Hindawi
Journal of Advanced Transportation
Volume 2022, Article ID 3099721, 17 pages
https://doi.org/10.1155/2022/3099721

Research Article

Improved F-DBSCAN for Trip End Identification Using Mobile Phone Data in Combination with Base Station Density

Haihang Jiang,[1] **Fei Yang**,[1] **Xin Zhu**,[2] **Zhenxing Yao**,[3] **and Tao Zhou**[4]

[1]*Department of Transportation and Logistics, Southwest Jiaotong University, Chengdu 610031, China*
[2]*China Mobile Group Sichuan Co., Ltd., Chengdu 610084, China*
[3]*College of Transportation Engineering, Chang'an University, Xi'an 710064, China*
[4]*Chongqing Transport Planning Institute, Chongqing 401147, China*

Correspondence should be addressed to Tao Zhou; taozhoucq@qq.com

Received 4 March 2022; Accepted 13 April 2022; Published 30 April 2022

Academic Editor: Wen Liu

Trip end identification based on mobile phone data has been widely investigated in recent years. However, the existing studies generally use fixed clustering radii (CR) in trip end clustering algorithms, but ignore the influence of base station (BS) densities on the positioning accuracy of mobile phone data. This paper proposes a new two-step method for identifying trip ends: (1) Genetic Algorithm (GA) is utilized to optimize the CRs of DBSCAN under different BS densities. (2) We propose an improved Fast-DBSCAN (F-DBSCAN) for two objectives. One is for improving identification accuracies; the parameter CRs for judging core points can be dynamically adjusted based on the BS density around each mobile phone trace. The other is for reducing time complexity; a fast clustering improvement for the algorithm is proposed. Mobile phone data was collected by real-name volunteers with support from the communication operator. We compare the identification accuracy and time complexity of the proposed method with the existing ones. Results show that the accuracy is raised to 85%, which is approximately 6% higher than the existing methods. Meanwhile, the median running time can be reduced by about 76% by the fast clustering improvement. Especially for noncommuting trip ends, the identification accuracy can be increased by 8%. The average identification errors of travel time and trip end coordinates are reduced by about 12 min and 321 m, respectively.

1. Introduction

With progress in new-generation wireless communication techniques, the spatial and temporal resolution of mobile phone data is gradually improved. Certain research accomplishments have been obtained in aspects of residents' trip pattern monitoring [1, 2], job-housing relationship analysis [3], and trip origin-destination identification [4–6] using mobile phone data. Due to an outbreak of COVID-19, mobile phone data attract extensive attention in epidemiological investigation and research [7, 8]. The technology for mobile phone data has been gradually diffused from academic research to practical application. However, the primary basis of relevant research and application is still the identification of individual trip ends. The accuracy and efficiency of trip end identification have a direct influence on large-scale residents' traffic information extraction.

Trip end identification refers to the extraction of the user's dwell location and time of each activity from the user's all-day mobile phone data. Since the beginning of the 20th century, some researchers have paid attention to mobile phone data for travel surveys because of its advantages of passive collection and wide sample coverage. As regards early mobile phone data from the 2G communication network, such as call detail records (CDR) data, the positioning frequency and accuracy are very low. A study found that the average service range of BS in 2G communication network was about 3 km^2 [9]. Another found the average interval time of the data is 260 min [10]. Some scholars extracted trip ends based on time features. For example, Pan et al. [11] directly took the location of mobile phone traces at night as the place of residence and the trace from 8:00 a.m. to 11:00 a.m. as the place of work. In some studies, the locations of BSs visited by users with the highest frequency were taken as their trip ends combined with historical data [2, 12, 13]. However, such an approach is inapplicable for extracting noncommuter and nonfrequent trip ends. Later, some studies proposed rule-based methods [6, 14, 15]. To be specific, if a sequence of mobile phone traces satisfies the following two judgment criteria, they can be identified as being at a trip end: (1) the maximum spatial distance of continuous traces is below the preset distance threshold, and (2) the time difference of the first and the last in the continuous traces is above the preset temporal threshold [16]. The rules mostly rely on common sense or prior knowledge. A variety of time and distance thresholds has been proposed. For example, time thresholds include 15 min [17, 18], 30 min [19, 20], and 60 min [21], and distance thresholds include 200 m [18, 19] and 1000 m [10, 15]. Despite simple implementation, the method tends to ignore short distance/time travel and lacks robustness, signifying that the results obtained are extremely vulnerable to outliers [22].

The generation upgrade of mobile communication networks and the Internet economy brings higher temporal-spatial resolution of mobile phone data. The sampling frequency of mobile phone data has rapidly increased to the level of minute intervals [23]. Some clustering analysis algorithms were applied for trip end identification [24]. Wang et al. [25] and Poonawala et al. [26] proposed temporal-spatial clustering to extract trip ends. The influence of noise and outliers presented in the dataset can be effectively reduced by carefully setting the thresholds [27]. Chen et al. [28] applied a model-based clustering method requiring a predetermined number of clusters. However, the method is sensitive to the spatial density of the mobile phone traces. Several faraway outliers may be clustered together, causing the resulting cluster to stray away [22]. Some studies applied the incremental clustering algorithm for extracting trip ends more steadily [22, 29, 30]. However, the clustering results are subject to clustering sequence, which easily lead to unreasonable clusters [22]. DBSCAN based on the density characteristic of mobile phone traces has been proved to be effective and obtain stable results [31–33].

However, the existing methods still have some deficiencies in identifying trip ends. Firstly, mobile phone data is expected for daily observation of large-scale residents' mobility patterns, signifying high demand for technical efficiency. The running time for DBSCAN is heavily dominated by finding neighbors or obtaining density for each data point with the time complexity $O(n^2)$ [34]. The clustering efficiency remains to be improved. Secondly, the influence of BS layout on the identification effect is ignored in existing studies. The positioning error of mobile phone traces depends on BS densities and varies from as little as a few hundred meters in metropolitan areas to a few kilometers in rural regions [35]. It means that the traces generated at different trip ends also differ in the spatial distribution, as shown in Figure 1 [23]. The parameter CR as the unit to measure the density of traces has a direct impact on identification results [24, 36]. Fixed CRs used in the traditional algorithm cannot be well applied to all trip ends at the same time [22, 37]. CRs applicable to high BS density areas are usually too small under low BS densities, which easily results in one trip end being misidentified as multiones, as shown in Figure 1(a). This result will give us the illusion that the user travels back and forth between several trip ends within a short time, which is named oscillation in some studies [38–40]. In contrast, CRs suitable for low BS density areas are rather large under high BS densities. In this case, more traces generated on a trip will be clustered into trip end clusters, which increases the identification errors of travel time and trip end coordinates, as shown in Figure 1(b). Especially if two trip ends are near enough in space, too large CRs will result in them being misidentified as one. Due to a whole trip chain of a traveler through different BS densities, fixed CRs cannot avoid the above problem regardless of careful setting. Therefore, if we can improve the algorithm by adjusting CRs dynamically based on BS densities in the clustering process, the identification result can be further enhanced.

This paper obtains real-name volunteers' mobile phone data with support from the communication operator. The actual travel information behind mobile phone data can be synchronously gathered as a data foundation for algorithm improvement and result validation. Given the above problem, this paper proposes a new method for trip end identification. Our contributions can be summarized as follows:

(1) Anonymous mobile phone data used in previous studies can only be validated by comparing with other aggregate data sources, such as household travel survey data which is not necessarily reliable [6, 29, 41]. This study constitutes one of the very first attempts that systematically validates the results at the individual level using the ground truth data.

(2) CR as the key parameter in DBSCAN was set largely dependent on subjective experience in the existing research without being optimized by considering the communication environment [23]. A CR optimization framework GA-DBSCAN is proposed for optimal CRs under different BS densities in this paper.

(3) This paper identifies trip ends by improving the traditional DBSCAN for two objectives. One is for enhancing identification effects. CRs optimized by GA-DBSCAN can be adjusted dynamically based on the BS density around each trace in the clustering process. The other is for increasing clustering efficiency. We reduce the time complexity of the algorithm from three aspects, namely, clustering sequence, unified processing of repeated traces, and that of traces around them. The improved F-DBSCAN is validated by comparison with existing methods.

The remaining parts of this paper are organized as follows: Section 2 describes the proposed trip end identification method. Section 3 presents the data collection experiment and characteristics of mobile phone data. Section 4 analyzes the identification results of the proposed methods. Section 5 concludes the study and reveals future research directions.

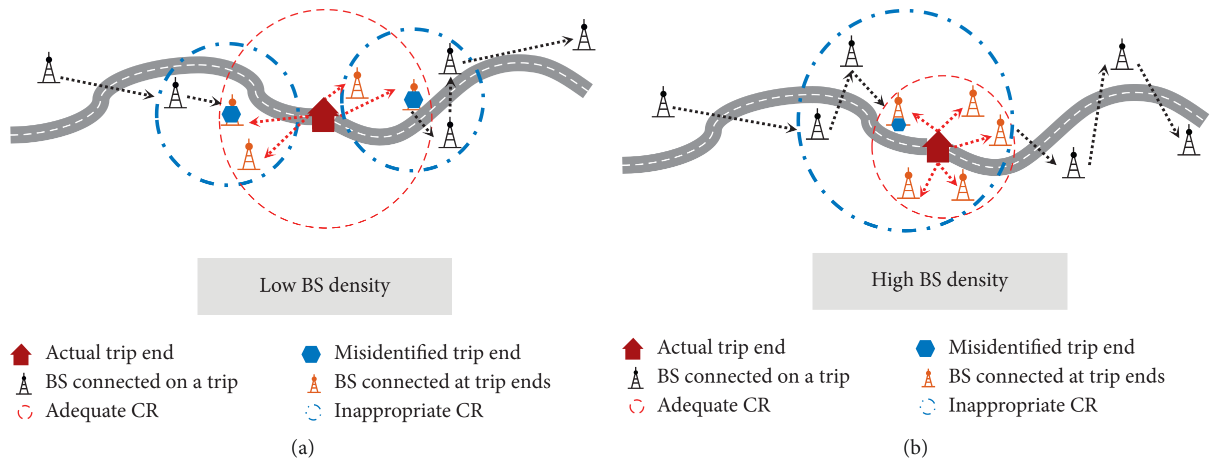


Figure 1: Inappropriate CRs under different BS densities. (a) Too small CR for low BS densities. (b) Too large CR for high BS densities.

2. Methodology

The BS density places significant influences on positioning errors of mobile phone data. In general, the positioning errors are comparatively small in areas where the BS density is high. As a consequence, spatial distribution ranges of traces produced at diverse trip ends vary, as shown in Figure 2. The traditional DBSCAN with a fixed CR is difficult to achieve good results under different BS densities. In this consideration, an equal-time-interval interpolation algorithm is firstly used to perform data preprocessing and balance time weights of mobile phone traces. Then, a GA-DBSCAN framework is built to optimize CRs under different BS densities. The functional relationship between optimal CRs and BS densities is acquired, that is, $R = \mathrm{Fun}(\text{density})$. On this basis, an improved F-DBSCAN is proposed and has the capacity of adjusting CRs dynamically with lower time complexity.

2.1. Data Preprocessing. Different from GPS data gathered in equal time intervals, mobile phone data are featured with time interval nonuniformity. This signifies that the time weights of different mobile phone traces vary. As shown in Figure 3(a), traces A and B are both at a trip end, without other traces in a period of T_1. Trace A represents not only its own position, but also the position during the period of T_1. By contrast, traces C and D on a trip can only represent the position at a certain moment. Therefore, trace A has a higher time weight. If mobile phone data occurs once per second, more traces will appear on the position of trace A, as shown in Figure 3(b). The equal-time-interval interpolation algorithm is used to estimate users' positions per second. It makes sure that high-density traces can be generated at trip ends, preventing trip ends ignored due to users' few communication behaviors, so that the identification result can be more stable.

The space-time three-dimensional coordinate of a trace is defined as (j, w, t) that represents longitude, latitude, and time. t is the second of the trace in a day. If the coordinates of two adjacent mobile phone traces are (j_1, w_1, t_1) and (j_2, w_2, t_2), the following two linear equations can be utilized to express the coordinates of traces at time t within the interval $[t_1, t_2]$:

$$
\begin{aligned}
j &= \frac{(t - t_1)(j_2 - j_1)}{t_2 - t_1} + j_1, \\
w &= \frac{(t - t_1)(w_2 - w_1)}{t_2 - t_1} + w_1.
\end{aligned}
\tag{1}
$$

After interpolation, mobile phone data turns into a per-second consecutive dataset. The higher the interpolation frequency is, the greater the computing amount and time cost of subsequent trip end identification will be. For this reason, the interpolation cycle F (unit: second) of traces should be adjusted according to computational power and timeliness need. In detail, on the basis of interpolation per second, a trace is repeatedly selected once every F seconds, while those not selected are deleted. Once F increases, it is more likely for identification errors of relevant information (e.g., travel time) to increase. In this paper, F is set at 10 seconds, signifying the time interval of traces after the data preprocessing is 10 s. Through the equal-time-interval interpolation algorithm, the number of traces can be used to represent dwell time, enabling the density of traces at trip ends to enormously rise.

2.2. CR Optimization. CR is the most important parameter in DBSCAN [42]. However, the existing setting for this parameter largely depends on subjective experience [23]. In this paper, a GA-DBSCAN framework is built for CR optimization. GA is a random search optimization algorithm based on the concepts of natural selection and genetics [43]. This CR optimization problem is solved mainly in the following two steps. Firstly, we need to determine the optimization goal of the target parameter CR, namely, the fitness function. Secondly, the clustering process of DBSCAN is integrated with the optimization flows of GA.

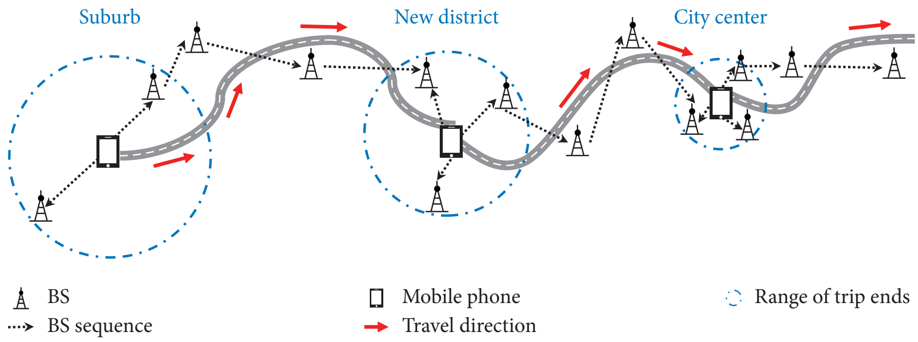


FIGURE 2: Spatial distribution of traces at trip ends under different BS densities.

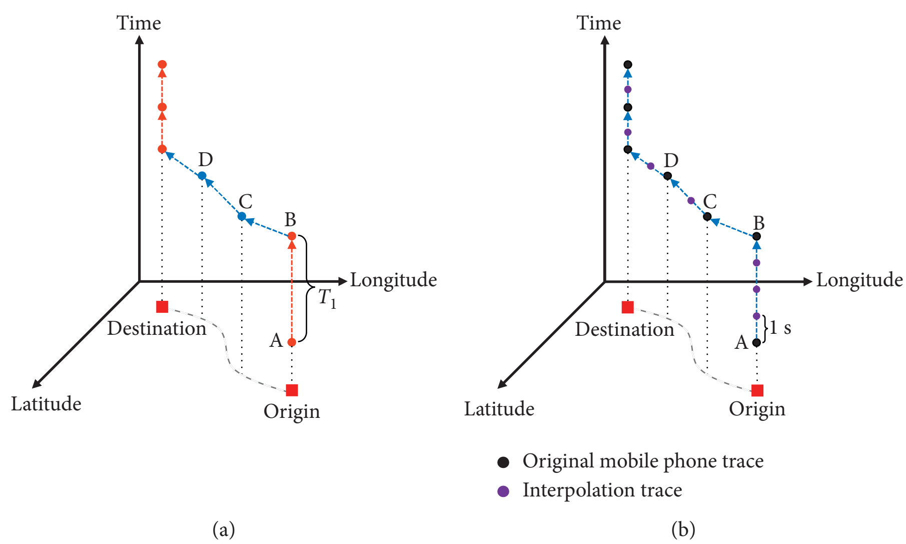


FIGURE 3: Schematic diagram of time weights of mobile phone traces. (a) Original mobile phone traces. (b) Traces after interpolation.

2.2.1. Fitness Function Construction. The fitness function of GA is the objective function of an optimization problem. In this problem, it reflects the proportion of correct identification. We rule out misidentification for getting the correctly identified proportion. There exist the following four categories of misidentification, as shown in Figure 4:

(1) Merged identification, where multiple trip ends are misidentified as one, as shown in Figure 4(a): specifically, suppose only one trip end is identified from a group of N_{M} actual ones. Then the number of misidentification samples is $N_{\mathrm{Mer}} = N_{\mathrm{M}} - 1$. It usually results from too large CR setting or a too short distance between different trip ends.

(2) Segmented identification, where N_{S} trip ends are falsely identified from one actual trip end, as shown in Figure 4(b): then, the number of misidentification samples under such circumstance is $N_{\mathrm{Seg}} = N_{\mathrm{S}} - 1$. It usually results from too small CR setting or drift data with large positioning errors caused by communication signal disturbance.

(3) Not identified, where an actual trip end is not identified from mobile phone data, as shown in Figure 4(c): it usually results from too short dwell time at the trip end.

(4) Additional identification, where an identified trip end consists of traces produced on a trip, as shown in Figure 4(d): it usually results from too long stay time on a trip caused by traffic jams, waiting at bus stations, and so on.

On this basis, two indexes of exact-identification accuracy (EIA) and extraidentification rate (EIR) are established to evaluate the above misidentification conditions.

$$\mathrm{EIA} = 1 - \frac{\sum_m N_{\mathrm{Mer}}^{(m)} + N_{\mathrm{Not}}}{N_{\mathrm{All}}},$$
$$\mathrm{EIR} = \frac{\sum_e N_{\mathrm{Seg}}^{(e)} + N_{\mathrm{Add}}}{N_{\mathrm{All}}}, \tag{2}$$

where N_{All} is the total number of the actual trip ends under the target BS density, m is the number of groups of merged identification, e is the number of groups of segmented identification, N_{Not} is the number of the trip ends not identified, and N_{Add} is the number of the trip ends of additional identification.

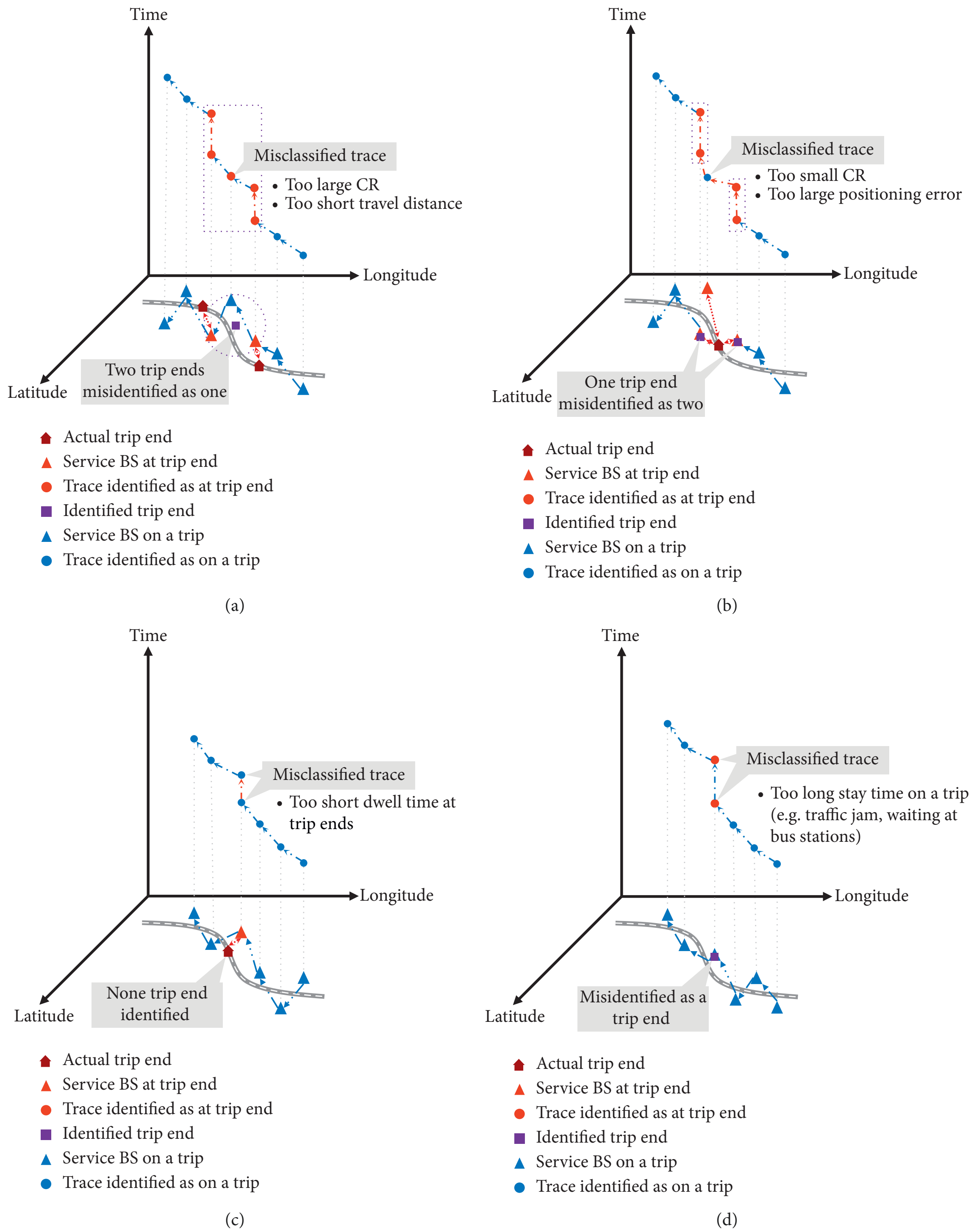


Figure 4: Diagrams of misidentification for trip ends. (a) Merged identification. (b) Segmented identification. (c) Not identified. (d) Additional identification.

The fitness function of a CR under the target BS density is constructed as follows. Firstly, DBSCAN with the fixed CR is used to identify trip ends from all mobile phone data. Secondly, the trip ends under the target BS density are screened out. Thirdly, the EIA and EIR of those trip ends are calculated. Finally, the difference between EIA and EIR, namely, FIT = EIA – EIR, is taken as the fitness function of GA.

2.2.2. Framework of GA-DBSCAN. The optimization process of the GA-DBSCAN framework is shown in Figure 5. The specific steps are presented below.

Step 1: we generate 30 binary CRs randomly. Each binary number is deemed as a chromosome of an individual

Step 2: the CRs are substituted into the DBSCAN algorithm. Then the FIT of each CR is calculated by the fitness function above

Step 3: we set the GA parameters, such as iterations, crossover probability, and mutation rate. Then the CRs are screened and generated by a classical genetic selection process, namely, Selection-Crossover-Mutation. The next generation CRs from the GA process are plugged into DBSCAN again unless the end condition is met.

Step 4: as Steps 2 and 3 are constantly iterated until meeting the end condition, the new generation CRs with higher FIT can be gradually screened out. Finally, the CR producing the maximum *FIT* is the optimal parameter under the target BS density.

Step 5: on the basis of the above process for optimizing the CR under the target BS density, the optimal CRs under different BS densities are searched out in a similar manner. Finally, a functional relation between optimal CRs and BS densities is obtained through function fitting, that is, $R = \text{Fun}(\text{density})$.

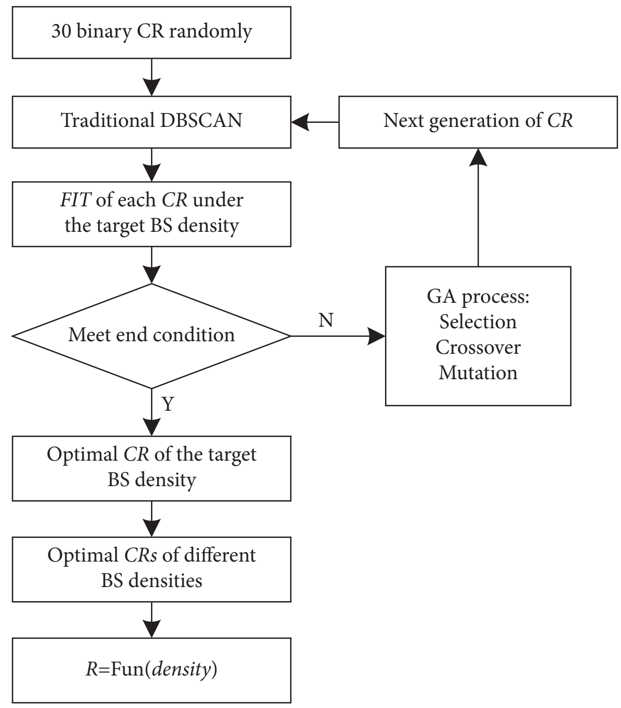


Figure 5: Flowchart of GA-DBSCAN framework.

2.3. Trip End Identification. DBSCAN is a clustering algorithm relying on the density characteristic of traces [44]. The preset parameters of DBSCAN are the density parameter M_{in} and the clustering radius R. The purpose of the algorithm is to detect all core points which are the traces with more than M_{in} other traces within the R-radius range [33]. In this paper, core points are deemed as traces generated at trip ends. Through the equal-time-interval interpolation algorithm, each trace stands for the dwell time of F seconds. Therefore, the number of traces within the range of a CR can represent the dwell time $T_{\text{stay}} = F \cdot M_{\text{in}}$.

Aiming at the deficiency of the traditional DBSCAN, this paper proposes an improved F-DBSCAN for two objectives, with the pseudocode shown in Figure 6.

One of the objectives is for increasing the identification accuracy of trip ends. The traditional fixed CR for judging core points is improved to a dynamic CR obtained from the function $R = \text{Fun}(\text{density})$ and the BS density around each trace. As shown in Figure 7, before judging whether a trace is a core point, the BS density surrounding this trace is firstly counted. In this paper, the number of BSs per square kilometer serves as the evaluation index of BS densities. Then, the corresponding optimal CR is selected by $R = \text{Fun}(\text{density})$. Finally, we count if the number of traces within the range of this CR is more than M_{in}. In areas with a high BS density, a small CR is adopted, while the CR selected is rather large in areas where BSs are sparsely distributed. The key steps of this improvement in the pseudocode refer to steps from 8) to 10) and from 21) to 23).

The other objective is for reducing the time complexity of the algorithm. The fast clustering improvement mainly depends on the characteristic that there are a large number of repeated traces with the same coordinates at trip ends in mobile phone data, which is from three aspects as follows.

(1) Unified processing of repeated traces: the repeated traces with the same coordinate are uniformly judged as whether they are core points instead of one by one in the traditional algorithm. In this way, although a large number of repeated traces are generated and interpolated at trip ends, these traces will not increase the running time of the algorithm. The key steps of this improvement in the pseudocode refer to steps 3), 7), 13), 20), and 26).

(2) Unified processing of traces around high repeated traces: if more than M_{in} repeated traces are at a certain position, these traces are clearly all core points. Without considering CR difference, the high repeated traces in this position make other traces within its CR range also become core points. Due to little change in BS densities around the traces in a short distance, if more than M_{in} repeated traces are

Input:
$D = \{x_1, \ldots, x_m\}$ - m traces to be clustered
$R = \mathrm{Fun}(density)$
M_{in} - density parameter
Output:
C - activity location cluster
Steps:
1) Mark all traces in D *'unvisited'* ;
2) Create an activity location cluster C, $C = \{\}$;
3) Define the set S_{x_m} of the traces with the same coordinate as x_m ;
4) The number of traces in S_{x_m} is defined as s_{x_m} ;
5) For select an *'unvisited'* trace x_p with the highest s_{x_p} :
6) Mark x_p *'visited'*;
7) Mark the traces in S_{x_p} *'visited'*;
8) Count BS density d_{x_p} around x_p ;
9) $R_{x_p} = \mathrm{Fun}(d_{x_p})$;
10) Count the set $D^*_{x_p}$ of the traces within a R_{x_p} -meter radius of x_p ;
11) If the number of the traces in $D^*_{x_p}$ is greater than M_{in} :
12) $C \leftarrow x_p$;
13) Add the traces in S_{x_p} to C , which is $C = C \cup S_{x_p}$;
14) If $s_{x_p} > M_{\mathrm{in}}$:
15) Mark the traces in $D^*_{x_p}$ *visited'*;
16) $C = C \cup D^*_{x_p}$;
17) Continue;
18) For each *'unvisited'* trace x^*_p in $D^*_{x_p}$:
19) Mark x^*_p *'visited'*;
20) Mark the traces in $S_{x^*_p}$ *'visited'*;
21) Count BS density $d_{x^*_p}$ around x^*_p ;
22) $R_{x^*_p} = \mathrm{Fun}(d_{x^*_p})$;
23) Count the set D^{**} of the traces within a $R_{x^*_p}$ -meter radius of x^*_p ;
24) If the number of the traces in D^{**} is greater than M_{in} :
25) $C \leftarrow x^*_p$;
26) $C = C \cup S_{x^*_p}$;
27) If $s_{x^*_p} > M_{\mathrm{in}}$:
28) Mark the traces in D^{**} *'visited'*;
29) $C = C \cup D^{**}$;
29) Continue;
30) else
31) Add the traces in D^{**} to $D^*_{x_p}$, which is $D^*_{x_p} = D^*_{x_p} \cup D^{**}$;
32) Output C ;

Figure 6: Pseudocode of the improved F-DBSCAN.

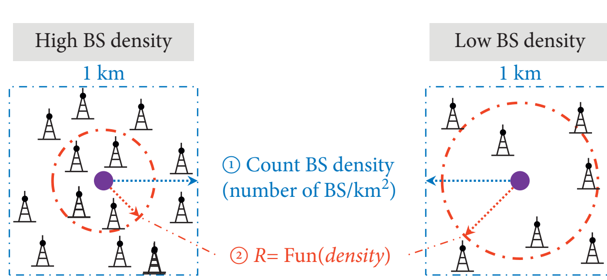


Figure 7: A diagram of adjusting CRs dynamically.

at a certain position, these traces and the traces within its CR range are directly judged as core points. In this way, most mobile phone traces around trip ends can be processed uniformly without judging one by one. The key steps of this improvement in the pseudocode refer to steps from 14) to 17) and from 27) to 29).

(3) Clustering sequence: the time complexity can be further reduced, if the high repeated traces and the traces around them are prioritized. Therefore, we first count the number of repeated traces at every position in mobile phone data. The number of repeated traces from high to low is taken as the clustering sequence, instead of selecting them randomly in the traditional algorithm. The key steps of this improvement in the pseudocode refer to steps 4) and 5).

At last, the traces x_v in the trip end cluster C as an algorithmic output are segmented by a time gap more than F. Define $C_{v\longrightarrow g} = \{x_v, x_{v+1}, \ldots, x_{v+g}\}$ as a time-continuous sequence of traces in C. Due to round trips, the traces with the same coordinate may be not in the same sequence. Therefore, the time difference T_g between the trace x_{v+g} and x_v needs to be further checked. If $T_g < T_{\text{stay}}$, $C_{v\longrightarrow g}$ is removed from C. Else if $T_g \geq T_{\text{stay}}$, the sequence $C_{v\longrightarrow g}$ is deemed as a trip end cluster. Moreover, the coordinate of the trip end can be expressed in $L(C_{v\longrightarrow g}) = 1/g + 1\sum_{k=v}^{v+g} c_k$, where c_k is the coordinate of a trace x_k.

3. Data Collection

3.1. Experimental Design. The existing literature using anonymous mobile phone data fails in obtaining actual travel information of users. As a consequence, it is rather difficult to evaluate the effects of the methods. In this paper, mobile phone data are derived from China Unicom with a large market share (around 30%). The operator provided not only anonymous mobile phone data from more than one million users in one month, but also mobile phone data provided by volunteers who have fulfilled real-name authentication and participated in the field travel experiment.

The data collection experiment was performed in Guiyang City which is densely populated. Within its administrative region, there are over 19,000 BSs of China Unicom. The average coverage radius of each BS is below 150 m. During the experiment, not only was mobile phone data collected as the research object, but also GPS data and travel log data were synchronously gathered for algorithm assessment. The mobile phone data was automatically collected from the smartphone where SIM cards of China Unicom have been installed. Besides, an APP independently developed for GPS data collection was also installed in the smartphone and remained activated throughout the whole course. The travel log was manually recorded by volunteers themselves, including the time of traffic jam and arrival and departure time at each trip end.

Three categories of trip ends were designed, that is, Work, Home, and Others (including entertainment and shopping). Between the trip ends, multiple trip modes were adopted, such as walking, buses, cars, and subways. The experimental design also gave full consideration to mobile phone data collection under different BS densities. BS densities in the city center, new district, and suburb of Guiyang City are shown in Figure 8. The points with different colors are the positions of BSs. The different colors represent the BS density around each BS, measured by the number of BSs per square kilometer. In such three regions, the average coverage radii of BSs turn out to be approximately 64 m, 128 m, and 276 m, respectively. From September to December 2019, 11.5 million GPS trajectory records were gathered from over 500 trips by dozens of volunteers. According to their SIM card information, the operator provided more than 180,000 mobile phone data records.

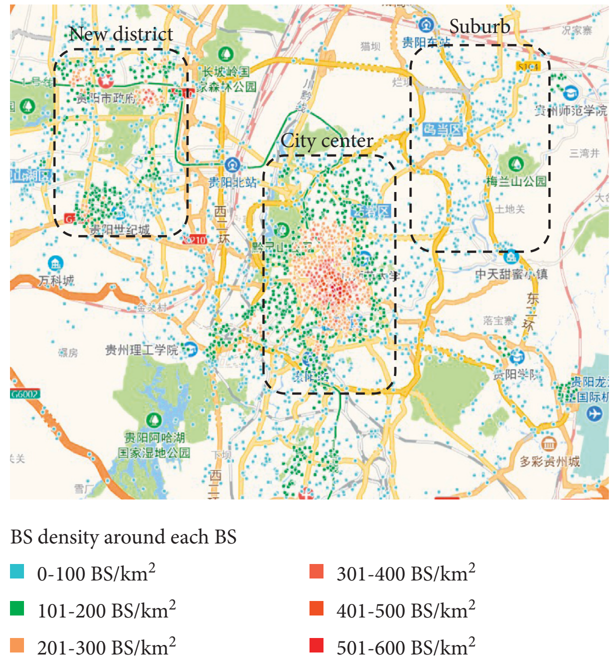


Figure 8: BS densities in different areas of Guiyang City.

3.2. Data Analysis. Mobile phone data directly records the coordinates of BSs connected with users when communication events take place. The communication events can be classified into two categories: (1) active events driven by users, such as calls, messages, or the Internet; (2) passive events driven by the communication network, such as handoff and location update. An example of mobile phone data is presented in Table 1. The spatial and temporal distribution characteristics of mobile phone data are analyzed as follows.

The temporal characteristic of mobile phone data is mainly reflected in the probability distribution of time intervals between adjacent data, as shown in Figure 9. The highest probability of time intervals lies in the range of 0~10 seconds, which is above 40%. As the time interval rises, the probability rapidly declines. The cumulative probability distribution shows that more than 90% of mobile phone data

Table 1: An example of mobile phone data.

Global identifier	Phone number	Device ID	LAC	Cell-ID
460 *** 38	130 *** 3477	869 **** 863	34050	167875275
460 **** 38	130 *** 3477	869 *** 863	34050	167967762
460 **** 38	130 *** 3477	869 *** 863	34050	167887754
Communication event	Start time (s)	End time (s)	Longitude (°)	Latitude (°)
103	12:21:03	12:21:03	106.7050	26.6582
103	12:21:04	12:21:04	106.8520	26.5981
103	12:21:09	12:21:09	106.7244	26.6692

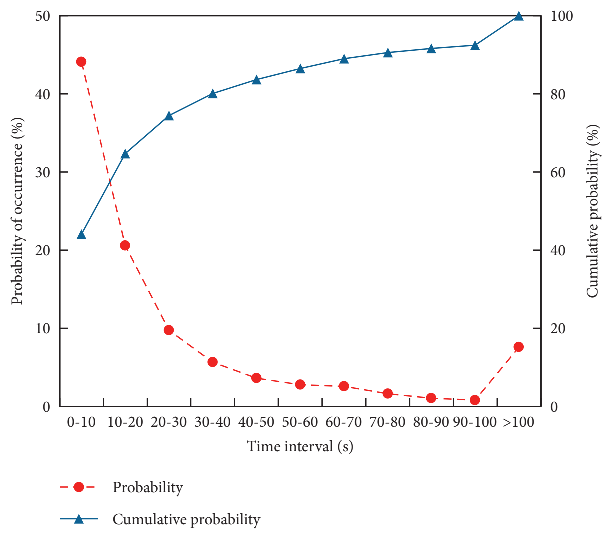


Figure 9: Probability distribution of time intervals between adjacent mobile phone data.

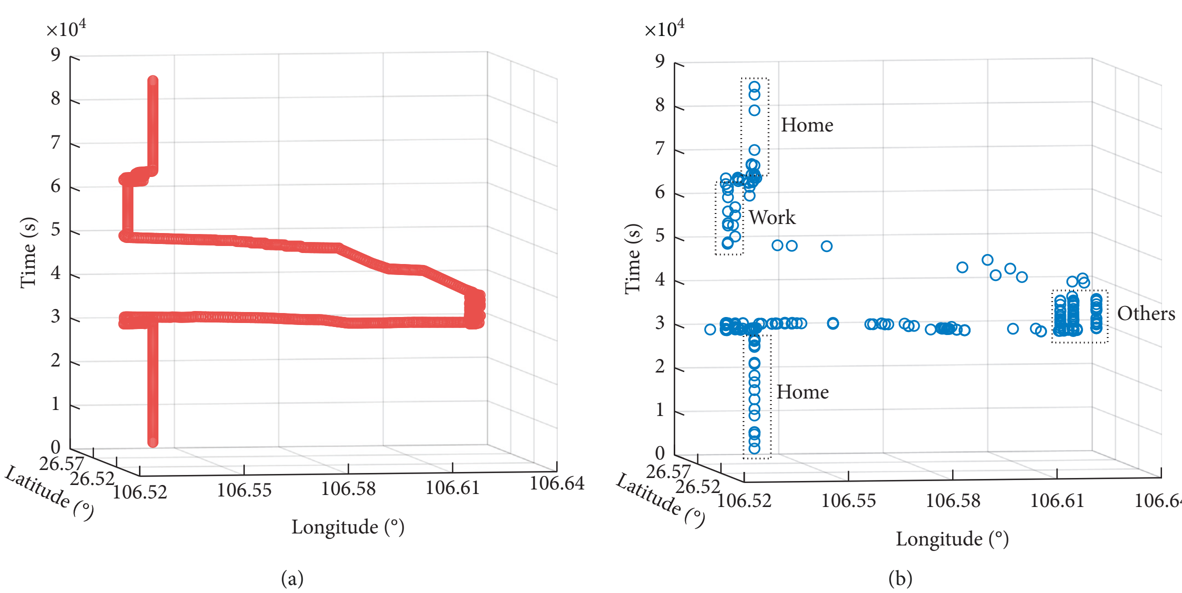


Figure 10: Spatiotemporal distribution of GPS and mobile phone data of a case sample. (a) GPS trajectories. (b) Mobile phone traces.

is generated within a time interval of 80 s. Less than 7% of the time intervals exceed 100 s, and the median is only 12 s. It signifies that mobile phone data can track the positions of users timely.

Spatiotemporal positioning distribution of mobile phone data is compared with that of GPS trajectory data, as presented in Figure 10. The horizontal and vertical axes contain longitudes and latitudes. The vertical axis represents time.

We can see that GPS trajectories are dense and continuous in time and space. As analyzed above, mobile phone data does not occur continuously in time. In space, there also exist some traces that dramatically deviated from the real positions. The corresponding reason is that the positioning errors of mobile phone data are affected by some communication environmental factors, such as BS densities. We further compare the traces of different types of trip ends in this case. The traces of Work and Home are concentratedly distributed in space, while those of Others are relatively spread.

We use the coordinates of GPS trajectories with positioning errors usually less than 10 m as reference data for measuring errors of mobile phone data. The errors under different BS densities are compared in Figure 11. BS densities are measured by the number of BSs per square kilometer. Due to insufficiency or deviations of the mobile phone data collected under some BS density environment, the average positioning errors are missing or fluctuate. While the average positioning errors tend to gradually decline overall as BS densities increase. When the number of BSs per square kilometer rises from 0–100 to 500–600, the average positioning errors reduce from 500–800 m to 0–200 m.

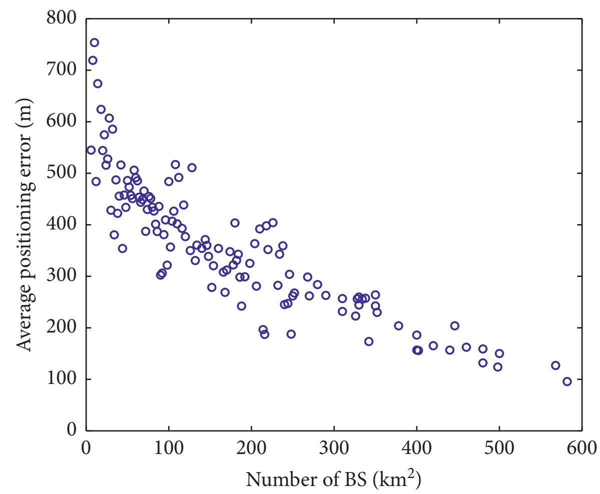


Figure 11: Average positioning errors of mobile phone data under different BS densities.

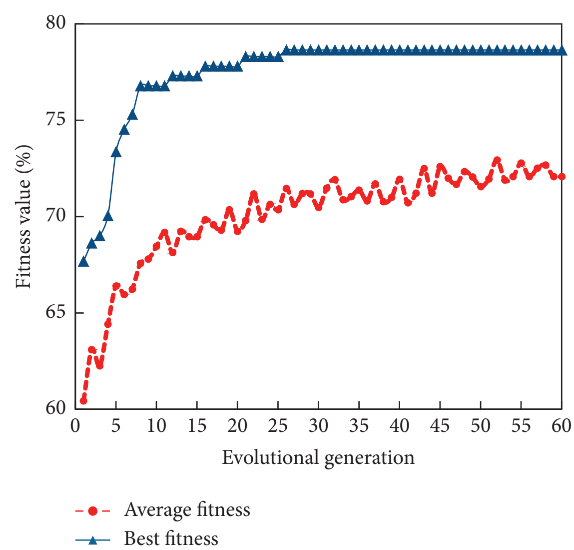


Figure 12: Variation curves of fitness values (BS density: 0–100/km^2).

4. Result Analysis and Discussion

4.1. Parameter Optimization and Case Study

4.1.1. Parameter Setting and Optimization. MATLAB is utilized to build and train the GA-DBSCAN framework for optimizing CRs under various BS densities. In this process, the maximum number of evolutional generations is set at 60, the crossover probability at 0.9, and the mutation probability at 0.03. In our experiments, we set the threshold of stay time T_{stay} to 20 min (i.e., 1200 s), which falls within the range of commonly accepted values for the typical minimum duration of a significant activity carried out by an individual at the same location [33, 45, 46]. As the interpolation cycle F of this paper is 10 s, $M_{in} = T_{stay}/F = 120$.

BS densities are divided into five groups in units of the number of BSs per square kilometer, namely, 0–100/km^2, 101–200/km^2, 201–300/km^2, 301–400/km^2, and 401–500/km^2. The CR of each group is optimized by the GA-DBSCAN framework. Figure 12 presents variations in the fitness values during optimization taking the group of 0–100/km^2 as an example. We can see that the fitness values gradually increase along with the evolutional generations. Although the average fitness values keep fluctuant due to the influence of random factors such as the mutation probability, the best fitness values converge to a stable value after the 26th generation.

The optimal CR of each BS density group is obtained by GA-DBSCAN, as shown in Figure 13. As can be observed, a rise in the BS densities is accompanied by a gradual decrease in the optimal CRs. The relationship of the two variables conforms to the power-law distribution. The median of each BS density group is selected as an independent variable of the optimal CRs. The relational expression is achieved in

$$R_{DBSCAN} = \mathrm{Fun}\left(d_{DBSCAN}\right) = 1554 \times d_{-0.3478}^{-0.3478}, \quad (3)$$

where R_{DBSCAN} is the optimal CR (unit: meter) and d_{DBSCAN} is the number of BSs per square kilometer. However, if the independent variable approaches 0, the power function will be positive infinity, making the function invalid. Therefore, when the number of BSs per square kilometer is below 50, we select the optimal CR to be the same as that when $d_{DBSCAN} = 50$, namely, 399 m.

4.1.2. Case Study. The example data in Figure 10 is identified by the proposed algorithm as a case study. Figure 14 shows the spatial and temporal distribution of the result. The red traces are the trip end clusters identified by the algorithm.

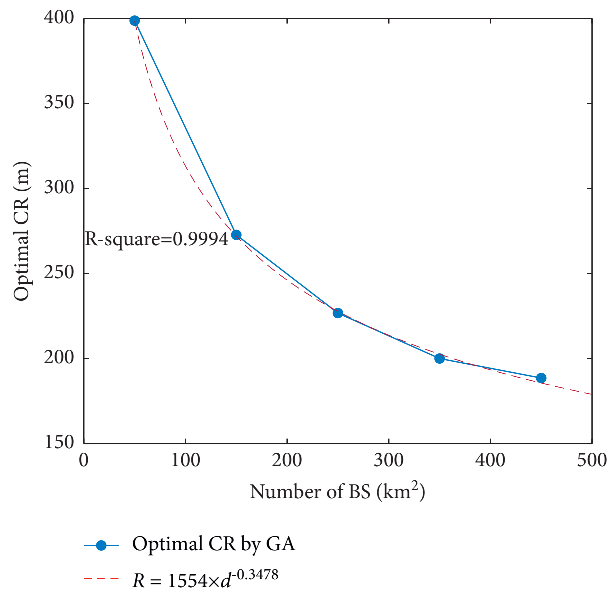
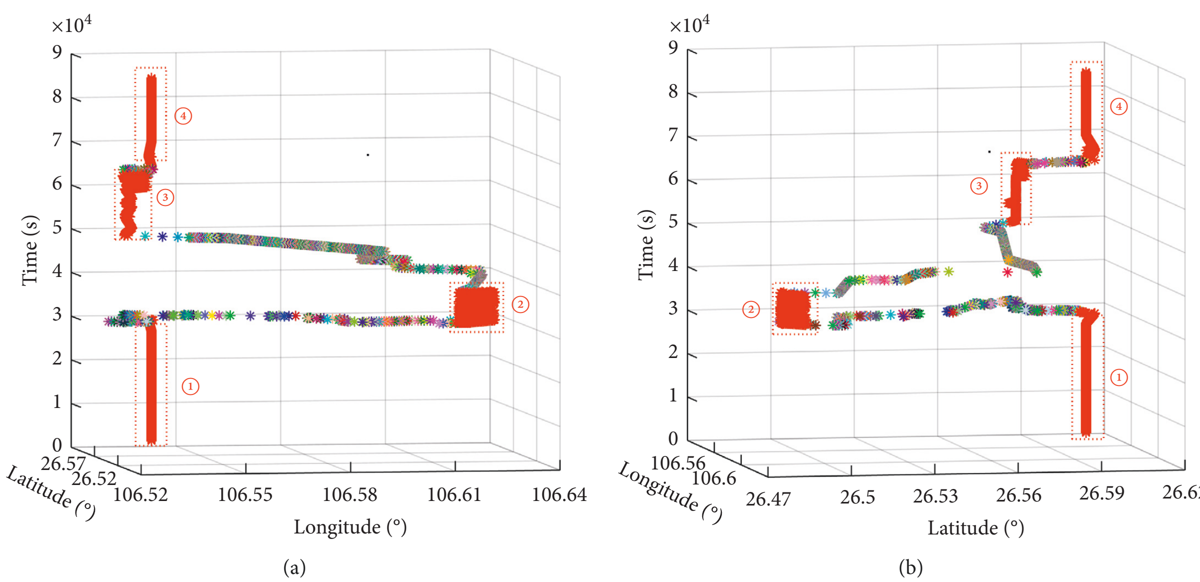
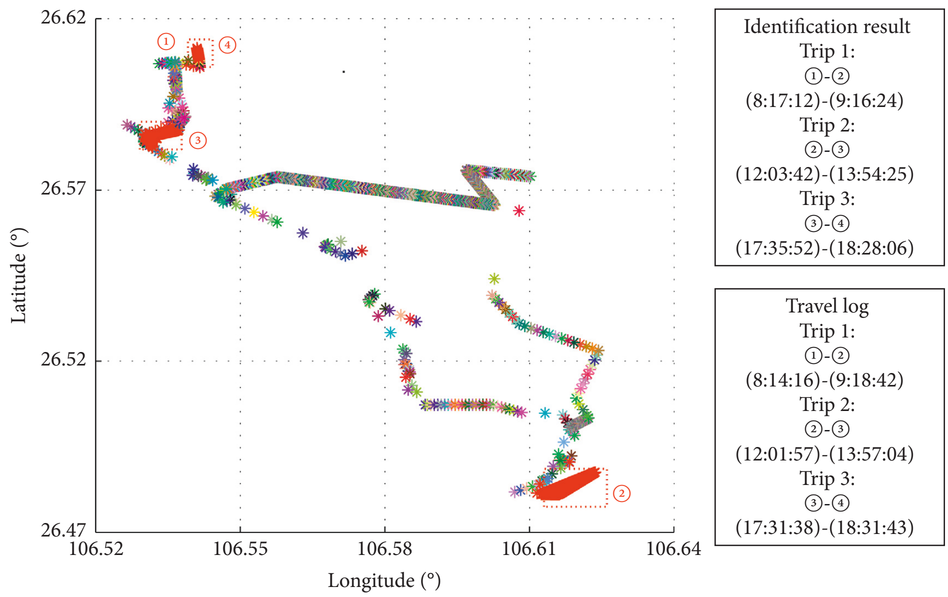


Figure 13: Optimal CRs under different BS densities.





(c)

Figure 14: An example of trip end identification results from three perspectives. (a) Spatiotemporal perspective: from north to south. (b) Spatiotemporal perspective: from west to east. (c) Spatial perspective.

The trace density at trip ends is significantly raised by the equal-time-interval interpolation algorithm compared with the distribution shown in Figure 10. It can be observed that the user traveled 3 times in total in the day, producing 4 trip ends. The trip ends 1 and 4 are in the same place, namely, the user's place of residence. We compare the identified travel time and trip end coordinates with the actual travel information. The time errors of the arrival/departure time lie within 4 min. The distance errors of the coordinates are all no greater than 140 m.

4.2. Comparative Analysis of Different Algorithms

4.2.1. Comparison of Identification Accuracy. The identification results of the improved F-DBSCAN are evaluated through comparison with the existing methods, as shown in Table 2. The methods to be compared include the traditional DBSCAN and the method proposed by Wang et al. [22]. In the traditional DBSCAN, three commonly used CRs are selected to be 200 m, 300 m, and 400 m [31, 33]. Wang et al. [22] firstly used an incremental clustering algorithm (ICA) to extract trip ends and preset the CR as 400 m. K-means clustering is subsequently adopted to perform post-optimization of results, where the number of clusters as the preset parameter is set to be the number of the trip ends identified by ICA.

As shown in Table 2, the EIAs of the traditional DBSCAN are all below 80%, which are approximately 6%–9% lower than that of the improved F-DBSCAN. Although the EIR of the traditional DBSCAN with a fixed 400 m CR is about 0.5% lower, its EIA is 10% below that of the improved F-DBSCAN due to merged identification caused by an excessively large CR. The EIA of ICA + K-means is close to that of the traditional DBSCAN with a fixed 300 m CR, which is also 5% lower than that of the improved F-DBSCAN. Although two clustering algorithms are combined, this method also uses fixed CRs, so that it is less likely to avoid inadequate applicability of fixed CRs caused by variations in BS densities. Given the above, the improved F-DBSCAN has a superior identification effect on the whole. The validity of the dynamic CR selection mechanism proposed in this paper is proved.

4.2.2. Comparison of Time Complexity. Reduction in time complexity can greatly facilitate the use of large-scale mobile phone data in daily traffic surveys, especially in million population cities. The improvement of improved F-DBSCAN consists of two parts, namely, dynamically adjust CRs and fast clustering. In order to evaluate the respective influence of the two parts on time complexity, we compare the running time of different DBSCAN algorithms, which are traditional DBSCAN, improved DBSCAN and improved F-DBSCAN. In traditional DBSCAN, the fixed CR is set as 300 m. Improved DBSCAN can only adjust CRs dynamically with higher identification accuracy, but without the fast clustering improvement. Because the time complexity of traditional DBSCAN and ICA is both $O(n^2)$ [34, 47], the method ICA + K-means with similar accuracies but higher time complexity is not added into the comparison.

Figure 15 is the boxplot of the running time for processing every user's daily mobile phone data using the different algorithms with the same computing hardware. We can see that the median running time of the improved DBSCAN is 0.55 s (about 42%) longer than that of the traditional DBSCAN. This is because the BS density and optimal CR around each trace are calculated in addition to

Table 2: Comparison of identification accuracy among different methods.

Method	CR (m)	EIR (%)	EIA (%)
ICA + K-means	400	4.1	78.2
	200	9.7	75.5
Traditional DBSCAN	300	4.9	79.5
	400	3.4	74.6
Improved F-DBSCAN	$R = \text{Fun (density)}$	3.9	85.3

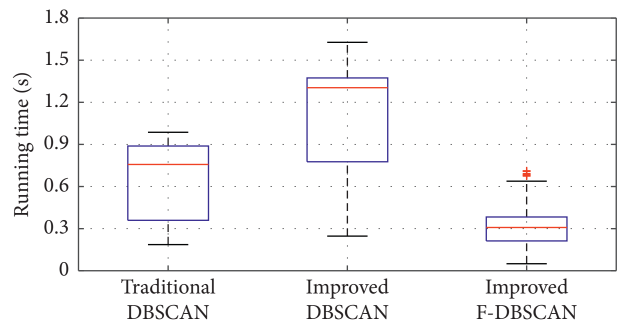


Figure 15: Comparison of running time among different DBSCAN algorithms.

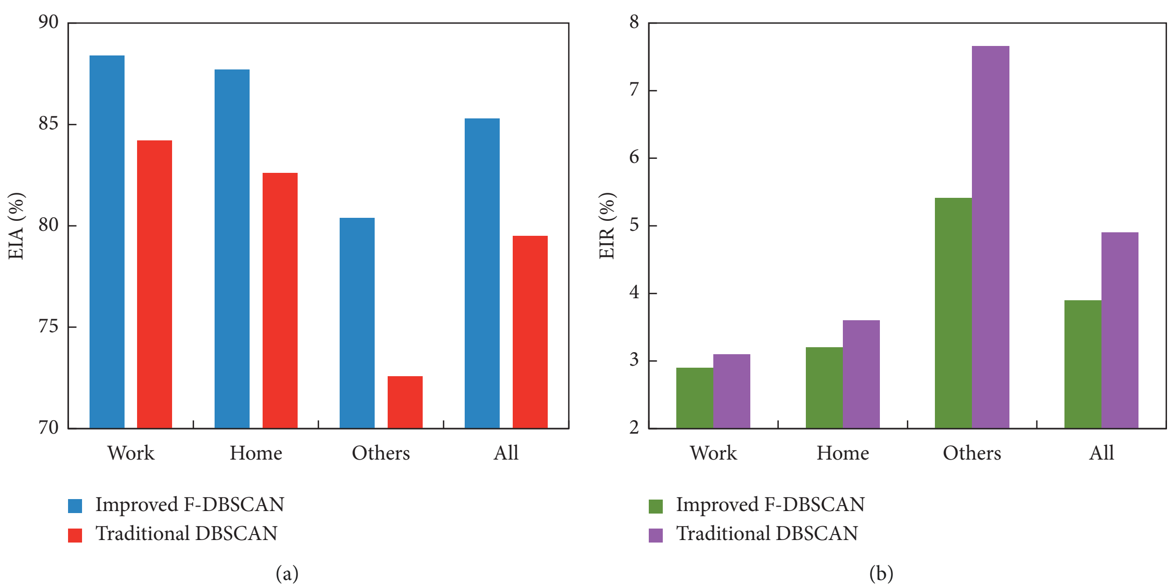


Figure 16: Identification accuracy of different types of trip ends. (a) EIA. (b) EIR.

the core point judgment of the traditional algorithm, so the time complexity increases. When calculating the BS density around a trace, we adopt the following simpler calculation method to reduce the computing amount. For every 0.01° difference in longitude and latitude, the distances are about 1000 m and 1112 m, respectively, according to the statistics in Guiyang City. This means that a distance of 500 m respectively corresponds to a difference of 0.005° in longitude and 0.0045° in latitude. When counting the number of BSs per square kilometer, we directly search out BSs with longitude and latitude differences within ±0.005° and ±0.0045° from the target trace coordinate, instead of computing the distance between their coordinates.

The median running time of the proposed improved F-DBSCAN is about 1 s (about 76%) lower than the improved DBSCAN. The average running time decreases from 1.11 s to 0.31 s, by about 72%. Even compared with the traditional DBSCAN, despite the computing amount for adjusting CRs dynamically in the improved F-DBSCAN, the median and average running time also decrease by about

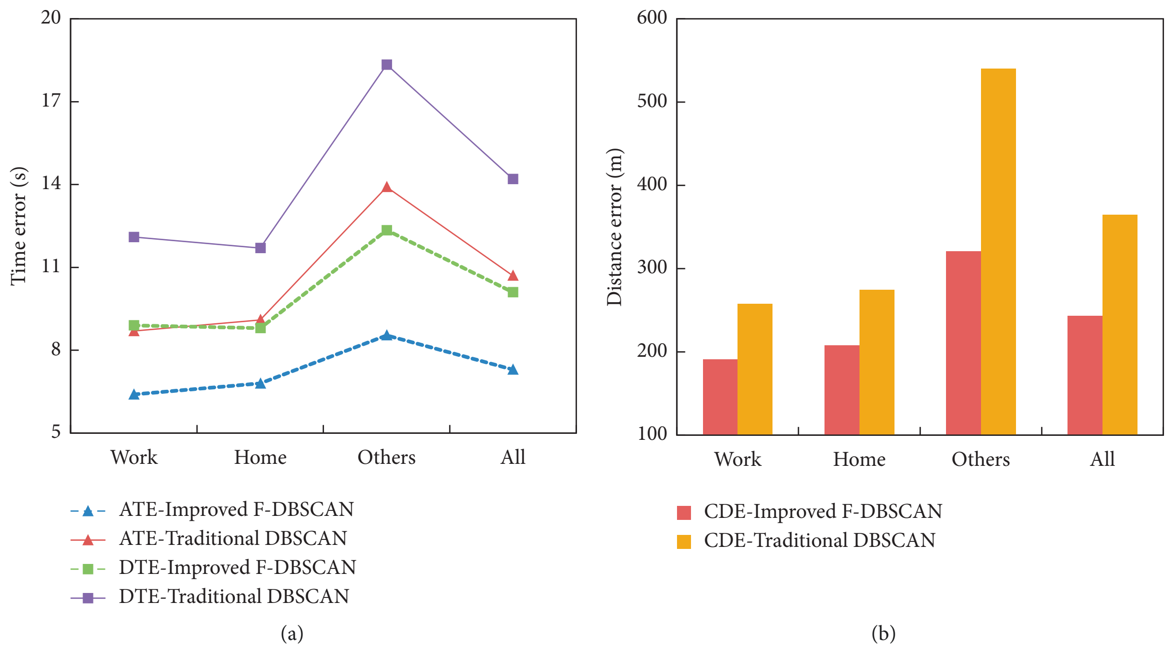


Figure 17: Identification errors of different types of trip ends. (a) ATE/DTE (unit: min). (b) CDE (unit: m).

59% and 55%, respectively. It is proved that the fast clustering improvement proposed in this paper has a great effect on reducing time complexity.

4.3. Result of Different Types of Trip Ends. Trip ends are usually divided into three types, namely, Work, Home, and Others [29, 48]. We further analyze their identification results and compare the traditional DBSCAN with a fixed 300 m CR and the improved F-DBSCAN. Two indexes described above, that is, EIA and EIR, are adopted to evaluate the identification accuracies of the trip ends. Besides, three average error indexes are utilized, including arrival time error (ATE), departure time error (DTE), and coordinate distance error (CDE), to assess the identification effects of travel time and trip end coordinates, where ATE/DTE is equal to an absolute value of the difference between the start/ending time in a trip end cluster correctly identified and the actual arrival/departure time at the trip end. CDE is the distance from the coordinates of the actual trip ends to those of the identified trip ends.

4.3.1. Identification Accuracy. A comparison of identification accuracies between the traditional and improved algorithms is presented in Figure 16. We can see that the difference in their overall EIA is approximately 6%. However, certain differences lie in optimization results of different types. The EIA in Work and Home produces a difference of about 4%, while in the type of Others including shopping and entertainment, the EIA of the improved F-DBSCAN is raised from 72.6% to 80.4%, by about 8%. Likewise, the reduction of the EIR in Others is greater than that in Work and Home. A reason is that the range of activity is rather small for users who are working or staying at home. Their connecting BSs are stable, so the mobile phone traces are comparatively dense even under low BS densities. In this condition, the traditional DBSCAN with a fixed CR can also obtain good identification results. By contrast, users usually move in an extensive range in supermarkets or parks where BSs are sparsely distributed. Their serving BSs are more likely to constantly change, resulting in their mobile phone traces being rather scattered. Consequently, it is difficult for fixed CRs to meet relevant clustering conditions. If the fixed CR is directly extended, other trip ends will be influenced, especially for those with a short distance easily merged identified. The improved F-DBSCAN is capable of dynamically adjusting CRs, so it is more suitable for identifying noncommuting trip ends under various BS densities.

4.3.2. Identification Error in Time and Coordinate. A comparison of identification errors in time and coordinates between the traditional and improved algorithms is presented in Figure 17. We can see that DTEs are generally about 1–5 min longer than ATEs. That is because when a user chooses to take a bus or taxi, the position where he/she waits for a bus or taxi is rather close to his/her actual trip end. In this case, no significant changes are incurred in the coordinates of their mobile phone data. This leads to misidentifying them still staying at the trip end before he/she gets on and leaves.

Then we compare the identification errors. It is demonstrated that the average ATE/DTE and CDE are respectively reduced by about 3 min and 67 m by the improved F-DBSCAN in Work and Home, but by about 5.5 min and 220 m in Others. Corresponding reasons are similar to those described above. That is, the range of activity is rather wide at Others trip ends, making it difficult for fixed CRs to be applied in diversified BS distribution.

The improvement effect on travel time identification of the improved F-DBSCAN is not obvious in numerical terms which is about 5 min. The two reasons are as follows. On one hand, the average ATE/DTEs are below 15 min and the error reduction proportions have been about 25%–39%. This means that the improvement space of the proposed method is limited. On the other hand, a traveler usually enters into the service range of a BS neighboring the trip end several minutes before their arrival. The mobile phone traces generated in this period are positioned into the trip end clusters in advance. When departing from a trip end, the mobile phone traces usually leave in a delayed manner. Therefore, the time error incurred from the data source itself cannot be easily removed by the method improvement.

The improvement effect on identifying the trip end coordinates is significant by the improved F-DBSCAN. The overall average CDE decreases from 364 m to 243 m, by about 33.3%. The average CDE in Others is further reduced from 541 m to 321 m, by about 40.6%. The scale of a traffic zone of the four-step model commonly used in the field of transportation planning is generally larger than $500 \times 500\ m^2$. Hence, the proposed method in this paper can obtain OD tables more accurately. Moreover, the method can assist in enhancing the precision of epidemiological investigation.

5. Conclusions

Trip end identification is fundamental in residents' travel information detection. It is still important to improve the identification effects of trip ends. Meanwhile, actual travel information for result evaluation is absent due to anonymous mobile phone data used in the existing literature. In this paper, mobile phone data is collected from real-name volunteers thanks to the support from the communication operator. We propose a new identification method that is improved based on the positioning characteristics of mobile phone data. Firstly, due to the influence of BS layout on the parameter setting ignored in current studies, we build a GA-DBSCAN framework to optimize CRs under different BS densities. On this basis, the traditional DBSCAN is improved to be able to adjust CRs dynamically based on BS densities, so that the identification accuracy can be raised. Secondly, considering that there are plenty of traces with the same coordinates in mobile phone data, we propose a fast clustering improvement for lower time complexity. On the premise of keeping the identification accuracy, the median running time can be reduced by over 76%. The improved F-DBSCAN can be more competent for large-scale travel surveys using mobile phone data.

Travel information data as ground truth can help us further explore supervised deep learning models for trip end identification. In future work, we will study the applicability of the Long Short Term Memory model for extracting travel characteristics, such as trip ends and transportation modes. We will also explore the data fusion method using multi-types of positioning datasets. On this basis, the accuracy of existing research topics based on mobile phone data, such as residents' trip pattern monitoring, can be further enhanced using our methods.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

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 Natural Science Foundation of China (Grant Nos. 52072313, 52002030, and 52002282); Science Program of Chongqing Municipal Planning and Natural Resources Bureau (Grant No. KJ-2021007); Humanities and Social Sciences Foundation of the Ministry of Education (Grant No. 20XJCZH011); Humanities and Social Sciences Foundation of Shaanxi Province (Grant No. 2020R035); Natural Science Foundation of Shaanxi Province (Grants Nos. 2020JM-222 and 2021JQ-256); and Fundamental Research Funds for the Central Universities CHD (Grants Nos. 300102219301 and 300102342105).

References

[1] C. Ratti, D. Frenchman, R. M. Pulselli, and S. Williams, "Mobile landscapes: using location data from cell phones for urban analysis," *Environment and Planning B: Planning and Design*, vol. 33, no. 5, pp. 727–748, 2006.

[2] R. Ahas, A. Aasa, S. Silm, and M. Tiru, "Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 1, pp. 45–54, 2010.

[3] X. Chen, X. Chen, C. Li, and J. Chen, "Sample Expansion model of household travel survey using Cellphone data," *Tongji Daxue Xuebao/Journal of Tongji University*, vol. 49, no. 1, pp. 86–96, 2021.

[4] F. Liu, D. Janssens, J. Cui, Y. Wang, G. Wets, and M. Cools, "Building a validation measure for activity-based transportation models based on mobile phone data," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6174–6189, 2014.

[5] 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, pp. 63–74, 2014.

[6] N. Caceres, L. M. Romero, and F. G. Benitez, "Exploring strengths and weaknesses of mobility inference from mobile phone data vs. travel surveys," *Transportmetrica: Transportation Science*, vol. 16, no. 3, pp. 574–601, 2020.

[7] H. Ghayvat, M. Awais, P. Gope, S. Pandya, and S. Majumdar, "Recognizing suspect and predicting the spread of Contagion Based on Mobile Phone location data (COUNTERACT): a system of identifying COVID-19 infectious and hazardous sites, detecting disease outbreaks based on the internet of things, edge computing, and artificial intelligence," *Sustainable Cities and Society*, vol. 69, Article ID 102798, 2021.

[8] B. Rostami-Tabar and J. F. Rendon-Sanchez, "Forecasting COVID-19 daily cases using phone call data," *Applied Soft Computing*, vol. 100, Article ID 106932, 2021.

[9] M. C. González, C. A. Hidalgo, and A. L. Barabási, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[10] F. Calabrese, G. D. Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Computing*, vol. 10, no. 4, pp. 36–44, 2011.

[11] C. Pan, J. Lu, S. Di, and B. Ran, "Cellular-based data-extracting method for trip distribution," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1945, no. 1, pp. 33–39, 2006.

[12] O. Järv, R. Ahas, and F. Witlox, "Understanding monthly variability in human activity spaces: a twelve-month study using mobile phone call detail records," *Transportation Research Part C: Emerging Technologies*, vol. 38, pp. 122–135, 2014.

[13] Y. Xu, S.-L. Shaw, Z. Zhao et al., "Another Tale of two Cities: Understanding human activity space using actively tracked Cellphone location data," *Annals of the Association of American Geographers*, vol. 106, no. 2, pp. 246–258, 2016.

[14] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: a case study in Rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

[15] F. Calabrese, M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti, "Understanding individual mobility patterns from urban sensing data: a mobile phone trace example," *Transportation Research Part C: Emerging Technologies*, vol. 26, pp. 301–313, 2013.

[16] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti, "Activity-aware map: identifying human daily activity pattern using mobile phone data," in *Human Behavior Understanding*, vol. 6219, pp. 14–25, LNCS, 2010.

[17] M.-H. Wang, S. D. Schrock, N. Vander Broek, and T. Mulinazzi, "Estimating dynamic origin-destination data and travel demand using cell phone network data," *International Journal of Intelligent Transportation Systems Research*, vol. 11, no. 2, pp. 76–86, 2013.

[18] Z. Yao, Y. Zhong, Q. Liao, J. Wu, H. Liu, and F. Yang, "Understanding human activity and urban mobility patterns from massive Cellphone data: Platform design and applications," *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 3, pp. 206–219, 2021.

[19] L. Ni, X. Wang, X. Chen, and X. M. Chen, "A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 86, pp. 510–526, 2018.

[20] Y. Yamada, A. Uchiyama, A. Hiromori, H. Yamaguchi, and T. Higashino, "Travel estimation using Control Signal Records in cellular networks and geographical information," in *Proceedings of the 2016 9th IFIP Wireless and Mobile Networking Conference*, pp. 138–144, WMNC, Colmar, France, July 2016.

[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, 2017.

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

[23] F. Yang, Y. Wang, P. J. Jin, D. Li, and Z. Yao, "Random forest model for trip end identification using cellular phone and points of interest data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2675, no. 7, pp. 454–466, 2021.

[24] Z. Wang, S. Y. He, and Y. Leung, "Applying mobile phone data to travel behaviour research: a literature review," *Travel Behaviour and Society*, vol. 11, pp. 141–155, 2018.

[25] H. Wang, F. Calabrese, G. D. Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records," in *Proceedings of the IEEE Conference on Intelligent Transportation Systems*, pp. 318–323, ITSC, Funchal, Madeira Island, Portugal, September 2010.

[26] H. Poonawala, V. Kolar, S. Blandin, L. Wynter, and S. Sahu, "Singapore in motion: Insights on public transport service level through farecard and mobile data analytics," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 589–598, CA, USA, August 2016.

[27] H. Huang, Y. Cheng, and R. Weibel, "Transport mode detection based on mobile phone network data: a systematic review," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 297–312, 2019.

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

[29] 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.

[30] 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.

[31] S. Park, Y. Xu, L. Jiang, Z. Chen, and S. Huang, "Spatial structures of tourism destinations: a trajectory data mining approach leveraging mobile big data," *Annals of Tourism Research*, vol. 84, Article ID 102973, 2020.

[32] C. Yang, Y. Zhang, X. Zhan, S. V. Ukkusuri, and Y. Chen, "Fusing mobile phone and travel survey data to model urban activity dynamics," *Journal of Advanced Transportation*, vol. 2020, Article ID 5321385, 17 pages, 2020.

[33] L. Bonnetain, A. Furno, N. E. E. Faouzi et al., "TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data," *Transportation Research Part C: Emerging Technologies*, vol. 130, Article ID 103257, 2021.

[34] Y. Chen, S. Tang, N. Bouguila, C. Wang, J. Du, and H. Li, "A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data," *Pattern Recognition*, vol. 83, pp. 375–387, 2018.

[35] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, 2016.

[36] M. Liang, R. W. Liu, S. Li, Z. Xiao, X. Liu, and F. Lu, "An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation," *Ocean Engineering*, vol. 225, Article ID 108803, 2021.

[37] F. Yang, H. Jiang, Z. Yao, and H. Liu, "Evaluation of activity location Recognition using cellular signaling data," *Xinan Jiaotong Daxue Xuebao/Journal of Southwest Jiaotong University*, vol. 56, no. 5, pp. 928–936, 2021.

[38] W. Wu, Y. Wang, J. B. Gomes et al., "Oscillation resolution for mobile phone cellular tower data to enable mobility

modelling,"vol. 1, pp. 317–324, in *Proceedings of the IEEE International Conference on Mobile Data Management*, vol. 1, pp. 317–324, IEEE, Brisbane, QLD, Australia, July 2014.

[39] W. Li, C. Wang, G. Xu, J. Luo, and X. Zhang, "Research of track resident point identification algorithm based on signaling data," *Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology*, vol. 42, no. 12, pp. 3013–3020, 2020.

[40] L. Qi, Y. Qiao, F. Ben Abdesslem, Z. Ma, and J. Yang, "Oscillation resolution for massive cell phone traffic data," in *Proceedings of the 1st Workshop on Mobile Data*, pp. 25–30, Association for Computing Machinery, Singapore, June 2016.

[41] D. Bachir, G. Khodabandelou, V. Gauthier, M. E. Yacoubi, and J. Puchinger, "Inferring dynamic origin-destination flows by transport mode using mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 254–275, 2019.

[42] R. W. Liu, M. Liang, J. Nie, W. Y. B. Lim, Y. Zhang, and M. Guizani, "Deep learning-Powered vessel trajectory Prediction for improving Smart traffic services in maritime internet of things," *IEEE Transactions on Network Science and Engineering*, vol. 14, no. 8, p. 1, 2022.

[43] H. Qu, L. Yin, and X. Tang, "An automatic clustering method using multi-objective genetic algorithm with gene rearrangement and cluster merging," *Applied Soft Computing*, vol. 99, Article ID 106929, 2021.

[44] R. W. Liu, J. Nie, S. Garg, Z. Xiong, Y. Zhang, and M. S. Hossain, "Data-driven trajectory quality improvement for promoting intelligent vessel traffic services in 6g-enabled maritime iot systems," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5374–5385, 2021.

[45] 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, 2016.

[46] M. Fekih, T. Bellemans, Z. Smoreda, P. Bonnel, A. Furno, and S. Galland, "A data-driven approach for origin-destination matrix construction from cellular network signalling data: a case study of Lyon region (France)," *Transportation*, vol. 48, no. 4, pp. 1671–1702, 2021.

[47] S. Balakrishna, M. Thirumaran, R. Padmanaban, and V. K. Solanki, "An efficient incremental clustering based improved K-Medoids for IoT multivariate data cluster analysis," *Peer-to-Peer Networking and Applications*, vol. 13, no. 4, pp. 1152–1175, 2020.

[48] Y. Wang, G. H. D. A. Correia, B. V. Arem, and H. J. P. Timmermans, "Understanding travellers' preferences for different types of trip destination based on mobile internet usage data," *Transportation Research Part C: Emerging Technologies*, vol. 90, pp. 247–259, 2018.