Transportation Research Part C 87 (2018) 58–74

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

On data processing required to derive mobility patterns from passively-generated mobile phone data

Feilong Wang, Cynthia Chen*

Department of Civil and Environmental Engineering, University of Washington, Seattle, United States

ARTICLE INFO

Keywords:
Human mobility trajectory
Locational uncertainty
Oscillation problem
Representativeness issue
Incremental clustering method
Time-window-based method
Statistical regularity

ABSTRACT

Passively-generated mobile phone data is emerging as a potential data source for transportation research and applications. Despite the large amount of studies based on the mobile phone data, only a few have reported the properties of such data, and documented how they have processed the data. In this paper, we describe two types of common mobile phone data: Call Details Record (CDR) data and sightings data, and propose a data processing framework and the associated algorithms to address two key issues associated with the sightings data: locational uncertainty and oscillation. We show the effectiveness of our proposed methods in addressing these two issues compared to the state of art algorithms in the field. We also demonstrate that without proper processing applied to the data, the statistical regularity of human mobility patterns—a key, significant trait identified for human mobility—is over-estimated. We hope this study will stimulate more studies in examining the properties of such data and developing methods to address them. Though not as glamorous as those directly deriving insights on mobility patterns (such as statistical regularity), understanding properties of such data and developing methods to address them is a fundamental research topic on which important insights are derived on mobility patterns.

1. Introduction

In the past decade, passively-generated mobile phone data (hereafter called “mobile phone data”) has risen as a viable data source for human mobility analysis and transportation applications. It is the by-product of the primary, often non-transportation related purposes such as billing and operations (Alexander et al., 2015a; Chen et al., 2014, 2016; Iqbal et al., 2014; Toole et al., 2015). Often location- and time-stamped, mobile phone data offer a tremendous opportunity to revolutionize the transportation field that goes beyond behavior analysis, traffic operations and safety analysis (Vlahogianni et al., 2015). Indeed, a search in the literature using keyword combinations of mobile phone data, mobility, and travel behavior resulted in more than 1000 articles published in journals across different disciplines (Ahas et al., 2010a; Becker et al., 2013; Calabrese et al., 2013; Candia et al., 2008; Chen et al., 2014, 2016; Gao et al., 2013; Song et al., 2010b; Wang et al., 2014). These articles cover a wide range of topics including, for example, estimating mobility patterns (Csáji et al., 2013; González et al., 2008; Song et al., 2010a), inferring OD matrix (Calabrese et al., 2011b; Iqbal et al., 2014), finding anchor locations (Dong et al., 2015; Isaacman et al., 2011), inferring activity types (Jiang et al., 2017; Widhalm et al., 2015) and travel modes (Qu et al., 2015; Wang et al., 2010).

Unlike household travel survey data that is generated by asking subjects record an entire day’s activities and travels, thus resulting in complete information on one’s spatio-temporal movement, mobile phone data generates one or more records when a user connects

* Corresponding author.
E-mail address: qzchen@u.washington.edu (C. Chen).

https://doi.org/10.1016/j.trc.2017.12.003
Received 18 March 2017; Received in revised form 19 November 2017; Accepted 7 December 2017

to a network (Chen et al., 2016; Wang et al., 2014). Each record typically includes the device ID, the timestamp and the location information. Two kinds of mobile phone data have been frequently used today: CDR (Call Detail Record) data and sightings data (Calabrese et al., 2013; Chen et al., 2014), where the former (CDR data) comprises a set of phone call records along with the time information and the location information of the cell tower that channels the call, and the sightings data contains sightings that are results of the triangulation of cell towers. Hereby we call either a record or a sighting as a trace. Neither data covers the full spectrum of a user's movement in a day. Since both rely on the phone usage, both are temporally sparse: for sightings data, the average time between two consecutive network-connection activities could be longer than one hour and for CDR data the interval could be even longer (Calabrese et al., 2013). This issue is in addition to their much coarser spatial resolution in location estimation compared with data from household travel surveys and GPS (Global Positioning System)-based surveys. The precision of location estimation is reported to be approximately 300 m on average in urban area for sightings data (Jiang et al., 2013) and between several hundred meters to several kilometers for CDR data, depending on the density of cellular towers (Chen et al., 2016; Järv et al., 2014). Another issue is purely signaling related: some traces are resulted from signaling activities, as opposed to users' movements. That is, when a user stays at one location, his/her phone may be handed over to different cell towers due to load-balancing or other operational purposes. This results in traces suggesting that the user bounces between multiple stations. This is known as oscillation (Lee and Hou, 2006; Qi et al., 2016).

Additionally, because a trace in mobile phone data represents either a phone call (for CDR data) or a sighting when a phone is seen on the network (for sightings data), a trace in the mobile phone data cannot be viewed directly as an activity location where a person spends time for an activity. Because of this and the previously noted issues such as uncertainty and oscillation, the mobile phone data must be processed first before it can be used for mobility analysis. The word "processed" here specificially refers to the process that must be applied to mobile phone data to address issues such as uncertainty and oscillation in order to infer activity and trip related information. This process is uniquely different from the cleaning process typically applied to household travel survey data whose records directly represent activities and trips conducted by the subjects. Clearly we do not need to emphasize the importance of this processing task: the results of processing directly affect the results of mobility analysis using such data (Horn et al., 2014). We demonstrate this in Section 5.3 of the paper.

Surprisingly, among the many studies that have used mobile phone data, only a few have reported the properties and issues associated with such data, and documented how they have processed the data and how the results compare to external data sources that may provide some validations (Palchykov et al., 2014; Serok and Blumenfeld-Lieberthal, 2015; Yin et al., 2017). This paper focuses on mobile phone sightings data, in particular, seeking to illustrate and address two prominent issues of the dataset: (1) locational uncertainty and (2) oscillation. The sightings data has now generated many papers in the field (Calabrese et al., 2011a, 2013; Jiang et al., 2013), is already being used by several planning agencies and departments of transportation in the US (Milone, 2015; Stabler and Sikder, 2014) and the data provider is now widely advertising its products (e.g., OD matrices) derived from such data (Milone, 2015). Thus, having a clear understanding of such data and the techniques that can be applied to process such data is not only important for researchers but also for practitioners who may be interested in using such data or their derived products. The first issue (locational uncertainty) is unique to the sightings data, as every location estimate contained in the data is unique, as the results of triangulation. Methods are designed to deal with this issue so that activity locations (where people spend time to conduct activities) can be identified from the sightings data. A common solution is to define a radius and a minimum duration corresponding to the positioning errors and the time spent on activities respectively (Jiang et al., 2013). Then an activity location is identified if a sequence of traces meets both the spatial and temporal constraints. The second issue—oscillation—exists in both sightings and CDR data. The phenomenon cannot be ignored as traces resulted from oscillation take up a substantial fraction of the total number of records: approximately 30% (Lee and Hou, 2006). Solutions, however, especially for sightings data, are lacking and are mostly discussed in computer science communities (Bayir et al., 2010; Lee and Hou, 2006; Shad et al., 2012). Two approaches have been commonly used to detect oscillation sequences in the traces: pattern-based and hybrid methods. The former examines trace sequences and the one that exhibits a specific switching pattern will be identified as the oscillation case (Lee and Hou, 2006). The latter extends pattern-based methods by utilizing temporal and/or spatial information to consider velocity or other measurements (Iovan et al., 2013). Our paper reviews these existing methods and proposes new solutions. We show that our proposed solutions exceed the performance of the existing methods in addressing these issues.

The rest of the paper is organized as follows. After reviewing related works in Section 2, Section 3 introduces the sightings data used in our study. In Section 4, we construct a data-processing framework that provides solutions for the two problems. The framework consists of two stages: the first is to address the issue of locational uncertainty by proposing a revised incremental clustering algorithm; and the second is to handle the oscillation problem with a time-window-based method in addition to a modified pattern-based method. Through analyzing the processed data, we show advantages and effectiveness of our framework. Section 5 covers this part of work. Lastly in Section 6, we provide a discussion on related questions that are not addressed within the scope of this paper. The representativeness issue, which is related to the issue of temporal sparsity, is discussed in this section. We also discuss how our developed framework and solutions can be applied to the CDR data. We hope our work will stimulate many more studies addressing this critical topic that will determine not only the accuracy of the results from using such data but also its wide applications in the future.

2. Related works

2.1. Understanding passively generated mobile phone data

As noted earlier, two types of mobile phone data have been commonly studied: Call Details Record (CDR) data (Ahas et al., 2010a;

Järv et al., 2014; Xu et al., 2015) and the sightings data(Calabrese et al., 2011b; Chen et al., 2016; Wang, 2014). In the former, each record represents a device-driven activity (e.g., phone calls, messaging services and Internet service) and contains information such as the initial time of the activity, the duration, and the ID and location information of the cell tower that channels the activity.

The sightings data differs with the CDR data in two aspects: (1) more records: in addition to just recording device-driven activities, data is also generated from other network-driven interactions, such as handovers when a device switches between two neighboring cell areas (Demissie et al., 2013; Hard et al., 2016; Iovan et al., 2013); and (2) higher spatial accuracy: instead of cell-level location information as in the CDR data, the location information contained in the sightings data is an estimation of the device's location through triangulation of multiple cells (Calabrese et al., 2011b; Chen et al., 2016).

Passively generated, both types are characterized with spatio-temporal uncertainties. From the temporal perspective, depending on device usage pattern, the temporal distribution of records can be irregular and heterogeneous (Iovan et al., 2013). Consequently, the time interval between two consecutive records is uncertain and could be too long to capture users' movements that took place during the long time gap. The interval for a sample CDR data is reported with a mean of as long as eight hours (González et al., 2008). And a median of slightly longer than one hour is reported for a sample sightings data (Calabrese et al., 2011b). One way to address the temporal issues is to apply a data-filtering process, such as only selecting highly active users (Calabrese et al., 2013; González et al., 2008; Song et al., 2010b; Zhao et al., 2016). However, results from those studies may be affected by the selected filtering process due to potential correlations between individual mobility and device usage, e.g., higher communication frequency is likely to lead to more trips derived (Iovan et al., 2013; Yuan et al., 2012). Though not directly addressed in the current study, we discuss it in Section 6.

From the spatial perspective, several uncertainty issues exist in the mobile phone data, which affect individual mobility derived from data. The first issue existing in both types of data is the relatively low "spatial accuracy". Depending on the density of cellular towers, the accuracy of location estimates varies. The accuracy is reported to be about 300 m on average in urban area for sightings data (Calabrese et al., 2011b; Hard et al., 2016; Jiang et al., 2013) and between several hundred meters to several kilometers for CDR data corresponding to the coverage of a tower/antenna (Järv et al., 2014). There are efforts to improve spatial accuracy, for example, via interpolation methods (Järv et al., 2017; Louail et al., 2014). The second uncertainty related issue that exists in both types of data is the oscillation phenomenon: due to dynamic factors including varying transmission conditions that influence received signal strength and load balancing policies, instead of connecting to the nearest cell tower, a mobile phone switches its connection between different towers (typically within a short period of time) even though the device itself is not moving (Calabrese et al., 2011b; Wu et al., 2014). The oscillation phenomenon generates a considerable number of records that do not reflect devices' actual movements (Lee and Hou, 2006). The third issue is specifically related to sightings data: location estimates for a stay at the same location change over time. This is related to the method of triangulation, as measurements on factors used for triangulation (e.g., number of surrounding cell towers, received signal strength) are subject to fluctuations, leading to distinct estimations for the same location (Calabrese et al., 2011a). Aggregation on these varying location estimates is required so that we can identify activity locations and calculate the time spent at those locations (Hard et al., 2016).

2.2. Addressing locational uncertainty by extracting activity locations

For CDR data with location information at the cell level, frequently visited locations (e.g. home, work places) by an individual are usually indicated by locations of network cells that were most frequently visited (Ahas et al., 2010b; Isaacman et al., 2011; Järv et al., 2014; Wang et al., 2012; Xu et al., 2016). By clustering cell towers that are spatially close, these locations could also be represented by clusters containing towers that frequently channel call activities (Jiang et al., 2017; Xu et al., 2016).

However, for sightings data, as the location estimation is the result of triangulation among multiple towers (Chen et al., 2016; Widhalm et al., 2015), each of them is unique. In other words, even when a user stays at the same location, locations recorded in the data differ from time to time (though they may scatter in close proximity) (Fig. 1). This means the methods for identifying activity locations from the CDR data, such as those relying on counting call activity frequency with cell towers, are not applicable for the sightings data. The common way to address this locational uncertainty and to reveal activity locations from the sightings data is to aggregate these distinct location estimates by applying a clustering algorithm (Calabrese et al., 2011b; Chen et al., 2016; Jiang et al., 2013). Typically, the centroid of the outputting clusters is used as representing the activity location.

Conventional clustering methods are usually not applicable mainly because parameters in these methods need to be pre-determined, which is infeasible in mobile phone data. For example, popular methods such as k-means and DBSCAN (Ester et al., 1996; Kanungo et al., 2002) require the number of clusters or density-related parameters be known, which cannot be specifically identified because of the variety of mobile phone usage and travel behaviors that exist across individuals (Chen et al., 2014).

Some new methods aggregate traces by segmenting one trajectory into several sequences of traces. Here, one trajectory of a user refers to the user's available traces of one day. In the trajectory-segmentation methods, an activity location is defined as a sequence of consecutive traces bounded by both temporal and spatial constraints (Calabrese et al., 2011a, 2013; Jiang et al., 2013). Given the trajectory $\{d_0, d_1, d_2, \ldots, d_k\}$, one cluster of traces $\{d_m, d_{m+1}, \ldots, d_n\}$ $(0 \le m, n \le k)$ will be defined as an activity location, if two conditions are met: (1) the maximal geographical distance between any two location estimates L_{di} and L_{dj} $(m \le i, j \le n)$ should not exceed R_c (here L_{di}, L_{dj} are location records of traces d_i and d_j respectively and R_c is the spatial constraint (e.g., 300 m); and (2) the time difference between the last and the first trace of this sequence (i.e. t_{dn}–t_{dm}) should be at least T_c (here t_{dn} and t_{dm} are timestamps of traces d_n and d_m respectively and T_c is the temporal constraint (e.g., 10 min)). Thus, in the trajectory-segmentation method, activity locations can be identified by applying these two conditions and segmenting the trajectories into sequences. This method is intuitive and easy to apply. But, the first condition, requiring any two traces in the sequence to satisfy the distance constraint, can

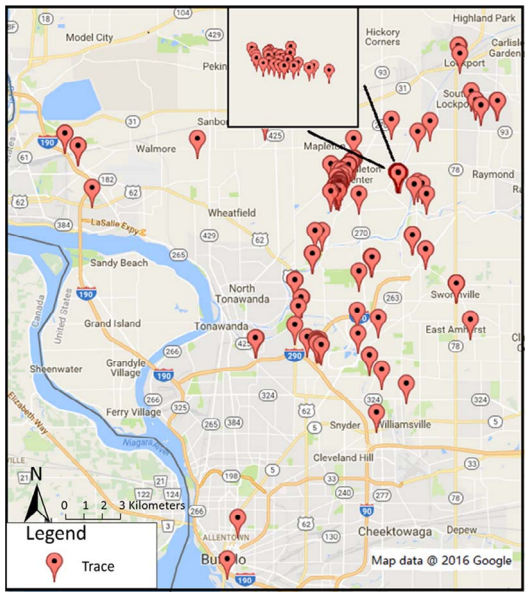

Fig. 1. Locational uncertainty. (The inset figure is the zoom-in of closely distributed location estimates of one activity location).

pose challenges if outliers exist. For example, given a sequence of traces $\{d_m, d_{m+1}, \ldots, d_n\}$ that are spatially close, the insertion of an outlier $d_{outlier}$ may result into three clusters: $\{d_m, d_{m+1}, \ldots, d_{m+k}\}$, $\{d_{outlier}\}$ and $\{d_{m+k+1}, \ldots, d_n\}$, consequently leading to biased estimation on the number and the types of activity places one has visited.

Chen et al. (2014) suggested a model-based clustering method. Rather than relying on a pre-determined threshold value for all devices, this method determines the optimal number of clusters for each device through a statistical model, more specifically, by searching through a finite set of possible numbers of clusters and finding the one that generates the highest Bayesian Information Criterion (BIC) (Fraley, 1998; Fraley and Raftery, 2002). The centroid of an identified activity cluster (defined as a sequence of traces that are considered to be activity places) is used as activity locations for subsequent modeling processes. The advantage of applying the model-based clustering is obvious: there is no requirement for pre-determined clustering parameters. However, there are limitations. This method is sensitive to the spatial density of the traces, which are often of low density. Additionally, the size of outputting clusters cannot be controlled, meaning that several far-away outliers may be clustered together, causing the resulting cluster to stray away (Fig. 2): in other words, a cluster without spatial constraint makes the clustering process less meaningful.

Another method to aggregate traces is the incremental clustering method (Alexander et al., 2015a; Hariharan and Toyama, 2004; Wang et al., 2015; Widhalm et al., 2015). Given traces $\{d_0, d_1, d_2, \ldots, d_k\}$ in one trajectory, the clustering is performed as follows: (1) starting from trace d_0, one new cluster C_0 is created and d_0 is the center; (2) each trace that is not clustered will be checked and the trace within a distance R_c to the center of C_0 is aggregated to the cluster; (3) every time the cluster grows, its center is updated; (4) if

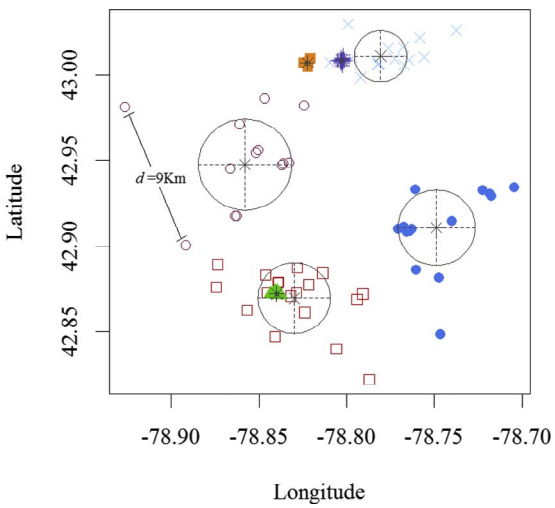

Fig. 2. A problematic classification example of model-based clustering. (Different clusters are shown in different symbols of different colors. The size of the ellipse overlaid on a cluster reflects the size of that cluster. Here locations as far as nine kilometers are aggregated into one cluster, making the clustering less meaningful.)

no trace could be aggregated in the current cluster, one new cluster will be created containing the non-clustered trace. This procedure repeats itself until all traces are clustered. After the clustering process, clusters with a time duration of no less than T_c will be considered as activity locations. The cluster duration is calculated as the largest time difference between any two traces in the cluster. This clustering method is relatively robust to outliers present among traces, as the mechanism of updating centers (step 3) improves its tolerance to outliers.

Most of above mentioned clustering methods are applied to individual days of traces. In data of multiple days, however, the same activity location may appear distinct in terms of their coordinates. This creates challenges for the identification of common activity locations, such as home and work locations (Xu et al., 2015), and observations on mobility patterns such as the regularity property (González et al., 2008). Therefore, it would be necessary to develop methods to identify common activity locations where users conduct activities on multiple days. By dividing the study area into small grids, activity locations on multiple days are regarded as a common location if they fall into the same grid (Andrienko et al., 2011; Zheng et al., 2010). Jiang et al. identified the common locations by aggregating activity locations of multiple days via the agglomerative clustering algorithm (Jiang et al., 2013).

2.3. Removing oscillation sequences

By applying the clustering algorithms, the issue of locational uncertainty is dealt with so that activity locations could be identified. However, the results can still be biased if traces generated from oscillations are not removed (Qi et al., 2016; Wu et al., 2014). When oscillation occurs, some devices may be observed switching between two (or more) far-away locations with high frequencies. In most cases, an oscillation pattern forms, such as L_0-L_1-L_0-L_1-L_0, where L_0 and L_1 are distinct location estimates. Table 1 provides an example oscillation case. The Euclidean distance between locations L_0 and L_1 is 2.7 km, but the user's location switches between them three times within four minutes. If the switching speeds are calculated, they are incredibly high, which should not reflect the actual movement of the user. Since oscillation is a common phenomenon, a data processing step must be applied.

A few heuristic rules were proposed for CDR data to detect oscillations among cellular towers. In general, they are based on the observation that oscillation generates trace sequences showing unique patterns (e.g., L_0–L_1–L_0–L_1) and usually lead to an illusion that users travel with an incredibly high speed (e.g. 1000 km/h) (Bayir et al., 2010; Lee and Hou, 2006; Shad et al., 2012; Yin et al., 2017).

Lee and Hou (2006) defined a sequence of consecutive location estimates with pattern L_0–L_1–L_0–L_1 or L_0–L_1–L_2–L_0 as an oscillation sequence. A similar method was introduced by Bayir et al. (2010). According to their definition, an oscillating sequence is detected if at least triple switches are found in one trip (One switch refers to a one-time switch between two locations). For example, the sequence of location estimates {L_0–L_1–L_0–L_2–L_3–L_2–L_1} is regarded as an oscillation sequence because of the triple switches between L_0 and L_1. This method was designed for a unique dataset collected by using an application installed on subjects' mobile phones and requesting subjects to put semantic tags to locations they visited. Shad et al. (2012) investigated the solution to oscillation using the same data. The similar pattern-based method they proposed relies on cellular towers' Location Area Code and radius information. For datasets where semantic tags and radius information are missing, it is hard to implement these methods. Meanwhile, relying on the unique patterns detected in the sequence of traces risks mistaking actual travels for occurrences of oscillation. In other words, an individual may indeed travel between two places multiple times in a typical day and traces generated from this actual travel behavior could be mistaken as oscillation traces.

A few methods have been designed to improve the detection precision by utilizing the temporal information available in the data (Iovan et al., 2013; Qi et al., 2016; Widhalm et al., 2015; Wu et al., 2014). In an oscillation sequence, the frequent switches between distinct locations give the illusion that people travel at an abnormally high speed. This problem is considered in the speed-based methods, where the switching speed is calculated. An oscillation sequence is identified if the switching speed exceeds a given threshold V_c. Iovan et al. (2013) set V_c as 200 km/h. In their work, they extended the definition with one more condition: for an oscillation case, the difference between the heading directions of two consecutive displacements should be 180 degrees, which corresponds to two switches (i.e. L_0–L_1–L_0) in the pattern-based methods. Thus, their work can be regarded as a hybrid method – a combination of pattern-based and speed-based method. But the criterion of two switches is too simple to detect complex patterns that involve more than two locations. Meanwhile, given the low accuracy of location estimation, the method involving the speed measurement could be misleading, especially if the time intervals of switches are small (Demissie et al., 2013).

Wu et al. elaborated several heuristic rules to detect oscillation traces (Wu et al., 2014). Working on CDR data, they begin with finding suspicious sequences that contain at least three traces from multiple cellular towers within a short time period (e.g., one minute). Then a suspicious sequence is confirmed as an oscillation sequence if it contains a circular event (i.e. a tour L_0–L_1–…–L_0). In

Table 1
An oscillation case.

Trace	Location	Time	Distance (km)	Switching speed (km/h)
d_0	L_0	12:21:48	\	\
d_1	L_1	12:22:01	2.7	748
d_2	L_0	12:25:20	2.7	49
d_3	L_1	12:25:39	2.7	512

(The switching speed of two consecutive traces is calculated as the division of the distance between two location estimates by the time interval of these two traces. A space-time visualization of this oscillation case is given in Fig. 8.)

this rule, the limit of a short time window is to account for the observation that oscillation is a short time event and the circular-event criterion is to rule out suspicious sequences resulting from fast movements.

Following the detection of oscillation sequences, a meaningful location needs to be identified to represent the device's location when oscillation occurs. This meaningful location could be the one, among all locations in the sequence, which is visited most frequently (Iovan et al., 2013). Wu et al. (2014) introduced a score for each location in the oscillation sequence by considering both the visiting frequency and the average distance to other locations. The one that receives the highest score is selected as the meaningful location.

Oscillation solutions for sightings data are limited (Alexander et al., 2015a; Dong et al., 2015; Palchykov et al., 2014; Yin et al., 2017). Calabrese et al. (2011b) and Widhalm et al. (2015) addressed two problems (both locational uncertainty and oscillation) simultaneously using clustering methods. Their clustering methods could eliminate some of oscillation traces, since oscillation traces falling into a same cluster are aggregated and replaced by the centroid of the cluster. The problem is: oscillation traces could be too faraway (e.g., several kilometers away) to be compressed into a constraint-specific cluster (see our results in Section 5.2). Additionally, Yin et al. (2017) mentioned the use of the pattern-based method to detect oscillation traces in their applied sightings data, but details (e.g., what parameters to use) are not present.

3. The sightings data

The mobile phone sightings data used in this study consist of sightings from about one million users in the Buffalo metropolitan area during the month of April 2014. Sightings are derived by an intermediate company from records generated by phone operators for operation and billing purposes so that phone users maintain connections to the network when needed (for calling, texting and activities that require accessing the internet) and are correctly billed. Each sighting reveals the trace of a user, containing an encrypted mobile identification number, a time stamp and a location estimate expressed in latitude and longitude (Table 2). The data include devices that were observed in the Buffalo area and served by the cellular network carriers contracted with the data provider in April 2014. According to the data provider, users in the data could be resident workers, home workers, inbound commuters, outbound commuters, short-term visitors, or long-term visitors, etc.

The location information of traces is estimated through the data provider's proprietary technology and its accuracy is reported to be about 300 m on average in the urban area. For each sighting, a variable called certainty radius is calculated, which is a measurement in meters representing the radius of a circular area where the actual location of the device can be found with a 90% probability. The center of the area gives the location estimate of each trace in the data in the format of latitude and longitude. Fig. 3 shows the spatial distribution of certainty radius. Area with smaller certainty radius is shown in lighter shade, indicating relatively higher spatial accuracy. We note that generally downtown area shows higher accuracy than the rest of the study area.

Fig. 4a shows the distribution of the number of days observed in the data. Half of the users are observed for fewer than seven days, and only 6797 users (less than 1% of all users) have at least one trace everyday in the entire month. As another test on the temporal sparsity of the data, we divide a day into 24 hourly slots and, for each trajectory of a device, we count, among 24 slots, the number of slots in which the device is sighted at least once (Widhalm et al., 2015) (Fig. 4b). The median is six slots, indicating that half of trajectories have no more than six hourly slots with locations revealed by the mobile phone network.

4. Methodology

4.1. An overview

In this section, we provide a framework to process sightings data. As noted earlier, the framework consists of two stages, corresponding to two issues of the data. In the first stage, the issue of locational uncertainty is addressed by aggregating locations into clusters using a revised incremental clustering algorithm. We show that the revised incremental clustering algorithm can maintain the spatial precision of the original data and is adaptive to the spatial density of traces in rural and urban area (see details in Section 5). In the second stage, oscillation sequences are detected by a time-window-based detection method and a modified pattern-based method,

Table 2
A sample of mobile phone sightings data.

ID	Time*	Location estimate**
36cc5**77ca	1,396,381,131	42.951554|−78.665607
36cc5**77ca	1,396,385,839	42.929892|−78.624486
36cc5**77ca	1,396,386,054	42.928625|−78.607994
36cc5**77ca	1,396,386,319	42.929892|−78.624436
36cc5**77ca	1,396,386,567	42.951554|−78.665607
36cc5**77ca	1,396,386,653	42.929892|−78.624436
36cc5**77ca	1,396,386,876	42.955625|−78.606594

* In Unix time – defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time, 1/1/1970.

** Location estimates in the table are modified to protect privacy. The same is applied as for Table 3.

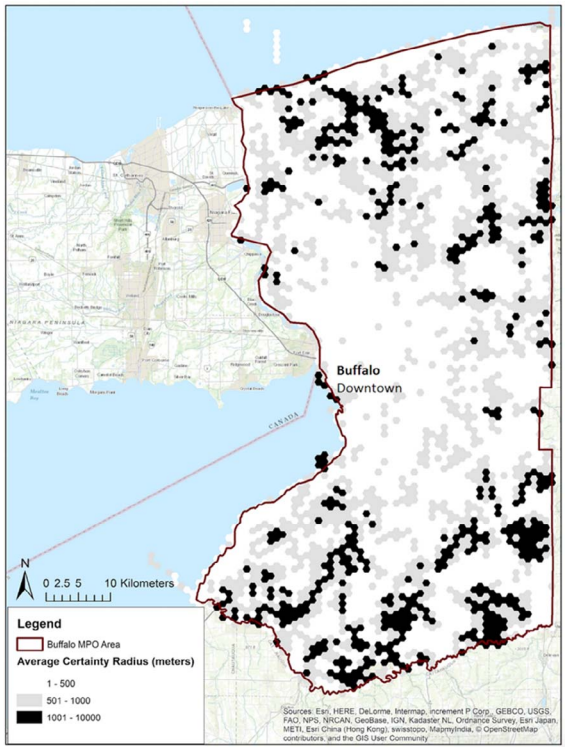

Fig. 3. Spatial distribution of certainty radius.

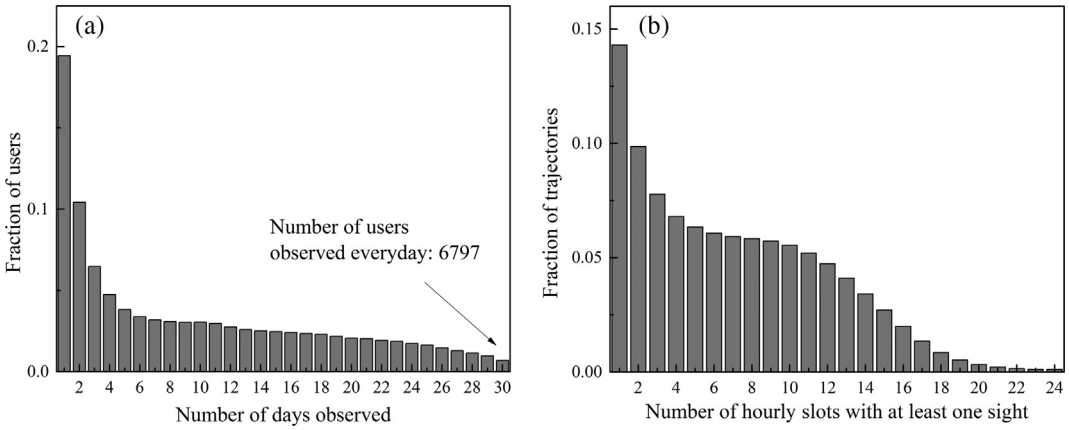

Fig. 4. (a) Distribution of the number of days observed. (b) Distribution of the number of hourly slots with at least one trace.

which are able to identify complex oscillation patterns and to reduce the risk of mistaking real trips. Here, the order of these two stages matters, which will be discussed at the end of this section. The framework is summarized in Fig. 5 and described in subsequent sections.

4.2. Addressing locational uncertainty toward extracting activity locations

Our method to deal with the locational uncertainty in the sightings data contains three steps: (1) applying the revised incremental clustering algorithm to aggregate traces; (2) finding a suitable parameter for the incremental clustering algorithm; and (3) applying the k-means clustering algorithm to deal with an order problem that is not addressed by the incremental clustering algorithm (see details below).

4.2.1. Incremental clustering algorithm

We first identify activity locations. As noted earlier, in sightings data, location estimates for a single activity location could be

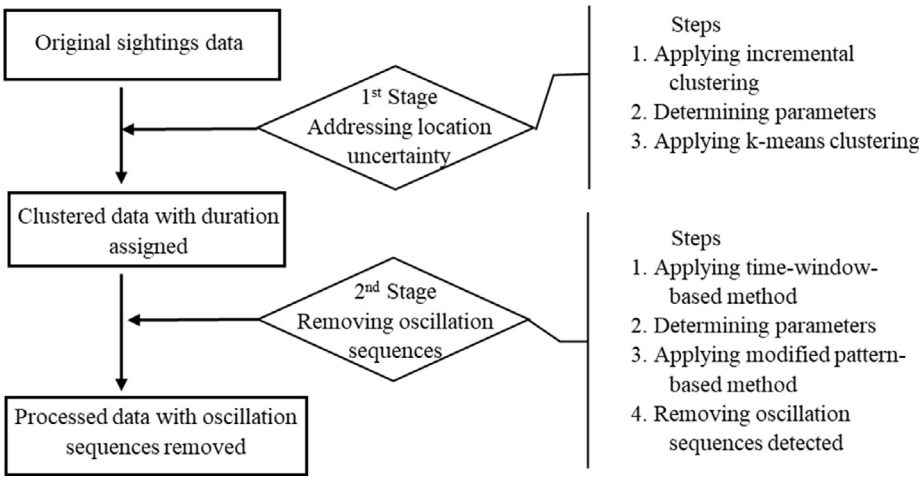

Fig. 5. Framework of processing sightings data.

unique because of the locational uncertainty. Thus, duration at a location cannot be directly inferred. In principle, these location estimates shall be closely distributed in space. Therefore, we could aggregate these closely distributed locations and represent them with one cluster so that the duration associated with each potential activity location could be calculated.

The other issue caused by the locational uncertainty is that the common activity location that a user visits on different days may appear distinct in their expressed format of longitude and latitude. If we put together all the available location estimates of a potential activity location, we will find that they scatter closely in a cluster as well. Thus, common activity locations could be revealed by putting traces of multiple days together and clustering them without regarding their time ordering.

In this study, we develop a revised incremental clustering algorithm to aggregate traces of multiple days for each user without considering their time ordering. This method enables us to identify common activity locations by aggregating traces that are close in space but may be far away in time (e.g., several days). Details of this algorithm are provided in Fig. 6.

4.2.2. Determining a suitable R_c via trial and error

The incremental clustering method requires a spatial constraint R_c as an input. Without information to choose a suitable R_c to cluster our data, we find it via trial and error. More specifically, we try different values of R_c and find the suitable one by counting the total number of activity locations identified in each trial and analyzing its relationship with R_c. To do this, we first need to identify activity locations, which is described bellow.

(1) Duration and activity locations Given a setting of R_c (e.g., 0.5 km), traces are aggregated into clusters without considering the temporal information. To determine the duration, we use individual trajectories. Through scanning every trajectory, we find

```
Input: multiple-days traces of one user d_list;
Output: A set of clusters of traces;
Steps:
  gather traces of multiple days to d_list;
  C_set ← {};
  create the first new cluster C_new and add any one trace d_0 to C_new;
  add C_new to C_set;
  C_current ← C_new;
  for other traces d_i in d_list not belonging to any cluster:
    if distance(d_i, C_current) < R_c:
      add d_i to C_current;
    else:
      C_current ← None;
      for all cluster C in C_set:
        if distance(d_i, C) < R_c:
          add d_i to C;
          C_current ← C;
          break;
      if C_current =None:
        create new cluster C_new;
        add C_new to C_set;
        C_current ← C_new;
  output C_set;
```

Fig. 6. Revised incremental clustering algorithm.

Table 3
A sample of data after clustering.

ID	Initial time	Duration (second)	Location estimate
36cc5**77ca	1,396,377,202	10,246	42.949429|−78.670337
36cc5**77ca	1,396,388,619	26	42.972285|−78.776146
36cc5**77ca	1,396,388,890	134	42.998380|−78.791654
36cc5**77ca	1,396,389,805	73	43.039335|−78.860423
36cc5**77ca	1,396,392,466	1310	42.949429|−78.670337

consecutive traces that belong to the same cluster and aggregate them into a newly created trace, which contains several fields (Table 3): the original user ID; the initial time given by the timestamp of the first trace of the cluster; the duration that is calculated as the difference of the first and last timestamp in the cluster; and the location information that is given by the centroid of the cluster. Then, an activity location is identified if its duration exceeds T_c. We set T_c as five minutes. This five-minute threshold also follows the rule used in many household travel surveys to define what counts an activity (Transportation Research Board, 2005). Additionally, it is an appropriate threshold for an activity location in the activity based modeling context (Yin et al., 2017).

(2) Determining a suitable R_c In Fig. 7, we show the average number of distinct activity locations visited per day per person n_{AL} as a function of R_c. If R_c is small, one actual activity location may be split into several clusters. With increasing R_c, small clusters start to merge into meaningful representations of actual activity locations. Fig. 7 suggests the suitable R_c appears at 1 km. Recall that the identification of activity locations is subject to both spatial and temporal constraints. Above 1 km, it is possible that some clusters may absorb far-away outliers (passing-by points), therefore lengthen their duration and meet the temporal constraint (i.e. 5 min), consequently leading to the growth of the number of activity locations. This is witnessed by the slight increase of n_{AL} and by the steep rise of cluster duration (per cluster) when R_c is above 1 km (Fig. 7).

4.2.3. K-means clustering algorithm

We notice an order problem that could not be handled by the incremental clustering algorithm alone—clustering results are subject to the order that how traces are clustered, which can result in unreasonable clusters. Fig. 8a gives one example, where fourteen traces are to be clustered. Following the incremental clustering algorithm, a new cluster is created at L_0, and consequently we obtain two strange clusters: one consists of locations $\{L_0, L_1, L_2, L_3, L_4\}$ and the other one contains other locations. For this case, however, a more reasonable output might be: one cluster contains L_0 alone and the other one compasses the remaining.

We apply the k-means algorithm to address this problem. This algorithm enables us to aggregate traces into k clusters in which each trace belongs to the cluster with the nearest distance to the centroid. We address the above order problem by initializing the k-means clustering using the results yielded from the incremental clustering algorithm. More specifically, the number of resulting clusters and their centroids are required for the initialization (Kanungo et al., 2002). In other words, the above order problem is resolved by applying the k-means clustering algorithm to revise the clustering results from the incremental clustering algorithm. We notice that the clustering error in the example is corrected (Fig. 8b). Following this revision, the duration and centroid of clusters are updated.

4.3. Removing oscillation sequences

In the second stage of data processing, we remove oscillation sequences. Our solution consists of four steps: (1) detecting oscillation sequences using a time-window-based method; (2) determining the suitable time window (required input parameter of previous step); (3) applying a modified pattern-based method to detect oscillation sequences with low switching frequency; and (4) removing oscillation sequences detected.

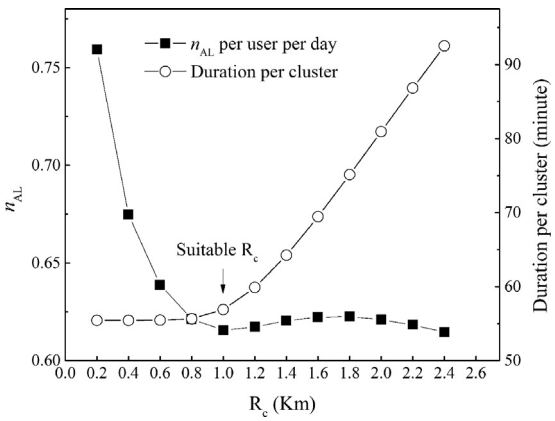

Fig. 7. Number of distinct activity locations per day and associated duration as a function of R_c.

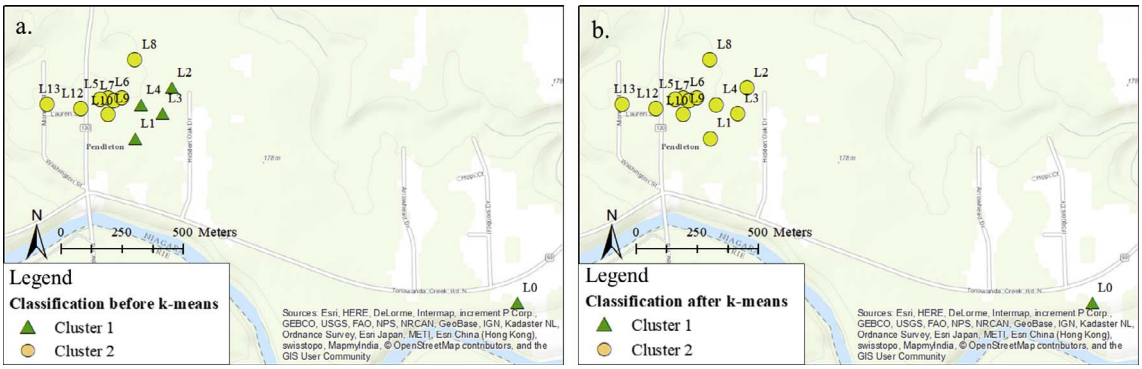

Fig. 8. Illustration of the order problem and the solution. (a) Classification without k-means algorithm. (b) Resulting classification after implementing k-means algorithm. (Following the incremental clustering algorithm, we obtain two strange clusters: one in triangle and the other one in circle. After applying k-means clustering, we have a more reasonable output.)

4.3.1. Time-window-based method

Inspired by the method proposed by Wu et al. (2014), we develop a time-window-based method to detect oscillation sequences. It scans trajectories with a short time window T_w (defined in the following section), which always starts at the ending time of a trace d_0. Here, the ending time of a trace is the time given by the sum of the initial time and duration of the trace (Table 3). A sequence of traces returns after the time window is applied (note that this sequence may contain a single trace). Among the sequences returned, we find the suspicious sequence which includes more than one distinct location estimates. And then oscillation sequences are confirmed as those containing at least one circular event. A circular event refers to a tour that one device is initially found at location L_0, later goes somewhere else (e.g., L_1, L_2), and at last returns to location L_0. Since the time window is defined quite short (e.g., several minutes), it is less likely for anyone to make a tour within such a short time period. Therefore, we consider that the oscillation sequences detected contain traces resulted from oscillation phenomenon. Fig. 9 illustrates how this time-window-based method works. The oscillation case shown in Table 1 is detected, because a circular event L_0-L_1- L_0 is found in the sequence $\{d_0, d_1, d_2, d_3\}$, which is returned after the time-window is applied.

4.3.2. Determining a suitable T_w via trial and error

To determine a suitable T_w, we test a set of time-windows of different lengths. Then we calculate the average oscillation ratio as a function of T_w (Fig. 10), where the oscillation ratio for a user is defined as the ratio of the number of detected oscillation traces over the total number of traces associated with that user. Fig. 10 shows that when T_w is small, the average oscillation ratio increases sharply; but beyond a certain value, the tangent of the curve remains stable, indicating five minutes as a reasonable choice to separate oscillation cases from real trips. More specifically, five-minute is chosen as a longer T_w does not yield much better performance on detecting oscillation traces. Instead, a longer T_w may be at the risk of mistaking more actual trips for oscillations. When T_w is five minutes, the estimated average oscillation ratio is 0.17, suggesting that on average approximately 17% of traces are generated due to

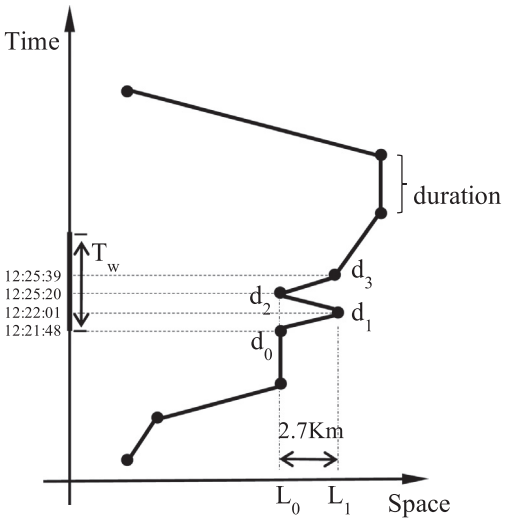

Fig. 9. Illustration of time-window-based detection method. (This figure corresponds to the oscillation case in Table 1. The implementation of the time-window returns a sequence of traces $\{d_0, d_1, d_2, d_3\}$, containing the circular event (L_0–L_1–L_0)).

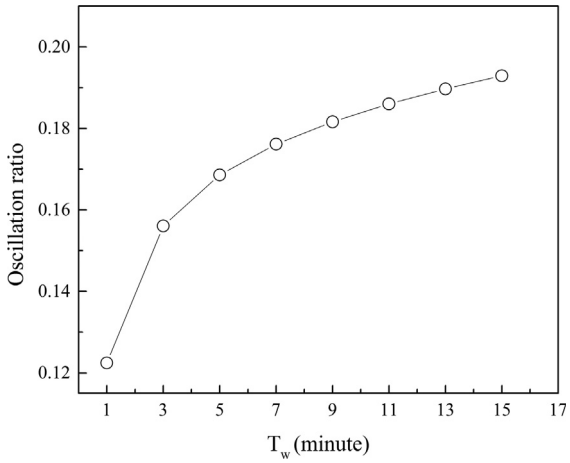

Fig. 10. Average oscillation ratio as a function of T_w.

pure signaling activities.

4.3.3. Modified pattern-based method

The time-window-based method is effective in identifying most oscillation sequences. But, we found that due to temporal sparsity, the time window may be too short to capture the circular event even though the oscillation occurs. Thus, we further process the result using a modified pattern-based method. We first detect sequences with a twice switches pattern, i.e. L_0–L_1–L_0–L_1. As the pattern-based method alone may mistake actual trips for oscillation, for each sequence detected, we look for additional evidence: only if a time interval, among all time intervals between any two consecutive traces in the sequence, is found shorter than T_w, we identify it as an oscillation sequence. In Section 5, we show that the simple two-switch criterion is effective, though more complex ones could be designed.

As the oscillation detection methods rely on identifying circular events, it is necessary to address the issue of locational uncertainty first. Without addressing the problem of locational uncertainty, location estimates in sightings data may be unique from time to time, preventing us from identifying circular events and thus the detection of oscillation.

4.3.4. Removing oscillation sequences detected

For each oscillation sequence detected, we remove oscillation traces by finding and replacing them with a meaningful location. An oscillation sequence may reoccur at different times of a day and/or on different days. If the meaningful location is determined from each oscillation sequence independently (Bayir et al., 2010; Wu et al., 2014), we probably yield different results for the same oscillation sequence that appears at different time. For example, we may obtain two different meaningful locations L_0 and L_1 for one oscillation sequence {L_0–L_1–L_0…} that occurs twice in one day (e.g., in the morning and evening). To address this issue, we determine the meaningful location by examining the traces of the entire month. Specifically, the meaningful location is identified as the one where the user spends most of time during the entire month. The time spent at one location is computed by summing up durations that is associated with the location (Table 3).

5. Results

5.1. Locational uncertainty

We first evaluate the effectiveness of our framework in terms of addressing the issue of locational uncertainty by comparing resulting clusters with those using commonly used agglomerative clustering method (Hariharan and Toyama, 2004; Jiang et al., 2013). Specifically, the radius of outputting clusters is computed and compared. Since location estimates belonging to one cluster are replaced by the cluster's centroid, the radius of the cluster could be a measure of the precision of new location estimates: the new location records generated from large clusters have lower precision. Fig. 11 shows the distribution of the radius of clusters produced by different methods using the same spatial constraint (R_c = 1 km). We notice that comparing with the uniform distribution of the radius of clusters obtained by agglomerative clustering, the radius of clusters resulting from our method is more heterogeneously distributed. In addition, approximately 80% of these clusters have a radius no longer than 500 m (compared with 50% in the output of agglomerative clustering method), meaning that our method performs better in terms of preserving location resolution while conducting the clustering to extract activity locations.

Interestingly, compared with the agglomerative clustering that outputs clusters with a clear cutoff on their size, our method produces some clusters whose size exceeds R_c (Fig. 11). This suggests that the clustering method proposed here is also adaptive to address cases where traces belonging to the same cluster can scatter far from each other, exceeding the given spatial constraint R_c. Note that this adaptive feature could be useful if traces of one activity location scatter broadly (i.e. the associated locational

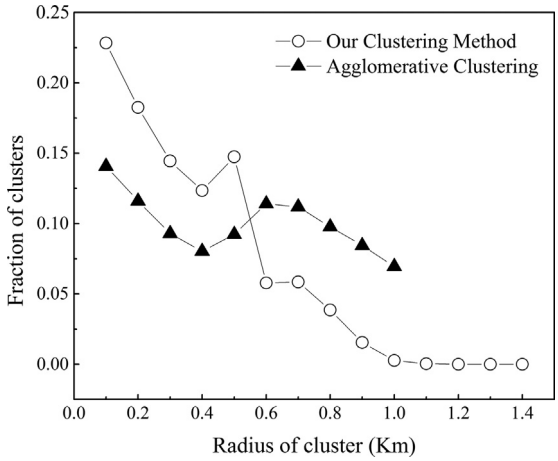

Fig. 11. Distribution of the radius of clusters.

uncertainty is large) because of the low density of cellular towers that are involved in estimating locations.

5.2. Oscillation

We then evaluate our solution to addressing the oscillation phenomenon. Since the oscillation sequences give us the illusion that a user travels back and forth within a short time, we may overestimate the trip rates of users whose records include oscillation traces. Here, one trip is defined as a sequence of traces starting and ending at two consecutive activity locations. Without removing oscillation sequences, we notice a large number of trajectories with very high trip rates. The Buffalo travel survey suggests that about 95% of population conduct no more than eight trips a day (Greater Buffalo-Niagara Transportation Survey, 2002). We thus assume if a trajectory includes more than eight trips, it is identified as a potentially abnormal trajectory because of possible presence of oscillation traces. We found that before removing oscillation sequences, the number of abnormal trajectories is high – about 12% of the total number of trajectories. After oscillation traces are removed, this value drops to about 0.7%. This resulting value is much smaller than 5% from the travel survey data, which could be attributed to the missing short-distance trips that are shorter than R_c and the related representativeness issues that are not dealt with in this paper but discussed in Section 6.

In Fig. 12, we plot the percentage of abnormal trajectories as a function of T_w. Before and after the implementing the modified pattern-based method, the substantial drop in the number of abnormal trajectories testifies the effectiveness of the method.

The time-window-based method is capable of identifying oscillation sequences with complex patterns that involve more than two locations. For example, it can detect oscillation sequences with pattern L_0–L_1–L_2–L_3–L_0–L_1–L_2–L_1–L_0, meaning a device is observed switching among four locations. Fig. 13a gives the distribution of different oscillation patterns this method detected in the data. Here, n locations involved in oscillation means devices are observed oscillating among n locations. We see that though in most of cases oscillation only involves two locations, mobile phone data contains complex oscillation patterns and they could be detected with our framework. Specifically, we notice that approximately 3% of oscillation sequences involve more than four locations. Meanwhile, Fig. 13b shows the distribution of distance between oscillation traces (detected using our method), indicating that some oscillation traces are farther apart than 1 km. This suggests that, as we noted earlier in Section 2.3, a clustering method with a specific spatial constraint may fail to detect faraway oscillation traces.

5.3. Trajectory

As a part of our evaluation, in Fig. 14 we visualize two trajectories of one user before and after the data is processed. The trajectories on the left are drawn from the original data and the right ones are from the processed data. We note that common places visited in multiple days are revealed and the trajectories become smoother after we remove oscillation sequences.

Previous studies show that individuals' trajectories are characterized by statistical regularity: preferable returns to a few locations and occasional explorations to other faraway locations (González et al., 2008; Pappalardo et al., 2015; Song et al., 2010a; Susilo and Axhausen, 2014). As the measurement of mobility regularity, the visiting frequency f of the k-th most visited locations is observed following a scaling law: $f_k \sim k^{-\xi}$. Here, as the indicator of mobility regularity, a larger exponential factor ξ suggests a higher level of regularity: less exploration into new places and more returns to familiar locations. The mobility analysis in the previous study, which is based on sightings data as well, shows $\xi = 1.04$ (Jiang et al., 2013). In Fig. 15, we test this scaling property with the datasets before and after removing oscillation sequences. We notice that though the regularity of mobility pattern could be revealed from both datasets, the exponential factors are different: the data processed by the first stage of our framework only (i.e. clustering algorithms) yields $\xi = 1.02$, which is consistent with Jiang et al.'s study (2013); while after removing the oscillation sequences, ξ decreases significantly to 0.88. Therefore, the oscillation phenomenon leads to the non-negligible overestimation of the regularity of individuals' mobility.

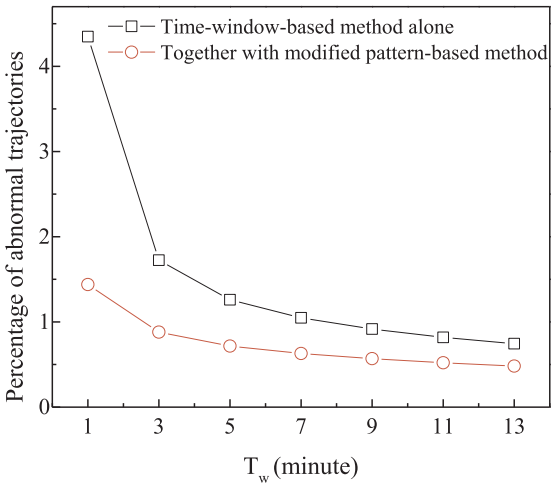

Fig. 12. Percentage of abnormal trajectories as a function of T_w.

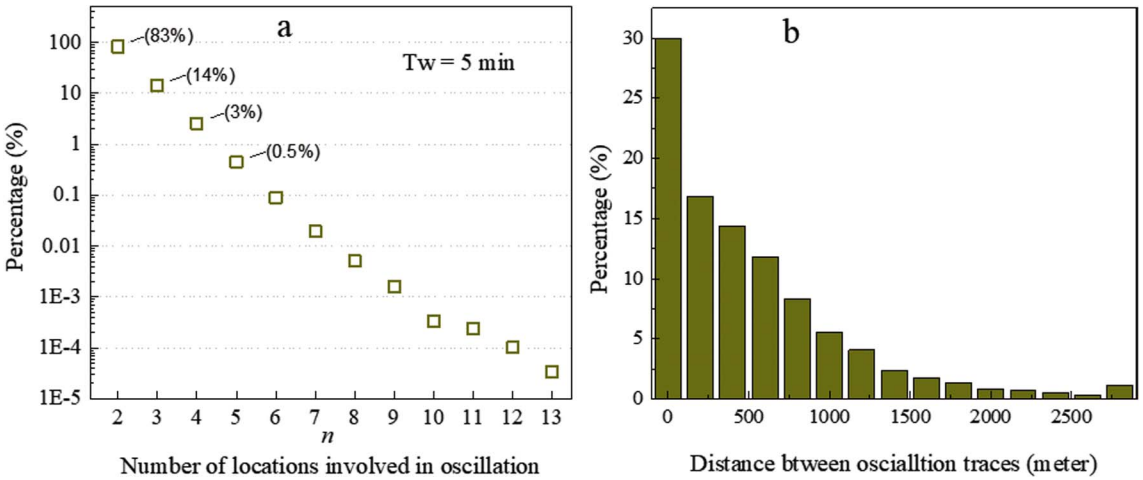

Fig. 13. (a) Distribution of different oscillation patterns detected. (b) The distribution of distances between oscillation traces detected (T_w = 5 min).

6. Discussions

As noted earlier, despite the rapid surge of studies using mobile phone data, our understanding on the nature of the data is very limited—few describes its issues and documents the procedures applied to such data to derive mobility trajectories. This paper is to fill this gap. Our study proposes a data-processing framework to specifically address two problems in the mobile phone sightings data: locational uncertainty and oscillation. For each problem, we extend existing works and propose our algorithms. These algorithms are simple to implement. We illustrate the effectiveness of the proposed framework by comparing to the results using the state of the art algorithms. As shown in Fig. 15, without applying those data processing steps, a fundamental characteristic of human mobility patterns—regularity (González et al., 2008; Jiang et al., 2013; Song et al., 2010a)—is overestimated. Chen et al. (2016) reviewed two lines of research for human mobility research—traditional travel behavior analysis using household travel surveys and human mobility analysis using big data such as mobile phone data. They noted that though regularity and variety are found in both lines of works, variety is much more pronounced in those studies using household travel surveys. Fig. 15 suggests one potential underlying cause of this discrepancy between the two lines of research may be due to our lack of understanding of the nature of mobile phone data and insufficient processing on them.

One known issue of mobile phone data that is not addressed in the paper relates to representativeness. Put it more plainly, since users in a mobile phone data are not the result of a probabilistic sample (as for household travel surveys) and all users voluntarily select into being the subscribers to certain service networks, there is an issue of selection bias (Arai et al., 2015; Wesolowski et al., 2013). We do not address it here as it itself is a complicated issue and addressing it at requires a complete separate study. We briefly discuss a number of complications here. First, the underlying population of mobile phone data is often unknown, as the data provider is often unwilling to release such information. Second, there is lack of ground truth information, an issue related to the unknown underlying population. Because of these two reasons, the usual means of validating against external, ground truth data is infeasible

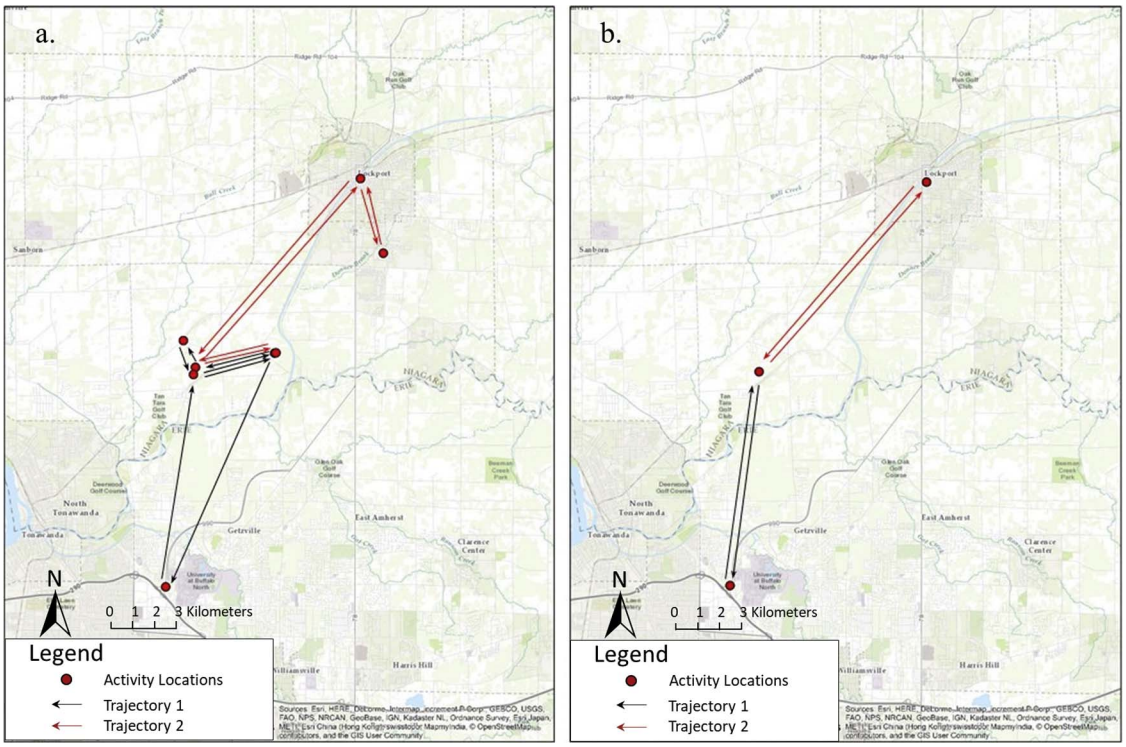

Fig. 14. Trajectories of one sample user. (a) before the data processing and (b) after the data processing. (We show two trajectories in different colors. For a clear observation, only traces with a duration (i.e. *duration* > 0) are shown here.)

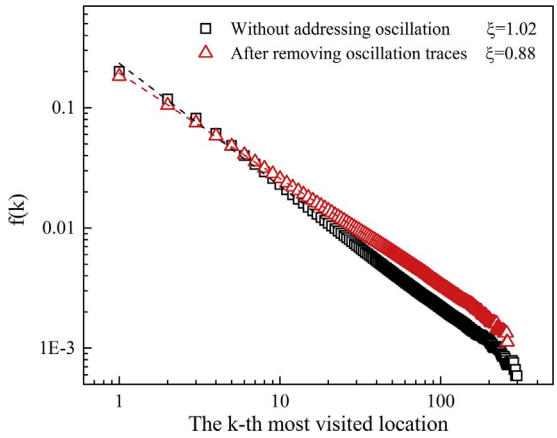

Fig. 15. The location visiting frequency (T_w = 5 min).

(Chen et al., 2014) and new methods and metrics must be developed for validation purposes. Third, there is a need to uncouple the strong correlation between trip rates and communication frequencies: more connections to communication networks, more sightings resulted and thus more trips derived (Iovan et al., 2013; Yuan et al., 2012). This correlation also suggests that for those studies relying only on active users, their results are likely affected by this issue (Alexander et al., 2015b; Wang et al., 2010). Lastly, because of the temporal sparsity and spatial uncertainty naturally associated with mobile phone data, there is a need for new thinking and new interpretations on the scale and meanings of representativeness (Lu et al., 2017; Zhao et al., 2016). In other words, a day of trajectory from mobile phone data may not be comparable to a trajectory derived from a household travel survey; rather, to be comparable to the latter, it may take trajectories of multiple days. Taking it further, since a key advantage of mobile phone data is its longitudinal nature, new metrics may be developed to capture how representative mobility patterns evolve over time (Chikaraishi et al., 2009, 2011).

Though the reported data-processing framework is developed for the sightings data, some of the steps within the framework can also be applied to the CDR data. As noted earlier and described in more detail in Chen et al. (2016), the location estimates in the CDR data are those of cell towers, rather than distinct coordinates for the devices in sightings data. This suggests that the clustering step as

described in our paper (Section 4.3.2) is not needed. However, retaining locations at the cell tower level may miss trips within a single cell area. In this study, after clustering the sightings data, the duration of clusters is then computed. This duration represents the time a user spends at the location. For CDR data, one would find the estimation of the activity duration is similar to the one in sightings data. The second stage of our framework (detecting and removing oscillation traces) can be applied to CDR data directly, as CDR data also has the oscillation issue. Given that the CDR data is temporally sparser than the sightings data (Chen et al., 2016), there may be more difficulty in differentiating oscillations from actual movements, as the detection of oscillations relies on the temporal distribution of traces. Therefore, a careful selection of time-window T_w is required.

It is our hope that the current study will spark many more studies discussing the nature of the data and its issues, documenting and reporting the various data processing procedures that are used, and addressing the representativeness issue. It is undoubted that such studies are of great importance—they are the foundations of those resulting trajectories on which new discoveries are made. This is particularly true if one is looking to use such data to inform our transportation policies.

Acknowledgement

The study is supported by the funding provided by the U.S. National Institute of Health (NIH) (1R01GM108731-01A1). The authors are grateful to Dr. Jae Hyun Lee's help in generating Fig. 3 as well as his participation in related discussions. Comments from three anonymous reviewers also greatly strengthen our paper. Views expressed in the paper do not represent those of NIH and the authors are responsible for all errors that may remain.

References

Ahas, R., Aasa, A., Silm, S., Tiru, M., 2010a. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. Transp. Res. Part C Emerg. Technol. 18, 45–54. http://dx.doi.org/10.1016/j.trc.2009.04.011.

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010b. Using mobile positioning data to model locations meaningful to users of mobile phones. J. Urban Technol. 17, 3–27. http://dx.doi.org/10.1080/10630731003597306.

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015a. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol. 58, 240–250. http://dx.doi.org/10.1016/j.trc.2015.02.018.

Alexander, L., Jiang, S., Murga, M., González, M.C., 2015b. Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp. Res. Part C Emerg. Technol., Big Data Transport. Traffic Eng. 58 (Part B), 240–250. http://dx.doi.org/10.1016/j.trc.2015.02.018.

Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S., Wrobel, S., 2011. From movement tracks through events to places: extracting and characterizing significant places from mobility data, in: Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on. IEEE, pp. 161–170.

Arai, A., Witayangkurn, A., Horanont, T., Shao, X., Shibasaki, R., 2015. Understanding the unobservable population in call detail records through analysis of mobile phone user calling behavior: a case study of greater Dhaka in Bangladesh. In: 2015 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 207–214. http://dx.doi.org/10.1109/PERCOM.2015.7146530.

Bayir, M.A., Demirbas, M., Eagle, N., 2010. Mobility profiler: a framework for discovering mobility profiles of cell phone users. Pervasive Mob. Comput. 6, 435–454. http://dx.doi.org/10.1016/j.pmcj.2010.01.003.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C., 2013. Human mobility characterization from cellular network data. Commun. ACM 56, 74–82. http://dx.doi.org/10.1145/2398356.2398375.

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C., 2011a. Real-time urban monitoring using cell phones: a case study in Rome. IEEE Trans. Intell. Transp. Syst. 12, 141–151. http://dx.doi.org/10.1109/TITS.2010.2074196.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Ratti, C., 2013. Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp. Res. Part C Emerg. Technol. 26, 301–313. http://dx.doi.org/10.1016/j.trc.2012.09.009.

Calabrese, F., Lorenzo, G.D., Liu, L., Ratti, C., 2011b. Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput. 10, 36–44. http://dx.doi.org/10.1109/MPRV.2011.41.

Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. J. Phys. Math. Theor. 41, 224015. http://dx.doi.org/10.1088/1751-8113/41/22/224015.

Chen, C., Bian, L., Ma, J., 2014. From traces to trajectories: how well can we guess activity locations from mobile phone traces? Transp. Res. Part C Emerg. Technol. 46, 326–337.

Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. Part C Emerg. Technol. 68, 285–299. http://dx.doi.org/10.1016/j.trc.2016.04.005.

Chikaraishi, M., Fujiwara, A., Zhang, J., Axhausen, K., 2009. Exploring variation properties of departure time choice behavior by using multilevel analysis approach. Transp. Res. Rec. J. Transp. Res. Board 2134, 10–20. http://dx.doi.org/10.3141/2134-02.

Chikaraishi, M., Fujiwara, A., Zhang, J., Axhausen, K., Zumkeller, D., 2011. Changes in variations of travel time expenditure. Transp. Res. Rec. J. Transp. Res. Board 2230, 121–131. http://dx.doi.org/10.3141/2230-14.

Csáji, B.C., Browet, A., Traag, V.A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., Blondel, V.D., 2013. Exploring the mobility of mobile phone users. Phys. Stat. Mech. Appl. 392, 1459–1473. http://dx.doi.org/10.1016/j.physa.2012.11.040.

Demissie, M.G., de Almeida Correia, G.H., Bento, C., 2013. Intelligent road traffic status detection system through cellular networks handover information: an exploratory study. Transp. Res. Part C Emerg. Technol. 32, 76–88. http://dx.doi.org/10.1016/j.trc.2013.03.010.

Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., Zhou, X., 2015. Traffic zone division based on big data from mobile phone base stations. Transp. Res. Part C Emerg. Technol., Big Data Transport. Traffic Eng. 58 (Part B), 278–291. http://dx.doi.org/10.1016/j.trc.2015.06.007.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 226–231.

Fraley, C., 1998. Algorithms for model-based Gaussian hierarchical clustering. SIAM J. Sci. Comput. 20, 270–281.

Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97, 611–631. http://dx.doi.org/10.1198/016214502760047131.

Gao, S., Wang, Y., Gao, Y., Liu, Y., 2013. Understanding urban traffic-flow characteristics: a rethinking of betweenness centrality. Environ. Plan. B Plan. Des. 40, 135–153. http://dx.doi.org/10.1068/b38141.

González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782. http://dx.doi.org/10.1038/nature06958.

Greater Buffalo-Niagara Transportation Survey 2002, 2003. < https://gbnrtc-mpo-osl2.squarespace.com/s/GBNRTC_Onboard_Transit_Study.pdf > .

Hard, E., Byron, C., Praprut, S., Steve, F., Darrell, B., Lisa, G., 2016. Synopsis of New Methods and Technologies to Collect Origin-Destination (O-D) Data (No. FHWA-HEP-16-083). Federal Highway Administration report FHWA-OR-15-01. < https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/ > .

Hariharan, R., Toyama, K., 2004. Project Lachesis: Parsing and Modeling Location Histories. In: Egenhofer, M.J., Freksa, C., Miller, H.J. (Eds.), Geographic Information

Science, Lecture Notes in Computer Science. In: Presented at the International Conference on Geographic Information Science, Springer, Berlin, Heidelberg. pp. 106–124. http://dx.doi.org/10.1007/978-3-540-30231-5_8.
Horn, C., Klampfl, S., Cik, M., Reiter, T., 2014. Detecting outliers in cell phone data. Transp. Res. Rec. J. Transp. Res. Board 2405, 49–56. http://dx.doi.org/10.3141/2405-07.
Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., Smoreda, Z., 2013. Moving and calling: mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In: Vandenbroucke, D., Bucher, B., Crompvoets, J. (Eds.), Geographic Information Science at the Heart of Europe. Springer International Publishing, Cham, pp. 247–265.
Iqbal, M.S., Choudhury, C.F., Wang, P., González, M.C., 2014. Development of origin–destination matrices using mobile phone call data. Transp. Res. Part C Emerg. Technol. 40, 63–74. http://dx.doi.org/10.1016/j.trc.2014.01.002.
Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., Varshavsky, A., 2011. Identifying important places in people's lives from cellular network data. In: International Conference on Pervasive Computing. Springer, pp. 133–151.
Järv, O., Ahas, R., Witlox, F., 2014. Understanding monthly variability in human activity spaces: a twelve-month study using mobile phone call detail records. Transp. Res. Part C Emerg. Technol. 38, 122–135. http://dx.doi.org/10.1016/j.trc.2013.11.003.
Järv, O., Tenkanen, H., Toivonen, T., 2017. Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. Int. J. Geogr. Inf. Sci. 31, 1630–1651. http://dx.doi.org/10.1080/13658816.2017.1287369.
Jiang, S., Ferreira, J., Gonzalez, M.C., 2017. Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans. Big Data. pp. 1–1. http://dx.doi.org/10.1109/TBDATA.2016.2631141.
Jiang, S., Fiore, G.A., Yang, Y., Ferreira, J., Jr., Frazzoli, E., González, M.C., 2013. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, UrbComp '13. ACM, New York, NY, USA, p. 2:1–2:9. http://dx.doi.org/10.1145/2505821.2505828.
Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24, 881–892. http://dx.doi.org/10.1109/TPAMI.2002.1017616.
Lee, J.-K., Hou, J.C., 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, MobiHoc '06. ACM, New York, NY, USA. pp. 85–96. http://dx.doi.org/10.1145/1132905.1132915.
Louail, T., Lenormand, M., Ros, O.G.C., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2014. From mobile phone data to the spatial structure of cities. Sci. Rep. 4, 5276. http://dx.doi.org/10.1038/srep05276.
Lu, S., Fang, Z., Zhang, X., Shaw, S.-L., Yin, L., Zhao, Z., Yang, X., 2017. Understanding the representativeness of mobile phone location data in characterizing human mobility indicators. ISPRS Int. J. Geo-Inf. 6, 7. http://dx.doi.org/10.3390/ijgi6010007.
Milone, R., 2015. Preliminary Evaluation of Cellular Origin-Destination Data as a Basis for Forecasting Non-Resident Travel. Presented at 15th TRB National Transportation Planning Applications Conferrence. Metropolitan Washington Council of Governments. < https://www.trbappcon.org/2015conf/Program.aspx > .
Palchykov, V., Mitrović, M., Jo, H.-H., Saramäki, J., Pan, R.K., 2014. Inferring human mobility using communication patterns. Sci. Rep. 4, 6174. http://dx.doi.org/10.1038/srep06174.
Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D., Giannotti, F., Barabási, A.-L., 2015. Returners and explorers dichotomy in human mobility. Nat. Commun. 6, 8166. http://dx.doi.org/10.1038/ncomms9166.
Qi, L., Qiao, Y., Abdesslem, F.B., Ma, Z., Yang, J., 2016. Oscillation resolution for massive cell phone traffic data. In: Proceedings of the First Workshop on Mobile Data, MobiData '16. ACM, New York, NY, USA. pp. 25–30. http://dx.doi.org/10.1145/2935755.2935759.
Qu, Y., Gong, H., Wang, P., 2015. Transportation mode split with mobile phone data. In: 2015 IEEE 18th International Conference on Intelligent Transportation Systems, pp. 285–289. http://dx.doi.org/10.1109/ITSC.2015.56.
Serok, N., Blumenfeld-Lieberthal, E., 2015. A Simulation model for intra-urban movements. Plos One 10, e0132576. http://dx.doi.org/10.1371/journal.pone.0132576.
Shad, S.A., Chen, E., Bao, T., 2012. Cell oscillation resolution in mobility profile building. ArXiv Prepr. ArXiv12065795.
Song, C., Koren, T., Wang, P., Barabási, A.-L., 2010a. Modelling the scaling properties of human mobility. Nat. Phys. 6, 818–823. http://dx.doi.org/10.1038/nphys1760.
Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010b. Limits of predictability in human mobility. Science 327, 1018–1021. http://dx.doi.org/10.1126/science.1177170.
Stabler, B., Sikder, S., Development of the Idaho Statewide Travel Demand Model Trip Matrices Using Cell Phone OD Data and Origin Destination Matrix Estimation. 2014 TREC Friday Seminar Series. < https://pdxscholar.library.pdx.edu/trec_seminar/30/ > .
Susilo, Y.O., Axhausen, K.W., 2014. Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl-Hirschman Index. Transportation 41, 995–1011. http://dx.doi.org/10.1007/s11116-014-9519-4.
Toole, J.L., Colak, S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C., 2015. The path most traveled: Travel demand estimation using big data resources. Transp. Res. Part C Emerg. Technol., Big Data Transport. Traffic Eng. 58 (Part B), 162–177. http://dx.doi.org/10.1016/j.trc.2015.04.022.
Transportation Research Board, 2005. Does the Built Environment Influence Physical Activity?: Examining the Evidence – Special Report 282. Institute of Medicine of the National Academies.
Vlahogianni, E.I., Park, B.B., van Lint, J.W.C., 2015. Big data in transportation and traffic engineering. Transp. Res. Part C Emerg. Technol., Big Data Transport. Traffic Eng. 58 (Part B), 161. http://dx.doi.org/10.1016/j.trc.2015.08.006.
Wang, H., Calabrese, F., Lorenzo, G.D., Ratti, C., 2010. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), pp. 318–323. http://dx.doi.org/10.1109/ITSC.2010.5625188.
Wang, M., 2014. Understanding Activity Location Choice with Mobile Phone Data (Thesis). < https://digital.lib.washington.edu/researchworks/handle/1773/26523 > .
Wang, M., Chen, C., Ma, J., 2015. Time-of-day dependence of location variability: application of passively-generated mobile phone dataset. Presented at the Transportation Research Board 94th Annual Meeting.
Wang, P., Hunter, T., Bayen, A.M., Schechtner, K., González, M.C., 2012. Understanding road usage patterns in urban areas. Sci. Rep. 2. http://dx.doi.org/10.1038/srep01001.
Wang, T., Chen, C., Ma, J., 2014. Mobile phone data as an alternative data source for travel behavior studies. Presented at the Transportation Research Board 93rd Annual MeetingTransportation Research Board.
Wesolowski, A., Eagle, N., Noor, A.M., Snow, R.W., Buckee, C.O., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. J. R. Soc. Interface 10, 20120986. http://dx.doi.org/10.1098/rsif.2012.0986.
Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M.C., 2015. Discovering urban activity patterns in cell phone data. Transportation 42, 597–623. http://dx.doi.org/10.1007/s11116-015-9598-x.
Wu, W., Wang, Y., Gomes, J.B., Anh, D.T., Antonatos, S., Xue, M., Yang, P., Yap, G.E., Li, X., Krishnaswamy, S., Decraene, J., Nash, A.S., 2014. Oscillation resolution for mobile phone cellular tower data to enable mobility modelling. In: 2014 IEEE 15th International Conference on Mobile Data Management, pp. 321–328. http://dx.doi.org/10.1109/MDM.2014.46.
Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., Li, Q., 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. Transportation 42, 625–646. http://dx.doi.org/10.1007/s11116-015-9597-y.
Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Lu, F., Chen, J., Fang, Z., Li, Q., 2016. Another tale of two cities: understanding human activity space using actively tracked cellphone location data. Ann. Am. Assoc. Geogr. 106, 489–502. http://dx.doi.org/10.1080/00045608.2015.1120147.
Yin, M., Qiao, S.M., Feygin S, Paiement J.-F, Pozdnoukhov, A., 2017. A generative model of urban activities from cellular data. In: IEEE Transactions in ITS, MobiData '16. ACM, New York, NY, USA. pp. 25–30. http://dx.doi.org/10.1145/2935755.2935759.

Yuan, Y., Raubal, M., Liu, Y., 2012. Correlating mobile phone usage and travel behavior – a case study of Harbin, China. Comput. Environ. Urban Syst., Special Issue: Geoinform. 36, 118–130. http://dx.doi.org/10.1016/j.compenvurbsys.2011.07.003.

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., Yin, L., 2016. Understanding the bias of call detail records in human mobility research. Int. J. Geogr. Inf. Sci. 30, 1738–1762. http://dx.doi.org/10.1080/13658816.2015.1137298.

Zheng, V.W., Zheng, Y., Xie, X., Yang, Q., 2010. Collaborative location and activity recommendations with GPS history data. In: Proceedings of the 19th International Conference on World Wide Web, WWW '10. ACM, New York, NY, USA, pp. 1029–1038. http://dx.doi.org/10.1145/1772690.1772795.