International Journal of Geo-Information

MDPI

Article

Understanding the Representativeness of Mobile Phone Location Data in Characterizing Human Mobility Indicators

Shiwei Lu [1,*], Zhixiang Fang [1,2], Xirui Zhang [3], Shih-Lung Shaw [1,2,4], Ling Yin [5], Zhiyuan Zhao [1] and Xiping Yang [1]

1 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China; zxfang@whu.edu.cn (Z.F.); sshaw@utk.edu (S.-L.S.); zhaozhiyuan@whu.edu.cn (Z.Z.); 0yangxiping0@163.com (X.Y.)
2 Collaborative Innovation Center of Geospatial Technology, 129 Luoyu Road, Wuhan 430079, China
3 Information Center of Urban Planning, Land & Real Estate of Shenzhen Municipality, 8007 Hongli West Road, Shenzhen 518040, China; xrzhangchn@gmail.com
4 Department of Geography, University of Tennessee, Knoxville, TN 37996, USA
5 Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Road, Shenzhen 518005, China; yinling@siat.ac.cn
* Correspondence: lusw@whu.edu.cn; Tel.: +86-27-6877-9889

Academic Editors: Norbert Bartelme and Wolfgang Kainz
Received: 8 November 2016; Accepted: 2 January 2017; Published: 6 January 2017

Abstract: The advent of big data has aided understanding of the driving forces of human mobility, which is beneficial for many fields, such as mobility prediction, urban planning, and traffic management. However, the data sources used in many studies, such as mobile phone location and geo-tagged social media data, are sparsely sampled in the temporal scale. An individual's records can be distributed over a few hours a day, or a week, or over just a few hours a month. Thus, the representativeness of sparse mobile phone location data in characterizing human mobility requires analysis before using data to derive human mobility patterns. This paper investigates this important issue through an approach that uses subscriber mobile phone location data collected by a major carrier in Shenzhen, China. A dataset of over 5 million mobile phone subscribers that covers 24 h a day is used as a benchmark to test the representativeness of mobile phone location data on human mobility indicators, such as total travel distance, movement entropy, and radius of gyration. This study divides this dataset by hour, using 2- to 23-h segments to evaluate the representativeness due to the availability of mobile phone location data. The results show that different numbers of hourly segments affect estimations of human mobility indicators and can cause overestimations or underestimations from the individual perspective. On average, the total travel distance and movement entropy tend to be underestimated. The underestimation coefficient results for estimation of total travel distance are approximately linear, declining as the number of time segments increases, and the underestimation coefficient results for estimating movement entropy decline logarithmically as the time segments increase, whereas the radius of gyration tends to be more ambiguous due to the loss of isolated locations. This paper suggests that researchers should carefully interpret results derived from this type of sparse data in the era of big data.

Keywords: era of big data; mobile phone location data; human mobility; representative issue

1. Introduction

Understanding human mobility is of crucial importance [1,2], with potential benefits for various fields such as mobility prediction [3,4], urban planning [5–7], transportation research [8,9], and human

health research [10]. With the rapid development of information and communication technology [11] in the past two decades, various types of massive digital footprints generated by humans such as smart card data, call detail records (CDRs), geo-tagged social media data, GPS tracking data, WiFi data, credit-card records data, and their concomitant analytics are used for human mobility research [2,12–18]. However, there is debate regarding the representativeness or inherent biases of the data. For example, previous studies demonstrate that mobile phone users are unevenly distributed in age, gender, and geography [19,20]. This type of bias also exists in social media data [21,22].

Unlike GPS tracking data that can have multiple records per minute [23], a main disadvantage of the data used in previous research, such as mobile phone location and social media check-in data, is that it is very '*sparsely*' sampled on a temporal scale. Thus, an individual's records can be distributed over a few hours a day, or week, or over just a few hours a month, due to the uneven distribution of peoples' phone activities in space and time, which is an issue that requires attention to the data [24]. Previous researchers have discussed how CDRs can introduce biases in human mobility research [25,26] and how the level of deviation is closely related to the ratio of sampled phone communication records in an individual's trajectory [26]. In addition, Sagarra et al. [27] proposed a supersampled model to assess the sampling biases of reduced data. The representativeness of different time segments has not been investigated comprehensively due to the lack of ground truth for trajectories. What is the representativeness of sparse mobile phone location data on estimations of human mobility? This question must be addressed before using data to study human mobility patterns and derive results.

In this paper, we quantitatively analyze the representativeness of mobile phone location data on estimations of individual human mobility patterns. CDRs usually capture individual footprints during phone communication, whereas the actively tracked mobile phone location data contains phone communication records and location records triggered by location update *strategies* such as periodic and regular updates and cellular handover. This study uses active tracking data to conduct the investigation. Figure 1 shows an individual's complete trajectory from our mobile phone location dataset over an entire day. The Voronoi tessellations were used to represent the service areas of cell phone towers. It is difficult to determine a real path because most cell phone towers had not been recorded even under active updating strategies. Therefore, the main research question of this study is to determine the effects of sparsely temporally sampled mobile phone location data on the evaluation of human mobility indicators.

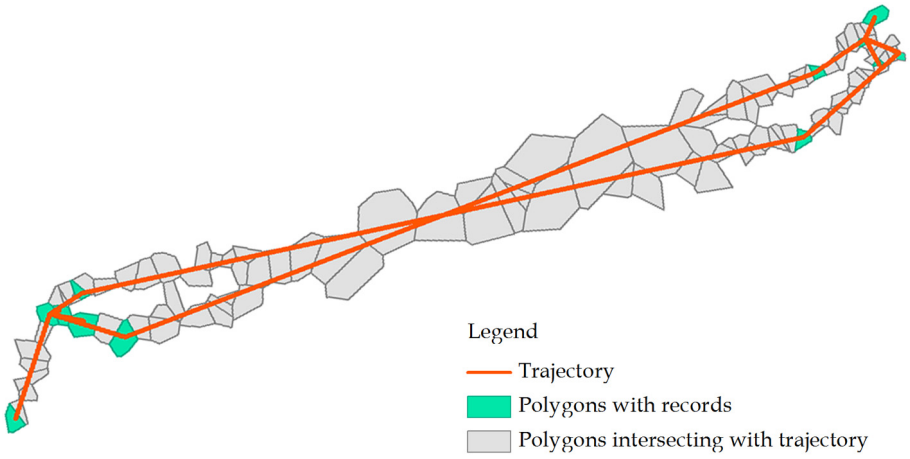

Figure 1. An example of one user's trajectory. Clearly, most of the Voronoi tessellations that intersect the trajectory are not recorded.

This paper investigates this question and provides several suggestions to select appropriate dataset to analyze human mobility. The findings of this research can also be used to evaluate the representativeness of other types of sparsely sampled data, such as geo-tagged social media data.

This paper is organized as follows: in the second section, we provide a review of studies related to this research. Section 3 introduces the active tracking mobile phone location dataset and the study area. Section 4 describes the method for evaluating the representativeness of sparse mobile phone data for measurement of human mobility indicators. Section 5 discusses the analysis results. The last section summarizes our findings and discusses future research directions.

2. Literature Review

This section presents relevant research in the following two areas; big data for human mobility research and representative issues of big data.

2.1. Mobile Phone Location Data for Human Mobility Research

Many valuable findings related to human mobility and interaction with urban environments have been reported in recent years with the advent of big data. These profound research studies can be used for mobility prediction [3,4], urban planning [5–7], transportation research [8,9,28], and other fields [10,29]. Among the datasets, mobile phone location data is very special data because mobile phones have an extremely high penetration rate and people usually take their cell phones with them, especially in Asian countries such as China. Some researchers view this type of data as a reasonable source to describe human mobility [30].

By using the sparsely sampled mobile phone location data, Kung et al. [31] explored the home-work commuting patterns of several cities in different countries and discovered that the commute time and average value distributions are independent of commute distance or country. Diao et al. [32] discovered the common laws governing an individual's activity participation and extracted the embedded information by presenting an activity detection model with travel diary surveys. Human footprints can also be used to analyze the spatial-temporal patterns of convergence and divergence in urban areas [33]. For transportation research, trip chain segments derived from mobile phone location data can be used to estimate the dynamic potential demand of bicycle trips in public transportation planning [34]. By estimating the dynamic origin-destination matrices, weekday and weekend travel patterns have been portrayed to analyze differences in travel demand over time [35]. However, how good are the subsample datasets in providing a good estimation of mobility patterns? The answer to this question is not simply yes or no, but investigations in the representativeness of the sparsely sampled location data may help to find some answers.

In addition, the human activity space and the mobility heterogeneity in this space are also the topics in many studies regarding human mobility research [2,36–39]. For instance, González et al. [2] found out that the radius of gyration for all individuals can be approximated with a truncated power-law. Yuan et al. [37] explored the relationships between phone usage and indicators of travel behavior characterized by movement entropy and radius. The absence of some outlying location points in the sparse mobile phone records may influence the calculation of movement radius in real scene. Calabrese et al. [30] compared the total trip length between mobile phone and vehicle data and demonstrated that using the Euclidean distance between cell phone towers to measure individual mobility could bring some downwards bias, but whether the sparse distribution of location records is also one of the reasons for this bias needs to be validated. Song and Barabási [39] and Gallotti [40] used entropy to predict individual mobility patterns. Moreover, Cuttone et al. [41] found out that there are also some relationships between the spatial and temporal resolution of the mobile phone data and the accuracy of predicting human mobility. The effectiveness of the sparse location data in the characterization of individual human mobility should be paid more attention.

Moreover, from the literature reviewed above, many indicators are used to characterize the human mobility patterns, such as the radius of gyration [2,38], movement entropy [37,39,40], and travel distance [26,30]. These indicators are usually used to characterize the travel distance, range of activity space, and heterogeneity of visitation patterns, which are three of the fundamental indicators

in human mobility. However, few studies have reported how representative the sparse location data is in the characterization of individual human mobility.

2.2. Representative Issues of Big Data

Despite the eager study of big data, there are also debates regarding privacy [42–44], data quality [45–48], and representative issues [25,26]. Previous studies demonstrate that mobile phone users are unevenly distributed in gender and geography [19,20] and population component [49]. This type of bias also exists in social media data [21,22]. The effects of spatial sampling and the granularity of sparse location data have also been studied [24,50].

Temporal sampling issues are of critical importance in using data to investigate human mobility patterns. GPS tracking data can have relatively fine granularity from both temporal and spatial perspectives [23,51]; however, the mobile phone location data and geo-tagged social media data used in previous studies are very sparsely temporally sampled due to the uneven distribution of peoples' phone activities in space and time, which is the main issue that requires attention [24]. An individual's mobile phone records or social media check-in data can be distributed over a few hours a day, or a week, or just a few hours a month. Goodchild [52] indicated that losses in quality control and rigorous sampling are characteristics of big data that can distinguish it from small data. Although previous studies have demonstrated that sparsely sampled CDRs introduce some biases to human mobility research [25,26] and that the level of deviation is closely related to the ratio of CDRs in an individual's complete trajectory [26], they do not describe how to obtain a more representative dataset if the complete trajectory is not available for comparison.

The incompleteness of temporally or spatially sampled location data is also a considerable factor leading to uncertainty issues in *GIScience* [53,54], raising concerns regarding how uncertainties could affect the findings [55,56]. Some researchers think that long periods of time help increase sample size; Jacobs [57] notes that these data are large numbers of repeated observations over time and/or space and may not get rid of the sparse issue. The critical question of 'how good are mobile phone location data at providing an accurate estimation of individual mobility indicators?' remains to be addressed before using data to investigate human mobility patterns and derive reasonable results.

Thus, this paper quantitatively evaluates the representativeness of sparse mobile phone location data in estimations of individual human mobility indicators. We not only focus on determining the effects of different time segments on human mobility characterization but also on providing a clear quantitative cognition of the representativeness of data.

3. Study Area and Dataset

The study area of this research is Shenzhen, one of the largest cities in China. This section provides background information on Shenzhen and the active tracking mobile phone location dataset collected there.

3.1. Study Area

The population of Shenzhen is greater than 15 million in an approximately 2000 square kilometer area, reflecting the highest population density among Chinese cities. Its annual gross domestic product (GDP) ranked fourth among all cities in China [58], after Shanghai, Beijing, and Guangzhou. Located on the south coast of China, Shenzhen is across the border from Hong Kong (Figure 2). Shenzhen has developed into an influential international city. The prosperous socioeconomic status of Shenzhen makes it a good choice for human mobility and business area analyses.

Figure 2. Location of Shenzhen (from OpenStreetMap).

3.2. Data

The mobile phone location data used in our research was collected by a very large mobile phone company that includes approximately 60% of the entire mobile phone market in Shenzhen. Approximately 16 million subscribers' location records were collected during a single workday. Table 1 shows the attributes of the mobile phone location data. For privacy concerns, the user ID is encrypted. Mobile communication carriers record the closest cell phone tower each time the subscriber uses his or her phone. Unlike call detail records data, the mobile phone location data records in this paper contain the following connection types:

(1) Making and receiving calls;
(2) Sending and receiving text messages;
(3) Regular location updates (triggered by moving from one cell phone tower to another), and
(4) Periodic location update (triggered by tower pinging if a subscriber has no phone activities for a specified time period).

The (3) and (4) are two active update strategies for this dataset. The connection types were not given in this dataset. Even under the active update strategies, we cannot determine the actual path because most of the cell phone towers had not been recorded (Figure 1).

Table 1. Example of individuals' cell phone records during a day.

User ID	Date	Time	Longitude	Latitude
User 1	2012/**/**	05:28:37	114. *****	22. *****
User 1	2012/**/**	11:07:52	114. *****	22. *****
User 1	2012/**/**	13:51:12	114. *****	22. *****
...	...	...	...	...
User 2	2012/**/**	02:28:16	114. *****	22. *****
...	...	...	...	...

The sign ***** ignores the minutes of a Longitude or a Latitude, and the sign **/** ignores the exact month and day due to privacy protection.

There are 5940 unique cell phone towers in this dataset. Figure 3 shows the spatial kernel density of the cell phone towers. The cell phone towers are unevenly distributed in the urban space.

Overall the cell phone towers are densely distributed in the center of the city and in highly populated areas, whereas the cell phone towers are sparsely distributed in suburban areas, resulting in a lower positioning accuracy. The average distance and maximum distance between adjacent cell phone towers is about 0.21 and 2.6 km, respectively.

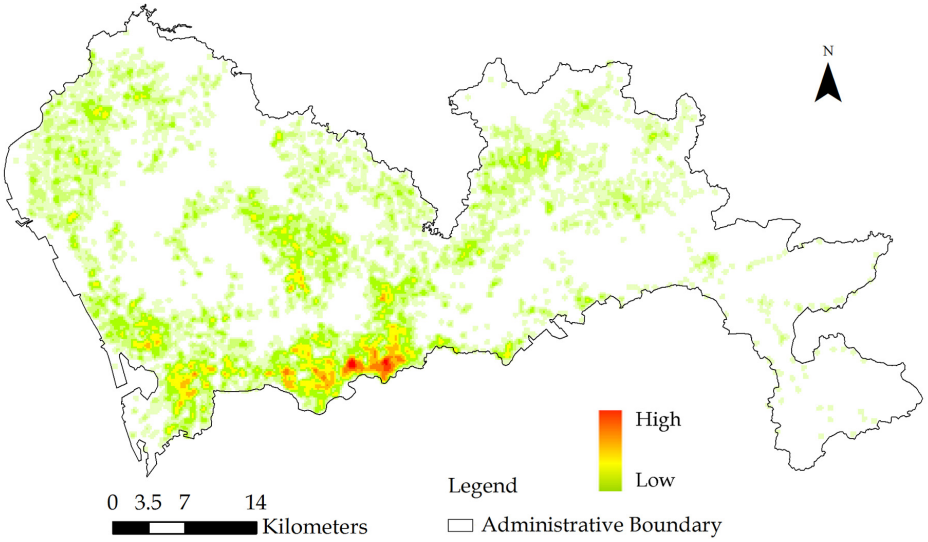

Figure 3. Cell phone tower spatial kernel density. Due to a dataset provider requirement, the point-based cell phone tower spatial distribution cannot be shown. The administrative boundary of Shenzhen is vectorized into shape-file according to the whitepaper of the Urban Planning, Land & Resources Commission of Shenzhen Municipality [59].

Since the focus of this paper is to investigate the representativeness of spare mobile phone location data in characterizing the human mobility patterns, the uneven distribution of people's phone activities in space and time is the main concern regarding to our research goal [24], while the dense distribution of cell phone towers across the urban area indicated that the spatial granularity at the cell phone tower level may not be a major drawback in this study area.

4. Methodology

This paper introduced the frequently used human mobility indicators. Thereafter, the method of evaluating the representativeness of mobile phone location data included three main steps. First, we divided the day into 24 hourly segments, extracted the subscribers whose records covered all the 24 hourly segments into a new dataset, and calculated their complete human mobility indicators as the benchmarks of this study. Then, we calculated the sampled human mobility indicators by selecting different numbers of time segments from the new dataset under random rules. Finally, a linear regression model was proposed to quantify the aggregated underestimation level between sampled and complete human mobility indicators in each random time.

4.1. Frequently used Human Mobility Indicators

There are many frequently used indicators to measure activity space, like maximum travel distance, radius of gyration, movement radius, total travel distance, movement entropy, visitation frequency, and so on. Mainly, these indicators could be classified into three categories, which are the range of activity space, the travel distance, and the heterogeneity of visitation patterns within the activity space. For instance, both of the movement entropy and visitation frequency are used to characterize the heterogeneity of visitation patterns. Thus, this paper used three of them to characterize human activity behavior. They are calculated based on a working day and defined as follows:

Total travel distance: The total travel distance is the sum of the Euclidian distance between each pair of consecutive records [26], which is a basic measure of individual mobility.

Movement entropy: A characterization of the heterogeneity of visitation patterns [37,38], calculated as

$$S = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (1)$$

where n is the number of distinct cell phone towers visited by a subscriber and p_i is the probability that location i is visited.

Radius of gyration: Describes how widely the subscriber travelled; one of the most frequently used measures to characterize the range of activity space [2,38], calculated as

$$R_g = \sqrt{\frac{1}{N}\sum_{j=1}^{N} |\vec{p_j} - \vec{p_{cm}}|^2} \quad (2)$$

where N is the number of time-sequenced cell phone towers visited by a subscriber, p_j is the *jth* tower that the subscriber visited, and p_{cm} is the center of all time-sequenced locations.

4.2. Extracting Valid Subscribers

After introducing the frequently used human mobility indicators, the method used to evaluate the representativeness of mobile phone location data in characterizing these indicators included three main steps, described below.

First, we extracted the subscribers whose records were sufficient for this research. We divided the day into 24 one-hour time segments; 00:00:00 to 00:59:59 (#0), 01:00:00 to 01:59:59 (#1), . . . , 23:00:00 to 23:59:59 (#23). In this paper, the number of time segments was used for describing the term of sparse sampled records from a temporal perspective. Then, we extracted subscribers whose records covered all 24 time segments.

Clearly, mobile phone location records of different subscribers are sparsely distributed in different numbers of time segments. The less time segments the subscriber's records are in, the sparser are the records from a temporal perspective. For example, about 3.37% of subscribers' records were just in one hour a day, and the percentage of users that have records in 24 temporal segments was 35.70%, which means that the records of almost 65% of users were distributed in less than 24 segments, as shown in Figure 4. Moreover, the records of approximately 13.18% of users were in 6 segments or less. Hence, it is questionable whether the mobility patterns of users can be properly characterized without covering enough temporal intervals.

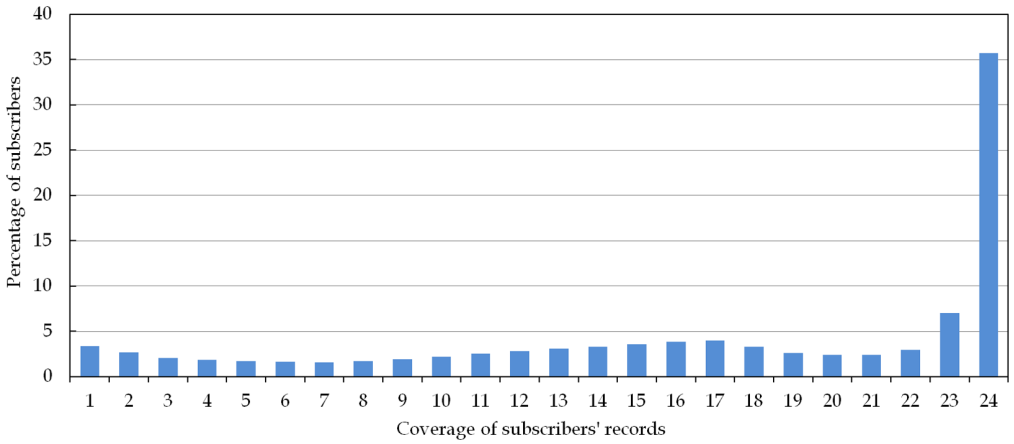

Figure 4. Temporal coverage of subscribers' records.

The data of 5.8 million subscribers were included in this research, thus it could be used to investigate the effects of different time segments in characterizing human mobility patterns. Perhaps these subscribers may habitually use their mobile phones more frequently than others. In addition, previous studies have demonstrated that mobile phone users are heterogeneously distributed in age, gender, and space [19,20]. Thus, mobile phone users in our subsample dataset may have different biases in these aspects, which need to be further explored in future work.

4.3. Random Rules

After the 5.8 million subscribers were extracted, we divided each subscriber's records into 24 time segments, as shown in Figure 5.

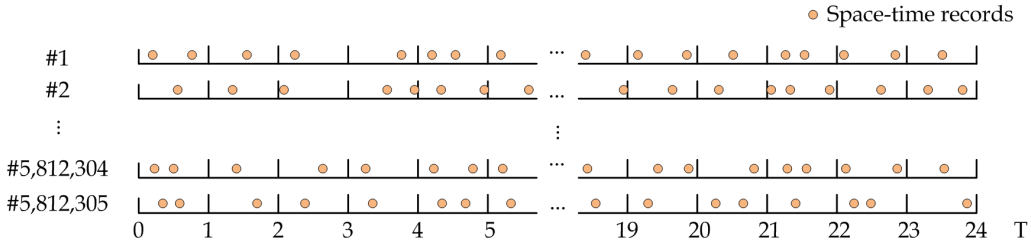

Figure 5. Dividing each subscriber's records into 24 time segments.

To investigate the representativeness of sparse mobile phone location data on estimation of individual human mobility indicators, this study varied the number of time segments selected from 2 to 23. For each number of time segments, the selection was randomized 100 times to ensure each time segment could be selected. For example, when the selected number of time segments was two, the time segments (#2, #5) or the segments (#3, #9) could be selected out; when the selected number of time segments was three, the segments (#4, #5, #21) or the segments (#2, #6, #17) could be selected out, as shown in Figure 6. In addition, the selected time segments were not repeated even if the number of time segments was the same. For instance, when the selected number of time segments was three, the segment combination (#4, #9, #22) was selected only once among all the 100 random times.

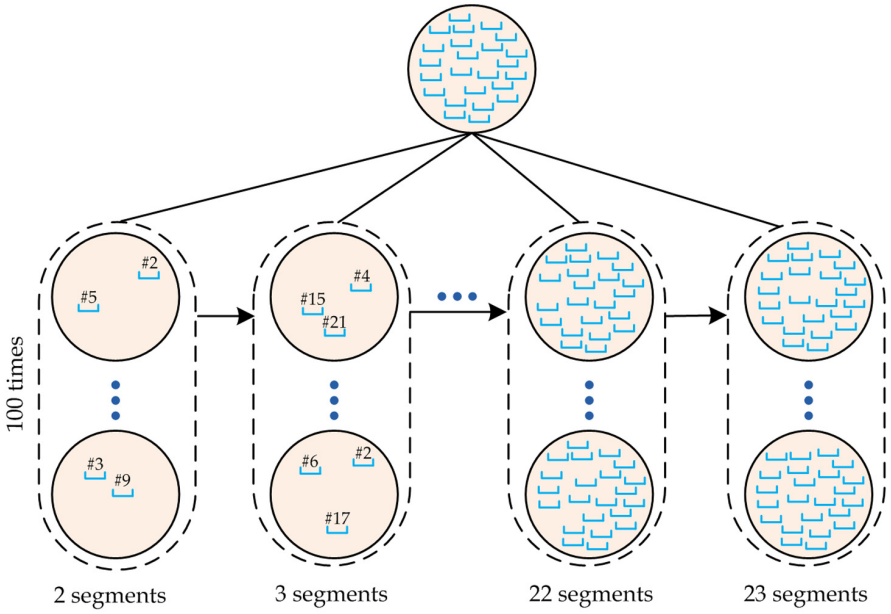

Figure 6. Random rules for selecting different numbers of time segments.

Moreover, for the same number of time segments, each segment should be selected at least five times among every 100 times. For example, in selecting two time segments, if the segments (#2, #5), (#2, #7), (#2, #16), (#2, #19), and (#2, #23) were selected, the #2 appeared five times but #5, #7, #16, #19, #23 appeared only once and the other 18 segments didn't appear. Thus, in the next 95 random times, more attention would be paid to 23 other segments in the random selections. This rule is designed to reduce the inequality in selecting each time segment.

When there are 23 time segments, there are only 24 choices from which to select the 23 segments. Each of the 5.8 million subscribers' randomly sampled mobility indicators were calculated by using all the mobile phone records in selected time segments at each random time.

4.4. Evaluating the Aggregated Underestimation Coefficient

For each random time, we calculated a set of sampled indicator values, the sampled total travel distance, the sampled movement entropy, and the sampled radius of gyration by using the sampled records in randomly selected time segments for all 5.8 million subscribers.

To quantify the aggregated underestimation level for sampled time segments in characterizing the human mobility indicators, a linear regression model was used [26].

$$y = ax + b \tag{3}$$

Here, for each random time, each of the mobility indicators calculated by using the complete records in the whole time segments are defined as the independent variable x, and the corresponding sampled mobility indicator are defined as the dependent variable y by using the records in randomly selected time segments. The coefficient a measures the relationship between sampled and complete indicators of all the 5.8 million subscribers. Here, b was set to 0 in the linear regression model because, when the mobility indicator in a complete benchmark dataset is 0, the mobility indicator in the selected dataset should also be 0. The coefficient a is calibrated by the least square regression method [60].

Thus, the aggregated underestimation coefficient (uc) is defined as follows:

$$uc = 1 - a \tag{4}$$

Clearly, the lower uc is, the lower the level of underestimation is and the more representative the randomly selected time segments are for characterizing human mobility indicators. For instance, when the selected time segments are (#3, #6, #7, #13, #19, and #21) and the coefficient between the sampled total travel distance and complete total travel distance is 0.25, the aggregated underestimation coefficient is 0.75. The uc is relatively high, which means the representativeness of these six time segments is low, because the total travel distance calculated by using records in these six time segments may be about 75% shorter than their total footprints in the study area.

5. Results

5.1. Measuring Mobility Indicators by Randomly Selecting Time Segments

This section analyzed the various differences between sampled mobility indicators and complete mobility indicators from the individual perspective. Then the quantitatively aggregated underestimation effects were explored from the average perspective.

5.1.1. Individual Perspective

This section focuses on evaluating the representativeness of sparse mobile phone location data in individual daily mobility pattern analysis. Examples of random mobility indicators and complete mobility indicators are shown in Table 2 and Figures 7–9.

The horizontal and vertical axes represent the mobility indicators from complete and random time segments, respectively. If random time segments are representative of the complete time segments,

the points on Figures 7–9 should be close to the light blue diagonal line from lower left to upper right. The representativeness of different time segments for estimations of individual mobility indicators are quite different, as the gray dots show. For example, when 10 time segments are used, the individual movement entropy is overestimated for 32.79% of subscribers and the individual radius of gyration is overestimated for 19.42% of subscribers.

Table 2. Mobility indicator statistics for different random time segments.

Time Segments		Total Travel Distance	Movement Entropy	Radius of Gyration
3 (#2, #14, #20)	Overestimation	0%	11.67%	17.16%
	Underestimation	100%	88.33%	82.84%
	Aggregated *uc*	0.86	0.49	0.48 (within 9 km)
	R^2 (*uc*)	0.291	0.943	0.901
10 (#5, #6, #7, #9, #11, #12, #14, #16, #17, #19)	Overestimation	0%	32.79%	19.42%
	Underestimation	100%	67.21%	80.58%
	Aggregated *uc*	0.52	0.18	0.34 (within 9 km)
	R^2 (*uc*)	0.894	0.986	0.882
23 (except #5)	Overestimation	0%	59.28%	8.94%
	Underestimation	100%	40.72%	91.06%
	Aggregated *uc*	0.05	0.01	0.29 (within 9 km)
	R^2 (*uc*)	0.995	0.999	0.882

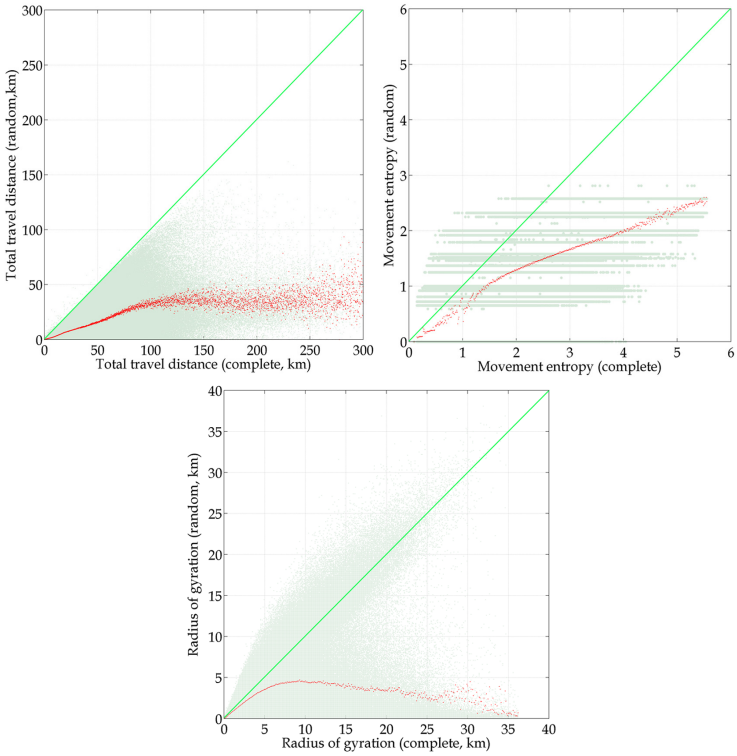

Figure 7. Human mobility indicators in 3 random segments (segment #2, #14 and #20). The light gray dots are the three random and complete mobility indicators for each subscriber. For the total travel distance and radius of gyration, the horizontal axis is 0.1 km bandwidth. For movement entropy, the horizontal axis is 0.01 bandwidth. The red dots are the average value of the gray dots in their corresponding bandwidth.

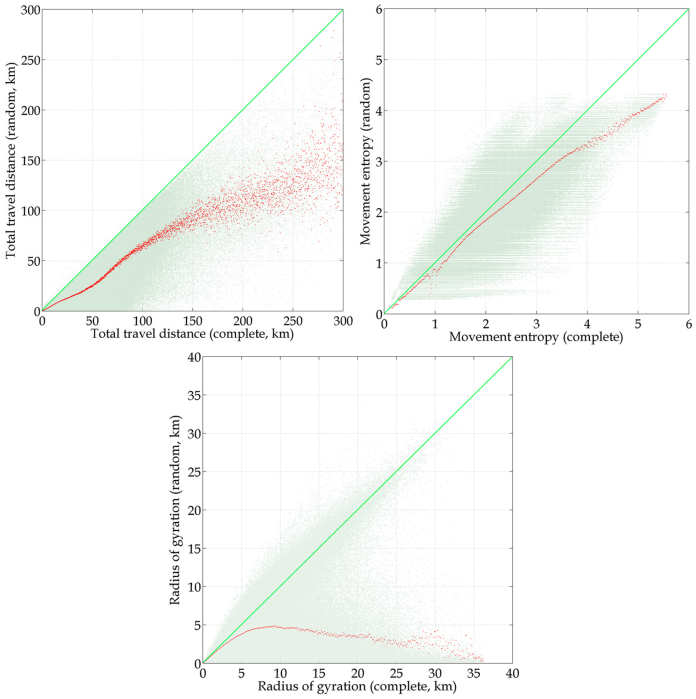

Figure 8. Human mobility indicators in 10 random segments (segment #5, #6, #7, #9, #11, #12, #14, #16, #17, and #19). The light gray dots are the three random and complete mobility indicators of each subscriber. For the total travel distance and radius of gyration, the horizontal axis is 0.1 km bandwidth. For movement entropy, the horizontal axis is 0.01 bandwidth. The red dots are the average of the gray dots in their corresponding bandwidth.

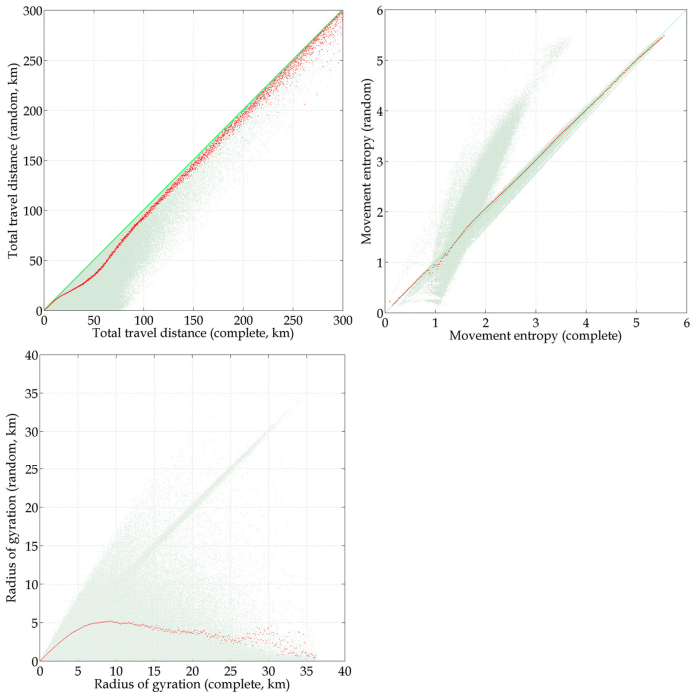

Figure 9. Human mobility indicators in 23 random segments (segment #0, #1, #2, #3, #4, #6, #7, #8, #9, #10, #11, #12, #13, #14, #15, #16, #17, #18, #19, #20, #21, #22, and #23). The light gray dots are the three random and complete mobility indicators for each subscriber. For the total travel distance and radius of gyration, the horizontal axis is 0.1 km bandwidth. For movement entropy, the horizontal axis is 0.01 bandwidth. The red dots are the average of the gray dots in their corresponding bandwidth.

The random total travel distance cannot be overestimated because fewer records lead to a shorter total travel distance due to the triangle principle. However, the movement entropy and radius of gyration could be overestimated or underestimated for different individuals due to the use of records from different time segments in the calculation. The average level is often used to characterize the distribution of the corresponding bandwidth [4,26,37,38]. The average level of the estimation was studied as described below.

5.1.2. Average Perspective

In Figures 7–9, there are deviations between the red dots and the blue diagonal line, which indicates that using fewer mobile phone location data time segments tends to underestimate the total travel distance, movement entropy, and radius of gyration from an average perspective, which can also be seen from Table 2.

From an average perspective, the underestimation coefficient of the total travel distance is 0.86 (R^2 = 0.291, goodness of fit [61]) when there are 3 time segments. When 10 time segments are used, the underestimation coefficient is 0.52 (R^2 = 0.894). The sampled total travel distance is not typically overestimated because fewer records lead to a shorter total travel distance. As the number of time segments increases, there are fewer deviations from the blue diagonal line for the total travel distance. Conversely, the variation in average total travel distance increases when the complete total travel distance increases. For example, when the complete total travel distance is 100 km for 10 time segments, the random travel distance is approximately 65 km but when the complete total travel distance is 200 km, the random total travel distance is between 70 km and 140 km. This was likely because the number of subscribers decreases rapidly as the total travel distance increases and because the location records in some time segments are distant from those in other time segments.

The total travel distance could be greater than 70 km, which was caused by subscribers such as taxi or bus drivers, package deliverers, and tourists. These subscribers account for less than 2.0% of the 5.8 million subscribers. Another interesting pattern is that the range of average total travel distance is supposed to be narrower when using 23 time segments. This is mainly because there are fewer random times and the selected records are very close to the total records for each individual.

The movement entropy could be overestimated or underestimated for different individuals due to calculation using records from different time segments. However, from an average perspective, the declining trend in the underestimation coefficient in estimating movement entropy can be observed from Figures 7–9. When 3 time segments are used, the underestimation coefficient of movement entropy is 0.49 (R^2 = 0.943), but when 10 time segments are used, the underestimation coefficient is 0.18 (R^2 = 0.986), which is very close to 0. Moreover, when 23 time segments are used, the points are close to the blue diagonal line and the underestimation coefficient is only 0.01, which means the records in these 23 time segments can represent the complete movement entropy entirely.

Unlike total travel distance and movement entropy, the distribution of the random average radius of gyration does not always increase with the complete radius of gyration. As shown in Figures 7–9, the random average radius of gyration increases until the complete radius of gyration is approximately 9 km. Then, although the complete radius of gyration increases, the average random radius of gyration declines. Therefore, the linear regression model is used within 9 km. To estimate the radius of gyration, incomplete mobile phone location records are probably good enough in most cases for analysis of subscriber travel within a normal daily activity range, i.e., less than 9 km.

In addition, for subscribers whose complete radius of gyration is greater than 9 km, the average random radius of gyration is often zero or very close to zero due to the loss of some long-distance locations. These subscribers account for less than 7.0% of all valid subscribers, usually travel in many different directions, and are likely to travel in a wide range. Thus, the lack of any time segments between 8 am and 8 pm may significantly affect the radius of gyration. The social identities of these subscribers may be taxi or Uber/Didi drivers, package deliverers, or tourists. Therefore, mobile phone location data might significantly underestimate the radius of gyration of subscribers whose activity

range is very wide (i.e., greater than 9 km). Moreover, the range of the random radius of gyration is supposed to be wider when the complete radius of gyration increases.

Most importantly, even the use of many time segments can generate a much smaller radius of gyration, which indicates that an incomplete trajectory remains questionable for deriving the range of daily activity space.

Using mobile phone location records from different numbers of time segments can generate very different results in the distribution of total travel distance, movement entropy, and radius of gyration, which indicate the distance, range, and heterogeneity of individual mobility patterns, respectively. Therefore, the representativeness of mobile phone location data should be addressed before using it to answer different research questions. Next, we provide a comprehensive comparison of the representativeness of different numbers of time segments and of the same number of time segments with different time slots using the underestimation coefficient from an average perspective.

5.2. Quantitative Analysis of the Total Travel Distance Underestimation Coefficient

To evaluate the representativeness of different numbers of time segments and of the same number of time segments with different slots, we varied the selected number of time segments from 2 to 23. For each number of time segments, we randomized the selection 100 times, except for when 23 time segments were used. For each random time, we can calculate the aggregated total travel distance underestimation coefficient. The distribution of the underestimation coefficients for estimating total travel distance is shown in Figure 10.

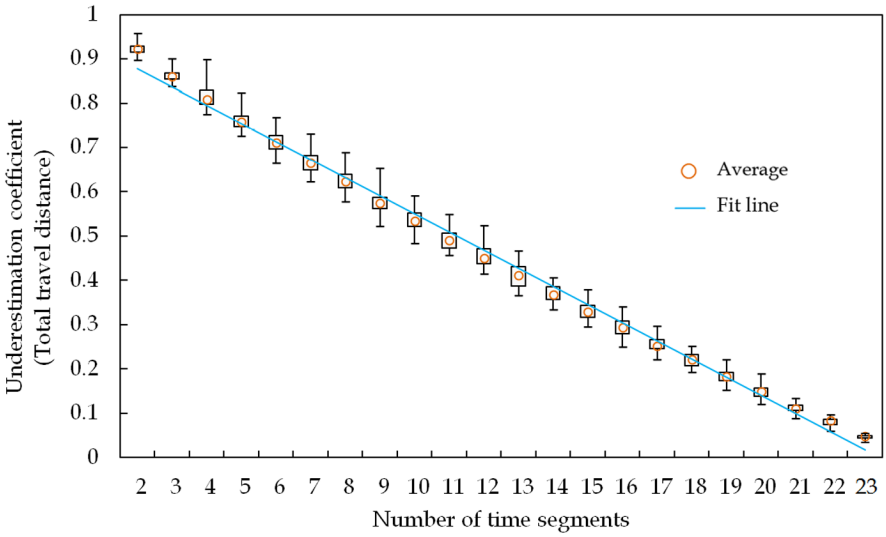

Figure 10. Distribution of aggregated underestimation coefficients for estimating total travel distance using different numbers of time segments.

First, it is obvious that, even with the same number of time segments, the underestimation coefficient can be quite different. For instance, when 4 time segments are used, the underestimation coefficient varies from 0.77 to 0.90 and when the 18 time segments are used, the underestimation coefficient is between 0.19 and 0.26. These patterns indicate that location records in different time segments have different representativeness for characterizing total travel distance in human mobility research. This is relatively easy to understand in the context of human activities: if the selected time segments are mainly related to home activity, the total travel distance tends to be shorter and the underestimation coefficient tends to be higher, but if the selected time segments cover home and work activity, the total travel distance tends to be larger, which leads to a lower underestimation coefficient.

Second, another interesting pattern is that, as the number of selected time segments increases, the underestimation coefficient tends to decline significantly. The average underestimation coefficient is 0.93 when the 2 time segments are used and declines to 0.04 when 23 time segments are used. By fitting

another linear regression model with an intercept, the declining trend is nearly linear (R^2 = 0.99) and n indicates the number of time segments.

$$\overline{uc}_d(n) = -0.04n + 0.92 \quad (5)$$

It is easy to determine how representative mobile phone location data is for estimating total travel distance using this model. For example, if each individual's records cover only eight time segments, the total travel distance may be approximately 60% shorter than their total footprint in the study area.

5.3. Quantitative Analysis of the Movement Entropy Underestimation Coefficient

Similarly, we can calculate a movement entropy underestimation coefficient for each random time. The distribution of the aggregated underestimation coefficients for estimating movement entropy is shown in Figure 11.

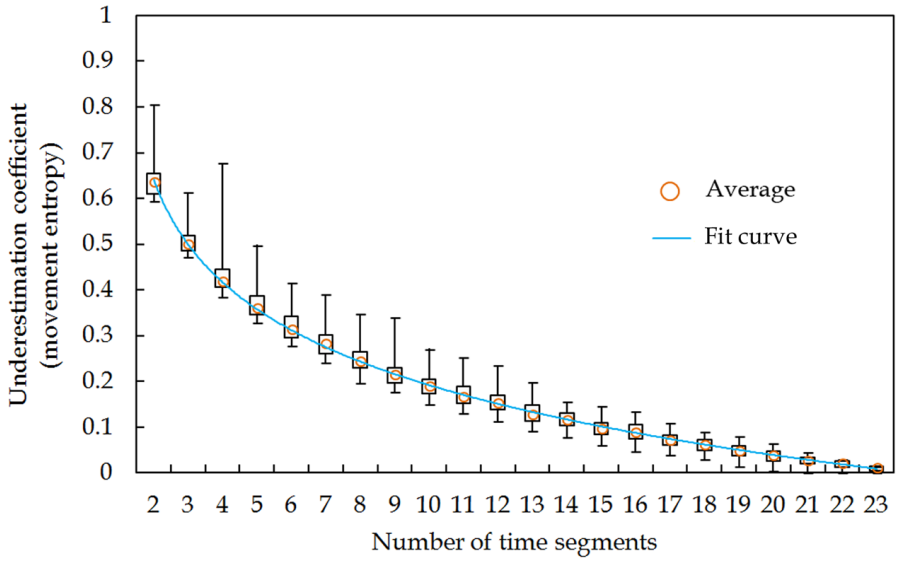

Figure 11. Distribution of aggregated underestimation coefficients for estimating movement entropy using different numbers of time segments.

As in the underestimation coefficient distribution for estimating total travel distance, it is evident that even with the same number of time segments, the underestimation coefficient can be quite different. For example, when 7 time segments are used, the underestimation coefficient varies from 0.24 to 0.39. This pattern is easy to understand as there may be new locations or the visiting frequency of some locations may change in different time segments. Moreover, the range of the underestimation coefficient is likely to be narrower as the number of time segments increases. For example, when 4 time segments are used, the underestimation coefficient varies from 0.39 to 0.68 and when 18 time segments are used, the underestimation coefficient is between 0.02 and 0.10. The average underestimation coefficient drops significantly from 0.64 to 0.20 when the number of time segments selected varies from 2 to 10.

Another interesting pattern is that, as the number of selected time segments increases, the underestimation coefficient tends to decline. The trend can be fitted by a logarithmic regression model with an intercept (R^2 = 0.99, n is the number of time segments).

$$\overline{uc}_s(n) = -0.20 \ln(n) + 0.64 \quad (6)$$

We can easily determine the representativeness of the mobile phone location data for estimating movement entropy using this model. For example, if each individual's records cover only eight time

segments, the underestimation coefficient is approximately 0.25, so the average movement entropy may be approximately 25% less than their total footprint in the study area.

5.4. Quantitative Analysis of the Radius of Gyration Underestimation Coefficient

Similarly, we can calculate an underestimation coefficient of the radius of gyration for each random time. The distribution of the aggregated underestimation coefficients for estimating radius of gyration is shown in Figure 12.

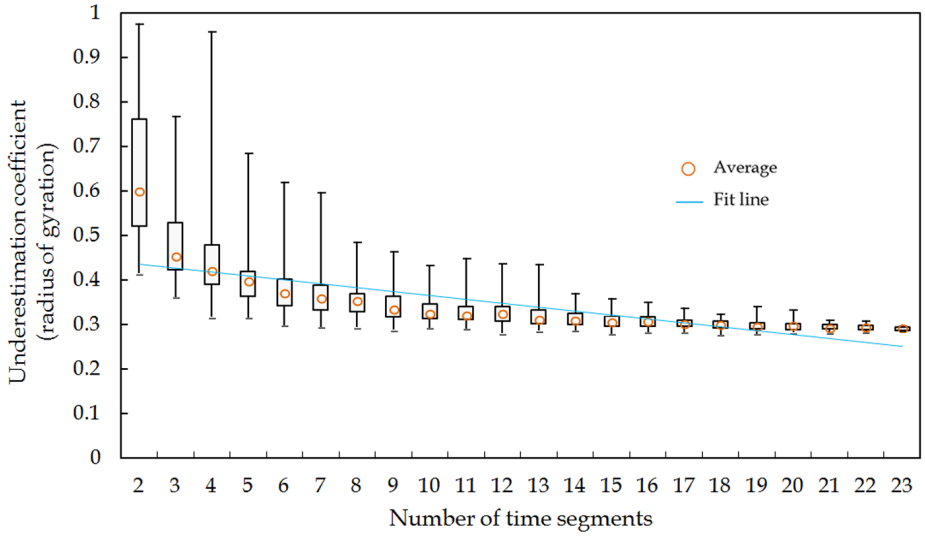

Figure 12. Distribution of aggregated underestimation coefficients for estimating radius of gyration using different numbers of time segments.

As interpreted in Section 5.1.2, to estimate the radius of gyration, incomplete mobile phone location records are probably good enough in most cases to analyze subscribers travel within a normal daily activity range, such as less than 9 km. Therefore, in this section, we mainly focus on radius of gyration less than 9 km.

Obviously, even with the same number of time segments, the underestimation coefficient can be quite different. For example, when 4 time segments are used, the underestimation coefficient varies from 0.31 to 0.97. This pattern is easy to understand, as there may be new locations due to different time segments. Moreover, the range of the underestimation coefficient is likely to be narrower as the number of time segments increases. For example, when 3 time segments are used, the underestimation coefficient varies from 0.36 to 0.77 and when there are more than 15 time segments used, the underestimation coefficient is between 0.28 and 0.35.

The declining trend could be fitted by a linear regression model with an intercept (R^2 = 0.63, n is the number of time segments), we can easily determine how representative the mobile phone location data is for estimating the radius of gyration within 9 km using this model. Unlike the total travel distance and movement entropy, the goodness of fit (R^2) is only 0.63.

$$\overline{uc}_r(n) = -0.009n + 0.44 \quad (7)$$

The radius of gyration is likely to be more uncertain with fewer selected time segments. As can be seen from Figure 12, the average underestimation coefficient is greater than 0.29 even when 23 time segments are used, which means that any number of sampled time segments could depict the range of daily travel as at least 29% shorter than their total footprint in the study area. In addition, as has been interpreted in Section 5.1.2, for subscribers whose activity range is greater than 9 km, the sampled radius of gyration could often be much lower due to the absence of outlying location point. This also

indicates that radius of gyration may not be the most appropriate measurement for characterizing the range of human mobility by using sparsely sampled location data, such as mobile phone location data. Thus, we suggest researchers use indicators cautiously to interpret results derived from sparely sampled location data.

Finally, based on the results and distribution of subscribers in Figure 4, if given a real sample of mobile-tracked individuals and supposing that the *uc* for 24 time segments is 0, the weighted underestimation levels of the total travel distance, the movement entropy, and the radius of gyration are about 23%, 11% and 21%, respectively, in this study area.

6. Conclusions

In this paper, we investigated the representativeness of sparse mobile phone location data in characterizing mobility indicators, which are used for measuring the range of activity space, the travel distance, and the heterogeneity of visitation patterns within activity space. The contribution of this study is threefold:

Firstly, the case study shows that the representativeness of estimations of human mobility indicators for each individual can lead to overestimation or underestimation. However, from an average perspective [4,26,37,38], when compared with all of the records, incomplete mobile phone location data tends to underestimate mobility indicators, such as average total travel distance and movement entropy. Moreover, the underestimation of the average radius of gyration becomes more significant. The representativeness of mobile phone data is also dependent on the records in different time segments.

Secondly, this study quantitatively assesses the representativeness of randomly selected time segments from the benchmark dataset in characterizing human mobility indicators. The aggregated underestimation coefficient results for estimating the total travel distance linearly decline as the number of time segments increases. For example, if each individual's records cover only 33% of the trajectory, the total travel distance may be approximately 60% shorter on average than their total footprints in the study area. The aggregated underestimation coefficient results for estimating movement entropy logarithmically declines as the number of time segments increases. For instance, if each individual's records cover only 33% of the trajectory, the aggregated underestimation coefficient is approximately 0.25, so the movement entropy may be approximately 25% less on average than their total footprint in the study area.

Lastly, the underestimation effects can be very significant for the radius of gyration, and the average underestimation coefficient is greater than 0.29 even when 23 time segments are used, which means incomplete mobile phone location data could depict an average of daily travel approximately 29% shorter than their total footprints in the study area. This may indicate that the radius of gyration should be used cautiously, because it is easily underestimated by using sparsely sampled location data, such as mobile phone location data. However, our findings may or may not be applicable to other cities due to different urban environments and mobile phone usage habits.

This study presents an alternative way to evaluate the representativeness of mobile phone location data for human mobility research. The method proposed in this paper can also be used for coarse data such as geo-tagged social media check-in data. Using the investigative approach here, researchers can understand the strengths and limitations of their data to help derive reasonable results. However we do note several limitations and challenges specific to sparsely sampled location data, such as:

(1) The mobile phone usage habits; Figure 4 shows that the temporal coverage of subscribers' records are mostly relatively low, which may be related to subscribers' mobile phone usage habits. So the underestimation coefficient may be higher in non-random sampled mobile phone location data if the subscribers travel a lot but rarely take their mobile phones.

(2) The bias of using subsamples instead of whole datasets; mobile phone users in subsample datasets may have different biases in gender, age, or geography [19,20]. We will further explore the effects of this bias in characterizing human mobility patterns in future study.

Acknowledgments: The authors would like to thank the valuable comments from anonymous reviewers. This study was jointly supported by the National Natural Science Foundation of China (Grants #41231171, #41371420, #41371377 and #41301511), the innovative research funding of Wuhan University (2042015KF0167), the Arts and Sciences Excellence Professorship and the Alvin and Sally Beaman Professorship at the University of Tennessee, and the International Science-technology Cooperation Project of Guangdong Province (2014A050503053).

Author Contributions: This research was mainly formulated and designed by Shiwei Lu and Zhixiang Fang. Ling Yin provided the dataset. Shiwei Lu, Xirui Zhang, Zhiyuan Zhao, and Xiping Yang performed the experiments and analyzed the data. Shiwei Lu, Zhixiang Fang, and Shih-Lung Shaw wrote the manuscript. Zhixiang Fang, Shih-Lung Shaw, and Ling Yin reviewed the manuscript and provided comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brockmann, D.; Hufnagel, L.; Geisel, T. The scaling laws of human travel. *Nature* **2006**, *439*, 462–465. [CrossRef] [PubMed]
2. González, M.C.; Hidalgo, C.A.; Barabási, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779–782. [CrossRef] [PubMed]
3. Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; Mascolo, C. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE* **2011**. [CrossRef]
4. Simini, F.; González, M.C.; Maritan, A.; Barabási, A.L. A universal model for mobility and migration patterns. *Nature* **2012**, *484*, 96–100. [CrossRef] [PubMed]
5. Ratti, C.; Frenchman, D.; Pulselli, R.M.; Williams, S. Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plan. B Plan. Des.* **2006**, *33*, 727–748. [CrossRef]
6. Gao, S.; Liu, Y.; Wang, Y.; Ma, X. Discovering spatial interaction communities from mobile phone data. *Trans. GIS* **2013**, *17*, 463–481. [CrossRef]
7. Pei, T.; Sobolevsky, S.; Ratti, C.; Shaw, S.L.; Li, T.; Zhou, C. A new insight into land use classification based on aggregated mobile phone data. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 1988–2007. [CrossRef]
8. Caceres, N.; Romero, L.M.; Benitez, F.G.; Del Castillo, J.M. Traffic flow estimation models using cellular phone data. *IEEE Trans. Intell. Transp. Syst.* **2012**, *13*, 1430–1441. [CrossRef]
9. Gao, H.; Liu, F. Estimating freeway traffic measures from mobile phone location data. *Eur. J. Oper. Res.* **2013**, *229*, 252–260. [CrossRef]
10. Dewulf, B.; Neutens, T.; Lefebvre, W.; Seynaeve, G.; Vanpoucke, C.; Beckx, C.; Weghe, N.V. Dynamic assessment of exposure to air pollution using mobile phone data. *Int. J. Health Geogr.* **2016**, *15*, 1–14. [CrossRef] [PubMed]
11. Lu, T.; Guo, X.; Xu, B.; Zhao, L.; Peng, Y.; Yang, H. Next Big Thing in Big Data: The Security of the ICT Supply Chain. *Int. Conf. Soc. Comput.* **2013**, *10*, 1066–1073.
12. Järv, O.; Ahas, R.; Witlox, F. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Trans. Res. Part C: Emerg. Technol.* **2014**, *38*, 122–135. [CrossRef]
13. Liu, Y.; Sui, Z.; Kang, C.; Gao, Y. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PLoS ONE* **2014**. [CrossRef] [PubMed]
14. Zhong, C.; Batty, M.; Manley, E.; Wang, J.; Wang, Z.; Chen, F.; Schmitt, G. Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. *PLoS ONE* **2016**. [CrossRef] [PubMed]
15. Lenormand, M.; Louail, T.; Cantúros, O.G.; Picornell, M.; Herranz, R.; Arias, J.M.; Barthelemy, M.; Miguel, M.S.; Ramasco, J.J. Corrigendum: Influence of sociodemographic characteristics on human mobility. *Sci. Rep.* **2014**. [CrossRef] [PubMed]
16. Zheng, Y.; Li, Q.; Chen, Y.; Xie, X.; Ma, W.Y. Understanding mobility based on GPS data. In Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, Korea, 21–24 September 2008.
17. Gallotti, R.; Bazzani, A.; Rambaldi, S.; Barthelemy, M. A stochastic model of randomly accelerated walkers for human mobility. *Nat. Commun.* **2016**. [CrossRef] [PubMed]
18. Wind, D.K.; Sapiezynski, P.; Furman, M.A.; Lehmann, S. Inferring Stop-Locations from WiFi. *PLoS ONE* **2016**. [CrossRef] [PubMed]
19. Wesolowski, A.; Eagle, N.; Noor, A.M.; Snow, R.W.; Buckee, C.O. Heterogeneous mobile phone ownership and usage patterns in Kenya. *PLoS ONE* **2011**. [CrossRef] [PubMed]

20. Wesolowski, A.; Eagle, N.; Noor, A.M.; Snow, R.W.; Buckee, C.O. The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface* 2013. [CrossRef] [PubMed]
21. Mislove, A.; Lehmann, S.; Ahn, Y.Y.; Onnela, J.P.; Rosenquist, J.N. Understanding the demography of Twitter users. In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011.
22. Hecht, B.; Stephens, M. A tale of cities: Urban biases in volunteered geographic information. In Proceedings of the Eighth International AAAI conference on Weblogs and Social Media, Ann Arbor, MI, USA, 2–4 June 2014.
23. Fang, Z.; Shaw, S.L.; Tu, W.; Li, Q.; Li, Y. Spatiotemporal analysis of critical transportation links based on time geographic concepts: A case study of critical bridges in Wuhan, China. *J. Trans. Geogr.* **2012**, *23*, 44–59. [CrossRef]
24. Becker, R.; Cáceres, R.; Hanson, K.; Isaacman, S.; Ji, M.L.; Martonosi, M.; Rowland, J.; Urbanek, S.; Varshavsky, A.; Volinsky, C. Human mobility characterization from cellular network data. *Commun. ACM* **2013**, *56*, 74–82. [CrossRef]
25. Ranjan, G.; Zang, H.; Zhang, Z.L.; Bolot, J. Are call detail records biased for sampling human mobility? *ACM Sigmobile Mob. Comput. Commun. Rev.* **2012**, *16*, 33–44. [CrossRef]
26. Zhao, Z.; Shaw, S.L.; Xu, Y.; Lu, F.; Chen, J.; Yin, L. Understanding the bias of call detail records in human mobility research. *Int. J. Geogr. Inf. Sci.* **2016**, *30*, 1738–1762. [CrossRef]
27. Sagarra, O.; Szell, M.; Santi, P.; Díaz-Guilera, A.; Ratti, C. Supersampling and network reconstruction of urban mobility. *PLoS ONE* **2015**. [CrossRef] [PubMed]
28. Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Fang, Z.; Li, Q. Understanding aggregate human mobility patterns using passive mobile phone location data: A home-based approach. *Transportation* **2015**, *42*, 625–646. [CrossRef]
29. Wesolowski, A.; Eagle, N.; Tatem, A.J.; Smith, D.L.; Noor, A.M.; Snow, R.W.; Buckee, C.O. Quantifying the impact of human mobility on malaria. *Science* **2012**, *338*, 267–270. [CrossRef] [PubMed]
30. Calabrese, F.; Mi, D.; Lorenzo, G.D.; Ferreira, J.; Ratti, C. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transp. Res. Part C Emerg. Technol.* **2013**, *26*, 301–313. [CrossRef]
31. Kung, K.S.; Greco, K.; Sobolevsky, S.; Ratti, C. Exploring universal patterns in human home-work commuting from mobile phone data. *PLoS ONE* **2013**. [CrossRef] [PubMed]
32. Diao, M.; Zhu, Y.; Ferreira, J.; Ratti, C. Inferring individual daily activities from mobile phone traces: A Boston example. *Environ. Plan. B Plan. Des.* **2015**, *43*, 1–25. [CrossRef]
33. Yang, X.; Fang, Z.; Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Zhang, T.; Lin, Y. Understanding spatiotemporal patterns of human convergence and divergence using mobile phone location data. *ISPRS Int. J. Geo-Inf.* **2016**. [CrossRef]
34. Xu, Y.; Shaw, S.L.; Fang, Z.; Ling, Y. Estimating potential demand of bicycle trips from mobile phone data—An anchor-point based approach. *ISPRS Int. J. Geo-Inf.* **2016**. [CrossRef]
35. Calabrese, F.; Lorenzo, G.D.; Liu, L.; Ratti, C. Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston metropolitan area. *IEEE Pervasive Comput.* **2011**, *10*, 36–44. [CrossRef]
36. Xu, Y.; Shaw, S.L.; Zhao, Z.; Yin, L.; Lu, F.; Chen, J.; Fang, Z.; Li, Q. Another tale of two cities: Understanding human activity space using actively tracked cellphone location data. *Ann. Assoc. Am. Geogr.* **2016**, *106*, 489–502.
37. Yuan, Y.; Raubal, M.; Liu, Y. Correlating mobile phone usage and travel behavior—A case study of Harbin, china. *Comput. Environ. Urban Syst.* **2012**, *36*, 118–130. [CrossRef]
38. Song, C.; Koren, T.; Wang, P.; Barabási, A.L. Modelling the scaling properties of human mobility. *Nat. Phys.* **2010**, *6*, 818–823. [CrossRef]
39. Song, C.; Barabási, A.L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021. [CrossRef] [PubMed]
40. Gallotti, R.; Bazzani, A.; Degli Esposti, M.; Rambaldi, S. Entropic measures of individual mobility patterns. *J. Stat. Mech. Theory Exp.* **2013**. [CrossRef]
41. Cuttone, A.; Lehmann, S.; González, M.C. Understanding predictability and exploration in human mobility. *e-Print: arXiv* **2016**.
42. Tene, O. Privacy in the age of big data: A time for big decisions. *Stanf. Law Rev. Online* **2012**, *20*, 42–56.

43. Smith, M.; Szongott, C.; Henne, B.; Von Voigt, G. Big data privacy issues in public social media. *IEEE Int. Conf. Digit. Ecosyst. Technol.* **2012**. [CrossRef]
44. Yin, L.; Wang, Q.; Shaw, S.L.; Fang, Z.; Hu, J.; Tao, Y.; Wang, W. Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PLoS ONE* **2015**. [CrossRef] [PubMed]
45. Haklay, M. How good is volunteered geographical information? A comparative study of Openstreetmap and ordnance survey datasets. *Environ. Plan. B Plan. Des.* **2010**, *93*, 3–11. [CrossRef]
46. Goodchild, M.F.; Li, L. Assuring the quality of volunteered geographic information. *Spat. Stat.* **2012**, *1*, 110–120. [CrossRef]
47. Fu, K.; Chau, M. Reality check for the Chinese microblog space: A random sampling approach. *PLoS ONE* **2013**. [CrossRef] [PubMed]
48. Haklay, M.; Basiouka, S.; Antoniou, V.; Ather, A. How many volunteers does it take to map an area well? The validity of Linus' law to volunteered geographic information. *Cartogr. J.* **2013**, *47*, 315–322. [CrossRef]
49. Arai, A.; Fan, Z.; Matekenya, D.; Shibasaki, R. Comparative perspective of human behavior patterns to uncover ownership bias among mobile phone users. *ISPRS Int. J. Geo-Inf.* **2016**. [CrossRef]
50. Liu, H.X.; Danczyk, A.; Brewer, R.; Starr, R. Evaluation of cell phone traffic data in Minnesota. *Transp. Res. Rec.* **2008**, *11*, 1–7. [CrossRef]
51. Chen, B.Y.; Shi, C.; Zhang, J.; Lam, W.H.; Li, Q.; Xiang, S. Most reliable path-finding algorithm for maximizing on-time arrival probability. *Transp. B Transp. Dyn.* **2016**. [CrossRef]
52. Goodchild, M.F. The quality of big (geo) data. *Dialogues Hum. Geogr.* **2013**, *3*, 280–284. [CrossRef]
53. Goodchild, M.F.; Gopal, S. *Accuracy of Spatial Databases*; Taylor and Francis: London, UK, 1989.
54. Zhang, J.; Goodchild, M. *Uncertainty in Geographic Information*; CRC Press: New York, NY, USA, 2002.
55. Jacquez, G. A research agenda: Does geocoding positional error matter in health GIS studies? *Spat. Spatio-Temporal Epidemiol.* **2012**, *3*, 7–16. [CrossRef] [PubMed]
56. Lu, S.; Fang, Z.; Shaw, S.L.; Zhang, X.; Yin, L. Quantitative analysis of the effects of spatial scales on intra-urban human mobility. *Geomat. Inf. Sci. Wuhan Univ.* **2016**, *41*, 1199–1204.
57. Jacobs, A. The pathologies of big data. *Commun. ACM* **2009**, *52*, 36–44. [CrossRef]
58. Shenzhen Statistical Yearbook 2012. Available online: http://www.sztj.gov.cn/nj2012/indexeh.htm (accessed on 12 September 2016).
59. Whitepaper of Urban Planning, Land & Resources Commission of Shenzhen Municipality 2015. Available online: http://www.szfdc.gov.cn/xxgk/ghjh/td/201508/t20150813_108651.html (accessed on 3 December 2016).
60. Fotheringham, A.S.; O'Kelly, M.E. *Spatial Interaction Models: Formulations and Applications*; Kluwer Academic Publishers: Boston, MA, USA, 1989.
61. Lemeshow, S.; Hosmer, D.W. A review of goodness of fit statistics for use in the development of logistic regression models. *Am. J. Epidemiol.* **1982**, *116*, 92–106.