




Research Article

Detecting Home and Work Locations from Mobile Phone Cellular Signaling Data

Yingkun Yang ^{1,2}, Chen Xiong ^{1,2}, Junfan Zhuo ^{1,2} and Ming Cai ^{1,2}

¹School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, China

²Guangdong Provincial Key Laboratory of Intelligent Transportation System, School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510006, China

Correspondence should be addressed to Chen Xiong; xiongch8@mail.sysu.edu.cn

Received 29 January 2021; Revised 19 February 2021; Accepted 2 March 2021; Published 12 March 2021

Academic Editor: Chi-Hua Chen

Copyright © 2021 Yingkun Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Obtaining the distribution of home and work locations is essential for city planning, as it defines the structure and mobility pattern of a city. With the development of telecommunication networks, mobile network data, having the advantages of large coverage and strong followability, have produced large amounts of information about human activities. Thus, it has become a popular research subject for human position detection. In this study, we proposed a new method to detect home and work locations based on the extraction of focal points in traces, identifying an individual's working and resting hours, and analyzing the characteristics of city grids using mobile phone cellular signaling data (CSD). At the individual level, we validated the algorithm on ground-truth volunteer data and achieved a small deviation of under 500 and 565 m for home and work location detection 85% of the time. At the aggregate level, we tested it on a city-wide anonymized CSD set and found a high Pearson correlation between our result and the census data of 0.93. Compared to existing studies, this study improved the granularity and location accuracy of home and work location detection, as well as validated the method using both individually labeled ground-truth data and aggregate data for the first time. Applying the algorithm in a city, we captured the population distribution, commuting patterns, and job-housing balance of the city and demonstrated the potential in using mobile network data for urban planning and policy formulation.

1. Introduction

Detecting home and work locations is of great importance to modern city and transportation planning, as it aids the understanding of the relationship of jobs-housing [1], the design of public transportation [2], and the optimization of urban land use [3]. Traditionally, this was accomplished by collecting survey and smart card data [4, 5]. However, survey data have a low update frequency, a small sample size, and a high implementation cost, while smart card data are confined to people using public transportation, which can possibly result in the sample being unrepresentative. To overcome these shortcomings, Global Positioning System (GPS) sensors were introduced to collect human mobility data. Compared to survey data, GPS data provided more accurate data with spatial and temporal details [6]. However, it required users to wear GPS loggers or actively allow GPS

tracking from dedicated applications on their mobile phones, which increased the cost and limited the collection scale.

With the development of telecommunications technology and the increase in the penetration rates of smartphones [7], smartphones have become ideal digital sensors to track human locations. Produced from the interactions between smartphones and telecommunication infrastructures, mobile network data are recorded by telecom operators automatically in the background. Therefore, they can record the carrier's spatiotemporal information at a massive scale with little effort from the carrier. Mobile network data can be classified as event-driven or network-driven data. The former, such as call detail records (CDRs), is produced with the usage of mobile services including calls and texts. The latter, which is generally called cellular signaling data (CSD), is produced from signaling events such as handovers, network

updates, periodic updates, and location area updates [8]. With the advantages, mobile network data have become a research interest in urban structure and human mobility studies [9–11]. Existing studies on home and work locations detection using mobile network data can be classified into two categories.

Studies in the first category defined two timeframes, working hours and resting hours, and selected the locations with the highest frequencies of phone usage or the longest staying times in the two timeframes. If the locations met additional requirements proposed in these studies, they would be identified as home and work locations. The location of a user was determined based on the base station that the user was connected to. The two timeframes were often set as the same for all users, although they could be adjusted based on researchers' knowledge. Kung et al. [12] set daytime and night-time timeframes with the thresholds of 8:00 and 20:00, respectively. In the two timeframes, the user's home and workplace were selected as the locations in which the user spent the maximum time as long as the staying time in such locations accounted for more than 50% of the timeframes. Yan et al. [13] identified the home and work locations as the most frequently visited base stations during the timeframes of 20:00–6:00 and 10:00–16:00, and then calculated the identification confidence and appearance days to determine potential commuters. Ahas et al. [14] developed an anchor point determination model to detect home, work, and multifunctional anchor points. Setting the boundary between the resting time and working time as 17:00, the average start time and the standard deviation of the start time of daily events were considered to distinguish the user's home and work points. The model was validated on a CDR data set in Estonia by comparing the home location identification results with the Estonian Population Register.

Studies in the second category discovered the user's mobility pattern and identified the meaningful places in the user's traces as home and workplace. Previous studies have proved that human trajectories could be mined by trajectory analysis and stay point detection from different kinds of data, including passive recording data like mobile network data [15]. Jiang et al. [16] extracted the stay locations of the user and identified the user's activity types (home, work, shop, etc.) by analyzing features in the trajectories including spatial and temporal regularities. Alexander et al. [17] converted records into clustered locations to identify origin-destination trips and inferred these locations to be home, work, or other considering observation frequency and time. Widhalm et al. [18] presented a filtering and clustering method to detect stay locations and enriched the location sequences with an activity type (work, home, shopping, leisure) inferring from land-use data and time of day. Isaacman et al. [19] clustered recorded locations using the Hartigan algorithm and trained a logistic regression model on CDR data from 18 volunteers to identify their important locations. The model considered the cluster features, including days, durations, and the number of events during working and resting hours. It was validated on data from 19 volunteers and achieved median errors of 0.9 and 0.83 miles

for home and work location detection, respectively. Inspired by this study [19], Zagatti et al. [20] also clustered the user's traces using the Hartigan algorithm. The clusters were then scored depending on the occurrence hours and days of events. The clusters were labelled as daytime clusters, evening time clusters, and undecided clusters depending on their scores, thus identifying the user's home and work clusters.

However, there were limitations in the previous studies. First, in most of the studies, the temporal information was selected as an important feature to detect home and workplace. They set the same working hours and resting hours for every user before identification despite the large differences in the working schedules of people in different occupations (for example, night workers), thereby introducing biases. As the basis of further analysis, defining timeframes with biases may introduce large errors in the identification process. In addition, in the majority of studies, base stations or clusters were selected as the home and work locations. However, the coverage area of base stations differed in urban and rural areas with different base station densities [21]. The spatial resolution of the identification results may range from less than 1 km to more than 10 km, depending on the density of the base stations and the clustering results, which creates concerns about the spatial accuracy of the detection algorithm. Lastly, most of the previous studies lacked validation both at the individual level and the aggregate level. Instead, some simply assessed the algorithms by comparing the identified results with aggregate data, like city-wide censuses or travel surveys, while some did not perform validation. Therefore, the performance of the methods on individual records was seldom reported, which limited their application at a high resolution.

In view of these limitations, this paper proposes a new algorithm to detect home and work locations with a finer resolution using CSD, as well as validates the algorithm at the individual level using ground-truth data and the aggregate level using census data. We first processed the raw CSD and extracted the focal points in active users' traces. Then, we adopted information entropy to measure the activity intensity and used ordered data binning to identify the unique working and resting hours of each user based on the variation of activity intensity of each user. By dividing the study area into grids, we used a regular-grid spatial tessellation to describe the coverages of the base stations, as well as analyzed their geographic, temporal and spatial features. Based on the features selected, multiple attribute decision-making was introduced to construct a selection model to detect work and home grids. We validated the algorithm at the individual level using volunteers' ground-truth data collected by a smartphone app that we developed, as well as at the aggregate level using a city-wide anonymized CSD set. Finally, the algorithm was employed to capture the population distribution, commuting patterns, and job-housing balance in a city, which demonstrated its potential in real practice for policy-makers and urban planners. In contrast to existing works, the main contributions of this work can be summarized as follows:

- (1) Investigating the unique schedule of each user by analyzing the variation of activity intensity of the user, thus distinguishing users with unusual working schedules and avoiding the biases from setting uniform timeframes
- (2) Improving the spatial accuracy of the home and work location detection by comprehensively analyzing the attributes of city grids
- (3) Evaluating the home and work location detection algorithm using both individually labelled ground-truth data and aggregate data for the first time

The rest of the paper is structured as follows. Section 2 presents an overview of the study area and data. Section 3 details the principle of the algorithm developed. Section 4 describes the verification work of the algorithm. Section 5 shows an example application of the model. Section 6 concludes the paper with a brief discussion.

2. Study Area and Data

2.1. Study Area. In this study, we chose Foshan as the study area. Foshan is a southern city in China and covers 3797.72 km². Bordering Guangzhou, which is one of the most-developed cities in China, Foshan is a well-known city of commerce and industry. According to the State Department, Foshan is composed of 5 districts, Chancheng, Nanhai, Shunde, Sanshui, and Gaoming, and it can be further divided into 32 subdistricts.

2.2. Data

2.2.1. Anonymized Cellular Signalling Data (CSD) Data Set. The data were an anonymized CSD set of Foshan and was provided by a large telecom operator in China. It covered the records of 4.9 million users on the operator’s network of 15 weekdays (the weekdays during July 9, 2018, and July 28, 2018). As shown in Table 1, each CSD record contained a user identification (ID) number, information about the user, timestamps of the events, and the connected base station ID number, according to which the longitude and latitude of the base station could be queried. The data set of the studied period contains a total of 8.6 billion interaction records. The number of base stations in the research area of Foshan is 7800. The overall base station density was 1.9 towers per square kilometer, but the spacing between them was not uniform across the city, as shown in Figure 1.

2.2.2. Ground-Truth Volunteer Data. The anonymized CSD set could not be matched to real users for algorithm validation. Therefore, we developed an app for Android phones to collect CSD from the volunteers that were recruited. After installing the app, the volunteer could mark his/her mobility traces and stay locations in the app, while their CSD was collected and uploaded to the database in the background automatically. After collation, the CSD was in the form shown in Table 1.

TABLE 1: Cellular signaling data (CSD) records.

User ID	Gender	Age	Datetime	Base station ID
4117***	M	30	20180715071145	23740
5701***	M	63	20180723053558	25815
5717***	M	20	20180620142020	19086

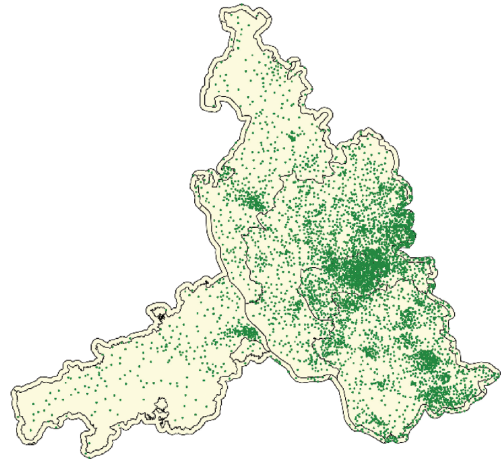


FIGURE 1: Distribution of the base stations.

In this study, a group of 41 volunteers was recruited, consisting of 22 males and 19 females, all of whom were adults aged 25–55 years. They lived and worked in Foshan, the study area, and had no unusual activities, such as moving or changing jobs, during the study period. They gave us permission to collect their CSD for five weekdays and to study the data for research purposes. In addition, they were also asked to provide us with the longitudes and latitudes of their true home and work locations with the help of a web map.

2.2.3. Point of Interest (POI) Data. The POI data consisted of information about the geographical points that represented a particular feature on the map. A POI could be anything closely related to human activities, such as an office building, bus station, or traffic sign. Therefore, the number and types of POIs in an area could reflect the main human activity in the area. The POI information used in this study was downloaded from the API (application programming interface) of a web map service provider. Each POI record contained key information, including the POI ID, type, longitude, and latitude, as shown in Table 2.

Since not all types of POIs are related to the home and workplace, we selected POIs of types closely related to the home and workplace for further analysis, as presented in Table 3. There were approximately 48,544 home-related POIs and 1,578,366 work-related POIs, and their distributions are presented in Figure 2.

3. Method

3.1. Focal Point Extraction. The focal points in a user’s traces were extracted to find the main activity space in his/her life

TABLE 2: POI information.

POI ID	Longitude	Latitude	Type
B02F5077VF	113.210716	22.875081	Culture and education services
B0FFL1FT5P	112.880293	23.180195	Enterprise
B0FFKUJWAO	113.002889	23.240415	Enterprise

TABLE 3: Home- and work-related POI types.

Home/work-related POI	POI type
Home-related POI	Commercial house
	Residential area
	Car services
	Restaurants
Work-related POI	Shopping services
	Life services
	Sports and recreation services
	Accommodation services
	Healthcare services
	Governmental organization
	Culture and education services
	Financial institutions
	Enterprises

for mobility pattern analysis. A typical characteristic of the CSD was that it contained a large amount of noise, and preprocessing was needed. Therefore, in extracting focal points, we aimed to (1) eliminate noise data, including abnormal data and redundant data, (2) filter sparse data to find active users, and (3) cluster spatially close base stations in a user's records.

3.1.1. Eliminating Noise Data. Removing the noise and redundancy in raw data is an essential issue in data preprocessing and help to show clearer information in further analysis. The noise data in the CSD contained abnormal data and redundant data. These are defined as follows:

- (i) *Abnormal Data.* This included records with missing values and records showing unusual behaviors. Records missing users' individual information were filled-in with the corresponding values in neighboring records to maintain data integrity while records missing key information including User ID, Datetime, and Base station ID were deleted. Unusual behaviors included alternate switching between several base stations in a short time and moving at a speed of 120 km/h or higher. In the former case, only records that connected to the most frequent base station in the series were kept. In the latter case, the records were deleted, as we considered it to be signal drifting.
- (ii) *Redundant Data.* As CSD was produced with a high frequency, a large number of records could be generated in a short time at the same place. Such records were merged by only keeping one record that contained the location of the base station, the first

connected time, and the staying time at the base station. In addition, some duplicate records were discarded.

3.1.2. Filtering Sparse Data. Temporal sparsity and inhomogeneous distributions of CSD records in a day could cause identification errors. Figure 3 presents the number of CSD records in each hour of one day from a user. This shows that the records were concentrated within 3 h and were lost for the rest of the day, which led to the incorrect conclusion that the user stayed at one place for a long time, such as between 3:00 and 13:00, thereby introducing errors. Therefore, we filtered out sparse data to obtain data that could reveal the user's mobility patterns. Referencing previous research [22], we divided a day into 48 30 min timeslots. If there were CSD records in at least 16 timeslots (8 h) in a day, the user's records for that day would be kept; otherwise, those records would be deleted. In this study, we selected active users with at least five days of data for further analysis.

3.1.3. Clustering Close Points. Although there were many points in a user's traces, only several points were needed to represent his/her main activity space. To obtain such focal points, we aimed to group together points that were spatially close and filtered out outliers in low-density areas, as they were not regular in the user's life. DBSCAN (density-based spatial clustering of applications with noise) [23], a density-based clustering algorithm, was applied to cluster the base stations connected to the user. It required two parameters: the distance threshold ϵ and the minimum number of points minpts to constitute a cluster. By experimenting with a range of combinations, we found that $\epsilon = 500$ m and minpts = 2 were the most suitable. Figure 4 shows the clustering result of a user's traces. Clusters are represented by blue circles, and their sizes are proportional to the user's staying time at a location. The red lines link the base stations and the clusters they belong to. It shows that the spatially close base stations in the user's traces are clustered and represented by the centroids of the clusters, thus extracting the focal points in the user's traces.

3.2. Schedule Identification. Identifying the schedule of the user helps to distinguish the resting and working hours, and thus detect the home and work location more accurately. Generally, the user's schedule is revealed by his/her unique mobility pattern, which is different in different times of a day and indicates the user's status.

Taking advantage of the characteristics of CSD, we analyzed the activity intensity of the user. In general, people tended to stay in more limited areas and moved less during resting hours, which could be distinguished from other time periods. We proposed adopting the information entropy from information theory to measure the user's activity intensity. Information entropy was first introduced by Shannon [24] to analyze the disorder degree of a system by measuring the uncertainty of information. Calculating the

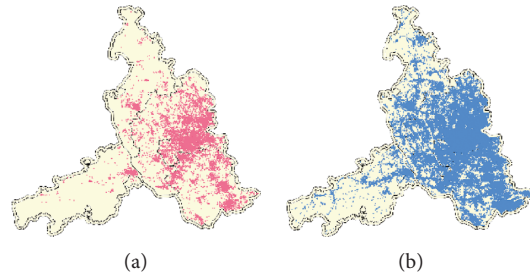


FIGURE 2: (a) Home-related POI distribution; (b) work-related POI distribution.

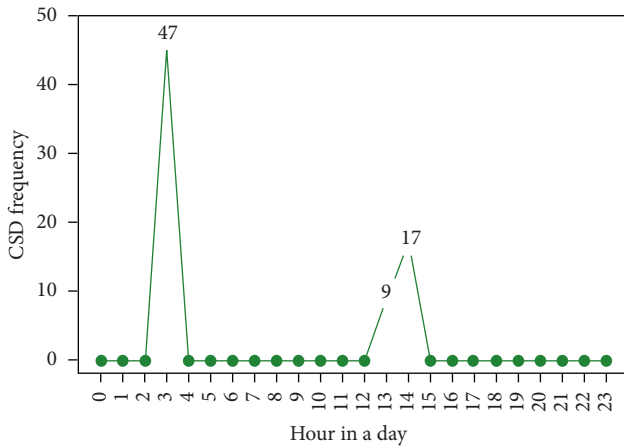


FIGURE 3: Example of the temporal distribution of the CSD.

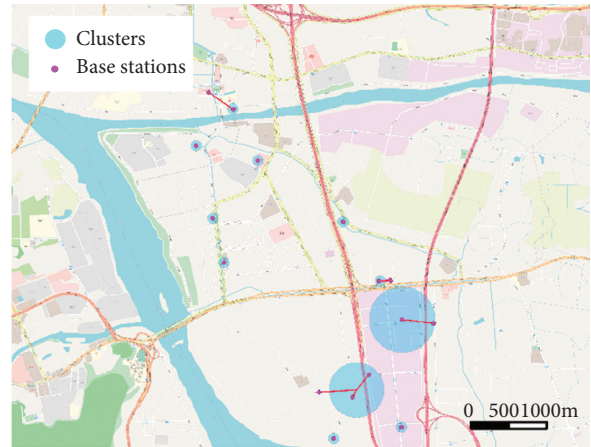


FIGURE 4: Clustering result of a volunteer's traces.

information entropy of a random variable is measuring the difference in the probability of events. The smaller the difference is, the more uncertain the information is, and the higher the information entropy is.

In this paper, the user's information entropy H_{period} during a given period is given by the following:

$$H_{\text{period}} = - \sum_{r=1}^R P_r \ln P_r, \tag{1}$$

$$P_r = \frac{st_r}{st_{\text{period}}},$$

where P_r is the proportion of the user's staying time st_r at location r to the total time st_{period} of the period and R is the number of the locations the user visited during this period.

H_{period} was measured by the difference in the probability that the user stayed at different locations during the period. The probability was represented by the proportion of the user's staying time at each location. A larger entropy means that the difference of the probability is small, thus the user spends his/her time more evenly in several places. On the contrary, a smaller entropy means that the difference of the probability is large, which resulted from the user contacting fewer base stations for longer time and moving less. When the user remains in one location, the information entropy will be 0.

Dividing a day into 24h, we calculated the user's information entropy in each time slot. Figure 5 shows an example of the information entropy in each hour in one day of a single user. The information entropy varied greatly from hour to hour according to the user's mobility pattern. Notably, the information entropy between 0:00 and 7:00 distinguishes from that during the daytime. This may indicate the user's resting time at home.

By measuring the information entropy in different times of a day, we could group time slots with similar activity intensity and divide a day into several timeframes. The time division based on the information entropy of each time slot was a data classification process. The user's day T was composed of 24h $\{t_1, t_2, \dots, t_{24}\}$ and for each hour t_j , its features were its information entropy in m days, which was denoted as $H_j = \{h_{1j}, h_{2j}, \dots, h_{mj}\}$. Based on this, we classified hours with similar features into p classes $\{T_1, T_2, \dots, T_p\}$. In particular, hours in a day could not be disordered in this case, which meant that the time slots in a class must be adjacent. Therefore, we introduced the Fisher-Jenks algorithm [25, 26], an ordered data binning algorithm, to split time slots into contiguous classes without scrambling the order. The key to Fisher-Jenks algorithm was to find the natural breaks in data that minimized the distance between the data points within various classes while maximizing the distance between the classes. It is supposed that the samples of a user could be described as follows:

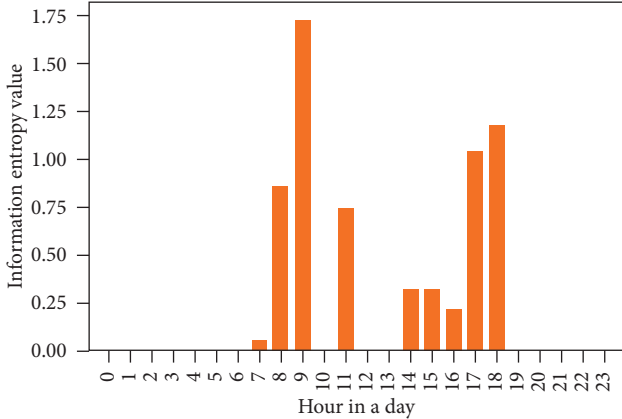


FIGURE 5: Example of the variation of the information entropy over one day.

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ h_{m1} & h_{m2} & \cdots & h_{mn} \end{bmatrix} = [H_1, H_2, \dots, H_n], \quad (2)$$

where h_{ij} is the information entropy in day i at hour j and H_n was the information entropy at hour n .

We chose the sum of squares of deviation to measure the data deviation in classes. The algorithm first calculates the distance for all sample pairs after normalizing the matrix H . Then, it computes the minimum deviations of classifying the samples into c classes by dynamically calculating the loss functions and finding the optimal result.

We experimented with a range of classification numbers and found that dividing the time into four classes worked well. The average entropy of a class T_G was calculated as

$$\overline{H}_{T_G} = \sum_{i=1}^m \sum_{j=a}^b h_{ij}. \quad (3)$$

We chose the class with the smallest average entropy as the resting hours of the user. Since most people went to work after the resting hours, we defined the start of the working hours as the end of the resting hours. Considering that the longest working time in a day in China is eight hours according to the labor law, the working time lasted for eight hours for users in our study. Figure 6 shows the information entropy of each hour for five days of one user, and a lighter color represents higher information entropy. According to the variation of information entropy in different times of the days, time was divided into four timeframes, as shown by the red dotted lines. Based on the results, the user was more inactive and remained stable in the first and last timeframe, while the user moved more in the other two timeframes. Using the method described above, the user's resting and working times were defined as 0:00–7:00 and 7:00–15:00.

3.3. Home and Work Grids Detection. Identifying the base stations or clusters as home and work locations resulted in

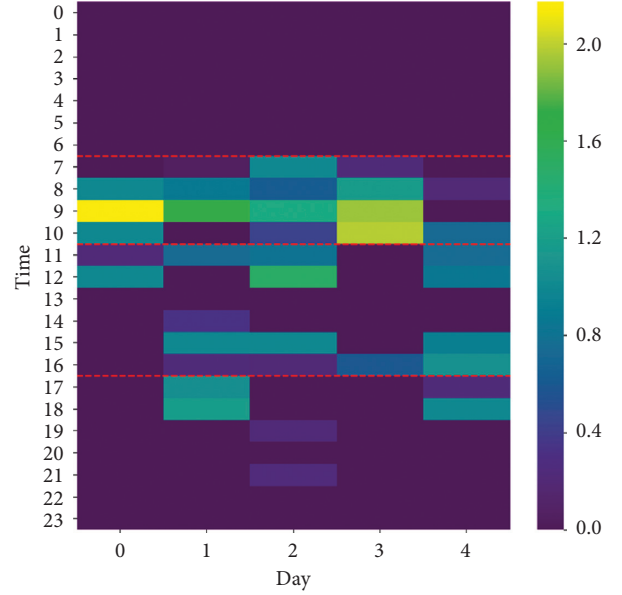


FIGURE 6: Time division using Fisher–Jenks algorithm.

spatial uncertainty of their positioning, as their coverage areas varied. We aimed to improve and stabilize the identification accuracy by extracting home and work grids based on their spatial, temporal, and geographic features.

3.3.1. Inferring Potential Grids. First, we divided the study area into grids and represented the coverage areas of the base stations by grids. As shown in Figure 7, a Voronoi tessellation technique [27] that divided the area into polygons according to the nearest neighbor rule was often used to simulate the coverage areas of base stations. Voronoi diagrams simulate the base station coverages based on the assumption that a mobile phone would connect to its closest base station and that the coverages do not overlap, which contradicts the actual behavior of mobile phones, which connect to the station with the highest signal strength [28]. This approach might underestimate the coverage of base stations. Also, the irregular diagrams might cause difficulties for further analysis in terms of calculation complexity and area segmentation. To solve these problems, we proposed using a regular-grid spatial tessellation to describe the coverages of the base stations. First, the study area was discretized into a mesh of grids with dimensions of $100\text{ m} \times 100\text{ m}$ considering the spatial distribution of the base stations, which achieved a balance between the grid refinement and the calculation efficiency. Then, we drew the circumcircle of each Voronoi polygon and defined the coverage of each base station as the grids that intersected its circumcircle. Figures 8(a) and 8(b) show an example of comparing the coverage areas of six base stations represented by Voronoi polygons and grids, respectively. The red points represent the base station locations and the blue areas represent their coverages. The grids cover more potential areas, and their regular shapes made them suitable for analysis.

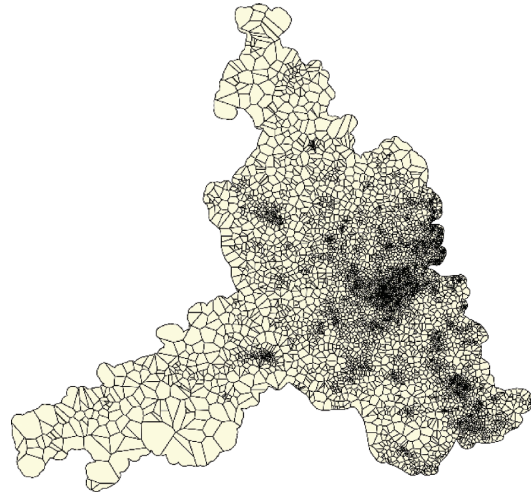


FIGURE 7: Voronoi diagram of the study area.

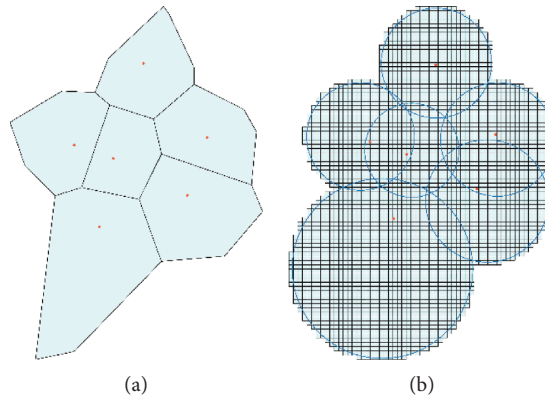


FIGURE 8: Base station coverages represented by (a) Voronoi diagram and (b) grids.

In Section 3.1, we classified the user's traces into different clusters. However, not all clusters were important in the lives of the users. To disregard transitional clusters in the user's traces, we selected the records in the user's resting and working hours, and calculated the staying time in each cluster during the relevant hours. For the two timeframes, we chose the cluster with the longest staying time as the important cluster in it. The grids covered by the base stations in the important clusters were inferred to be potential home or work grids.

3.3.2. Multiple Attribute Decision-Making. For each user, there were many potential grids, but normally, only one of them contained the home or workplace of the user. In this section, we describe how to determine which grid was the home or work grid. Studying the characteristics of the grids, we found that the following three observable factors were key to determining their importance.

(i) *Number of Home-Related or Work-Related POIs.* In general, a user's home and work locations, such as residential areas, apartment buildings, business

buildings, and schools, were recorded as POIs. Since POIs in an area indicated what people tend to do in the area, a grid with home- or work-related POIs with a higher density could more likely cover the user's home or workplace. Instead, if there were no relevant POIs in the grid, it may be located in an irrelevant area, such as a park or a road, and thus, it was less likely to cover the user's home or workplace.

(ii) *Staying Time at the Nearest Base Station.* Without knowing other conditions, such as the building occlusion effect and signal strength, we assumed that the distance between the user's position and base stations was the dominant factor that determined which base station the user connected to. Therefore, the longer the user connected to a base station, the more likely the user was to stay in the nearby grids, and the more likely it was that the grids cover his/her home or workplace.

(iii) *Average Distance to Base Stations in an Important Cluster.* During resting and working hours, the user might have connected to several base stations. This

may have been due to the user moving or the signal drifting inside his/her living and working areas. Therefore, the home and work locations are more likely to be at the center than at the edges of the cluster. This factor reveals the grid's position in the cluster by considering the distribution of the base stations. A grid with a shorter average distance was more likely to cover the home or workplace.

With the factors described above, we constructed a grid selection model based on multiple attribute decision-making to comprehensively analyze the attributes of the grids and select the home and work grids from the potential grids. A multiple attribute decision-making process made optimal decisions from several alternatives depending on their attributes and the relative weights of the attributes. Designing the weights of the attributes was one of the most important parts in the process as it would have a deep influence on the results. To enhance the objectivity of the result, we applied the entropy weight method (EWM) [29] instead of any subjective weighting models to determine the attribute weights. Basically, the EWM measures the importance of an attribute by the amount of useful information it contains. Equal attribute values for all samples does not provide any useful information for differentiating the home or work grid from other grids and such attribute should be given lower importance. We calculated the information entropy of an attribute. The lower the information entropy is, the higher the degree of differentiation of the attribute is, and the more useful it is to the evaluation. Thus, the attribute with lower information entropy will be given a higher weight and vice versa.

In this case, there were three factors and s samples in the evaluation, and y_{ij} was the i^{th} attribute of the j^{th} sample. First, the attributes were standardized as y'_{ij} . The entropy E_i of the i^{th} attribute was calculated as

$$E_i = -\frac{1}{\ln s} \sum_{j=1}^s p_{ij} \ln p_{ij}, \quad (4)$$

where $p_{ij} = y'_{ij} / \sum_{j=1}^s y'_{ij}$.

Larger weights should be given to attributes with higher entropy. The weights are calculated as follows:

$$w_i = \frac{1 - E_i}{\sum_{i=1}^3 (1 - E_i)}. \quad (5)$$

The final score Z_j for grid j is

$$Z_j = \sum_{i=1}^3 y'_{ij} w_i. \quad (6)$$

A user's home or work location was identified as the potential grid with the highest score.

Finally, we checked if the user was a regular resident or worker in the study area. The clusters that covered the home (work) grid in the user's traces were extracted. If they appeared on more than 3/5 of the days of the study period, which meant that the user visited the location on most of the days, the user would be considered to be a regular resident (worker). Otherwise, he/she would be

labelled as not having a regular home (workplace) in the study area.

4. Validation

In this section, we validated the proposed method with a ground-truth volunteer data set and a large-scale anonymized CSD set from the study area, as well as reported the evaluation results.

4.1. Validation on Volunteer Data Set. The algorithm was validated on the volunteer data set to evaluate the algorithm at the individual level. As described in Section 2.2.1, since we had the ground-truth data of individuals, we were able to evaluate the accuracy of the algorithm by measuring the position deviation between the identified locations and the true locations. To comprehensively analyze the result, we also compared our algorithm with two other identification algorithms.

In the first algorithm, called the TimeAccumulation algorithm [12], the resting and working hours were defined as 0:00 to 6:00 and 10:00 to 16:00, based on the typical timeframes of Foshan. A user's home and work locations were identified as the base station that the user was connected to for the longest time during resting and working hours. The second algorithm, called the HomeWorkCluster algorithm [20], clustered the base stations that had interacted with the user. Then, the clusters were scored according to the time of records in the clusters. Records between 8:00 and 17:00 were assigned a score one and those between 19:00 and 7:00 were assigned a score of minus one. The center of the cluster with the highest score was identified as the workplace while the center of the cluster with the lowest score was identified as home. The two algorithms could be classified into the two categories described in Section 1.

Figure 9 shows the cumulative distribution function (CDF) of the deviation between the true home locations and the home locations detected by the algorithm proposed in this paper, the TimeAccumulation algorithm, and the HomeWorkCluster algorithm. For the new algorithm, 85% of the home location deviations were under 500 m. Furthermore, as shown in Table 4, with a mean error of 246 m, the new algorithm outperformed the other two models, whose mean errors were 604 and 600 m, respectively.

Figure 10 shows the deviation between the true work locations and the work locations detected by the three algorithms, and a local enlargement of the figure. For the new algorithm, 85% of the work location deviations were less than 565 m. Although the work deviations were slightly higher than the home location deviations, the new algorithm also achieved a smaller mean error of 566 m, while the mean errors of the other two algorithms were 770 and 868 m, respectively, as shown in Table 4.

4.2. Validation on Anonymized Large-Scale Cellular Signaling Data. To evaluate the performance of our algorithm at the aggregate level, we applied it on the anonymized CSD set of Foshan city and compared the result with the sixth national

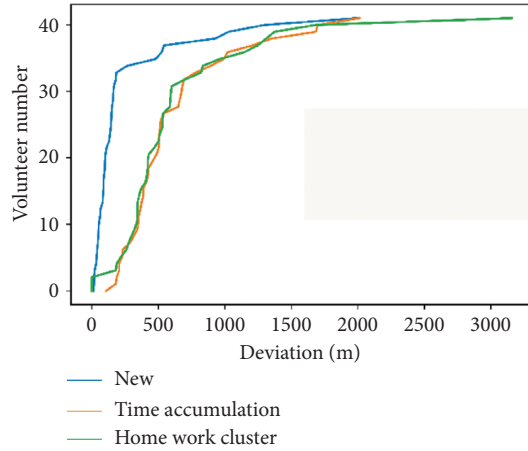


FIGURE 9: Home location deviations of the proposed (new), TimeAccumulation, and HomeWorkCluster algorithms.

TABLE 4: Mean home/work location deviations of the three algorithms.

	New algorithm	TimeAccumulation algorithm	HomeWorkCluster algorithm
Mean deviation of home location	246.47	604.32	599.98
Mean deviation of work location	566.29	769.64	867.74

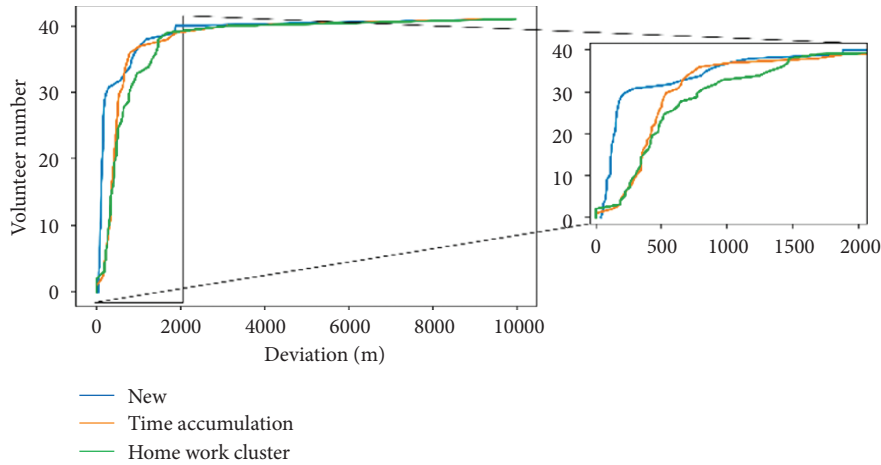


FIGURE 10: Work location deviation of the proposed (new), TimeAccumulation, and HomeWorkCluster algorithms.

census [30] in 2010. Both data sets contained population information of the study area, Foshan. As Foshan could be divided into 32 subdistricts, we inferred users’ home locations from the CSD set using the algorithm and aggregated them in the subdistricts. Figure 11 shows a comparison between the output of our algorithm and the number of residents by subdistrict from the census data. There was a linear relationship between them, indicating that the distribution of residents inferred from the CSD was consistent with the census population, while a few mismatches may have resulted from the different data collecting years and the change of the town borders during the eight-year gap. To further estimate their consistency, we calculated the Pearson correlation coefficient between our result and the census

data, which was $r = 0.93$. The high linear correlation shows that the output of our algorithm reflected the distribution of the population fairly well at the subdistrict level and was reliable for location identification.

5. Case Study

To demonstrate the application of the algorithm in practice, we present a case study in Foshan using the anonymized CSD set. Urban planners pay attention to the commuting between districts and depend on it to analyze land use and transportation connections in the city. Using the algorithm presented in this paper, the home and work locations of the users in Foshan were determined. To analyze the commuting

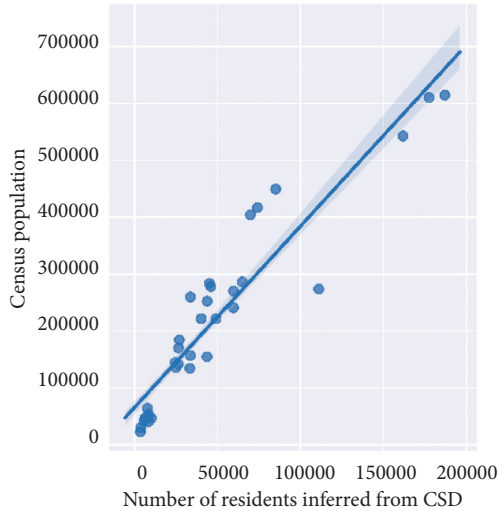


FIGURE 11: Comparison between our inferred residential population from the CSD and the population from the census data.

patterns between districts inside Foshan, we summed up the home and work locations by district and drew the commuting OD (origin-destination) desire lines between the five districts according to the home and work location distribution, as shown in Figure 12. This is a visualization of the commuting matrix of Foshan. The desire lines were drawn between the centroids of the districts and illustrate the flows of commuting people between them. The thicknesses of the lines depend on the number of commuting trips. Figure 12 shows that the majority of interdistrict commuting trips inside Foshan were generated between the Chancheng, Nanhai, and Shunde districts, which made up 86% of the interdistrict commuting trips. This agreed with their role in city planning as the most-developed part of Foshan. In addition, the east side of Foshan borders the heart of Guangzhou, which creates large numbers of job opportunities and attracts residents. In contrast, the other two districts are less economically developed, with lower GDPs (gross domestic products), and the commuting flows into and out of the districts are smaller accordingly.

To further analyze the job and housing situation inside the districts, we narrowed the scope into each subdistrict and calculated the density distribution of homes and workplaces, as presented in Figures 13(a) and 13(b), respectively. The distribution indicates the land-use characteristics in different parts of the city. Figure 13(a) suggests that the home locations are distributed broadly in the city, and the home density was higher on the east side, especially in the city center of the Chancheng district and its surroundings. The highest density reached 3,434 users per square kilometer in the Zumiao subdistrict. Figure 13(b) suggests that the workplace density is relatively high in the eastern portion of Foshan as well, especially in the Zumiao subdistrict, with the highest density of 3,527 users per square kilometer, as Zumiao is the heart of the Chancheng district and a transportation hub of Foshan, where many enterprises, shopping malls, and stations are located.

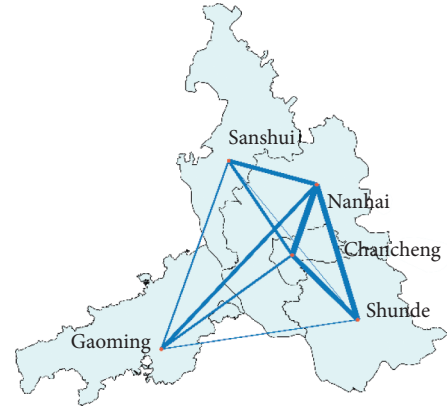


FIGURE 12: Commuting origin-destination (OD) desire lines between districts.

Furthermore, analyzing the traffic demand is essential for transportation planning in cities, and commuting travel is a significant part of the daily traffic demand. From the home and work locations, we could analyze the characteristics of the commuting demand. Considering the reduction of the CSD positioning error as well as the spatial accuracy of the algorithm, as evaluated in Section 4.1, we discretized the city into grids with dimensions of $500\text{ m} \times 500\text{ m}$. We then selected users whose home and workplace were both inside Foshan and allocated their home and work locations into the grids. The commuting distance of each user was calculated as the distance between the centers of their home/work grids. To show the complete distribution of the commuting distance in Foshan, the distribution is presented in double logarithmic coordinates, as shown in Figure 14(a). It shows that with the growth of commuting distance, the percentage of users falls down quickly and the majority of people have relatively small commuting distance. Focusing on this group of people, the pie chart, as presented in Figure 14(b), shows the percentage of users in different commuting distance intervals. About 62% of users' commuting distances were less than 2 km. For this group of people, it was possible to travel in nonmotorized mode. However, almost 5.36% of the users were long commuters whose homes were over 10,000 m from their workplaces, and these groups of people were more likely to travel by car. Overall, the mean commuting distances of the users who lived and worked inside Foshan was about 2,936 m, and about 80% of the users live within 4,000 m from their workplaces. The characteristics of the commuter trips in Foshan showed that people had relatively small commuting distances and might rely more on nonmotorized travel mode or public transport than private cars in their daily commuting. Therefore, for transportation planners and policy-makers of Foshan, more importance should be given to nonmotorized travel infrastructures and public transportation design, including bike-sharing services, bus line planning, and bus scheduling in peak hours.

From the above analysis, we inferred that Chancheng district was the key area of Foshan. Therefore, we focused on analyzing the characteristics of Chancheng, which is located

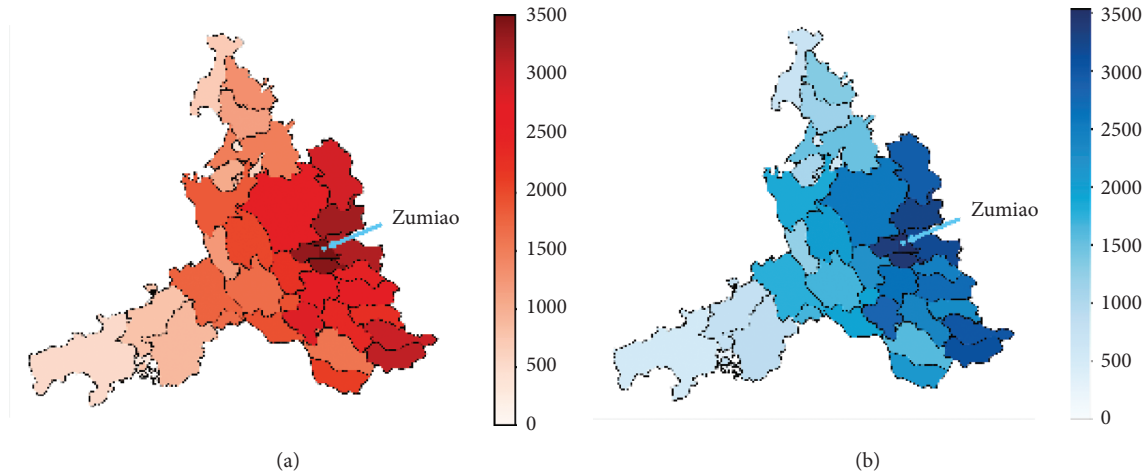


FIGURE 13: (a) Home density distribution in Foshan; (b) workplace density distribution in Foshan.

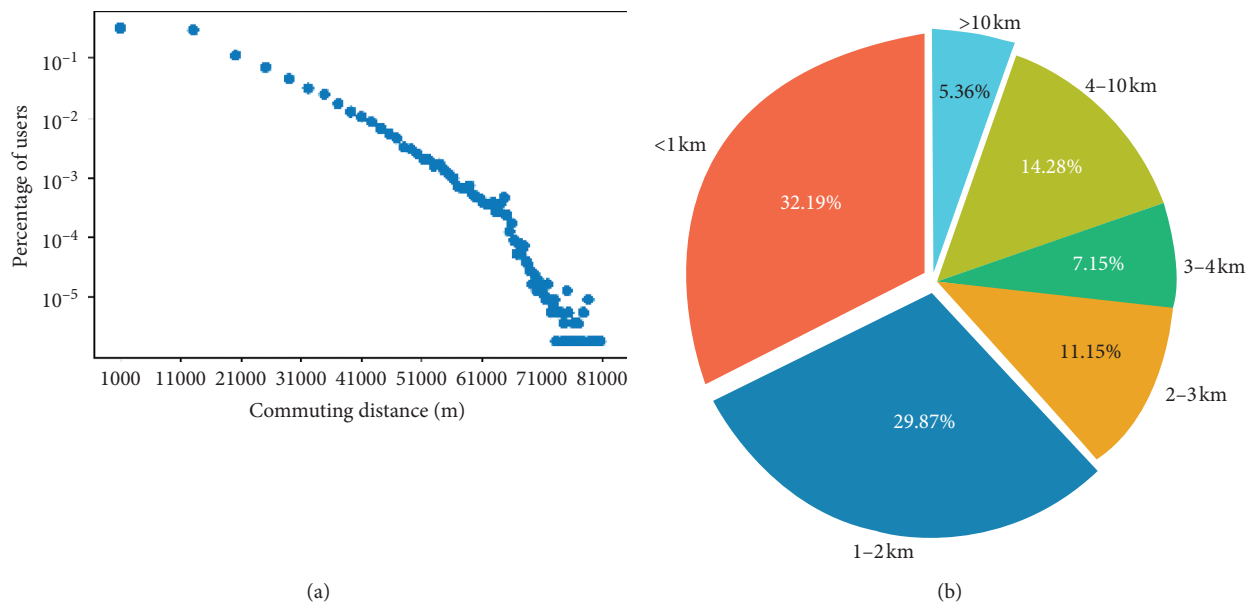


FIGURE 14: (a) The commuting distance distribution in double logarithmic coordinates; (b) the percentage of users in different commuting distance intervals.

at the center of Foshan, covering 154 square kilometers. It is the political, commercial, and cultural center of Foshan and a mixture of office, residential, and industrial areas. First, we focused on the spatial organization of homes and workplaces in Chancheng, as presented in Figures 15(a) and 15(b), respectively. We could see that both the living and working centers were in the east side of the district, while the home and workplace density in the west portion were relatively lower.

To assess the planning and development of the district, we introduced three indices that are often used to evaluate the equitability of job and housing distributions in a region. The first index is the employment self-sufficiency (ESS) [31], which refers to the number of people working and living locally out of the number of local workers. It describes the

self-containment of a region from the supply-side, and a higher ESS indicates that fewer people travel to the region for work. The second index is the employment self-containment (ESC) [31], which refers to the number of people working and living locally out of the number of local residents. It describes the self-containment of a region from the demand-side, and a higher ESC shows that fewer local residents travel to other regions for work. The third index is the job-housing ratio (JHR) [32], which represents the number of workers over the number of residences inside the region. It measures the matching degree between the number of jobs and the number of residences in the region. Table 5 presents the three calculated indices for the Chancheng district.

The ESS of 0.77 showed that 23% of workers were attracted from other districts, and the ESC showed that about

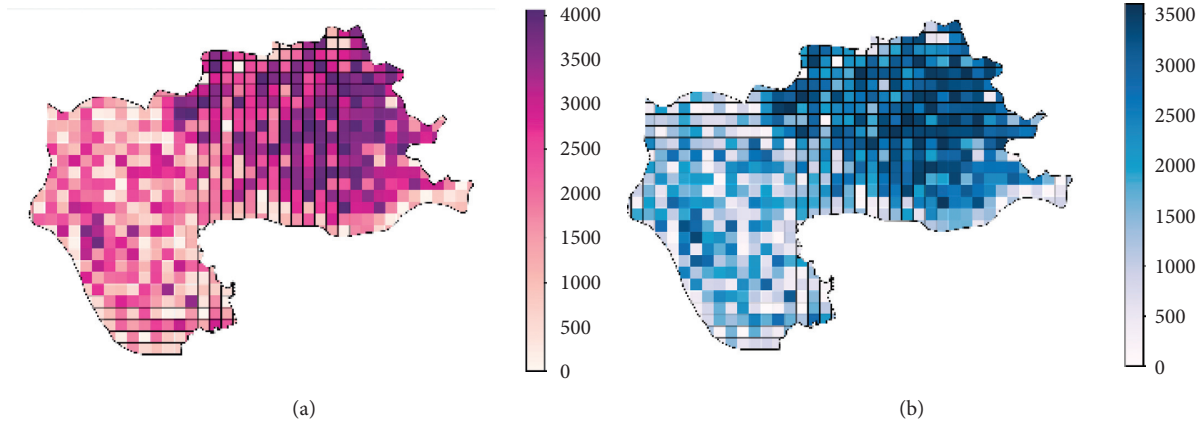


FIGURE 15: (a) Home location distribution in Chancheng; (b) workplace location distribution in Chancheng.

TABLE 5: Employment self-sufficiency (ESS), employment self-containment (ESC), and job-housing ratio (JHR) for the Chancheng district.

Indices	Chancheng district
ESS	0.77
ESC	0.78
JHR	1.01

78% of residents worked inside the district. This indicated that Chancheng, as the heart of Foshan, has great employment appeal for people all over the city and is a one-way attraction for other districts. Furthermore, while the commuting space for local residents is mainly concentrated inside the district, there is a large amount of interdistrict commuting demand to Chancheng that should not be ignored. However, the JHR of Chancheng was 1.01. According to Cervero [32], a value of the JHR between 0.8 and 1.2 indicates a relatively high match between the number of jobs and number of residences, and thus, the Chancheng district reaches a balance between providing job opportunities and housing.

6. Conclusion

Urban and transportation planning relies significantly on obtaining the home and work locations of people. This paper presented a method that could process massive CSD to detect the home and work locations of anonymized phone users. Considering the irregular temporal sampling and uncertain spatial accuracy characteristics of the data, the proposed method analyzed the variation of activity intensity of each user to investigate their schedules and comprehensively considered geographic, temporal, and spatial features of city grids to identify the home and work location. Using ground-truth data from volunteers and census, the study showed that the algorithm had high precision and could be used to detect the home and work locations of people on a large scale with small deviation. At the individual level, the validation results showed that the algorithm could detect users' home and work locations to within 500 and 565 m, respectively, 85% of the time, while at the

aggregate level, the Pearson correlation between our results and the census data was 0.93.

Compared to previous studies, this study is a significant step forward to use mobile phone data for home and work location detection in terms of granularity, location accuracy, and distinction of people with different working schedules. Also, to the best of our knowledge, this is the first study to evaluate the home and work location detection algorithm using both individually labelled ground-truth data and aggregate data.

As the literature suggests, in view of the development of cities, the expansion of the population, and the changing of urban structures and working schedules, the living and working behaviors of people have become more complex. Due to its small sample size and high collection cost, traditional survey data may fail to support the work in city planning and management, and reliable new data sources and methods are needed. Because of their characteristics of wide coverage, large sample size, and strong followability, mobile network data have made it possible to analyze human mobility pattern on a large spatiotemporal scale. By presenting a case study, it was demonstrated that applying our method can help obtain the distribution of home and work locations, extract commuting demand, and assess the job-housing balance in a big city using massive mobile network data. This shows the ultimate goal of the study to help urban planners and policy-makers to derive a new understanding of the city from big data, thus aiding their work in transportation planning for targeted areas, public facility construction to improve service quality, and policy formulations for future urban development.

Data Availability

The data used to support the findings of this study have not been made available because these are anonymized data and are confidential due to volunteer agreement and privacy policies.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2020YFB1600400) and the Fundamental Research Funds for the Central Universities (Grant No. 19lgpy290).

References

- [1] T. Li, Y. Chen, Z. Wang, Z. Liu, R. Ding, and S. Xue, "Analysis of jobs-housing relationship and commuting characteristics around urban rail transit stations," *IEEE Access*, vol. 7, pp. 175083–175092, 2019.
- [2] G. Qiu, R. Song, S. He, W. Xu, and M. Jiang, "Clustering passenger trip data for the potential passenger investigation and line design of customized commuter bus," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 9, pp. 3351–3360, 2019.
- [3] Z. Yang, X. Zhu, and D. Moodie, "Optimization of land use in a new urban district," *Journal of Urban Planning and Development*, vol. 141, no. 2, 2015.
- [4] Y. Ji, Y. Cao, Y. Liu, W. Guo, and L. Gao, "Research on classification and influencing factors of metro commuting patterns by combining smart card data and household travel survey data," *IET Intelligent Transport Systems*, vol. 13, no. 10, pp. 1525–1532, 2019.
- [5] J. Huang, D. Levinson, J. Wang, and Z. Wang, "Tracking job and housing dynamics with smartcard data," in *Proceeding of the National Academy of Sciences of the United States of America*, pp. 12710–12715, USA, 2018.
- [6] F. Zhao, F. Pereira, R. Ball et al., "Exploratory analysis of a smartphone-based travel survey in Singapore," *Transportation Research Record*, vol. 2494, Article ID 45e56, 2015.
- [7] M. S. Iqbal, C. F. Choudhury, P. Wang, and M. C. González, "Development of origin-destination matrices using mobile phone call data," *Transportation Research Part C: Emerging Technologies*, vol. 40, no. 1, pp. 63–74, 2014.
- [8] F. Calabrese, L. Ferrari, and V. Blondel, "Urban sensing using mobile phone network data: a survey of research," *ACM Computing Surveys*, vol. 47, no. 2, pp. 1–20, 2014.
- [9] X. Song, Y. Ouyang, B. Du, J. Wang, and Z. Xiong, "Recovering individual's commute routes based on mobile phone data," *Mobile Information Systems*, vol. 2017, 11 pages, 2017.
- [10] M. Ghahramani, M. Zhou, and C. T. Hon, "Mobile phone data analysis: a spatial exploration toward hotspot detection," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 1, pp. 351–362, 2019.
- [11] Z. Duan, Z. Lei, M. Zhang, W. Li, J. Fang, and J. Li, "Understanding evacuation and impact of a metro collision on ridership using large-scale mobile phone data," *IET Intelligent Transport Systems*, vol. 11, no. 8, pp. 511–520, 2017.
- [12] K. Kung, G. Kael, S. Stanislav, and C. Ratti, "Exploring universal patterns in human home-work commuting from mobile phone data," *Plos One*, vol. 9, no. 6, Article ID e96180, 2014.
- [13] L. Yan, D. Wang, S. Zhang, and D. Xie, "Evaluating the multi-scale patterns of jobs-residence balance and commuting time-cost using cellular signaling data: a case study in Shanghai," *Transportation*, vol. 46, no. 3, pp. 777–792, 2019.
- [14] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of Urban Technology*, vol. 17, no. 1, pp. 3–27, 2010.
- [15] Y. Zheng, "Trajectory data mining," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 3, pp. 1–41, 2015.
- [16] S. Jiang, J. Ferreira, and M. C. Gonzalez, "Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 208–219, 2017.
- [17] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240–250, 2015.
- [18] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. C. González, "Discovering urban activity patterns in cell phone data," *Transportation*, vol. 42, no. 4, pp. 597–623, 2015.
- [19] S. Isaacman, R. Becker, R. Cáceres et al., "Identifying important places in people's lives from cellular network data," in *Proceedings of the 9th International Conference on Pervasive Computing*, pp. 133–151, Heidelberg, Germany, 2011.
- [20] G. A. Zagatti, M. Gonzalez, P. Avner et al., "A trip to work: estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR," *Development Engineering*, vol. 3, pp. 133–165, 2018.
- [21] C. Horn, H. Gursch, R. Kern, and M. Cik, "QZTool—automatically generated origin-destination matrices from cell phone trajectories," *Advances in Intelligent Systems and Computing*, vol. 484, pp. 823–833, 2016.
- [22] C. M. Schneider, V. Belik, T. Smoreda, and M. C. González, "Unravelling daily human mobility motifs," *Journal of The Royal Society Interface*, vol. 10, no. 84, Article ID 20130246, 2013.
- [23] M. Daszykowski and B. Walczak, "Density-based clustering methods," in *Comprehensive Chemometrics*, pp. 635–654, Elsevier, Amsterdam, The Netherlands, 2010.
- [24] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [25] W. D. Fisher, "On grouping for maximum homogeneity," *Journal of the American Statistical Association*, vol. 53, no. 284, pp. 789–798, 1958.
- [26] G. F. Jenks and F. C. Caspall, "Error on choroplethic maps: definition, measurement, reduction," *Annals of the Association of American Geographers*, vol. 61, no. 2, pp. 217–244, 1971.
- [27] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," *ACM Computing Surveys*, vol. 23, no. 3, pp. 345–405, 1991.
- [28] M. Kjaergaard, "Location-based services on mobile phones: minimizing power consumption," *IEEE Pervasive Computing*, vol. 11, no. 1, pp. 67–73, 2011.
- [29] K. Qiu, *Management Decision and Applied Entropy*, Machine Press, Beijing, China, 2002.
- [30] China Statistics Press, *Tabulation on the 2010 Population Census of the People's Republic of China by Township, Population Census Office under the State Council*, China Statistics Press, Beijing, China, 2012.
- [31] K. Martinus and S. Biermann, "Strategic planning for employment self-containment in metropolitan sub-regions," *Urban Policy and Research*, vol. 36, no. 1, pp. 35–47, 2018.
- [32] R. Cervero, "Jobs-housing balance revisited: trends and impacts in the san francisco bay area," *Journal of the American Planning Association*, vol. 62, no. 4, pp. 492–511, 1996.