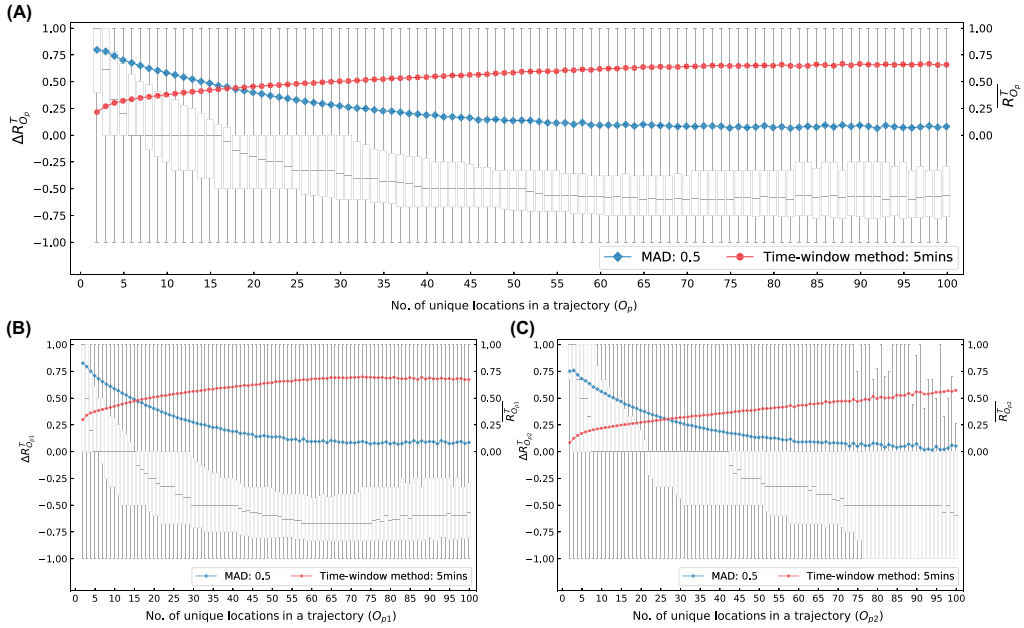

Figure 10: Detection ratio of oscillation pairs in a user daily trajectory when combining or splitting type 1 and type 2 oscillation pairs. Different from Figure 9, horizontal axes here denote total number of unique locations in a trajectory.

Next, we evaluate the impact of parameter choice on detection ratio. For the time-window based method, increasing the window size will remove more oscillations 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, i.e., the 5-min and 10-min window size.

As showin in Figure 11, the MAD approach with the first four parameter settings ($\Delta M = 0.01$, $\Delta M = 0.1$, $\Delta M = 0.25$, $\Delta M = 0.5$) tend to outperform the time-window based method with the 5-min threshold (red curve). However, when ΔM are set to higher values, 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 trajec-

tory is small (e.g., less than 10). This indicates that without carefully calibrating the value of ΔM, the MAD approach might produce unsatisfactory output when oscillations are sparse in a trajectory. One the other hand, the MAD approach shows 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 inter-event time is greater than the window size. But these highly frequented oscillations are detected and revmoed properly by the MAD approach. 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. However, the time-window based method ignores the inherent structures in oscillation patterns, thus failing to remove part of the highly frequented oscillations.

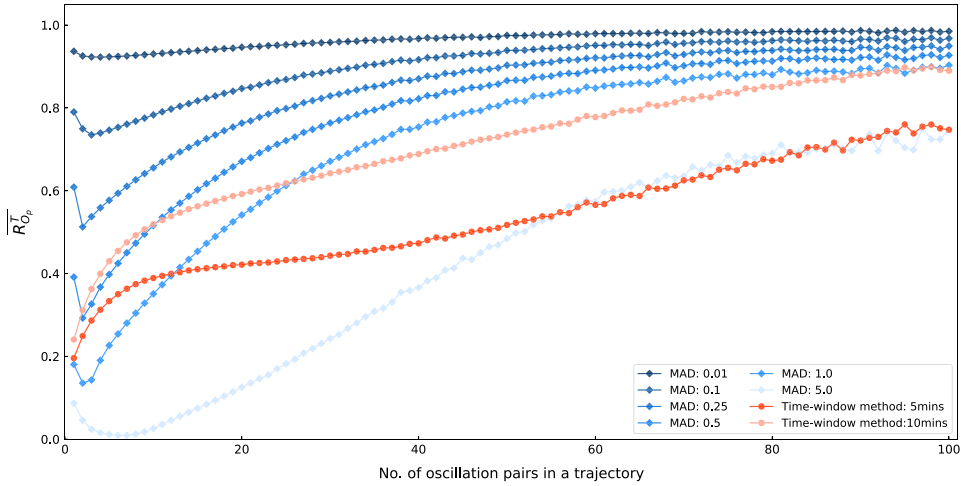

Figure 11: Average detection ratio of trajectories ($\overline{R^T_{op}}$) of the two methods under different parameter settings. Horizon axis denotes total number of oscillation pairs in a user daily trajectory.

5.3 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 daily trajectory, we collect the outputs generated from them, i.e., 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 daily trajectory, we first derive all the stay activities with a duration above a threshold, e.g., 10 minutes. Then, origin-destination trips are derived from consecutive stays. Note that if two consecutive stays for an individual corresponds to the same location, we don't 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 $OD_{clustering}$, $OD_{timewin}$, 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 behaviour studies to detect travelers' meaningful stays (Alexander et al. 2015; Jiang et al. 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-min or 30-min threshold). A large value indicates that the user's daily activities tend to distribute across a variety of activity locations. We compare this number across the three methods, denoted as $A_{clustering}$, $A_{timewin}$, and A_{mad}, respectively.

3. *Total stay time* is defined as the total amount of time that a user stays across all the activity locations, i.e., $\sum duration(l_i)$. 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 $S_{clustering}$, $S_{timewin}$, 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 $H_{clustering}$, $H_{timewin}$, 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. Performing the time-window based method, as shown in Figure 12B, results in a slight increase of 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., $OD_{timewin} - OD_{clustering} = 0$) for 96.6 percent of the trajectories, while for the rest, the time-window based method always gives a higher estimation.

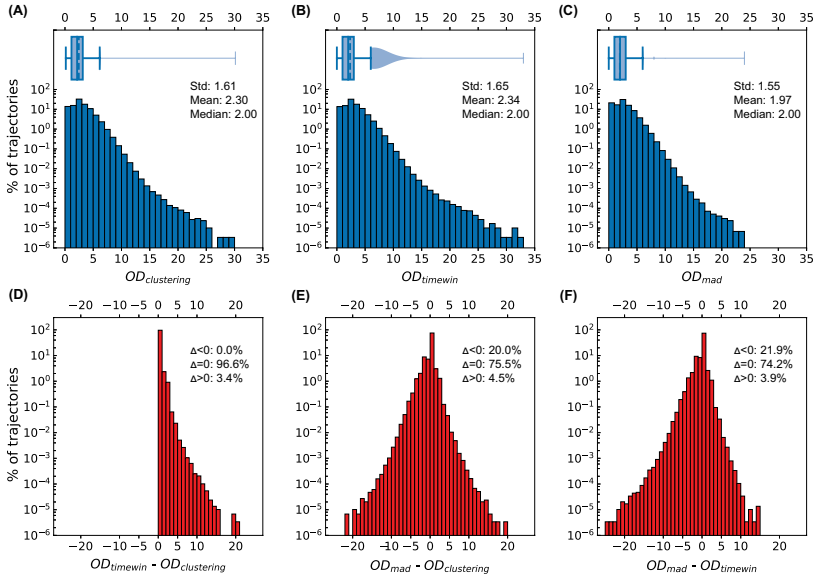

Figure 12: (A-C) Distribution of $OD_{clustering}$, $OD_{timewin}$, and OD_{mad}; (D-F) Pair-wise comparison of the three methods. OD trips are generated based on stay duration of 10 minutes.

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, however, the MAD approach gives higher estimations for 4.5 percent of the trajectories, but lower estimations for 20.0 percent of the cases. Similar conclusion can be made 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, 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 Appendix). In sum, if viewing result of two-stage clustering as the baseline, the MAD approach tends to produce lower estimations of OD trips, while the time-window based method has a negligible impact. This is because for the MAD approach, certain cells are merged towards the "focal point" during oscillation removal. In other words, the movements that are part of these likely oscillations are not considered as valid OD trips. The result suggests that in what 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 a mean 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). However, the results of time-window based method and two-stage algorithm closely resemble each other (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 down-sampling process of mobile phone trajectories. (Readers could refer to Figure A.2 in Appendix for comparative results based on stay duration of 30 minutes).

Regarding the total stay time, the output of two-stage clustering produces a mean and median value of 831.05 and 835.65 minutes, respectively (Figure 14A). The time-window based method produces slightly larger values (Figure 14B). However, the MAD approach 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). Since the MAD approach is able to detect highly frequented oscillations, when these oscillations are removed, or more precisely speaking, merged towards the focal points, the observed stay time at these focal points will increase. These focal points, which

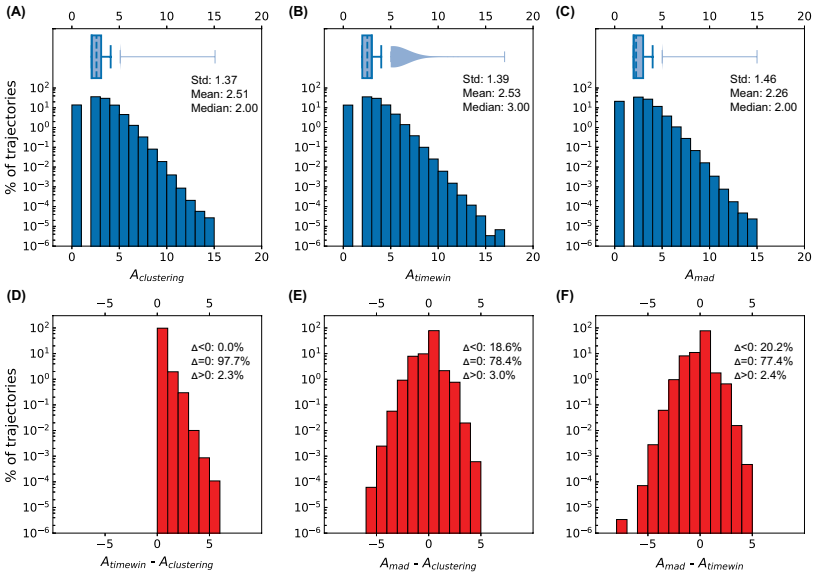

Figure 13: (A-C) Distribution of $A_{clustering}$, $A_{timewin}$, and A_{mad}; (D-F) Pair-wise comparison of the three methods. Results are generated based on stay duration of 10 minutes.

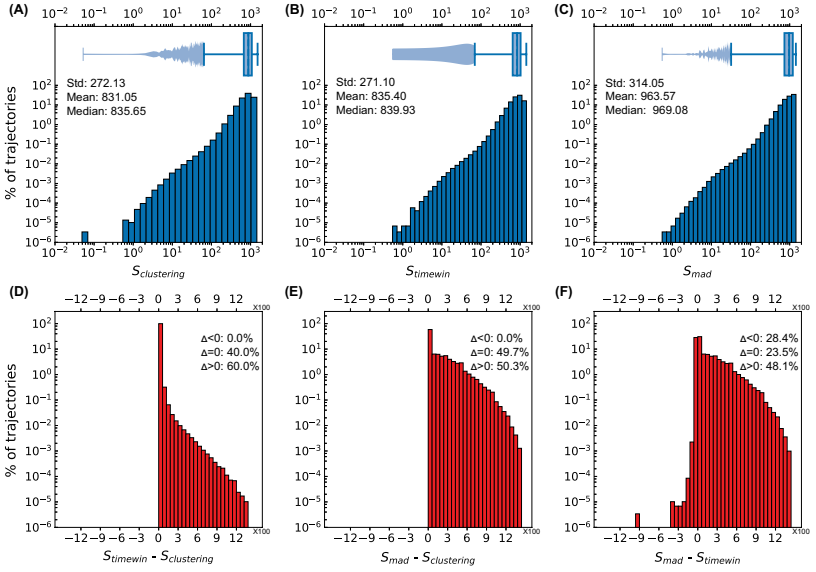

Figure 14: (A-C) Distribution of $S_{clustering}$, $S_{timewin}$, and S_{mad}; (D-F) Pair-wise comparison of the three methods.

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 while the distributions of the other two methods are relatively more similar.

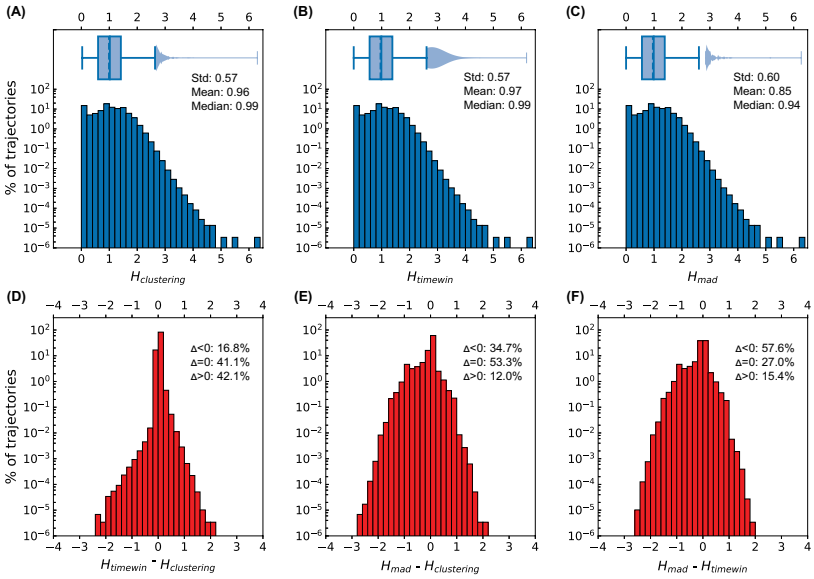

Figure 15: (A-C) Distribution of $H_{clustering}$, $H_{timewin}$, and H_{mad}; (D-F) Pair-wise comparison of the three methods.

6 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 the validity of research that involves the usage of big data. Without paying attention to this issue, the knowledge generated from the data, as claimed by (Kwan 2016), will have the risk of becoming artifact of the algorithms used. In this study, we aim to reflect on this issue through the analysis of a large-scale mobile phone dataset. Our results demonstrate that the choice of data preprocessing methods could lead to changes of the data characteristics. Such changes, which are non-trivial, will further affect the characterization of human mobility patterns.

By applying a two-stage clustering algorithm over the mobile signaling dataset (MSD), we highlight the effectiveness of this step in tackling the location uncertainty issues. Meanwhile, we find that some issues, primarily the celltower 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 of 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 af-

ter 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 oscillations, we find that the oscillation issue is partially addressed. Despite 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 inter-event time is larger than the window size. This makes the time-window based method problematic. Since the inter-event time of oscillations vary notably from each other (Figure 7), the time-window based method, depending on the choice of window size, becomes a down-sampling process that removes oscillations in a somewhat random way.

We then propose an approach based on the notion of mean absolute deviation (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 repeat 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 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. However, the optimal value or range of ΔM should be further evaluated when ground truth data of human movements is available. We believe that the choice of ΔM is jointly affected by the spatial distribution of cellphone towers in the study area as well as the characteristics of mobile signaling data. Testing the proposed approach across different datasets 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 that each cellphone tower is associated with oscillations (using $\Delta M = 0.5$). We summarize such information at the level of 1km*1km grid. As shown in Figure 16, the oscillations are observed more frequently

in the core part of Shanghai, which is also the area where cellphone 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 makings on urban design and management are frequently needed (e.g., infrastructure investment, disease control, transport planning).

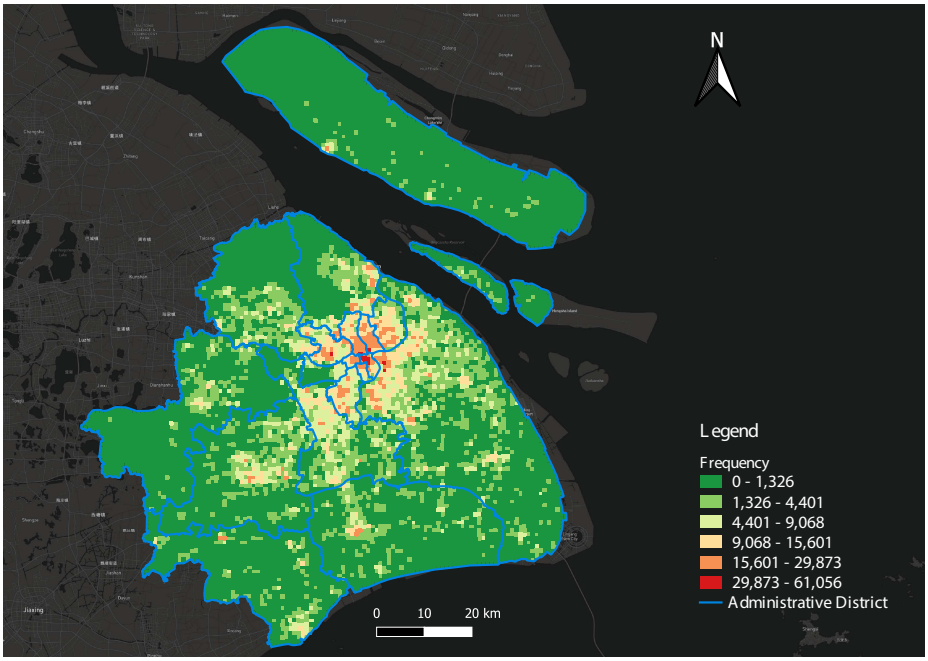

Figure 16: The number of occurrences that a cellphone tower is associated with oscillations (using $\Delta M = 0.5$). The numbers are aggregated and summarized at the level of 1km*1km grid.

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, 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 the 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 "down-sampling" nature of time-window based method when it is used to detect oscillations. Second, performing 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 into 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 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 impacts geographic knowledge discovery is case-dependent. For instance, when estimating OD trips from mobile phone data, different methods might produce quite different estimations at individual level. But when these estimations are further aggregated — 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. However, this is not always the case. Recently, many studies have performed big data analytics to understand the relationship between mobility patterns and socioeconomic status of travellers (Xu et al. 2018; Pappalardo et al. 2016; Smith-Clarke et al. 2014; Frias-Martinez et al. 2013; Almaatouq et al. 2016; Blumenstock et al. 2015; Wu et al. 2019). In these studies, a collection of individual mobility (and/or sociality) indicators are derived and used to correlate with, or to predict personal socioeconomic status. In this case, the variations of 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), the ways 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*[2]. 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 the proposed method in this research — once calibrated when ground truth becomes available — could help improve the estimations of human mobility patterns (e.g., origin-destination trips, daily activity locations, and dwelling time at these locations) to support applications in transportation planning and location-based services. Since 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 et al. 2010)), the proposed method is useful to other datasets 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. However, it is possible that some oscillations removed by the MAD approach refer to the actual movements of phone users, while some others that are not removed could be "fake" movements. A dataset 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, as 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 dataset. How the findings would generalize in a broader sense is worth further investigations. In this study, we already reveal the varying impacts of preprocessing methods on data characteristics. We believe this veracity issue

[2]https://sgsup.asu.edu/sparc/RRWorkshop

is not unique to the dataset used in this study, but exists in other mobile phone datasets. In the future, we intend to enlarge the research scope by repeating the experiments over multiple datasets and across different study areas. Nevertheless, we hope 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 to geographic knowledge discovery.

Funding

This research is 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.

References

L. Alexander, S. Jiang, M. Murga, and M. C. González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation research part c: emerging technologies*, 58:240–250, 2015.

A. Almaatouq, F. Prieto-Castrillo, and A. Pentland. Mobile communication signatures of unemployment. In *International conference on social informatics*, pages 407–418. Springer, 2016.

M. A. Bayir, M. Demirbas, and N. Eagle. Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing*, 6(4):435–454, 2010.

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

A. Birenboim and N. Shoval. Mobility research in the age of the smartphone. *Annals of the American Association of Geographers*, 106(2):283–291, 2016.

V. D. Blondel, A. Decuyper, and G. Krings. A survey of results on mobile phone datasets analysis. *EPJ data science*, 4(1):10, 2015.

J. Blumenstock, G. Cadamuro, and R. On. Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076, 2015.

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

F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. 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, 2011a.

F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti. Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PloS one*, 6(7): e20814, 2011b.

F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira Jr, and C. Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, 26:301–313, 2013.

C. Chen, H. Gong, C. Lawson, and E. Bialostozky. 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):830–840, 2010.

C. Chen, L. Bian, and J. Ma. From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C: Emerging Technologies*, 46:326–337, 2014.

C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, 68:285–299, 2016.

E. Cho, S. A. Myers, and J. Leskovec. 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*, pages 1082–1090. ACM, 2011.

B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel. Exploring the mobility of mobile phone users. *Physica A: statistical mechanics and its applications*, 392(6):1459–1473, 2013.

Z. Fan, T. Pei, T. Ma, Y. Du, C. Song, Z. Liu, and C. Zhou. Estimation of urban crowd flux based on mobile phone location data: A case study of beijing, china. *Computers, Environment and Urban Systems*, 69:114–123, 2018.

V. Frias-Martinez, C. Soguero-Ruiz, E. Frias-Martinez, and M. Josephidou. Forecasting socioeconomic trends with cell phone records. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, page 15. ACM, 2013.

S. Gao, Y. Liu, Y. Wang, and X. Ma. Discovering spatial interaction communities from mobile phone d ata. *Transactions in GIS*, 17(3):463–481, 2013.

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

M. F. Goodchild. Uncertainty: the achilles heel of gis. *Geo Info Systems*, 8(11):50–52, 1998.

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

S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252. Acm, 2012.

A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2551–2572, 2015.

S. Jiang, Y. Yang, S. Gupta, D. Veneziano, S. Athavale, and M. C. González. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113(37):E5370–E5378, 2016.

S. Jiang, J. Ferreira, and M. C. González. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3 (2):208–219, 2017.

M.-P. Kwan. Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge. *Annals of the American Association of Geographers*, 106 (2):274–282, 2016.

J.-K. Lee and J. C. Hou. 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*, pages 85–96, 2006.

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

Z. Li, L. Yu, Y. Gao, Y. Wu, G. Song, and D. Gong. Identifying temporal and spatial characteristics of residents' trips from cellular signaling data: Case study of beijing. *Transportation research record*, page 0361198118793495, 2018.

R. B. McMaster and E. L. Usery. *A research agenda for geographic information science*. crc Press, 2004.

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

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

C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, 33(5):727–748, 2006.

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

S. Silm and R. Ahas. 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):542–559, 2014.

C. Smith-Clarke, A. Mashhadi, and L. Capra. Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 511–520. ACM, 2014.

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

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

P. R. Stopher and S. P. Greaves. Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, 41(5):367–381, 2007.

J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, 2015.

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

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. Acm, 2011.

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

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

L. Wu, L. Yang, Z. Huang, Y. Wang, Y. Chai, X. Peng, and Y. Liu. Inferring demographics from human trajectories and geographical context. *Computers, Environment and Urban Systems*, 77:101368, 2019.

W. Wu, Y. Wang, J. B. Gomes, D. T. Anh, S. Antonatos, M. Xue, P. Yang, G. E. Yap, X. Li, S. Krishnaswamy, et al. Oscillation resolution for mobile phone cellular tower data to enable mobility modelling. In *2014 IEEE 15th International Conference on Mobile Data Management*, volume 1, pages 321–328. IEEE, 2014.

Y. Xu, S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation*, 42(4):625–646, 2015.

Y. Xu, S.-L. Shaw, Z. Zhao, L. Yin, F. Lu, J. Chen, Z. Fang, and Q. Li. 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, 2016.

Y. Xu, A. Belyi, I. Bojic, and C. Ratti. How friends share urban space: An exploratory spatiotemporal analysis using mobile phone data. *Transactions in GIS*, 21(3):468–487, 2017.

Y. Xu, A. Belyi, I. Bojic, and C. Ratti. Human mobility and socioeconomic status: Analysis of singapore and boston. *Computers, Environment and Urban Systems*, 72:51–67, 2018.

Y. Xu, A. Belyi, P. Santi, and C. Ratti. Quantifying segregation in an integrated urban physical-social space. *Journal of the Royal Society Interface*, 16(160):20190536, 2019.

L. Yan, D. Wang, S. Zhang, and D. Xie. Evaluating the multi-scale patterns of jobs-residence balance and commuting time–cost using cellular signaling data: a case study in shanghai. *Transportation*, pages 1–16, 2018.

F. Yang, P. J. Jin, Y. Cheng, J. Zhang, and B. Ran. Origin-destination estimation for non-commuting trips using location-based social networking data. *International Journal of Sustainable Transportation*, 9(8):551–564, 2015.

Y. Yuan and M. Raubal. 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):1594–1621, 2016.

Z. Zhao, S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9):1738–1762, 2016.

Author Biography

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. Email: 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 spatio-temporal 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 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. E-mail: luf@lreis.ac.cn. His re-

search interests involve spatial data modeling, trajectory data mining, complex network analysis, knowledge graph, 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. E-mail: yinling@siat.ac.cn. Her research interests include human dynamics, spatial epidemic models, space-time GIS, and GIS for transportation.

BIYU 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 geographic information systems (GIS) for transportation, transport geography, and spatiotemporal big data analytics.

Appendices

A Impact of three methods on mobility estimations

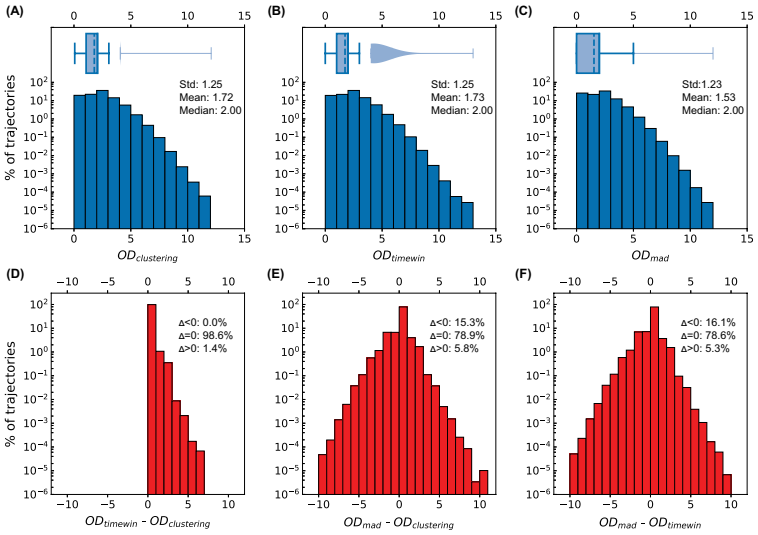

Figure A.1: (A-C) Distribution of $OD_{clustering}$, $OD_{timewin}$, and OD_{mad}; (D-F) Pair-wise comparison of the three methods. OD trips are generated based on stay duration of 30 minutes.

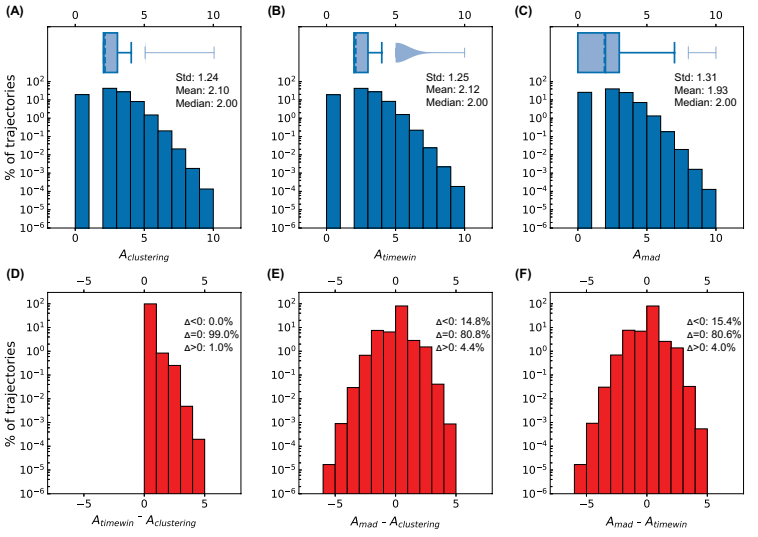

Figure A.2: (A-C) Distribution of $A_{clustering}$, $A_{timewin}$, and A_{mad}; (D-F) Pair-wise comparison of the three methods. Results are generated based on stay duration of 30 minutes.