IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 9, 2019, accepted December 16, 2019, date of publication December 26, 2019, date of current version January 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2962536

Research on Urban Resident Activity Patterns and Hotspot Area Based on GPS Floating Car Data

YONGFA LI [1], XIAOQING ZUO [1], AND FANG YANG [2]
[1]School of Land Resources Engineering, Kunming University of Science and Technology, Kunming 650093, China
[2]Kunming Surveying and Mapping Institute, Kunming 650051, China

Corresponding author: Xiaoqing Zuo (zuoxqing@163.com)

This work was supported by the National Natural Science Foundation of China for through ''Urban Resident Travel Pattern Recognition Based on Mobile Location Data'' under Grant 41061043/D010703. The project mainly provides the experimental data and experimental equipment (computer, etc.) for this article.

ABSTRACT With the rapid development of urbanization, the living standard has been improved on a continued basis for urban residents and their activities have become increasingly frequent. Therefore, it is of massive significance to study the hot spots of urban residents' activities and enforce effective planning and decision-making for urban and traffic departments. In this paper, the data is preprocessed in the first place. Then, the passengers' pick-up points and travel track points are extracted, and the statistical analysis method is employed to analyze the travel length and travel time of urban residents. Finally, an improved FCM algorithm is proposed. The conventional Fuzzy c-means (FCM) clustering algorithm is classed as a local optimal algorithm, and the number of clustering is made uncertain. In view of the shortcomings as mentioned above, an improved (FCM) clustering algorithm is suggested in this paper, which adopts adaptive distance norm and adds its own norm induction matrix to each cluster in order to ensure global optimization. The partition coefficient (PC), classification entropy (CE) and (XB) index are introduced to assist in determining the optimal number of clustering. According to the statistical analysis of GPS track data, the morning peak and evening peak in a day is 7:00-10:00 and 17:00-20:00, respectively. Cluster analysis is carried out for each time period and the whole day using the model proposed in this paper. The results show that the number of hot spots in each time period and the whole day is 12, 10, 8, 13, 13, respectively. The hot spots are distributed in the business center, office and residential areas, which are consistent with the actual situation. It shows that the model proposed in this paper can effectively and accurately mine the hot spots of urban residents' activities. It plays an important role in urban planning and commercial layout.

INDEX TERMS GPS data, improved FCM clustering algorithm, activities of urban residents, hot spots.

I. INTRODUCTION

Urban residents represent a significant part of the city, and their travel activities have been regarded as a crucial research content in such fields as accessibility research, traffic planning research and location-based Service (LBS) research [1]. Traffic trips in urban traffic are initiated by people, and urban residents are the main crowd. Kunming is a relatively suitable city for people to reside in. With the fast-paced economic development, both the population and the number of motorized vehicles are also on the rise. As a result, the traffic pressure faced by Kunming is substantial and the circulation is poor, which puts more strain on the work and life of urban residents while hindering economic development for the city as whole. To this end, the relevant urban planning departments in Kunming have made plenty of efforts, such as adding 14 new bus lines and 923 new buses in 2015, and increasing the passenger carrying rate of public transport to 54 [2]. In addition, subway construction across the city has been accelerated. Nevertheless, the number of motorized vehicles in the city has yet to show a declining trend, and the issue of traffic congestion remains severe during the peak hours. In order to improvement to the current traffic situation in Kunming, it is necessary to study the activity patterns of urban residents in Kunming. By the analysis of the activity patterns exhibited by urban residents, we can gain understanding as to the travel conditions for urban residents, predict the travel behavior and activity patterns of residents in the future, and

The associate editor coordinating the review of this manuscript and approving it for publication was Dalin Zhang.

provide the basis for improvement to the traffic situation in Kunming. This plays a significant role in the rational planning of traffic for the relevant government agencies.

The traditional approach to information acquisition about urban residents' activities is reliant on questionnaires and log activities, primarily through visits and telephone interviews. However, there are various problems arising from this, such as long time, subjectivity, less information, inaccurate survey results [3]. In recent years, many scholars have made use of various positioning data to study the patterns of residents' activity and hot spots, as a result of which many Spatiotemporal data mining methods have been devised. To meet the transportation demand for taxis, mitigating the mismatch between taxi supply and demand, an urban taxi fleet size calculating model based on GPS tracking data is proposed by [4]. Authors of [5] customer journey time metrics for New York City bus service using big data. In order to realize the efficient cooperative organization of passenger flow and vehicles, a set of complete scheduling models is developed by [6]. To assess energy conservation and emission reduction effect on traffic and transportation at macro level more scientifically, the evaluation system based on the improved osculating value method for transportation is established by [7]. The above studies indicate that most of the existing studies in this field are conducted from the macroscopic aspect, while few are conducted from the microscopic aspect. Therefore, this paper proposes an improved FCM clustering algorithm. By researching and analyzing the daily travel time, activity rules, and hotspots of urban residents in Kunming, the research results can provide references for related departments, such as traffic department or the bus company can according to the resident travel time and hotspots constantly improve and update the scheduling system, and urban planning departments can allocate and improve infrastructure construction according to the hot spots of residents' activities, such as the location and number of shared bicycles.

Fuzzy C-means clustering is one of the most popularly algorithms in cluster analysis [8]. It is widely used in various fields, such as intelligent transportation [9], image segmentation [10], dynamic risk assessment [11], bearing fault diagnosis [12] and others. But the number of clusters needs to be pre-set by humans. It has great limitations in practical applications. For different application fields, scholars have continuously improved the FCM algorithm to better meet people's needs. Concerning the problem that general Fuzzy C-Means (FCM) and its improved algorithm are sensitive to noise in the samples clustering and clustering boundary is not accurate enough, an improved FCM clustering algorithm based on spatial correlation is proposed by [13]. As FCM algorithm usually affected by the selection of initial values and noise data, and can easily fall into local optimal value. To overcome these drawbacks of FCM, the algorithm of FCM based on multi-chain quantum bee colony algorithm (MQBC-FCM) is proposed by [8]. FCM algorithm is easy to be affected by fuzzy parameters, initial clustering, noise, and because of the single iteration path, the phenomenon of local extreme can be generated, an improved search strategy method is proposed by [14]. At present, the improvement of FCM algorithm mainly focuses on the FCM's local optimal problem caused by initial clustering and noise influence. There are few studies on the uncertainty of cluster number, but the number of clusters is significance in practical applications. Therefore, this paper mainly studies the uncertainty of the number of clusters of FCM algorithm.

FCM algorithm is a clustering algorithm that classifies logarithm data points according to the membership degree [15], [16]. Prior to clustering, this algorithm is required to manually determine the number of clustering and is sensitive to parameters and noise, as a result of which it is a kind of local optimal algorithm. In order to address the shortage of the C- means algorithm, in this paper, the improved FCM clustering algorithm, based on each cluster to add their own induced matrix norm for the purpose of ensuring the global optimal, and by introducing the partition coefficient (PC), classification entropy (CE) and (XB)index of three indicators to determine the optimal clustering number. It is effective in resolving the shortage of traditional FCM clustering algorithm, which requires manual determination of the number of clustering. The specific technical route of this paper is as follows: Firstly, the data is preprocessed, including map matching, eliminating invalid and erroneous data. Secondly, analyze the travel time and length of residents via statistical analysis; finally, mines the hotspots of urban residents' activities via improved clustering algorithm.

II. DATA COLLECTION AND PRE-PROCESSING

The original data obtained by taxi GPS positioning can not be directly used as experimental data, as it needs to be preprocessed, including data source and structure analysis, coordinate transformation and abnormal data processing. The original data is transmitted to the taxi management center on a regular basis by the taxi GPS device and stored in the database after being processed by the data processing module of the system center. The data primarily includes taxi number, data acquisition time, longitude and latitude of the location, driving speed, driving direction, passenger status and other information. The detailed data information is shown in table 1, showing only some of the data.

Due to the influence exerted by surrounding objects, GPS failure and other factors in the process of driving, the taxi records the data outside the research area, including wrong data and abnormal data. Therefore, GPS floating car data is supposed to be processed for improvement to the effectiveness and accuracy of the research results. The treatment method is detailed as follows:

(1) The coordinate of GPS floating car data and the coordinate of the research area are compared to determine whether the floating car data obtained by GPS is in the research area. If it is not in the research area, this part of data will be removed.

(2) Invalid data. If multiple data with the same latitude and longitude are returned at different times, this part of the data

TABLE 1. Partial data information.

stime	Latitude	Longitude	speed	Orientati	State	GpsTime
2010-10-22 17:32:36	25031521	102686415	5	107	1	2010-10-22
2010-10-22 17:32:52	25031465	102686325	0	0	1	2010-10-22
2010-10-22 17:33:04	25031343	102686170	13	111	0	2010-10-22
2010-10-22 17:33:07	25031250	102686075	17	109	1	2010-10-22
2010-10-22 17:33:22	25031031	102685821	0	0	1	2010-10-22
2010-10-22 17:34:04	25031031	102685821	0	0	0	2010-10-22
2010-10-22 17:35:04	25030968	102685378	11	134	0	2010-10-22
2010-10-22 17:35:04	25030968	102685378	11	134	1	2010-10-22
2010-10-22 17:35:19	25031385	102684673	14	155	1	2010-10-22
2010-10-22 17:35:35	25031573	102684341	0	0	1	2010-10-22
2010-10-22 17:36:04	25031573	102684341	0	0	0	2010-10-22
2010-10-22 17:36:23	25031806	102683898	27	152	1	2010-10-22
2010-10-22 17:36:38	25032623	102684235	35	14	1	2010-10-22
2010-10-22 17:36:53	25033520	102684560	13	25	1	2010-10-22
2010-10-22 17:37:08	25033660	102685933	44	42	1	2010-10-22
2010-10-22 17:37:23	25033725	102687403	7	44	1	2010-10-22
2010-10-22 17:37:38	25033645	102687830	0	67	1	2010-10-22
2010-10-22 17:37:53	25033435	102687913	14	94	1	2010-10-22

is treated as invalid. The main reason is that the floating car is located at the traffic light intersection and in a traffic jam state during this period, so this part of data ought to be deleted.

(3) Wrong data. Multiple records of the same data were recorded, and the duplicate records were deleted by writing a program.

Map matching technology [16], [17] provides a commonly used technical approach to the research on GPS track [18] positioning. The positioning point obtained by using GPS positioning technology gives rise to certain error, and a situation will occur that the positioning point does not match the road. The purpose of map matching is to use GPS. When the electronic map is accurate, the position displayed on the electronic map is the correct position of the vehicle after the map is matched. There are two ideas of map matching. The other is to match GPS points to the road. In either case, the most significant thing is to find the current driving path, which is crucial to map matching algorithm [19], [20]. Based on the needs of research and the situation of GPS positioning information, the road buffer based map matching algorithm is applied in this paper for map matching.

Then for the road buffer analysis, based on the accuracy of GPS points in the experiment, the road network includes 10 meters of buffer. If the anchor point is located in a certain section of the buffer, put the anchor point out. If the GPS anchor point is within a certain section of buffer, then the anchor point is possibly related to the match, and a section of it can improve the efficiency of computation significantly. Each section to buffer the GPS inside anchor point is matching with the corresponding road. In the process of matching by using dot to line matching method, consideration shall be given to two factors, distance and direction. With the distance factor considered, it is needless to say that under the premise that there are multiple possible sections within a certain range, it is a necessity to consider which section is closest to the GPS point direction.

III. METHODOLOGY

The research in this paper is based on a general case, without considering special circumstances, so this paper is carried out under the following assumptions:

(1) Assume that people's travel habits remain unchanged during the study period, and the integrated transportation system has not changed significantly;

(2) Assume that the taxi operating price is basically stable during the study period, and there will be no population transfer due to economic factors;

(3) Assume that the number of people taking a taxi is one person at a time.

Based on the above assumptions, the improved FCM algorithm is used to study and analyze the hotspots of urban residents' activities via the statistical analysis method and improved FCM algorithm.

A. STATISTICAL ANALYSIS METHOD

1) CALCULATION OF TRAVEL LENGTH

The travel length of residents is calculated by using the extracted trajectory data, and the so-called travel length refers to the space, a trajectory data packet that contains the starting point, ending point and trajectory point. In this case, the form is $\{(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_n, y_n, t_n)\}$, then the corresponding coordinate sequence on the trajectory is $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$. According to the distance formula between two points in mathematics, the distance can be obtained by using the following formula:

$$D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} + \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2} + \ldots + \sqrt{(x_n - x_{n-1})^2 + (y_n - y_{n-1})^2} \quad (1)$$

where x and y respectively represent the horizontal and vertical coordinates of the locus in the trajectory.

2) CALCULATION OF TRAVEL TIME

The method of calculation for the time that residents spend on travel is similar to that for travel length. When calculating travel time, only time parameters can be used. If one track of residents' travel is expressed as $\{(x_1, y_1, t_1), (x_2, y_2, t_2), \ldots, (x_n, y_n, t_n)\}$, the travel time that the resident spends can be expressed as $t_n - t_1$, and the time unit is finally unified as hours.

B. IMPROVED FCM ALGORITHM

Traditional FCM is known as a local optimal algorithm, which has a greater sensitivity to the initial clustering center [10], [15]. The number of clustering needs to be determined manually prior to clustering. If the initial value fails to be selected properly, the clustering effect will be reduced significantly [17]. Therefore, an improved FCM clustering algorithm based on the traditional FCM algorithm is proposed in this paper. This algorithm adopts the adaptive distance norm [21], for the purpose of finding cluster classes with different set shapes in the data set, as a result of which each cluster class has its own norm induction matrix A_i [22], [23]. This algorithm removes the need to determine the number of clusters manually. However, it requires the introduction of the corresponding index to determine the best number of clusters.

It is assumed that there are N data samples $X = (x_1, x_2, x_3, \ldots, x_N)$, C indicates the number of clusters, $V = (v_1, v_2, v_3, \ldots, v_N)$ denotes the clustering center vector of each class, and its distance norm [24] is defined as:

$$D_{ikA}^2 = (x_k - v_i)^T A (x_k - v_i), \quad 1 \le i \le c, 1 \le k \le N \quad (2)$$

where D_{ikA}^2 represents the square distance norm, i is clustering center, k is cluster sample, the following meanings are the same here. The matrix A_i means the optimization variable in the average function. Therefore, each cluster class is allowed to adapt the distance norm to the data of local topology structure. Let $A = (A_1, A_2, .., ,A_c)$ represent c elements in the matrix induced norm, than the objective function J of the improved FCM clustering algorithm is defined as:

$$J(U, V, A) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{ik})^m D_{ikA}^2 \quad (3)$$

where u_{ik} represents the membership of sample x_i to category A_k, and the objective function (3) is incapable to be minimized directly by A_k, which is because it's linear in A_k, m is the fuzzy weight index. In order to obtain a workable solution, A_k needs to be constrained in some way, and the usual way to do this is to constrain the determinant of A_k, while allowing the matrix A_k to change with its determinant, and corresponding to the shape of the optimized cluster when its volume remains unchanged:

$$||A_i|| = \rho_i, \quad \rho > 0 \quad (4)$$

where ρ_i is fixed for each cluster, and the expression of A_i obtained by Lagrange multiplier method [25] is as follows:

$$A_i = [\rho_i \, det\,(F_i)]^{\frac{1}{n}} F_i^{-1} \quad (5)$$

where F_i denotes the fuzzy covariance matrix of the i cluster [26], which is defined as follows:

$$F_i = \frac{\sum_{k=1}^{N} (u_{ik})^m (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{N} (u_{ik})^m} \quad (6)$$

The steps of the improved FCM clustering algorithm are as follows:

Step 1: use (7) to calculate the cluster center of the cluster;

$$v_i = \frac{\sum_{k=1}^{N} (u_{ik})^m x_k}{\sum_{k=1}^{N} (u_{ik})^m}, \quad 1 \le i \le c \quad (7)$$

Step 2: use (8) to calculate the covariance matrix of the cluster;

$$F_i = \frac{\sum_{k=1}^{N} (u_{ik})^m (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{N} (u_{ik})^m}, \quad 1 \le i \le c \quad (8)$$

Step 3: use (9) to calculate the distance;

$$D_{ikA_i}^2(x_k, v_i) = (x_k - v_i)^T \left[(\rho_i \, det(F_i))^{\frac{1}{n}} F_i^{-1} \right] \times (x_k - v_i), \quad 1 \le i \le c, \; 1 \le k \le N \quad (9)$$

Step 4: update the partition matrix with (10);

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ikA_i}(x_k, v_i)}{D_{jk}(x_k, v_j)} \right)^{\frac{2}{(m-1)}}}, \quad 1 \le i \le c, \; 1 \le k \le N \quad (10)$$

Clustering algorithms are in constant attempt to determine a fixed optimal cluster number and parameterized cluster shape. Despite this, it does not necessarily mean that the optimal fitting is meaningful. If the data can be grouped in a meaningful way, the cluster number has a potential to be wrong or the discovered cluster shape possibly does not correspond to the data set. The optimal number of clusters can be determined by following the two steps below.

The first step is to define a validity function to evaluate a complete partition. It is necessary to estimate the upper limit c_{max} of the number of clustering, and the algorithm is required to run each $c \in \{2, 3, \ldots, c_{max}\}$. For each partition, the validity function provides a value that compares with the results of the analysis.

The second step is to define the validity function to evaluate the clusters. Thirdly, c_{max} must be estimated and cluster analysis must be conducted for c_{max}. The clusters are compared with each other on the basis of validity function. Similar clusters are classed into one cluster, and very bad clusters are removed, as a result of which the number of clusters is reduced. The process can be repeated until a bad cluster emerges. Therefore, multiple validity indexes are proposed to measure the clustering algorithm.

(1) Partition coefficient (PC): measure the amount of "overlap" between clusters, as defined below:

$$PC = \frac{1}{N} \sum_{i=2}^{c} \sum_{j=2}^{N} (u_{ij})^2 \quad (11)$$

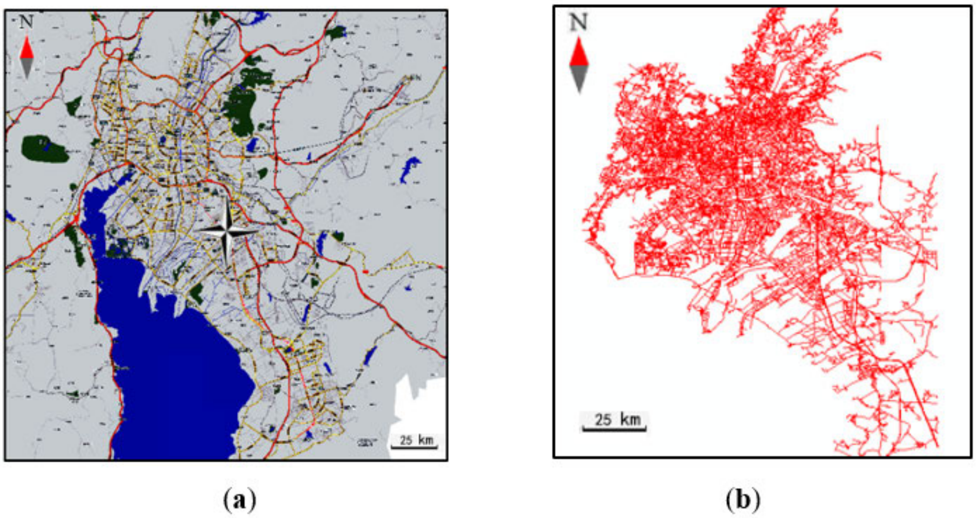


FIGURE 1. Map (a) and road network data (b) of the Kunming city.

where u_{ij} indicates the membership degree of data point j in cluster i. The disadvantage of PC lies in a lack of direct connection with some attributes of the data itself. When the value of PC reaches its maximum, the corresponding clustering number is taken as the optimal clustering number of the data set.

(2) Classification entropy (CE): it measures the fuzzy partition of a cluster, which is similar to the partition coefficient.

$$CE = \frac{1}{N}\sum_{i=2}^{c}\sum_{j=2}^{N} u_{ij}log(u_{ij})^2 \tag{12}$$

(3) Xie and Beni's index (XB): it is aimed at quantifying the change in the cluster and the ratio of cluster separation. The optimal number of clusters is supposed to minimize the value of XB.

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(u_{ij})^m ||x_j - v_i||^2}{N_{min_{i,k}} ||x_j - v_i||^2} \tag{13}$$

IV. EXPERIMENT AND ANALYSIS

A. STUDY AREA

Kunming is situated in the middle of Yunnan-Guizhou plateau in southwest China, with Yunnan lake in the south, mountains on three sides, Sichuan and Chongqing in the north, Vietnam, Laos and Cambodia in the south, and Myanmar, India and Pakistan in the west. Kunming is located in east longitude 102 ° 10 '∼ 103 ° 40', north latitude 24 ° 23 '∼ 22 ° 26', with its city center positioned in north latitude 25 ° 02 '11'', east longitude 102 ° 42' 31''. Well-known as a spring city, Kunming represents a national tourist destination for China that attracts a large number of tourists every year. Therefore, traffic congestion in the city is severe. The electronic map of the main urban area and the road network data are presented in Figure 1.

B. RESSIDENTS' PICK-UP AND DROP-OFF POINTS AND TRAVEL TRAJECTORY EXTRACTION

In this paper, the GPS positioning data on Kunming taxis is used, which contains the operation data collected from 195 taxis for 6 consecutive months (from October 22, 2010 to April 22, 2011). The original data is transmitted to the taxi management center on a regular basis by the taxi GPS device and stored in the database after being processed by the data processing module of the system center. The data primarily includes taxi number, data acquisition time, longitude and latitude of the location, driving speed, driving direction, passenger status and other information.

It is an intuitive reflection of the total amount of travel for urban residents. The more boarding and boarding points, the more times residents spend on travel, which indicates the more frequent activities of residents. In addition, through the statistics in relation to residents' boarding and boarding points, the hot spots for the activities of urban residents can be identified. The GPS floating car data records the passenger carrying state of taxi. Under normal circumstances, there is either 0 or 1 passenger onboard. 0 means no passenger carrying state and 1 means passenger carrying state. When the passenger carrying state changes from 0 to 1, it indicates the boarding point of passengers. When the passenger carrying state changes from 1 to 0, it indicates the boarding point of passengers. The GPS locating point in the middle of the boarding and boarding points is the trajectory point. During the study, 195 taxi positioning data were categorized into 5 groups, with each group made up of 39 taxi GPS positioning data. The results of extracting the pick-up and drop-off points of urban residents are shown in Figure 2.

C. STATISTICAL ANANLYSIS OF TRAVEL LENGTH AND TRAVEL TIME OF RESIDENTS

By studying the patterns for residents to travel, the travel characteristics of residents is summarized macroscopically,

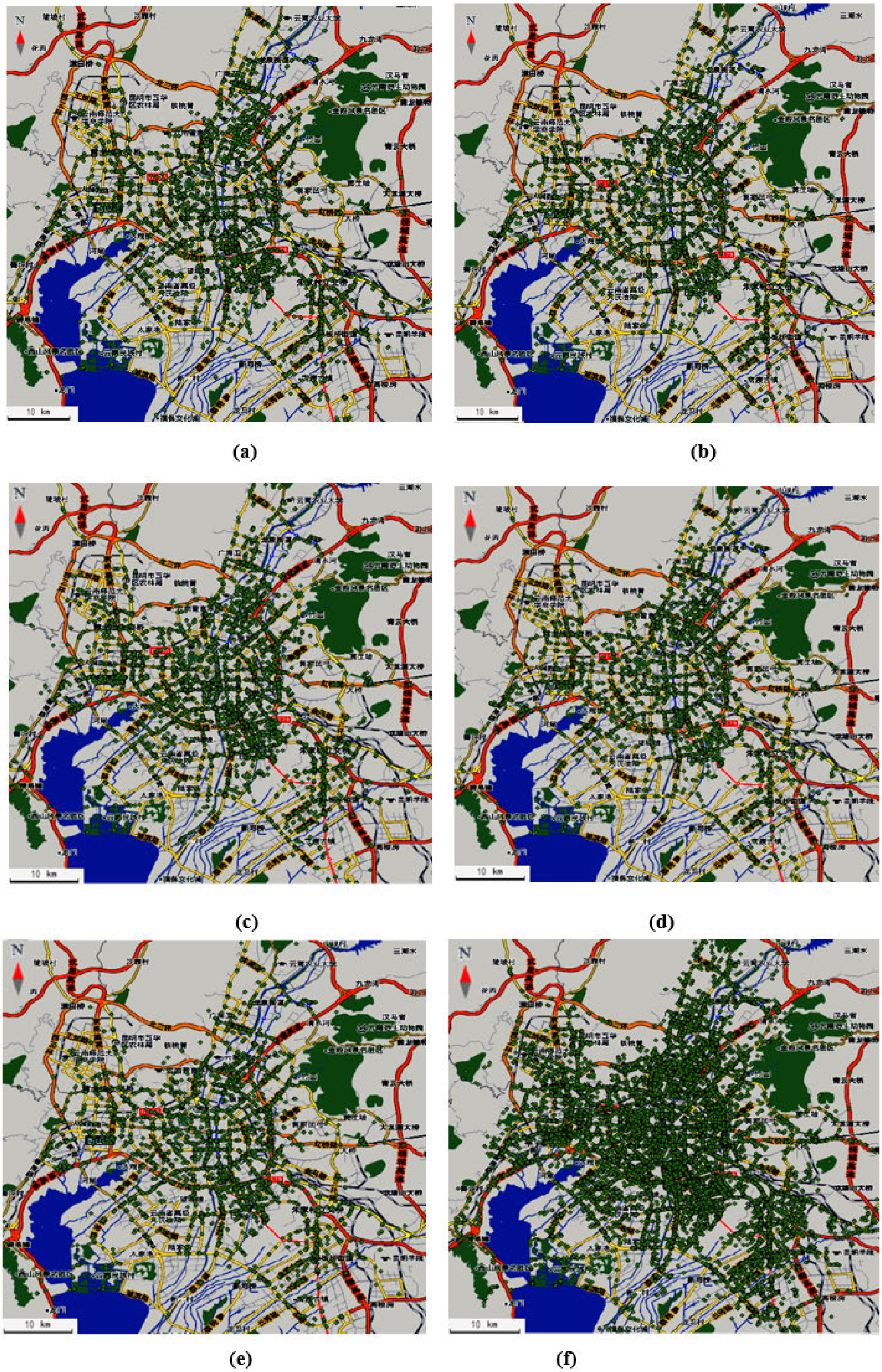


FIGURE 2. The location of boarding points of passengers (a)(b)(c)(d)(e) is the boarding and boarding points.

and the prediction of how residents behave under specific conditions is made possible, which is of great significance to gaining in-depth understanding of the patterns of residents to travel and the establishment of travel behavior models in line with human beings. Therefore, in this section, statistical research is conducted

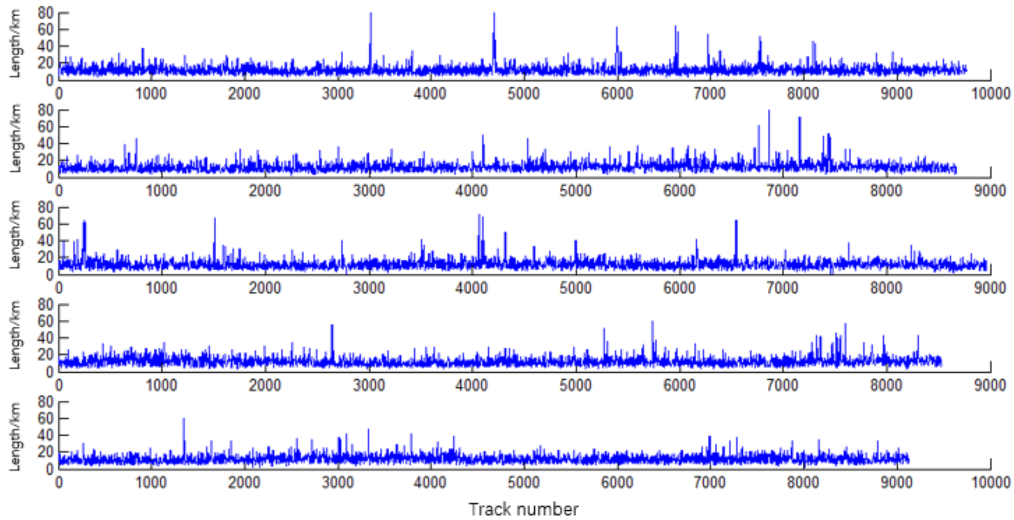


FIGURE 3. Length of residents' travel trajectory.

from two perspectives, travel length rule and travel time rule.

1) LENGTH OF TRAVEL OF RESIDENTS

As taxis in the city covers a massive area, including the entire urban traffic network, and the travel purpose of taxi passengers is clear, it gives rise to certain activities need, as a result of which taxi travel activities can better reflect the characteristics of resident travel. Moreover, taxi management is made more standardized, which improves the quality of GPS data. In this section, the extracted trajectory data are counted and the length of each trajectory is calculated.

After the length of each track is calculated by using equation (1), the track length of each group of GPS positioning data is calculated, as shown in Figure 3.

As indicated by Figure 3, the trajectory length of each of the five sets of GPS positioning data is subject to a similar rule, that is, the majority of residents' travel trajectory length ranges between 2 and 10 kilometers, and the travel length varies between 2 and 20 kilometers, accounting for more than half of the total. A majority of the remainder are between 20 and 40 kilometers, with only a fraction in excess of 40 kilometers. Therefore, it can be concluded that the travel distance of urban residents is mostly short, within (0-10km). A significant proportion of them travel within (10-20km), a small minority of them travel within (20-40km), and only a tiny fraction of them travel beyond 40 km.

2) TRAVEL TIME OF RESIDENTS

Consistent with the travel length of residents, the travel time of urban residents is calculated by calculating the travel time of each trajectory, and the travel time that urban residents spend is calculated as shown in Figure 4.

3) STATISTICAL ANALYSIS OF RESIDENTS' TRAVEL IN DIFFERENT PERIODS

In order to have better understanding as to the travel situation of urban residents in each time period of the day, statistical analysis of residents' travel situation in each hour of the day is conducted.

A day is comprised of 24 hours. In order for better understanding as to the travel situation of residents per hour, the GPS positioning data on 195 taxis for six consecutive months are split into 24 hours to facilitate the calculation of the passenger's boarding and disembarking situation respectively. The passenger's boarding and disembarking represent a complete trip. The statistical results are indicated in Table 2.

As revealed by the statistical results, the distribution map of resident travel volume per hour is drawn. From Figure 5, it can be seen that based on the different time periods, the resident travel situation changes. The resident travel is shown to concentrate in two time periods, which are 7:00-9:00 in the morning and 19:00-21:00 in the evening. These two periods are morning peak and evening peak respectively. In comparison, the resident travel volume exhibits a declining trend starting from 1:00 to 5:00 in the morning, reaching 5:00 in the morning. The time was the lowest, began to show an increasing trend, and then peaked at 9:00 a.m., before a gradual decline to the relatively stable state.

4) CLUSTERING ANALYSIS OF BOARDING AND BOARDING POINTS OF PASSENGERS IN DIFFERENT TIME PERIODS

The day is split into four different time periods, including 7:00-10:00, 11:00-14:00, 17:00-20:00, and 20:00-23:00. The improved FCM clustering algorithm is applied to conduct clustering analysis for each time period, for the purpose of identifying the number and distribution of hot spots in each time period. Then all the data are clustered to determine the number and distribution of hot spots. In chapter 3, the data

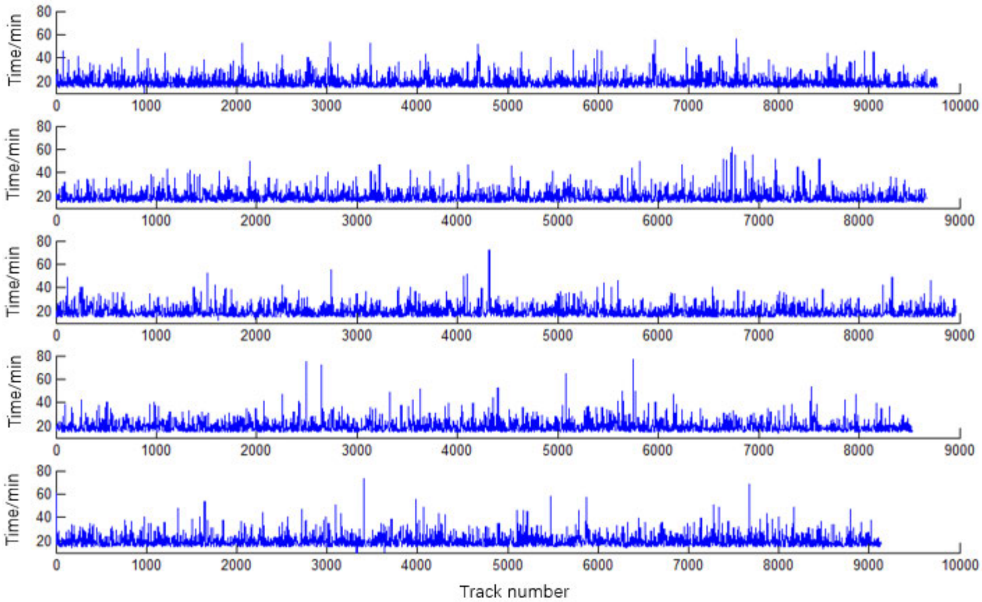


FIGURE 4. Statistical chart of travel time of residents.

TABLE 2. Hourly travel of residents.

Time interval	Number of trains	Number of trips	Time interval	Number of trains	Number of trips
0:00-1:00	17550	17443	1:00-2:00	9678	9653
2:00-3:00	8723	8796	3:00-4:00	8443	8487
4:00-5:00	8157	8167	5:00-6:00	9625	9284
6:00-7:00	13246	13250	7:00-8:00	28253	28384
8:00-9:00	33425	33420	9:00-10:00	26682	26603
10:00-11:00	23251	23139	11:00-12:00	25364	25198
12:00-13:00	26643	26593	13:00-14:00	28542	28620
14:00-15:00	25365	25297	15:00-16:00	24332	24257
16:00-17:00	31327	31258	17:00-18:00	33447	33498
18:00-19:00	35431	35339	19:00-20:00	36764	36783
20:00-21:00	37543	37486	21:00-22:00	28986	28901
22:00-23:00	24364	24301	23:00-24:00	18673	18584

TABLE 3. Residents' travel situation in each time period.

Period of time	Get on the bus number	Out of the car number
7:00-10:00	88360	88407
11:00-14:00	80549	80411
17:00-20:00	105642	105620
20:00-23:00	90893	90688

of boarding and disembarking points for passengers in each time period have been extracted, as indicated in Table 3.

Firstly, formula (11), (12) and (13) are applied to calculate the partition coefficient (PC), classification entropy (CE) and (XB) index of data in each time period respectively. The calculation results are presented in Figure 6. The abscissa in the figure represents the number of clusters, and in the ordinate, the (PC) represent the overlapping between clusters,

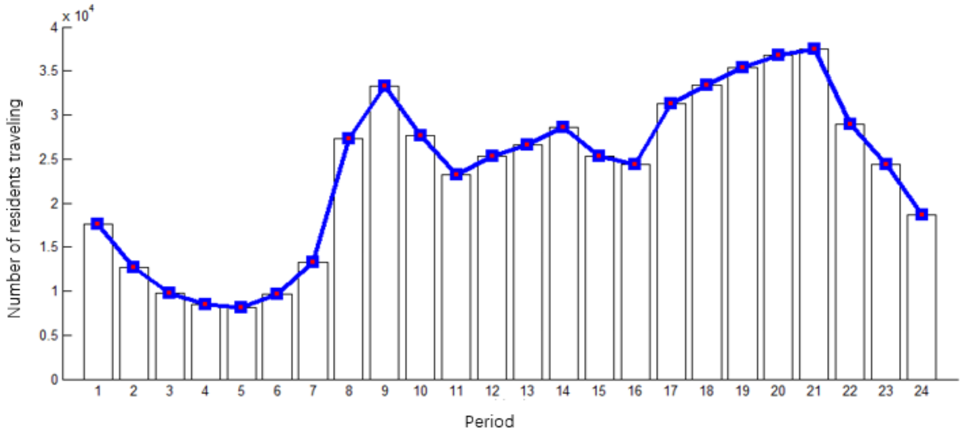


FIGURE 5. Hourly travel statistics.

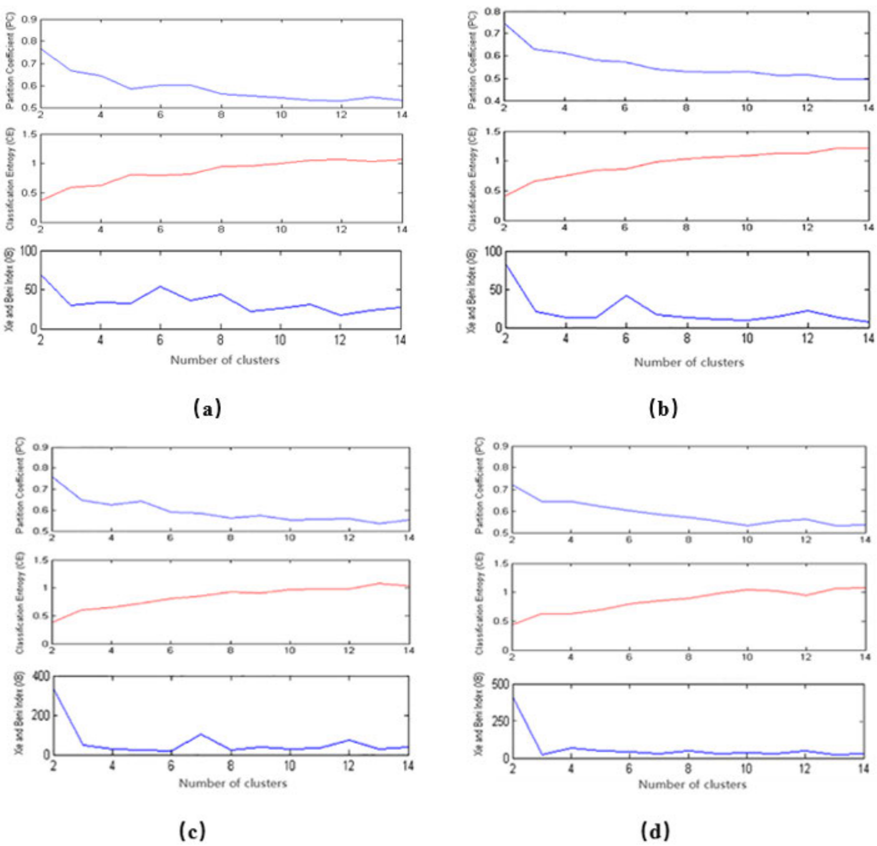


FIGURE 6. Data partitioning coefficient, classification entropy and index value in each time period. (a) is the time 7:00-10:00; (b) is the time 11:00-14:00; (c) is the time 17:00-20:00; (d) is the time 20:00-23:00.

the (CE) represent the fuzzyness of clusters and the (XB) represent the ratio of the total variation within clusters and the separation of clusters.

The index value is the best classification number when it reaches the minimum. According to Figure 6, the index value reaches the minimum when the clustering number is from 7:00 to 10:00. Subsequently, both the partition coefficient (PC) and the classification entropy (CE) tend to be flat, as a result of which it is most appropriate to cluster the data in the period from 7:00 to 10:00 into 12 categories. Similarly, for the time period 11:00-14:00, when the clustering number, or the index value reaches the minimum, the partitioning coefficient (PC) and classification entropy (CE) tend to flatten subsequently, as a result of which the number of clustering

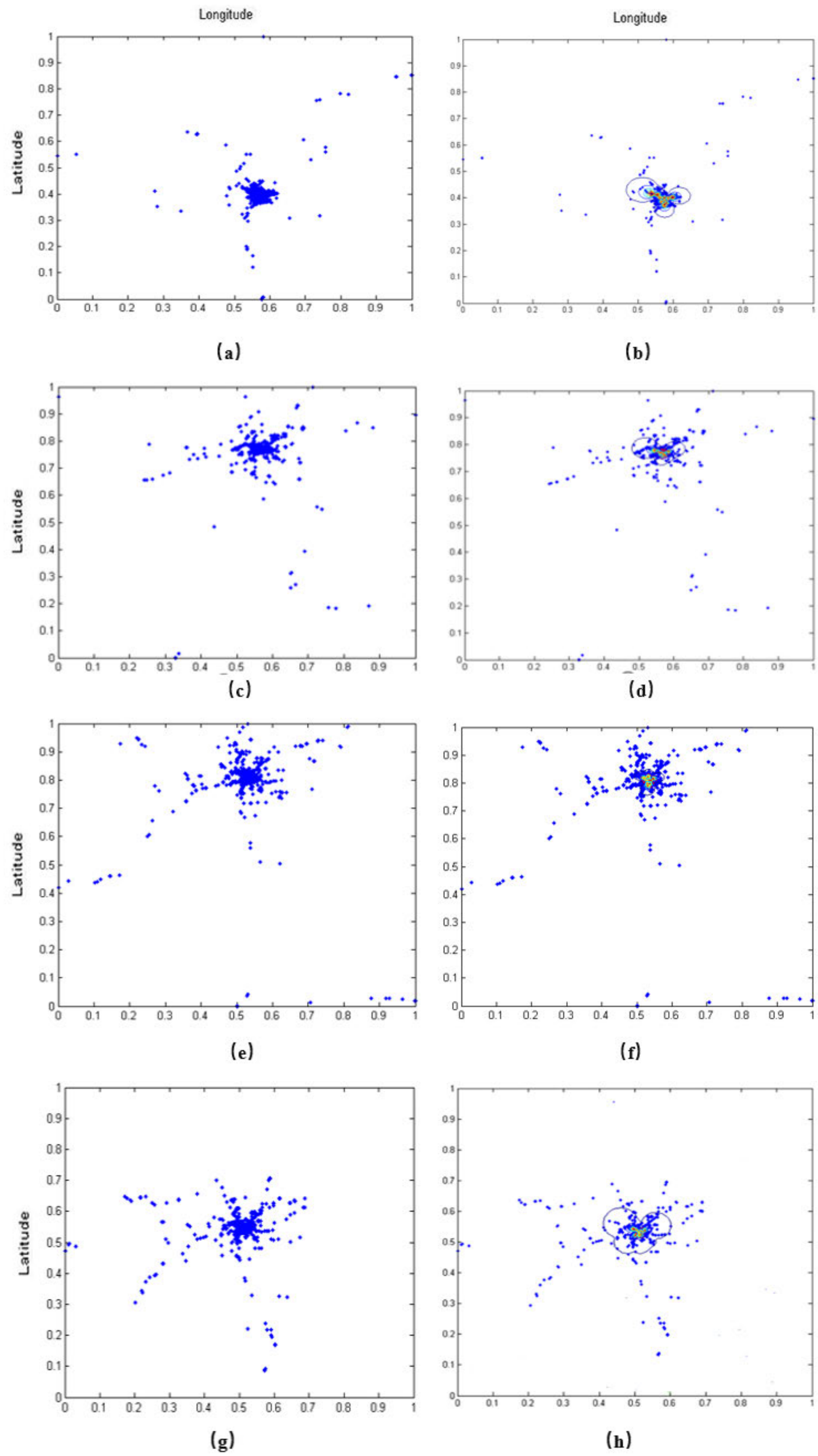


FIGURE 7. Distribution of boarding points and clustering results of passengers in each time period. (a)-(b) is the time 7:00-10:00; (c)-(d) is the 11:00-14:00; (e)-(f) is the time 17:00-20:00; (g)-(h) is the time 20:00-23:00.

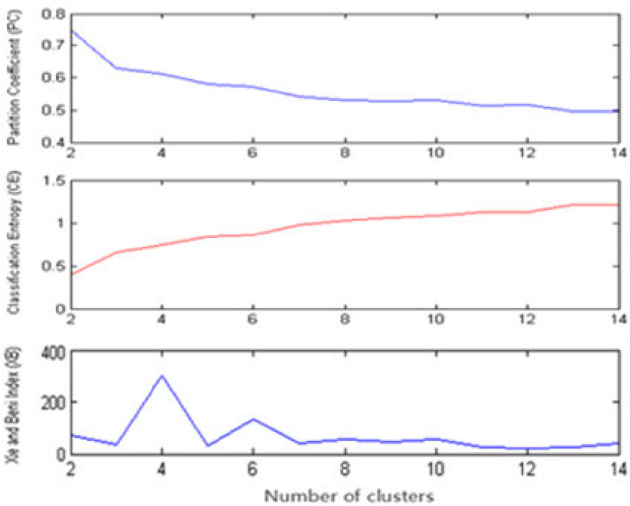


FIGURE 8. Partition coefficient, classification entropy and index.

in the time period 11:00-14:00 is determined as 10. For the time period 17:00-20:00, when the clustering number or, the index value reaches the minimum, but after partition coefficient (PC) and classification entropy (CE) show significant change, while after partition coefficient (PC) and classification entropy (CE) tend to be flat, as a result of which the optimal clustering number for the time period 17:00-20:00 is supposed to be 8. For the time period 20:00-23:00, when the (XB) index value reaches the minimum, but the partition coefficient (PC) and classification entropy (CE) show noticeable change in the future, and only when, the partition coefficient (PC) and classification entropy (CE) tend to be flat, therefore, the optimal clustering number for the time period 20:00-23:00 is supposed to be 13.

After the number of clustering is determined, the improved FCM clustering algorithm was applied to conduct clustering analysis of GPS data in each time period, and the hot spots of urban residents' activities were mined. The clustering results are shown in Figure 7. Because detailed coordinates are involved, the coordinates of Figures 7 and Figures 9 are not real coordinates.

D. CLUSTERING ANALYSIS OF PASSENGER BOARDING POINTS THROUGHOUT THE DAY

The data on boarding and disembarking points of urban residents as collected by 195 taxis were taken out for 6 consecutive months, and cluster analysis was conducted. Firstly, the partition coefficient (PC), classification entropy (CE) and the value of the (XB) index are taken into full consideration so as to obtain the optimal number of clustering. The calculation results of partition coefficient (PC), classification entropy (CE) and the value of the (XB) index are presented as follows.

As demonstrated by Figure 8, when the number of clustering, the index value of the minimum, but in between and partition coefficient (PC), can cause a notable change to the classification entropy (CE), and after the partition coefficient (PC), classification entropy (CE) change is insignificant, leveling off, and when the (XB) index value is small, therefore, in consideration of the partition coefficient of entropy (PC), classification (CE) and (XB) index based on the value of the option, as the most appropriate optimal clustering number, the hot issue within the area of resident activity is split into 13. After the number of clustering is determined, the improved FCM clustering algorithm was applied to perform clustering analysis of the boarding and boarding points of 195 taxis in Kunming for 6 consecutive months, and the hot spots for urban residents' activities were identified. The clustering results are shown in Figure 9.

In the clustering process of passengers getting on and off the point, the formula (6) is applied to calculate the cluster center of each cluster, and then the cluster center is taken as the center to conduct search for the number of passengers

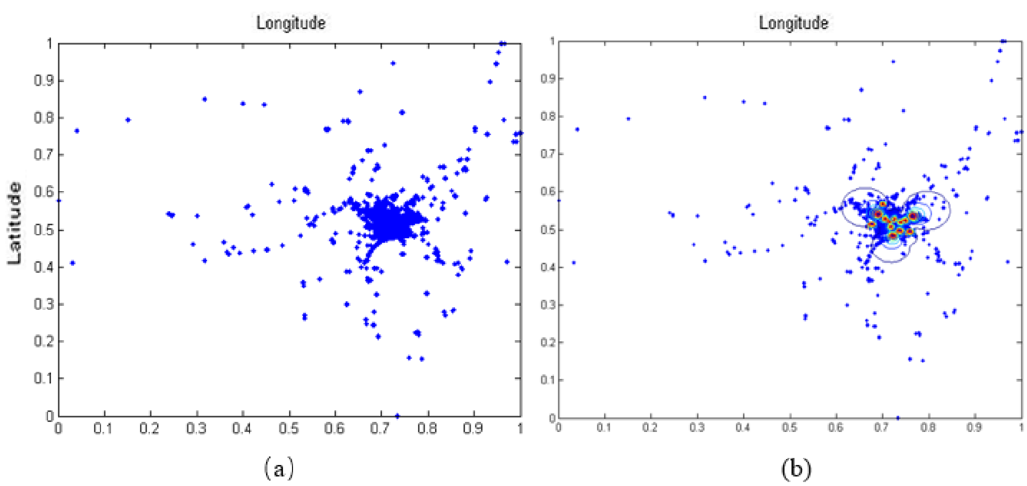


FIGURE 9. Distribution and clustering results of boarding and boarding points of passengers and residents in Kunming. (a) is distribution of boarding and disembarking points of Kunming passengers (b) is clustering results of boarding and disembarking points of Kunming passengers.

FIGURE 10. Distribution map of hotspots in residents' activities in Kunming. (a) is the result of the method in this paper, (b) is the result obtained in the literature [27].

TABLE 4. Distribution statistics of hot spots of residential activities in Kunming throughout the day.

Period of time	Number of hot spots	Distribution of hot spots
All day	13	Hot spot area A: Wanhua Road (near Jinkang Park)
		Hot spot area B: Kunshi Road (near Xiaoximen)
		Hot spot area C: Renmin Middle Road (Daguan Commercial City)
		Hot spot D: Youth Road (near Bailian Square)
		Hot spot E: Youth Road (near Golden Eagle Shopping Center)
		Hot spot area F: Huancheng North Road (near Camellia Hotel)
		Hot spot area G: Bailong Road (near the Golden Garden)
		Hotspot area H: Near Dashuying Overpass
		Hot spot area I: Jinbi Road (near Dongsi Street)
		Hot spot area J: near the boat house
		Hot spot K: Kunming Jinjiang Hotel (near Kunming Railway Station)
		Hot spot area L: Guomao Road (International Exhibition Center)
		Hot spot area L: Guomao Road (International Exhibition Center)

entering and departing from the surrounding 0.5 km, and then it is used. The passengers getting on and off points are combined with the buffer analysis tool in Arcgis software to create a grading buffer. In order to verify the reliability and effectiveness of the method used in this paper, the results of this paper are compared with the research results in the literature [27] using mobile phone positioning data, is shown in figure 10. The results show that the two data travel periods are completely consistent, and the active hotspot areas are basically the same. There are only a few deviations. The main reason is that the positioning data types and time are different, so there will be deviations.

As shown in Figure 10, by conducting research and analyzing the statistics of various time periods in Kunming and the hotspots of all-day residents, it is known that the hotspots include the most populated areas, the youth roads (small gardens) and the hotspots around Xiaoximen. There are widespread regional distributions, followed by residential areas, such as the boat house community, Jinxing Community, Huangtupo and Palm Tree Camp, and again adjacent to the school and hospital, such as Yi Er Yi Street, Longquan Road near Yunnan University of Finance and Economics, Kun the hospital is attached to the vicinity of the Second Hospital.

According to Figure 10, the distribution map of residential activity hotspots in Kunming was extracted with the assistance of the improved FCM algorithm, and the distribution of hot spots in Kunming throughout the day was statistically analyzed. The statistical results are indicated in Table4.

V. CONCLUSION

As the society develops on a continued basis, the activities of urban residents are increasingly frequent, and the mode of travel has also undergone changes. An increasing number of residents prefer taxis as a means of transport, which increase pressure on urban traffic networks, as a result of which the study of urban residents' activities is deemed necessary. It is of great significance for the relevant government agencies to carry out traffic planning in a more effective way to reduce the strain put on traffic. As science and technology advance, various positioning data are increasingly diversified, thus creating favorable conditions for the study on the patterns of residents' activities, such as mobile phone positioning data, vehicle GPS positioning data, bus IC card data, POI check-in data. In this paper, a study is performed on the patterns of urban residents' activities in Kunming by analyzing GPS positioning data collected from vehicles.

Research indicates that the travel length of most residents is within 2-10 km, and the travel formation time is within 10-30 minutes. Then, the day was split into 24 hours. The time period 7:00-10:00 is the morning peak, and the time period 17:00-20:00 is the evening peak, this result is completely consistent with the results of the literature [27].

Cluster analysis is carried out for each time period and the whole day using the model proposed in this paper. The results show that the number of hot spots in each time period and the whole day is 12, 10, 8, 13, 13, respectively. This result is basically consistent with reference [27], and there are differences in a few hot spots. The reason is that the two results use different types and times of data. This article uses taxi positioning data in 2010, and mobile phone positioning data in 2011 used in reference [27]. Traffic distribution data. In addition, the number of hotspot areas clustered by the two is also different. As a result, minor differences have emerged. But it is consistent with the actual situation, which indicates that the improved clustering algorithm can accurately determine the clustering number of clusters, and optimize the clustering results of the whole research area, so as to effectively mine the hotspots of urban residents' activities, for urban planning and Relevant departments to make decision-making services.

There are also some shortcomings in this paper. For example, there is no comparative analysis experiment, and the algorithm has not been verified by specific quantitative indicators. I hope that I can continue to improve and in-depth study in future research.

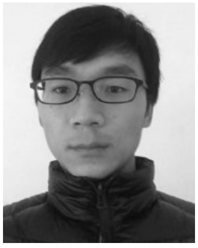
ACKNOWLEDGMENT

The authors would like to sincerely thank all the authors and reviewers for the tremendous efforts towards the success of this Special Section. They would also like to thank to the Editor and Reviewer, for their help. They provided many valuable comments to the paper.

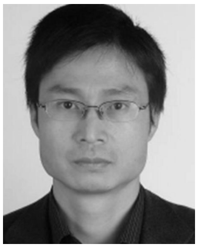
REFERENCES

[1] H. Yu and S. Shaw, "Exploring potential human activities in physical and virtual spaces: A spatio-temporal GIS approach," *Int. J. Geograph. Inf. Sci.*, vol. 22, no. 4, pp. 409–430, Apr. 2008.
[2] *Kunming City Economic and Social Development Statistical Bulletin 2015*. Accessed: Mar. 24, 2019. [Online]. Available: http://mt.sohu.com/20160524/n451124386.shtml
[3] S. Sirisup, S. U-Ruekolan, "Conversion of volunteer-collected GPS diary data into travel time performance measures: Literature review, data requirements, and data acquisition efforts," *Palliative Med.*, vol. 22, nos. 3–4, pp. 445–470, 2005.
[4] Y. Yang, Z. Yuan, X. Fu, Y. Wang, and D. Sun, "Optimization model of taxi fleet size based on GPS tracking data," *Sustainability*, vol. 11, no. 3, p. 731, Jan. 2019.
[5] E. Graves, S. Zheng, L. Tarte, B. Levine, and A. Reddy, "Customer journey time metrics for New York city bus service using big data," *Transp. Res. Rec.*, Jan. 2019, Art. no. 0361198118821632.
[6] Y. Yang, Z. Z. Yuan, J. Y. Li, Y. H. Wang, and W. C. Wang, "Multi-mode public transit OD prediction and scheduling model," *Adv. Transp. Stud.*, vol. 3, pp. 133–146, Dec. 2018.
[7] Y. Yang, Z. Yuan, J. Chen, and M. Guo, "Assessment of osculating value method based on entropy weight to transportation energy conservation and emission reduction," *Environ. Eng. Manag. J.*, vol. 16, no. 10, pp. 2413–2423, 2017.
[8] Z. Cai, L. Deng, D. Li, X. Yao, D. Cox, and H. Wang, "A FCM cluster: Cloud networking model for intelligent transportation in the city of Macau," *Cluster Comput*, vol. 22, no. S1, pp. 1219–1228, Jan. 2019.
[9] Y. Feng, H. Lu, W. Xie, H. Yin, and J. Bai, "An improved fuzzy C-means clustering algorithm based on multi-chain quantum bee colony optimization," *Wireless Pers. Commun.*, vol. 102, no. 2, pp. 1421–1441, Sep. 2018.
[10] J. Zheng, D. Zhang, K. Huang, and Y. Sun, "Adaptive image segmentation method based on the fuzzy c-means with spatial information," *IET Image Process.*, vol. 12, no. 5, pp. 785–792, May 2018.
[11] A. Jamshidi, D. Ait-Kadi, A. Ruiz, and M. L. Rebaiaia, "Dynamic risk assessment of complex systems using FCM," *Int. J. Prod. Res.*, vol. 56, no. 3, pp. 1070–1088, Feb. 2018.
[12] C. Li, M. Cerrada, D. Cabrera, R. V. Sanchez, F. Pacheco, G. Ulutagay, and J. V. D. Oliveira, "Some preliminary results on the comparison of FCM, GK, FCMFP and FN–DBSCAN for bearing fault diagnosis," in *Proc. Int. Conf. Sens., Diagnostics, Prognostics, Control (SDPC)*, Aug. 2017, pp. 41–46.
[13] X. Mansheng, X. Zhe, W. Zhicheng, and Z. Liqian, "Improved FCM clustering algorithm based on spatial correlation and membership smoothing," *Dianzi Yu Xinxi Xuebao/J. Electron. Inf. Technol.*, vol. 39, no. 5, pp. 1123–1129, 2017.
[14] Z. Tianyuan, "Improved fuzzy clustering algorithm based on intelligent computing," in *Proc. Int. Conf. Robots Intell. Syst. (ICRIS)*, Huai'an, China, Oct. 2017, pp. 161–164.
[15] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, Jan. 1984.
[16] H. Sun, S. Wang, and Q. Jiang, "FCM–based model selection algorithms for determining the number of clusters," *Pattern Recognit.*, vol. 37, no. 10, pp. 2027–2037, Oct. 2004.
[17] C. E. White, D. Bernstein, and A. L. Kornhauser, "Some map matching algorithms for personal navigation assistants," *Transp. Res. C, Emerg. Technol.*, vol. 8, nos. 1–6, pp. 91–108, Feb. 2000.
[18] K. Cao and J. J. Tang, "Intelligent map-matching algorithm using Frechet distance measure based," *Comput. Eng. Appl.*, vol. 24, no. 5, pp. 79–83, 2007.
[19] L. Tang, X. Yang, Z. Kan, and Q. Li, "Lane–level road information mining from vehicle GPS trajectories based on Naïve Bayesian classification," *ISPRS Int. J. Geo-Inf.*, vol. 4, no. 4, pp. 2660–2680, Nov. 2015.
[20] M. A. Quddus, W. Y. Ochieng, and R. B. Noland, "Current map-matching algorithms for transport applications: State-of-the art and future research directions," *Transp. Res. C, Emerg. Technol.*, vol. 15, no. 5, pp. 312–328, Oct. 2007.
[21] J. Yu and M.-S. Yang, "Optimality test for generalized FCM and its application to parameter selection," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 1, pp. 164–176, Feb. 2005.
[22] F. Moghimi, A. Nasri, and R. Schober, "Adaptive L_p—Norm spectrum sensing for cognitive radio networks," *IEEE Trans. Commun.*, vol. 59, no. 7, pp. 1934–1945, Jul. 2011.

[23] V. Choulakian, "Matrix Factorizations based on induced norms," *Statist., Optim. Inf. Comput.*, vol. 4, no. 1, pp. 1–14, 2016.
[24] V. Chellaboina and W. M. Haddad, "Is the Frobenius matrix norm induced?" *IEEE Trans. Autom. Control*, vol. 40, no. 12, pp. 2137–2139, Dec. 1995.
[25] Z. Lin, M. Chen, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," Sep. 2010, *arXiv:1009.5055*. [Online]. Available: https://arxiv.org/abs/1009.5055
[26] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE Conf. Decis. Control Including 17th Symp. Adapt. Processes*, Jan. 1978, pp. 761–766.
[27] X. Zuo and Y. Zhang, "Detection and analysis of urban area hotspots based on cell phone traffic," *J. Comput.*, vol. 7, no. 7, pp. 1753–1760, 2012.



YONGFA LI received the M.Sc. degree in GIS from the Kunming University of Science and Technology, in 2017, where he is currently pursuing the Ph.D. degree with a focus on data mining, intelligent traffic processing, and InSAR theory and technology.



XIAOQING ZUO received the M.Sc. degree in GIS from the Kunming University of Science and Technology, in 2001, and the Ph.D. degree in GIS from Wuhan University, in 2004. He is currently a Professor with the Kunming University of Science and Technology. His main research interests are data mining, spatio-temporal big data mining and analysis, and remote sensing image processing and analysis. He was named Yunnan Young and Middle-Aged Academic and Technical Leaders Reserve Talent. He is also the Deputy Director of the GIS Professional Committee of the Yunnan Provincial Surveying and Mapping Society.

FANG YANG received the M.Sc. degree in GIS from the Kunming University of Science and Technology, in 2016. She is currently with the Kunming Institute of Surveying and Mapping. Her research focuses on geographic information systems and data processing.