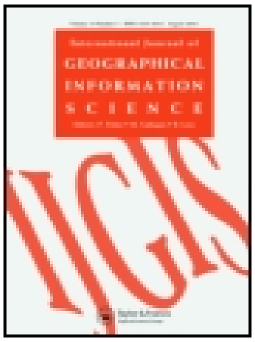


Taylor & Francis
Taylor & Francis Group

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

Understanding the bias of call detail records in human mobility research

Ziliang Zhao, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen & Ling Yin

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE, 2016
http://dx.doi.org/10.1080/13658816.2015.1137298

Understanding the bias of call detail records in human mobility research

Ziliang Zhao[a], Shih-Lung Shaw[a,b], Yang Xu[a], Feng Lu[c], Jie Chen[c] and Ling Yin[d]

[a]Department of Geography, University of Tennessee, Knoxville, TN, USA; [b]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China; [c]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; [d]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

ABSTRACT

In recent years, call detail records (CDRs) have been widely used in human mobility research. Although CDRs are originally collected for billing purposes, the vast amount of digital footprints generated by calling and texting activities provide useful insights into population movement. However, can we fully trust CDRs given the uneven distribution of people's phone communication activities in space and time? In this article, we investigate this issue using a mobile phone location dataset collected from over one million subscribers in Shanghai, China. It includes CDRs (~27%) plus other cellphone-related logs (e.g., tower pings, cellular handovers) generated in a workday. We extract all CDRs into a separate dataset in order to compare human mobility patterns derived from CDRs vs. from the complete dataset. From an individual perspective, the effectiveness of CDRs in estimating three frequently used mobility indicators is evaluated. We find that CDRs tend to underestimate the total travel distance and the movement entropy, while they can provide a good estimate to the radius of gyration. In addition, we observe that the level of deviation is related to the ratio of CDRs in an individual's trajectory. From a collective perspective, we compare the outcomes of these two datasets in terms of the distance decay effect and urban community detection. The major differences are closely related to the habit of mobile phone usage in space and time. We believe that the event-triggered nature of CDRs does introduce a certain degree of bias in human mobility research and we suggest that researchers use caution to interpret results derived from CDR data.

ARTICLE HISTORY
Received 21 July 2015
Accepted 21 December 2015

KEYWORDS
Mobile phone location data; call detail records; human mobility; urban dynamics

1. Introduction

The advent of so-called 'big data era' offers many new opportunities to study human mobility using various types of massive digital footprints, such as geo-tagged social media data (Batty 2010). Despite those exciting discoveries that reveal the pulse of the city, there have been debates regarding the biases that come with the data. For

CONTACT Shih-Lung Shaw ✉ sshaw@utk.edu 🖃 Department of Geography, University of Tennessee, 304 Burchfiel Geography Building, Knoxville, TN 37996, USA

instance, studies report that distribution of social media users is predominantly uneven in terms of geography, gender, and race/ethnicity (Mislove *et al*. 2011, Hecht and Stephens 2014).

Mobile phone location data, collected by mobile network operators (MNOs), have also been an appealing data source, given the unprecedented scale of digital footprints it offers. The type of mobile phone location data used in the most existing studies is referred to as call detail records (CDRs), which are generated by phone communication activities (i.e., make/receive a phone call, send/receive a text message). For billing purposes, CDRs keep track of the relevant information (e.g., caller/callee, time, duration) of each event, plus a unique identifier of the cell tower that handles the communication.

Many valuable findings regarding human activity and their interactions with the urban environment have been reported since CDRs became prevalent in the research community during recent years (e.g., González *et al*. 2008, Song *et al*. 2010a, 2010b). However, most previous studies did not discuss how representative their data (i.e., CDRs) were and the applicability of their analysis results to the entire population. Also, few studies have examined the questions that CDRs might not be appropriate to address (see Kang *et al*. 2012 for an example). Are we overly optimistic about the usefulness of CDRs and the validity of our conclusions? Like what have been discussed, debated, and acknowledged in the social media community, the representativeness of CDRs also needs to be carefully examined.

As pointed out by Becker *et al*. (2013), CDRs are coarse in space and sparse in time. In large cities, the spatial granularity at the cell tower level may not be a major drawback as cell towers are usually densely distributed across an urban area. What really matter are the uneven distribution of people's phone communication activities in space and time. On one hand, people are more likely to contact others at certain places, such as home or work place, and it is highly possible that these locations account for only a fraction of all visited locations. On the other hand, depending on how actively one engages in phone communication, the total number of CDRs each subscriber generates varies significantly. The dataset used in this research reveals that the population size drops with the increased intensity of phone-related activities (Figure 1). About 17% subscribers in our dataset have two or fewer CDR records in a day and over 38% subscribers have fewer than seven CDR records in a day. Hence, whether the mobility pattern of subscribers without heavy phone usage can be properly characterized is indeed questionable. One may argue that this problem can be solved by collecting CDRs over a longer period of time, such as a week, a month, or even longer. Although this workaround does help increase sample size, the uneven spatiotemporal distribution of digital footprints caused by people's habit of mobile phone usage is not addressed. The 'quiet minority' who rarely make use of mobile device remain underrepresented.

This research takes a first step to evaluate the representativeness of CDRs in human mobility characterization, using a mobile phone location dataset that includes both CDRs and non-CDR footprints. The non-CDR footprints are generated by events irrelevant to phone communication, such as moving out of the service area of a cell tower, active pinging from cell tower, etc. By extracting the CDR records into a separate dataset, we are able to quantitatively evaluate the effectiveness of CDRs in human mobility analysis, from both an individual perspective and a collective perspective. The findings of this research not only facilitate a better understanding of CDR data but also

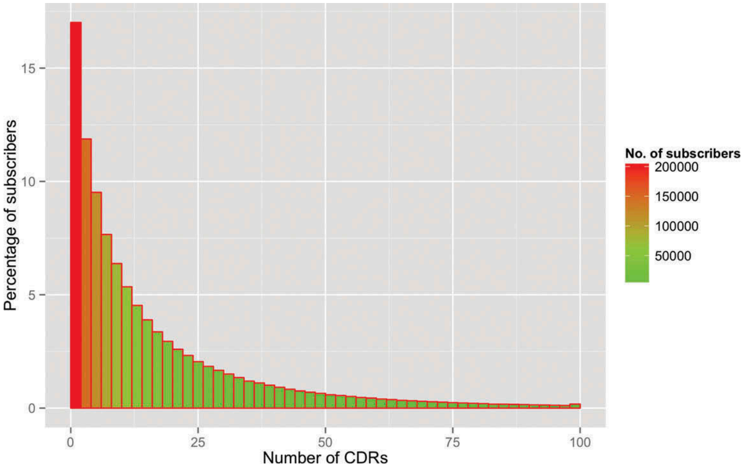


Figure 1. Distribution of subscribers under different intensity levels of phone communication.

prompt us to rethink some conclusions of human mobility patterns derived from CDR data that have been reported in the literature.

The remainder of this article is organized as follows. The next section discusses the existing research related to this study. Section 3 presents the study area and the mobile phone location data used in this research. In Section 4, we adopt an individual perspective and evaluate the effectiveness of CDRs in estimating some most frequently used mobility indicators. We then take a collective approach in Section 5 and examine the performance of CDRs in distance decay effect analysis and urban community detection. We conclude and discuss this research in Section 6.

2. Relevant research

This section discusses some relevant research in the following three areas: (1) CDRs and human mobility, (2) CDRs and urban dynamics, and (3) uncertainty issue.

2.1. *CDRs and human mobility*

CDRs have helped enhance our knowledge of individual human mobility considerably in recent years. A large body of literature focuses on individual activity space, which denotes the spatial extent of people's daily activities (Golledge and Stimson 1997). Understanding individual activity space has profound implications in the real world, such as accessibility to healthcare facilities (Sherman *et al*. 2005), environmental exposure (Perchoux *et al*. 2013), etc. Note that the term of activity space is related to several other concepts, for instance, awareness space (Brown and Moore 1970), action space (Horton and Reynold 1971), and space-time prism (Hägerstrand 1970).

Activity space can be characterized from individual trajectories which reflect a person's movement in space over time. A number of measures have been used to describe spatiotemporal characteristics of individual trajectories, such as the daily range of travel, movement radius, and movement entropy (e.g., Yuan *et al*. 2012). Two dimensional measures, such as standard deviational ellipse (SDE), can be used to describe the range and direction of one's activity space (e.g., Zenk *et al*. 2011). Based on these mobility indicators, statistical analyses have been performed to compare activity space of people in different social groups (e.g., age, gender, see Kang *et al*. 2010, Yuan *et al*. 2012) or people at different locations (Becker *et al*. 2013). In addition, individual trajectories collected over a long term allow us to extract meaningful anchor points of one's activity space, such as home or work locations (e.g., Calabrese *et al*. 2010a, Ahas *et al*. 2010), and to examine people's activity patterns around anchor points (e.g., Xu *et al*. 2015).

Besides activity space research, CDRs have been utilized by physicists to gain new insights into the nature of human travel. In the past, many studies assumed that human movements were associated with a large degree of randomness and could be explained by the random walk model or the Lévy flight model (Brockmann *et al*. 2006, Rhee *et al*. 2011). However, analysis results from CDR trajectories have suggested that human movements often follow reproducible patterns (González *et al*. 2008) and are highly predictable (Song *et al*. 2010a, 2010b).

Despite substantial progress reported in the literature, this study argues that the representativeness of individual trajectories derived from CDRs are strongly influenced by people's habit of using mobile phones at certain locations and time in a day. For example, CDR trajectories of a traveling salesperson who talks to his/her customers on a mobile phone regularly may well depict his/her daily movements, whereas CDR trajectories of a person who uses his/her phone occasionally should not be used to understand his/her mobility pattern in space and time. As a result, it is important to investigate to what extent we can trust the mobility indicators derived from CDR trajectories and the conclusions drawn from these indicators. It should be noted that using CDRs collected over a long period of time as a workaround cannot address this issue as people who rarely engage in phone communication remain underrepresented.

2.2. *CDRs and urban dynamics*

Instead of focusing on individual trajectories, many studies adopt a collective approach to uncover varying mobility patterns. Some frequently used indicators include Erlang value (i.e., the total call traffic volume in one hour), number of phone calls/text messages, number of active subscribers, etc. For instance, CDRs have been used to quantitatively measure different levels of popularity in New York City in terms of the density and distribution of aggregate phone calls (Girardin *et al*. 2009). Distinct patterns of mobility variation throughout different time periods in a day, or different days in a week also can be extracted and compared using methods such as K-means clustering (Reades *et al*. 2007), eigen decomposition (Calabrese *et al*. 2010b), dynamic time warping (Yuan and Raubal 2012), etc. In addition to mobility pattern analysis, aggregate population flows among cell towers serve as an indication of human interactions in urban space, which enable us to detect urban communities with strong internal interactions (Gao *et al*. 2013). Moreover, some recent studies utilize the characteristics of people's phone communication activities and develop innovative methodologies to address problems that are usually tackled by other approaches. For example, Pei *et al*. (2014) developed a new method for urban land use classification based on normalized hourly call volume and the total call volume.

Similar to individual human mobility research, many CDR-based urban dynamics studies also make an implicit assumption that phone communication records can serve as a direct indication of human activity intensity, which is debatable. A careful evaluation of the representativeness of CDRs can help us answer this type of questions.

2.3. *Uncertainty issue*

Uncertainty has been an important research topic in GIScience (Goodchild and Gopal 1989, Zhang and Goodchild 2002). It is associated with several related concepts, such as accuracy, precision, consistency, completeness, to name a few (Veregin 1999). Considerable efforts have been made to visualize and analyze spatiotemporal uncertainties (Pang 2001, MacEachren *et al*. 2005, Delmelle *et al*. 2014). Many critical concerns have been raised regarding how uncertainties could influence our findings (e.g., Griffith *et al*. 2007, Zinszer *et al*. 2010, Jacquez 2012) and the risk level in a decision making process (Gollege and Stimson 1997).

The issue of uncertainty is often examined in the field of environmental modeling (Refsgaard *et al*. 2007, Ascough II *et al*. 2008). Despite limited discussions in the literature, uncertainties embedded in mobile phone location data with respect to their spatiotemporal granularity should be examined. From the spatial perspective, spatial resolution of CDR data often is limited to the cell tower level (Becker *et al*. 2013). In urban areas where cell towers are sparsely distributed, geographic positioning becomes less precise (Bengtsson *et al*. 2011). Based on a mobile phone location dataset obtained from AirSage, Liu *et al*. (2008) evaluated the accuracy of derived travel speed against observed data collected by loop detectors and concluded that the consistency between two datasets varied to certain extent. Interestingly, Bar-Gera (2007) reported that speed calculated from the mobile phone location data is acceptable for practical applications. Another major issue that generates spatial uncertainty is the occurrence of signal jump, which occurs when a mobile device switches back and forth among a set of neighboring cell towers due to similar intensity of signal strength (Iovan *et al*. 2013). Xiong *et al*. (2012) took this issue into account by dividing a study area into non-overlapping regions and mark each region in terms of the possibility of signal jump. From the temporal perspective, the locations of a subscriber between two phone communication events are uncertain. Within a two-hour period, the potential area that a subscriber can visit may cover the entire city. With CDR data, the interval between two phone communication activities is often longer than 2 hours, which leads to a large degree of uncertainty in human mobility analysis. The temporal granularity of CDR data can be improved with additional data such as active pinging collected by mobile network operators to reduce uncertainties between two phone communication records. Considering both the spatial and the temporal domains, Couronne *et al*. (2011) proposed an indicator to assess mobility as well as uncertainty.

The uncertainty issue itself cannot be fully prevented. Our challenges are to understand how uncertainties could result in imperfect knowledge and recognize ‘which cannot be known’ (Couclelis 2003). This is the fundamental objective of this article.

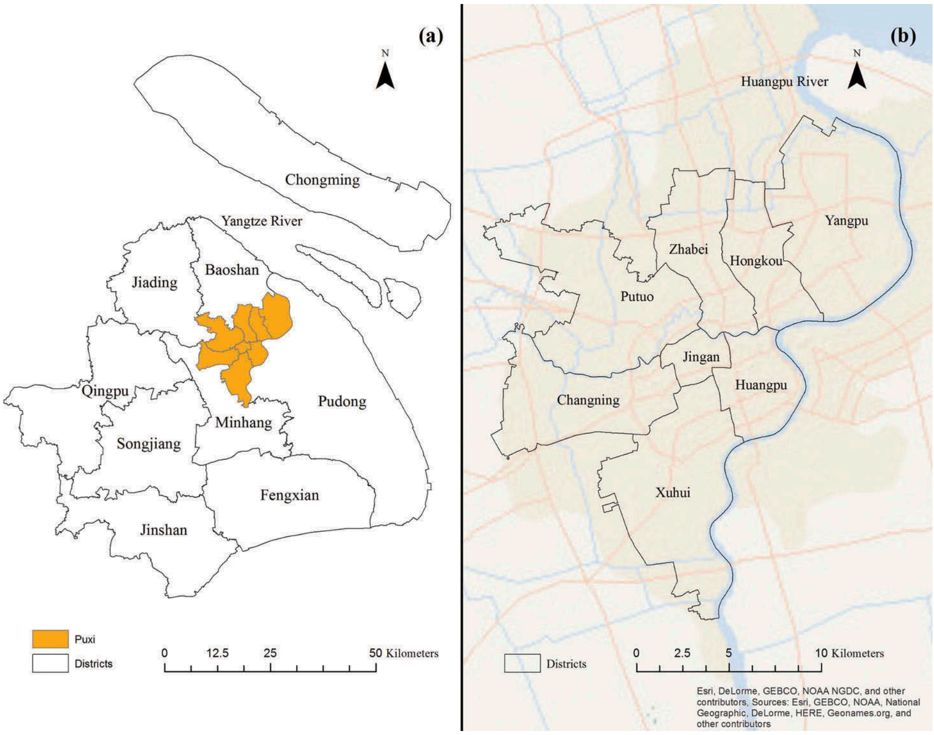
3. Data

Our study area is Shanghai, one of the largest cities in China. In this section, we introduce some background information of Shanghai and the mobile phone location dataset collected in this city.

3.1. *Area of study*

Shanghai has a resident population of 23.8 million as of 2012 (Shanghai Municipal Statistics Bureau 2012), which makes it the largest city in China by population. Shanghai is one of the global financial centers and the busiest container port in the world (World Shipping Council 2013). Its annual gross domestic product (GDP) also ranks No.1 in China in 2012 (National Bureau of Statistics of China 2012).

Located in the central east coast of China, Shanghai has a total area of 6,340.5 square kilometers (Shanghai Municipal Statistics Bureau 2012). It consists of 16 administrative districts and the Chongming County (Figure 2a). Among those districts, eight of them on the west bank of the Huangpu River (i.e., Huangpu, Xuhui, Jingan, Changning, Yangpu, Hongkou, Putuo, and Zhabei), also known as Puxi, are referred to as the downtown area of Shanghai (Figure 2b). Over the past two decades, the economy of the Pudong District, situated on the east bank of the Huangpu River, has been growing rapidly, with its famous zone of Lujiazui being widely considered as the financial center of Shanghai.



Esri, DeLorme, GEBCO, NOAA NGDC, and other contributors, Sources: Esri, GEBCO, NOAA, National Geographic, DeLorme, HERE, Geonames.org, and other contributors

Figure 2. (a) Shanghai and its administrative districts. The orange areas represent the “Puxi” region, the downtown area of Shanghai. (b) Eight administrative districts in the “Puxi” region. “Puxi” and the Pudong districts are divided by the Huangpu River.

3.2. *Dataset*

The mobile phone location dataset used in this study is collected by a major MNO in China. It is obtained through a joint research collaboration. It includes the records generated by 1,252,797 subscribers on a workday in 2012. Different from a CDR dataset, this dataset contains both CDRs and actively generated logs, recorded with seven event codes listed in Table 1. This particular MNO operated over 33,000 cell towers in Shanghai and the cell tower ID associated with each record indicated the approximate location where each event took place. It should be pointed out that, to protect individual privacy, we do not have access to any personal information (e.g., age, gender, phone number) and the spatial granularity is restricted at the cell tower level.

Figure 3 illustrates the total number of each event recorded during every hour. Given the way different events are triggered, those numbers vary differently throughout the day. In general, except periodic updates (PU), very few records are generated between midnight and 6:00. At 6:00–7:00, the volumes of regular updates (RU) and cellular handover (CH) grow significantly due to population travel. As a result, the number of active pinging from towers, recorded as PU, declines accordingly. The peaks of RU at 8:00–9:00 and 17:00–18:00 correspond to the morning and evening rush hour, respectively. Similar to RU, the numbers of IN and OT events start to increase from 6:00–7:00. ON and OF events together account for a very small portion of the data as turning mobile phone on and off frequently is not a common practice.

3.3. *Data processing*

Various types of events recorded in this dataset offer a unique opportunity to understand the bias of CDRs in human mobility analysis. For the purpose of comparisons, we extract all CDRs (i.e., IN and OT events) from every subscriber and store them in a separate dataset. Therefore, each subscriber has two sets of data: CDRs only and the

Table 1. Summary of event codes.

Code	Event	Description	Avg. no. of records per subscriber
RU	Regular update	Regular update triggered by moving from the service area of a cell tower to that of another tower.	12.51
PU	Periodic update	Periodic update triggered by tower pinging if a subscriber has been 'silent' (i.e., no other events listed in this table is detected) for a certain time period. However, the specific condition (e.g., duration of silence) that triggers periodic update is irregular. In addition, mobile phones which are turned off or disconnected from the cellular network do not receive pinging signals from the cellular network.	4.88
OT	Phone communication (outbound)	Subscriber makes a phone call or sends a text message.	4.45
ON	Power on	Mobile phone is turned on and connected to cellular network.	0.62
OF	Power off	Mobile phone is turned off and disconnected from cellular network.	0.39
IN	Phone communication (inbound)	Subscriber receives a phone call or a text message.	14.67
CH	Cellular handover	Transfer of an ongoing phone call from one cell tower to another due to a subscriber's movements.	5.45

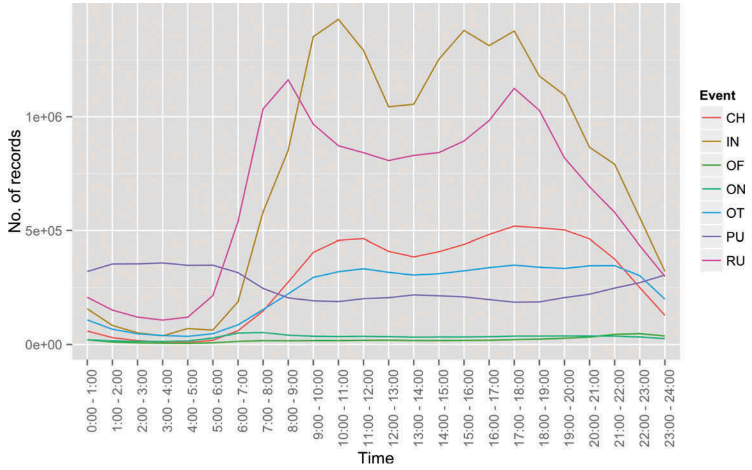


Figure 3. Temporal variation of the total number of each event.

entire set of records. In the remainder of this article, we call these two datasets as the CDR group and the complete group, respectively.

As a subset of the data, the temporal variation of the total number of records in the CDR group mirrors that in the complete group (Figure 4), although the former does not reveal a striking upsurge during 17:00–18:00. For each subscriber, the CDR ratio (i.e., number of CDRs/number of total records) is calculated. The average CDR ratio is 43.09%, with a median value of 41.18%. However, depending on how actively one engages in phone communication activities, this number varies significantly (Figure 5).

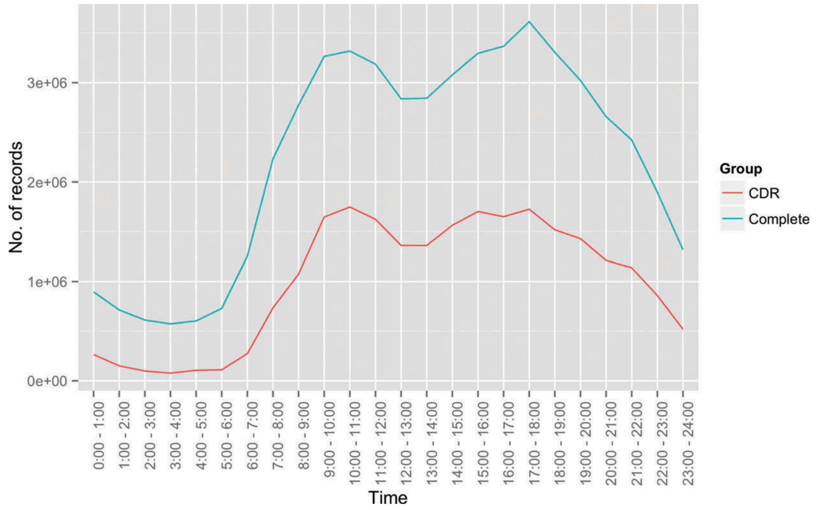


Figure 4. Temporal variation of the total number of records in the CDR group and the complete group.

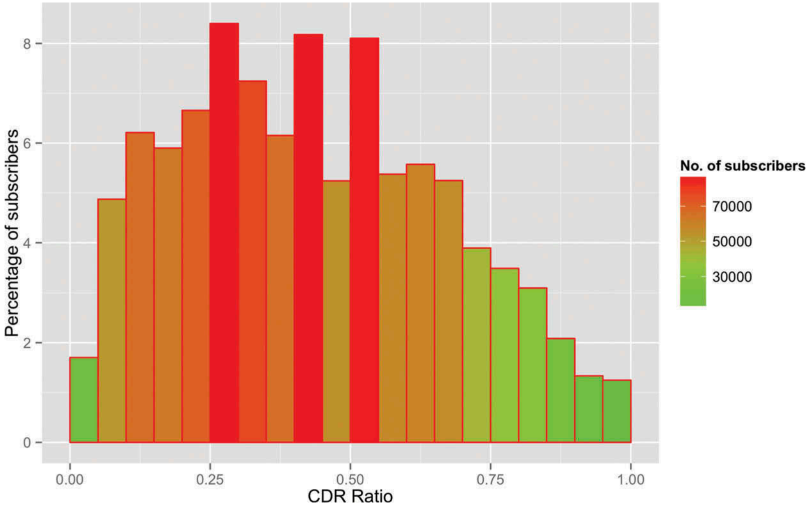


Figure 5. Distribution of subscribers under different ratio of CDRs.

4. Individual human mobility

This section focuses on evaluating the representativeness of CDRs in individual daily mobility pattern analysis. We aim to answer the question: compared with the complete set of footprints, how well do CDR footprints characterize one's daily mobility pattern? We focus on three basic properties of human mobility: distance, range, and heterogeneity. Hence, the following three frequently used mobility indicators are chosen in this study: (1) the total travel distance, (2) radius of gyration, and (3) movement entropy. In the evaluation process, the complete group is considered as the control group. To tackle our research question, two additional data processing steps are performed.

First, if the footprints generated by a subscriber in a day do not provide sufficient temporal coverage (e.g., people who kept their mobile phone turned off most of the time), this subscriber should be removed from the complete dataset that will be used as a benchmark to evaluate the representativeness of CDR footprints. We divide the day into four six-hour periods (0:00–6:00, 6:00–12:00, 12:00–18:00, and 18:00–24:00) and only those subscribers with at least one footprint in each six-hour period are included in this study. After this step, a total of 686,642 subscribers remain in the complete group.

Second, as discussed in Section 3.3, the range of CDR ratios varies significantly among the subscribers and the CDR ratio could be a critical factor in the evaluation process. For people who use their mobile phone to communicate frequently, mobility indicators derived from the CDR group and the complete group should be very close. On the contrary, for people who travel a lot but rarely use their phone to communicate, their CDRs are likely to yield biased mobility indicators. In order to understand how the CDR ratio influences the estimation of mobility indicators, we further group the 686,642 subscribers into four classes by their CDR ratio (Table 2).

Table 2. Four subscriber classes based on their CDR ratio.

Class	CDR ratio (%)	Number of subscribers
A	75–100	60,519
B	50–75	173,940
C	25–50	251,187
D	0–25	200,996

4.1. *Total travel distance*

The total travel distance is a basic measure of individual mobility. It is calculated as the sum of the Euclidian distance between each pair of consecutive footprints. For each subscriber, we compute two values of the total travel distance, D_{CDR} and D_{complete}, based on the CDR group and the complete group, respectively. The results of all four subscriber classes are shown in Figure 6. The horizontal axis and the vertical axis represent the complete group and the CDR group, respectively. In this figure, the horizontal axis is binned with a bandwidth of 0.1 km. Subscribers are grouped in terms of (1) class assignment based on CDR ratio (Table 2), and (2) the 0.1-km bin that D_{complete} falls in. Then, for all subscribers in each bin, the average value of D_{CDR} is computed and plotted. Figure 6 allows us to examine the representativeness of CDRs via visual inspection. If CDRs are representative of the complete group, points on Figure 6 should be close to the diagonal line from lower-left to upper-right. On the contrary, a large deviation from the diagonal line leads to an indication that CDRs tend to underestimate the total travel distance. For this mobility indicator, an overestimation is not possible since D_{CDR} cannot be greater than D_{complete}.

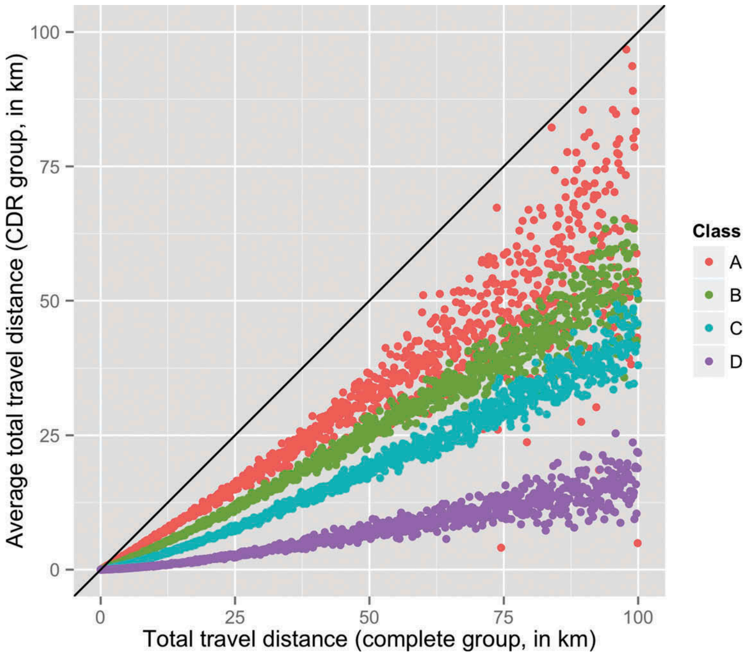


Figure 6. Total travel distance (complete group) vs. average total travel distance (CDR group).

Table 3. Correlation between the total travel distance (complete group) and the average total travel distance (CDR group).

Class	Pearson correlation	Spearman correlation
A	0.941	0.960
B	0.987	0.992
C	0.989	0.994
D	0.955	0.976

Several interesting findings regarding the effectiveness of CDRs in estimating the total travel distance are revealed in Figure 6. First, all four classes suggest that D_{CDR} and D_{complete} have a very high positive correlation. This is confirmed by both Pearson correlation coefficient and Spearman correlation coefficient (Table 3). Second, as the CDR ratio declines (from Class A to Class D), points in Figure 6 deviate more from the diagonal. To quantify the level of underestimation, we fit points in each subscriber class with a linear regression model: $y = \alpha + \beta x$, using D_{complete} as the independent variable x and D_{CDR} as the dependent variable y. Provided that D_{CDR} must have a value of 0 if $D_{\text{complete}} = 0$, we enforce α to be 0 in the linear regression model. The regression coefficient β indicates the relationship between D_{CDR} and D_{complete}. Since the regression model does not have an intercept, $1 - \beta$ can be interpreted as the level of underestimation, which implies how well CDRs can estimate one's total travel distance. It is evident that CDRs tend to significantly underestimate the total travel distance even for subscribers in Class A, whose CDRs account for at least 75% of all footprints. On average, D_{CDR} of Class A subscribers is 35.3% shorter than his/her D_{complete} (Table 4). This regression coefficient β turns out to be smaller in Class B and Class C, which indicates that CDRs become more and more biased in estimating the total travel distance when CDR ratio drops. For subscribers in Class D, their CDRs on average underestimate the total travel distance by 84.6%. Figure 6 also suggests that the variation of average D_{CDR} becomes larger when the value of D_{complete} grows. This pattern of heteroscedasticity is relatively easy to understand under the context of human travel: if one's daily travel distance is longer, the range of estimated travel distance based on his/her CDRs is expected to be wider. Another possible reason is that the size of subscribers drops rapidly as the total travel distance increases. It also could result in a wider range of average D_{CDR}.

4.2. *Radius of gyration*

The radius of gyration is one of the most frequently used measures of activity space. It is defined as the root mean squared distance between a set of visited locations up to time t and the center of mass:

Table 4. Linear regression results between the complete group and the CDR group for the total travel distance.

Class	Regression coefficient (β)	Level of underestimation $(1 - \beta)$%
A	0.647	35.3
B	0.529	47.1
C	0.401	59.9
D	0.154	84.6

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} \left(\overrightarrow{r_i^a} - \overrightarrow{r_{cm}^a}\right)^2}, \tag{1}$$

where $\overrightarrow{r_i^a}$ represents the $i = 1, \ldots, n_c^a(t)$ location of subscriber a and $\overrightarrow{r_{cm}^a} = \frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} \overrightarrow{r_i^c}$ defines the center of mass (González *et al.* 2008). The radius of gyration reflects the range of activity space, typically around the center of home and work locations for commuters.

Similar to Section 4.1, we compute two values of the radius of gyration for each subscriber based on the CDR group and the complete group, denoted as R_{CDR} and R_{complete}, respectively. Figure 7 uses the horizontal axis to represent the complete group with a 0.1-km bandwidth and the vertical axis to represent the average R_{CDR} of subscribers in the same 0.1-km bin. Again, the consistency between two groups can be inferred by the closeness of data points to the diagonal. Note that unlike the total travel distance, R_{CDR} could be larger than R_{complete} if CDR footprints spread more widely than non-CDR records.

The effectiveness of CDRs in estimating radius of gyration is noteworthy. First, for Classes A, B, and C, R_{CDR} are strongly correlated with R_{complete} (Table 5). However, both Pearson correlation coefficient and Spearman correlation coefficient show a significant drop in Class D, although they still suggest a positive correlation. In addition, for Class D subscribers whose R_{complete} is larger than 25 km, their average values of R_{CDR} are often zero or very close to zero. Therefore, CDRs might significantly underestimate the radius

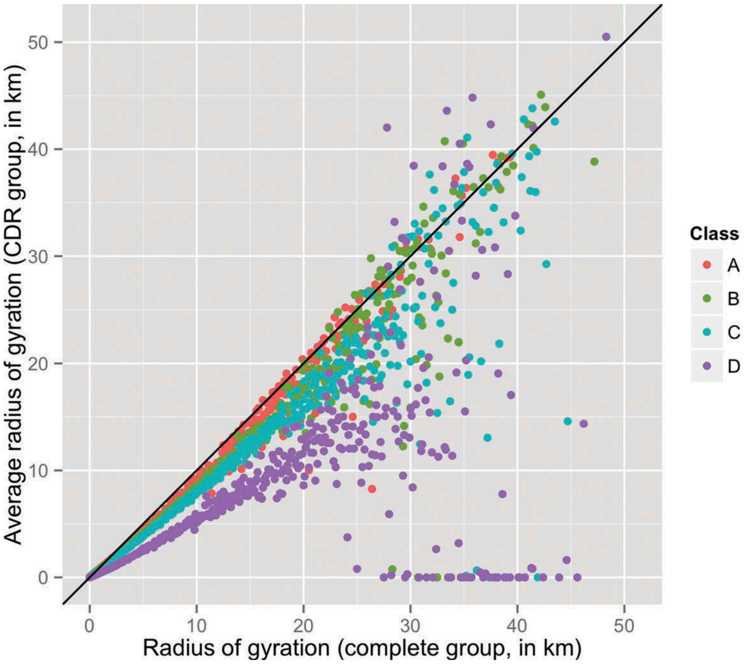


Figure 7. Radius of gyration (complete group) vs. average radius of gyration (CDR group).

Table 5. Correlation between the radius of gyration (complete group) and the average radius of gyration (CDR group).

Class	Pearson correlation	Spearman correlation
A	0.980	0.982
B	0.936	0.939
C	0.860	0.882
D	0.532	0.521

of gyration of people who commute over long distance and rarely use mobile phone. Second, Figure 7 also reveals a pattern of heteroscedasticity: the range of estimated radius of gyration based on CDRs is supposed to be wider if one's radius of gyration derived from the complete group grows. From Class A to Class D, such a pattern of heteroscedasticity turns out to be more obvious as CDRs become a smaller part of one's footprints.

By fitting a linear regression model to each class we can quantify the effectiveness of CDRs for estimating this mobility indicator. The regression coefficient is very high for Class A and Class B (Table 6). It suggests that CDRs could depict the range of daily travel very well for subscribers whose 50% or more footprints are collected from phone communication events. Adding other non-CDR footprints makes very limited difference on the derived radius of gyration. For subscribers in Class C, CDRs on average underestimate their radius of gyration by 20.6%. Depending on specific applications, this margin of error may be acceptable. However, the small regression coefficient (0.491) in Class D indicates that CDRs fail to provide a good estimate for subscribers whose fraction of CDRs is below 25%. Many subscribers in this group engage none, or very few phone communications in a day. Others who make some use of mobile phones also travel a lot and leave numerous digital footprints (RU event). As a result, this group of CDRs remains questionable for deriving daily activity space.

4.3. *Movement entropy*

The movement entropy measures the heterogeneity of visitation patterns (Song *et al*. 2010a, Yuan *et al*. 2012). It can be calculated using the following equation:

$$E = -\sum_{i=1}^{n} p_i log_2 p_i, \tag{2}$$

where n is the number of distinct locations (i.e., cell towers) visited by a subscriber and p_i is the probability that location i is visited. Mathematically, the value of movement

Table 6. Linear regression results between the complete group and the CDR group for the radius of gyration.

Class	Regression coefficient (β)	Level of underestimation $(1-\beta)$%
A	0.940	6.0
B	0.887	11.3
C	0.794	20.6
D	0.491	50.9

entropy grows with a more heterogeneous visitation pattern. Consider the following examples:

(1) If a subscriber stays at a single location, $E = -(1.0 \times \log_2 1.0) = 0$;
(2) If a subscriber visits Location A one time and Location B four times, $E = -(0.2 \times \log_2 0.2 + 0.8 \times \log_2 0.8) \approx 0.72$;
(3) If a subscriber visits Location A five times and Location B five times, $E = -(0.5 \times \log_2 0.5 + 0.5 \times \log_2 0.5) = 1$;
(4) If a subscriber visits Locations A, B, C, and D, two times each, $E = -(0.25 \times \log_2 0.25 + 0.25 \times \log_2 0.25 + 0.25 \times \log_2 0.25 + 0.25 \times \log_2 0.25) = 2$.

For this mobility indicator, E_{CDR} and E_{complete} are calculated for each subscriber. The correlation coefficients all indicate a very positive correlation between E_{CDR} and E_{complete} (Table 7). Unlike the other two selected mobility indicators, we cannot identify an evident pattern of heteroscedasticity when $E_{\text{complete}} < 5$ (Figure 8). We find that CDRs can estimate the movement entropy very well for subscribers in Class A given the high regression coefficient (0.876, see Table 8). This coefficient declines for Class B and Class C, which implies that as the CDR ratio decreases, it is more likely that some non-CDR footprints are collected at other visited locations where subscribers do not engage phone communications. This is likely the most reasonable explanation for the low regression coefficient (0.328) associated with Class D. Apparently, CDRs underestimate the movement entropy by far (67.2%) for subscribers in Class D. Moreover, data points in Class D suggest some abnormal drops of average E_{CDR} when $E_{\text{complete}} > 6$ (Figure 8), which is not the case for the other three classes. We believe that it is also caused by a low likelihood of making phone communications at some visited locations.

In this section, we evaluate the representativeness of CDRs based on mobility indicators that measure activity space from three aspects: distance, range, and heterogeneity. We reveal some important findings by answering 'whether CDRs can provide a good estimate of individual mobility patterns'. We have indicated that the answer is not simply yes or no. Perhaps the question should instead be phrased as, 'how good are CDRs in providing a good estimate of individual mobility patterns'. According to our

Table 7. Correlation between movement entropy (complete group) and average movement entropy (CDR group).

Class	Pearson correlation	Spearman correlation
A	0.998	0.999
B	0.996	0.999
C	0.991	0.997
D	0.900	0.934

Table 8. Linear regression results between the complete group and the CDR group for movement entropy.

Class	Regression coefficient (β)	Level of underestimation ($1 - \beta$) (%)
A	0.876	12.4
B	0.751	24.9
C	0.625	37.5
D	0.328	67.2

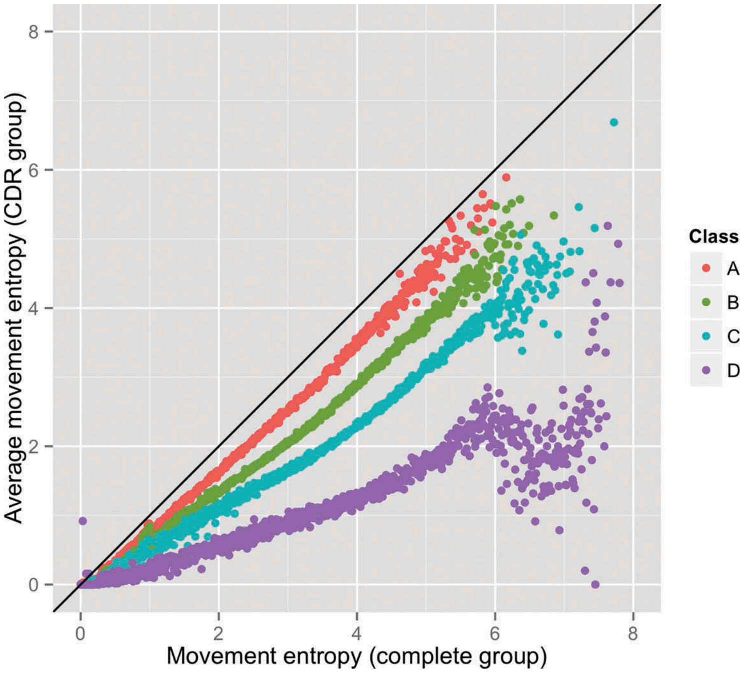


Figure 8. Movement entropy (complete group) vs. average movement entropy (CDR group).

analysis, the effectiveness of CDRs in individual mobility study depends on the research question and the mobility measure selected to address that question. To estimate radius of gyration, CDRs in most cases are probably good enough for subscribers who (1) make at least some phone communications throughout the day, and (2) travel within normal daily activity range (e.g., less than 25 km in Shenzhen). On the contrary, one needs to be cautious when using CDRs to study other problems such as travel distance or heterogeneity of human mobility. To a large extent, the validity of analysis result is subject to how actively subscribers engage in phone communications. Therefore, in many cases we may bear the risk of underestimating mobility indicators of interest.

5. Collective human mobility

Many researchers approach human mobility study from a collective perspective and pay attention to data aggregated from the individual level. In this section, we evaluate the representativeness of CDRs from this perspective. The distance decay effect and urban community detection are selected for the evaluation because they have been examined in several CDR-based human mobility studies (e.g., González *et al.* 2008, Walsh and Pozdnoukhov 2011, Gao *et al.* 2013).

5.1. *Distance decay effect*

The existing studies reveal that human movements can be modeled by a Lévy flight (Brockmann *et al.* 2006), while the power law distribution of displacements is an indication of the distance decay effect (Liu *et al*. 2012, Gao *et al*. 2013). The notion of distance decay has a close relationship with *The First Law of Geography*: 'everything is related to everything else, but near things are more related than distant things' (Tobler 1970, p. 236). Although highly developed urban areas offer various means of transportation, human activity remains to be restricted by a number of factors such as distance and accessibility. Many researchers argue that the 'death of distance' hypothesis is premature even with today's technologies (Wang *et al*. 2003, Rietveld and Vickerman 2004).

Massive CDR data offer some new opportunities to validate and/or adjust our understanding of the frictional effect of distance. For instance, González *et al*. (2008) and Gao *et al*. (2013) report distance decay parameters of 1.75 and 1.60, respectively. However, as discussed earlier, CDRs are generated only upon phone communication activities and most people do not use their mobile phone at all places they visit. Therefore, displacements between CDR footprints can only represent movements between phone communications. Taking advantages of the various event types recorded in the mobile phone location dataset used in this study, we are able to compare the distance decay effect observed from CDR data against that derived from the complete set of footprints.

We derive 4,992,719 displacements from the CDR group and 27,686,129 displacements from the complete group. Figure 9 shows cumulative distribution function (CDF) curves of these two groups. The cumulative distribution curves indicate that most displacements are short, with about 90% displacements in the CDR group below 5 km and roughly 90% of displacements in the complete group under 2.5 km. These displacements can be approximated by a power law distribution in the following form:

$$P(d) \propto d^{\beta}, \tag{3}$$

where β is the distance decay parameter (Gao *et al*. 2013). A large value of β indicates that distance is a strong deterrent to interaction, whereas a small value of β implies a relatively weak influence of distance.

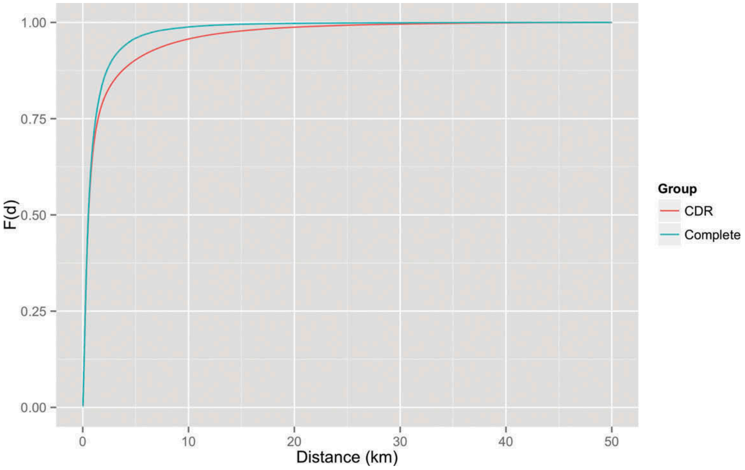


Figure 9. Cumulative distribution function (CDF) of displacements.

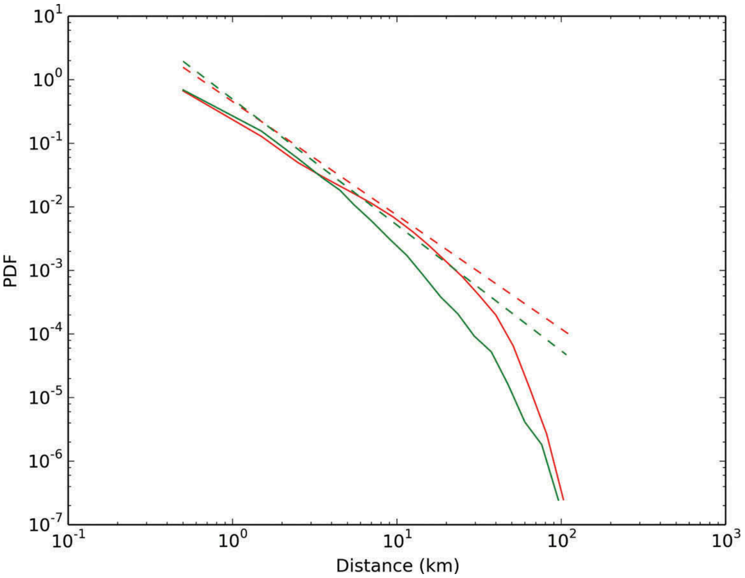


Figure 10. Probability density function (PDF) and the fitted power law distribution. The green line and red line represent the probability distribution of the displacements derived from the CDR group and the complete group, respectively. The dashed green line and the dashed red line are the fitted power law distributions for the CDR group and the complete group, with a decay parameter of 1.79 and 1.98, respectively.

Figure 10 shows the probability density function (PDF) in log-log scale and the fitted power law distribution. The fitted decay parameters are $\beta_1 = 1.79$ for the CDR group and $\beta_2 = 1.98$ for the complete group. Note that β_1 is very close to 1.75 reported by González *et al*. (2008), which indicates that a similar mechanism that drives the frictional effect of distance is captured. As we would expect based on the CDF plot in Figure 9, β_2 should be larger than β_1 since the complete group captures more short-distance displacements. For this reason, we believe that CDR data slightly underestimate the distance decay effect in Shenzhen. A possible explanation is related to the habit of mobile phone usage that most people do not contact others by phone or text at every visited location. On average, displacements between phone calls (or text messages) are longer than those between consecutive locations people visit. Although it is true that many long-distance trips may be missing in the CDR database, the amount of short-distance trips CDRs do not capture could be substantially larger, which results in a less steep curve on the CDF plot for $x < 20$ km and a smaller value of distance decay parameter.

5.2. *Community detection*

Various types of network (e.g., social network) in a city often establish a structure of communities that are more tightly connected internally and structurally distinct from others (Girvan and Newman 2002). Identifying communities in a network can help us understand the internal structure of a city that is shaped by human interactions as opposed to pre-defined administrative boundaries. Cell towers operated by MNOs can be considered as nodes in a large cellular network. In recent years, some urban dynamics studies have used CDR data to detect urban communities (e.g., Walsh and Pozdnoukhov 2011, Gao *et al.* 2013). It again is important to examine if digital footprints based on phone communication logs introduce a bias to the outcome of community detection.

Community detection aims at partitioning a network into communities that consist of densely connected nodes. The quality of partition is often evaluated by modularity. In a weighted network, it is defined as:

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta\left(c_i, c_j\right), \tag{4}$$

where A_{ij} denotes the weight of edge between two nodes i and j. $k_i = \sum_j A_{ij}$ represents the sum of weight of all edges towards node i. c_i denotes the community to which node i is assigned. $\delta\left(c_i, c_j\right)$ has a value of 1 if node i and node j belong to the same community and a value of 0 otherwise. m is half of the total edge weight in the entire network and $m = \frac{1}{2}\sum_{i,j} A_{ij}$ (Blondel *et al.* 2008). Many algorithms have been proposed to improve partition quality by maximizing modularity (e.g., Clauset *et al.* 2004, Newman 2004).

Edges of a cellular network are usually weighted by the intensity of human interaction (e.g., volume of population flow). Given the size of our network (33,044 cell towers), we adopt the Louvain method (Blondel *et al.* 2008) which takes a heuristic approach to optimize modularity of a large network efficiently. Using population movement volumes among all cell towers in the entire day, the Louvain method is used to detect urban communities based on the CDR group and the complete group, respectively. Table 9 shows the result of community detection results with high modularity scores. For visualization purpose, we also create a Voronoi diagram based on cell tower locations and assign a unique color to Voronoi cells in the same community.

Figure 11 shows the 20 detected communities using data from the CDR group. At the urban scale, the following findings of communities derived from CDR data are noteworthy:

(1) Natural barriers play an important role in community separation. Two examples in Shanghai include the Yangtze River and the Huangpu River. The former separates the three islands of Chongming County from the other areas of Shanghai, while the latter divides Pudong and Puxi. As suggested by the results, although bridges and ferries provide means to transport people from one side to another, naturally separated regions remain sparsely connected in terms of the intensity of human interaction.

Table 9. Summary of community detection results.

Group	No. of subscribers	Edges	No. of detected communities	Modularity
CDR	811,330	1,724,465	19	0.754
Complete	1,185,383	2,707,959	21	0.809

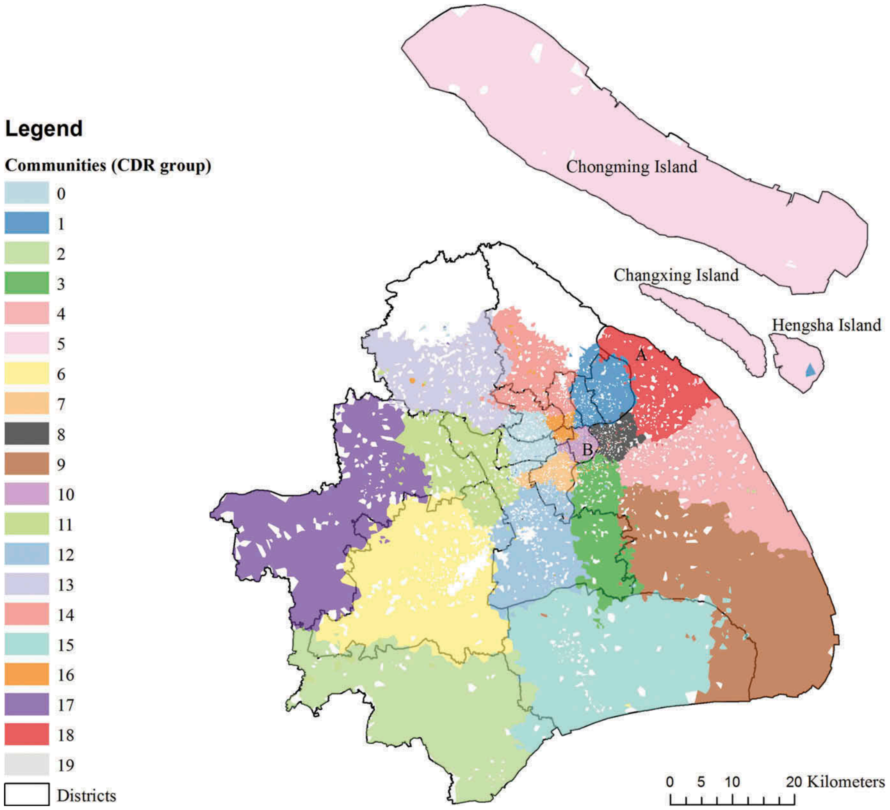


Figure 11. Detected communities based on the CDR group.

(2) In many regions, administrative boundaries appear to possess a similar influence on movements as natural barriers do given that the boundaries of identified communities match very well with the administrative boundaries. It implies that human movements within the administrative districts are much more intense than cross-boundary movements. In other words, CDRs reveal that human interactions in Shanghai are largely affected by political boundaries.

(3) Communities detected in Lujiazui area and Puxi, situated at the east bank and west bank of the Huangpu River (Figure 2), cover much smaller areas than other communities do. Compared with suburban districts (e.g., Jinshan, Songjiang, etc.), land use patterns in Lujiazui and Puxi, the most developed and most populated region in Shanghai, are highly mixed. Therefore, typical activities in this region do not require long-distance travel, resulting in smaller activity space on a workday.

Using data from the complete group, two more communities are identified. While the overall detection result resembles the one derived from the CDR group in terms of the number of communities and their boundaries, we highlight and discuss some major differences:

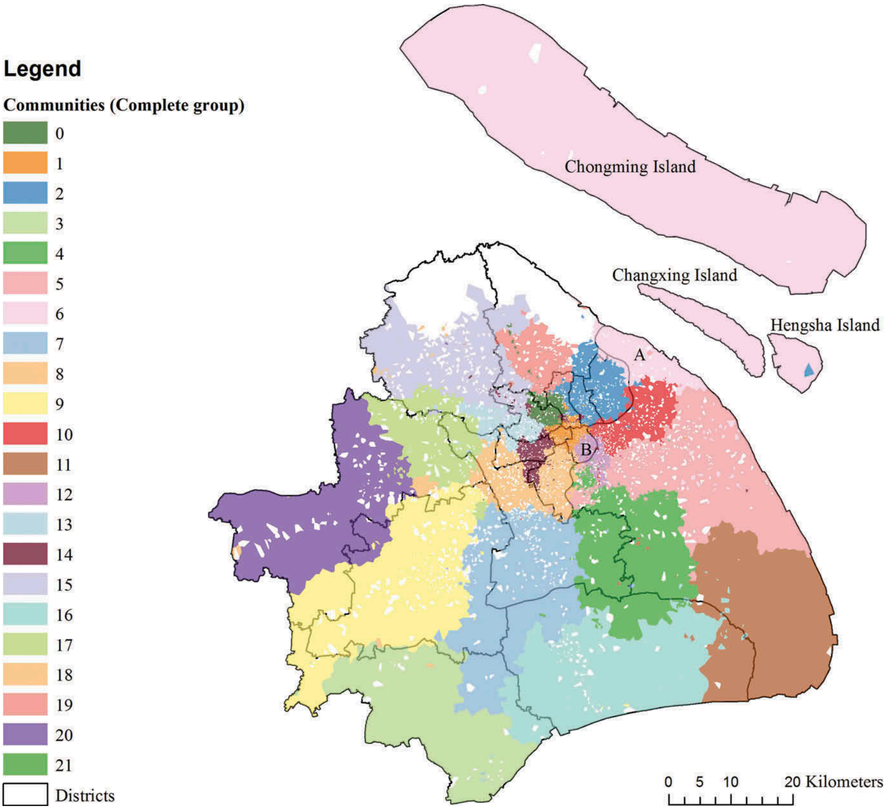


Figure 12. Detected communities based on the complete group.

(1) Natural barriers are not a major factor in community separation for the complete group. For instance, Figure 12 shows that Region A, on the west bank of the Yangtze River, is closely connected to the three islands of the Chongming County (Chongming, Changxing, Hengsha, see Figure 12). Apparently, human movements recorded in the complete group better capture population interactions between Region A and the Changxing Island through ferries and a major arterial called the Changjiang Tunnel. Serving as critical components of the Shanghai Port, industries related to port operation (e.g., container ports, shipyards, shipping companies) are agglomerated in Region A and Changxing Island. Region B, which covers both sides of the Huangpu River in the southern part of the downtown area, presents another example. The Nanpu Bridge, one of the main bridges over the Huangpu River, connects the two sides of the river to overcome the natural barrier. This observation is not present among the communities derived from the CDR group since a large portion of people who travel between two regions separated by natural barriers tend to use their mobile phones more frequently only in one of the regions.

(2) Similar to natural barriers, political boundaries turn out to be less important in community separation based on the complete group. For example, with CDRs, communities in southern Shanghai are divided by administrative boundaries of

Qingpu, Songjiang, Minhang, Jinshan, and Fengxian. With the complete group, the community in Minhang clearly crosses the administrative boundaries. Similar examples can be found in other places (e.g., the communities that cross Qingpu and Songjiang, Minhang and Pudong, Jiading, Baoshan, etc.). This finding suggests that the influence of administrative boundaries on interaction patterns may be exaggerated in CDR data. Similar to the previous finding, we speculate the main reason to be the biased spatial distribution of calling/texting activities. For many subscribers, their primary phone communication activities may be limited to certain places, probably within the same administrative district.

(3) By conducting a close visual inspection, we notice that in most cases communities have clear boundaries and they are mutually exclusive. Voronoi cells that belong to one community seldom appear inside the territory of another community. However, a number of exceptions are observed at different locations of the city. Data from the complete group suggest that a large percentage of residents travel to work outside the boundary of the community where their home is located, making the connection between the residential area and the destination community closer (see Figure 13 for an example). On the contrary, communities detected from CDR data are more mutually exclusive, with very few Voronoi cells found inside a different community. This finding further confirms that CDRs only partially reflect the intensity of human interactions over space. Tight connections across adjacent neighborhoods may not be detected from CDR data.

In this section, we take a collective approach to compare digital footprints from the CDR group and the complete group. Based on the aggregate spatial displacements, we

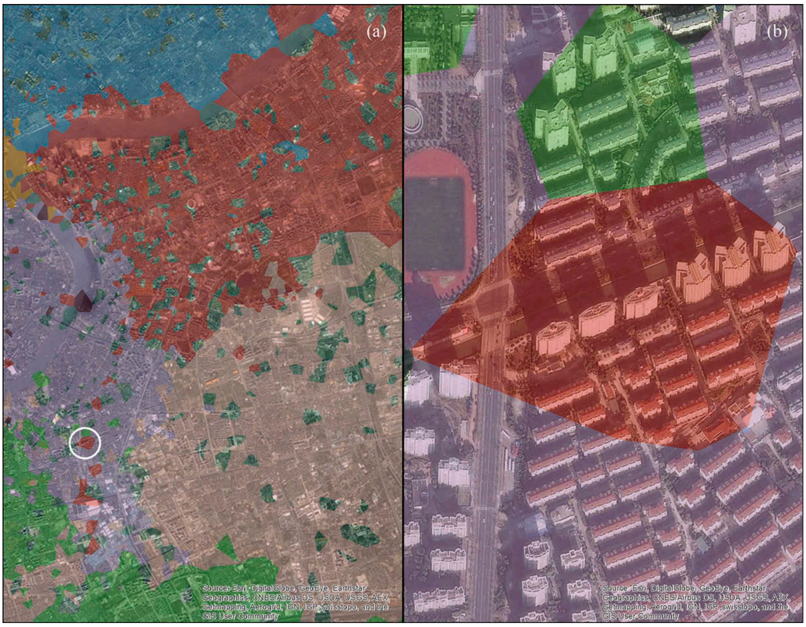

(a)
(b)
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 13. Decreased level of mutual exclusiveness in community detection using data from the complete group. (a) Many Voronoi cells are located outside the community boundary. (b) Taking the Voronoi cell circled in the left figure as an example, we notice that most Voronoi cells outside the community boundary cover high-density residential area.

investigate the distance decay effect and indicate how CDR data could lead to a biased understanding of urban dynamics. We also use a community detection method to identify the spatial structure of Shanghai in terms of interaction patterns. Although the results from the CDR group and the complete group do not differ significantly, some important differences regarding how urban areas are separated by human interactions are identified. Hence, we reach a conclusion that aggregate urban interaction patterns we uncover from CDR data also could be biased and misleading.

6. Conclusion and discussion

CDRs have been considered as a useful data source in support of human mobility research. However, given the uneven distribution of people's communication activities in space and over time, can we rely on CDR data to understand human movement and interaction patterns? This study takes the first step to assess the bias of CDR data for human mobility research. Based on our findings, we cannot answer the question with a simple yes or a no. First, it depends on what research question we want to answer. For instance, CDRs tend to underestimate mobility indicators such as the total distance and the movement entropy. On the contrary, for other indicators such as individual activity range, CDRs may be able to provide reasonable estimates. Second, the effectiveness of CDRs is closely related to the habit of mobile phone usage. How frequently one uses mobile phone to contact others and when and where those communications occur largely determine the representativeness of CDRs to reflect true mobility characteristics. In summary, we believe that the event-triggered nature of CDR data does introduce some biases to human mobility research and we suggest researchers to use CDR data with caution.

We do not attempt to deny the value of CDRs in human mobility research. Our objective is to have a fair assessment of what CDR data can and cannot do well in support of human mobility research. At present, CDRs remain as a useful data source with its wide coverage of population in many parts of the world and relatively low extra costs to collect such data. No datasets are perfect. As far as we understand the strengths and limitations of CDR data, we can gain many insights into human behaviors from CDR data. In the meantime, it is worth thinking about potential approaches to reduce/correct the biases embedded in CDR data. For instance, spatial and temporal interpolation of CDR footprints could possibly help us gain insights into data biases and develop solutions to address the issue. Applying post-hoc adjustments could be another promising workaround. In Section 4 of this article, we use linear regression to assess to what degree CDRs underestimate particular mobility indicators. The regression coefficient could be used to adjust the mobility indicator of interest. For instance, if we know CDRs normally lead to a 50% underestimation on the movement entropy of people who rarely use mobile phone, we can double the estimated movement entropy values accordingly to improve the accuracy level. However, our findings may or may not be applicable to other cities due to different urban environments and habits of mobile phone usage. A thorough study of local mobile phone usage patterns is required for post-hoc adjustments.

This article presents a particular way of assessing the validity of using CDR data for human mobility research. Future research could develop additional approaches/methods to analyze CDR data in a systematic manner. Knowledge of the effectiveness of CDRs in answering different research questions can be beneficial to many application fields ranging from urban design, transportation planning to air pollution and smart cities in this big data era.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was jointly supported by the National Natural Science Foundation of China [41231171, 41271408, and 41571431], Arts and Sciences Excellence Professorship and Alvin and Sally Beaman Professorship at the University of Tennessee, Guangdong Natural Science Foundation [2014A030313684], and Shenzhen Fundamental Research Funding [JCYJ20140610151856728].

References

Ahas, R., *et al.*, 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology*, 17 (1), 3–27. doi:10.1080/10630731003597306

Ascough II, J.C., *et al.*, 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision making. *Ecological Modeling*, 219 (3–4), 383–399. doi:10.1016/j.ecolmodel.2008.07.015

Bar-Gera, H., 2007. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from Israel. *Transportation Research Part C: Emerging Technologies*, 15, 380–391. doi:10.1016/j.trc.2007.06.003

Batty, M., 2010. The pulse of the city. *Environment and Planning B: Planning and Design*, 37 (4), 575–577. doi:10.1068/b3704ed

Becker, R., *et al.*, 2013. Human mobility characterization from cellular network data. *Communication of the ACM*, 56 (1), 74–82. doi:10.1145/2398356.2398375

Bengtsson, L., *et al.*, 2011. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS Medicine*, 8 (8), e1001083. doi:10.1371/journal.pmed.1001083

Blondel, V., *et al.*, 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10), 1–12. doi:10.1088/1742-5468/2008/10/P10008

Brockmann, D., Hufnagel, L., and Geisel, T., 2006. The scaling laws of human travel. *Nature*, 439, 462–465. doi:10.1038/nature04292

Brown, L.A. and Moore, E.G., 1970. The intra-urban migration process: a perspective. *Geografiska Annaler. Series B, Human Geography*, 52 (1), 1–13. doi:10.2307/490436

Calabrese, F., Reades, J., and Ratti, C., 2010a. The geography of taste: analyzing cell-phone mobility and social events. *In*: P. Floréen, A. Krüger, and M. Spasojevic, eds. *Pervasive computing*. LNCS, 6030. Berlin: Springer Berlin Heidelberg, 22–37.

Calabrese, F., Reades, J., and Ratti, C., 2010b. Eigenplaces: segmenting space through digital signatures. *IEEE Pervasive Computing*, 9 (1), 78–84. doi:10.1109/MPRV.2009.62

Clauset, A., Newman, M.E.J., and Moore, C., 2004. Finding community structure in very large networks. *Physical Review E*, 70 (6), 066111. doi:10.1103/PhysRevE.70.066111

Couclelis, H., 2003. The certainty of uncertainty: GIS and the limits of geographic knowledge. *Transactions in GIS*, 7 (2), 165–175. doi:10.1111/tgis.2003.7.issue-2

Couronne, T., Olteanu, A.-M., and Smoreda, Z., 2011. Urban mobility: velocity and uncertainty in mobile phone data. *In*: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*. Boston, MA: IEEE, 1425–1430.

Delmelle, E., *et al.*, 2014. Visualizing the impact of space-time uncertainties on dengue fever patterns. *International Journal of Geographical Information Science*, 28 (5), 1107–1127. doi:10.1080/13658816.2013.871285

Gao, S., *et al.*, 2013. Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, 17 (3), 463–481. doi:10.1111/tgis.12042

Girardin, F., *et al.*, 2009. Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructure Research*, 4, 175–200.

Girvan, M. and Newman, M.E.J., 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (12), 7821–7826. doi:10.1073/pnas.122653799

Gollege, R.G. and Stimson, R.J., 1997. *Spatial behavior: A geographic perspective*. New York, NY: The Guilford Press.

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

Goodchild, M. and Gopal, S., 1989. *Accuracy of spatial databases*. London: Taylor and Francis.

Griffith, D.A., *et al.*, 2007. Impacts of positional error on spatial regression analysis: a case study of address locations in Syracuse. *Transactions in GIS*, 11 (5), 655–679. doi:10.1111/j.1467-9671.2007.01067.x

Hägerstrand, T., 1970. What about people in regional science? *Papers of the Regional Science Association*, 24 (1), 6–21. doi:10.1007/BF01936872

Hecht, B. and Stephens, M., 2014. A tale of cities: urban biases in volunteered geographic information. *In*: *Proceedings of the International Workshop on Web and Social Media*. Ann Arbor, MI: AAAI Publications, 197–205.

Horton, F. and Reynolds, D.R., 1971. Effects of urban spatial structure on individual behavior. *Economic Geography*, 47 (1), 36–48. doi:10.2307/143224

Iovan, C., *et al.*, 2013. Moving and calling: mobile phone data quality measurements and spatio-temporal uncertainty in human mobility studies. *In*: D. Vandenbroucke, B. Bucher, and J. Crompvoets, eds. *Geographic information science at the heart of Europe*. Cham: Springer, 247–265.

Jacquez, G., 2012. A research agenda: does geocoding positional error matter in health GIS studies? *Spatial and Spatio-Temporal Epidemiology*, 3 (1), 7–16. doi:10.1016/j.sste.2012.02.002

Kang, C., *et al.*, 2010. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. *In*: *Proceedings of the 18th International Conference on Geoinformatics*, Beijing, China. Piscataway, NJ: IEEE, 1–7.

Kang, C., *et al.*, 2012. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19 (4), 3–21. doi:10.1080/10630732.2012.715479

Liu, Y., *et al.*, 2008. Evaluation of cell phone traffic data in Minnesota. *Transportation Research Record: Journal of the Transportation Research Board*, 2086, 1–7. doi:10.3141/2086-01

Liu, Y., *et al.*, 2012. Understanding intra-urban trip patterns from taxi trajectory data. *Journal of Geographical Systems*, 14 (4), 463–483. doi:10.1007/s10109-012-0166-z

MacEachren, A.M., *et al.*, 2005. Visualizing geospatial information uncertainty: what we know and what we need to know. *Cartography and Geographic Information Science*, 32 (3), 139–160. doi:10.1559/1523040054738936

Mislove, A., *et al.*, 2011. Understanding the demography of Twitter users. *In*: *Fifth International AAAI Conference on Weblogs and Social Media*.

National Bureau of Statistics in China, 2012. *Annual GDP of Major Cities in China*. Available from: http://data.stats.gov.cn/workspace/index?m=csnd [Accessed 25 June 2015].

Newman, M.E.J., 2004. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69 (6), 066133. doi:10.1103/PhysRevE.69.066133

Pang, A., 2001. Visualizing uncertainty in geo-spatial data. *In*: *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology*. Washington, DC: National Academies Committee of the Computer Science and Telecommunications Board.

Pei, T., *et al.*, 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographic Information Science*, 28 (9), 1988–2007. doi:10.1080/13658816.2014.913794

Perchoux, C., *et al.*, 2013. Conceptualization and measurement of environmental exposure in epidemiology: accounting for activity space related to daily mobility. *Health & Place*, 21, 86–93. doi:10.1016/j.healthplace.2013.01.005

Reades, J., *et al.*, 2007. Cellular census: explorations in urban data collection. *IEEE Pervasive Computing*, 6 (3), 30–38. doi:10.1109/MPRV.2007.53

Refsgaard, J.C., *et al.*, 2007. Uncertainty in the environmental modelling process – a framework and guidance. *Environmental Modelling & Software*, 22 (11), 1543–1556. doi:10.1016/j.envsoft.2007.02.004

Rhee, I., *et al.*, 2011. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)*, 19 (3), 630–643. doi:10.1109/TNET.2011.2120618

Rietveld, P. and Vickerman, R., 2004. Transport in regional science: the "death of distance" is premature. *Papers in Regional Science*, 83 (1), 229–248. doi:10.1007/s10110-003-0184-9

Shanghai Municipal Statistics Bureau, 2012. *Land area, resident population, and population density of districts and counties (2012)*. Available from: http://www.stats-sh.gov.cn/tjnj/nj13.htm?d1=2013tjnj/C0202.htm [Accessed 25 June 2015].

Sherman, J.E., *et al.*, 2005. A suite of methods for representing activity space in healthcare accessibility study. *International Journal of Health Geographics*, 4, 24. doi:10.1186/1476-072X-4-24

Song, C., *et al.*, 2010a. Limits of predictability in human mobility. *Science*, 327 (5968), 1018–1021. doi:10.1126/science.1177170

Song, C., *et al.*, 2010b. Modelling the scaling properties of human mobility. *Nature Physics*, 6, 818–823. doi:10.1038/nphys1760

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit Region. *Economic Geography*, 46, 234–240.

Veregin, H., 1999. Data quality parameters. *In*: P.A. Longley, *et al.*, eds. *Geographic information systems*. New York, NY: Wiley, 177–189.

Walsh, F. and Pozdnoukhov, A., 2011. Spatial structure and dynamics of urban communities. *In*: *Proceedings of the 2011 Workshop on Pervasive Urban Applications (PURBA*), 12–15 June, San Francisco, CA, 1–8.

Wang, Y., Lai, P., and Sui, D., 2003. Mapping the Internet using GIS: the death of distance hypothesis revisited. *Journal of Geographical Systems*, 5 (4), 381–405. doi:10.1007/s10109-003-0117-9

World Shipping Council, 2013. *Top 50 World Container Ports*. Available from: http://www.worldshipping.org/about-the-industry/global-trade/top-50-world-container-ports [Accessed 25 June 2015].

Xiong, H., *et al.*, 2012. Predicting mobile phone user locations by exploiting collective behavioral patterns. *In*: *Proceedings* of the *9th International Conference on Ubiquitous Intelligence & Computing and 9th International Conference on Autonomic & Trusted Computing (UIC/ATC*). Fukuoka: IEEE, 164–171.

Xu, Y., *et al.*, 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation*, 42 (4), 625–646. doi:10.1007/s11116-015-9597-y

Yuan, Y. and Raubal, M., 2012. Extracting dynamic urban mobility patterns from mobile phone data. *In*: N. Xiao, *et al.*, eds. *GIScience 2012*. LNCS, 7478. Columbus, OH. Berlin: Springer, 354–367.

Yuan, Y., Raubal, M., and Liu, Y., 2012. Correlating mobile phone usage and travel behavior – a case study of Harbin, China. *Computers, Environment and Urban Systems*, 36 (2), 118–130. doi:10.1016/j.compenvurbsys.2011.07.003

Zenk, S.N., *et al.*, 2011. Activity space environment and dietary and physical activity behaviors: a pilot study. *Health & Place*, 17 (5), 1150–1161. doi:10.1016/j.healthplace.2011.05.001

Zhang, J. and Goodchild, M., 2002. *Uncertainty in geographic information*. New York, NY: CRC Press.

Zinszer, K., *et al.*, 2010. Residential address errors in public health surveillance data: a description and analysis of the impact on geocoding. *Spatial and Spatio-Temporal Epidemiology*, 1 (2–3), 163–168. doi:10.1016/j.sste.2010.03.002