Hindawi
Journal of Advanced Transportation
Volume 2021, Article ID 5814879, 25 pages
https://doi.org/10.1155/2021/5814879

Research Article

Mining Taxi Pick-Up Hotspots Based on Grid Information Entropy Clustering Algorithm

Shuoben Bi, Ruizhuang Xu, Aili Liu, Luye Wang, and Lei Wan

School of Geographical Sciences, Nanjing University of Information Science and Technology, Nanjing 210044, China

Correspondence should be addressed to Shuoben Bi; bishuoben@163.com

Received 28 June 2021; Revised 1 November 2021; Accepted 23 November 2021; Published 27 December 2021

Academic Editor: David Rey

In view of the fact that the density-based clustering algorithm is sensitive to the input data, which results in the limitation of computing space and poor timeliness, a new method is proposed based on grid information entropy clustering algorithm for mining hotspots of taxi passengers. This paper selects representative geographical areas of Nanjing and Beijing as the research areas and uses information entropy and aggregation degree to analyze the distribution of passenger-carrying points. This algorithm uses a grid instead of original trajectory data to calculate and excavate taxi passenger hotspots. Through the comparison and analysis of the data of taxi loading points in Nanjing and Beijing, it is found that the experimental results are consistent with the actual urban passenger hotspots, which verifies the effectiveness of the algorithm. It overcomes the shortcomings of a density-based clustering algorithm that is limited by computing space and poor timeliness, reduces the size of data needed to be processed, and has greater flexibility to process and analyze massive data. The research results can provide an important scientific basis for urban traffic guidance and urban management.

1. Introduction

With the development of GPS positioning, wireless communication, and other related technologies in recent years, mobile terminals equipped with GPS can be used for a wide variety of applications. These mobile positioning devices can record both the spatial locations of people's daily trips and provide the corresponding time-axis information, thus making it possible to analyze the spatiotemporal trajectory data obtained through positioning in three-dimensional (3D) space. The trajectory data contain the spatial position (such as longitude and latitude) of the moving object, the current time, the instantaneous speed, the passenger-carrying state, and other important information. Trajectory data can be analyzed through data analysis and data mining, thereby determining the movement patterns of the research objects [1, 2], mining the hotspots where residents go to and the corresponding time and space characteristics [3, 4], and applying the obtained valuable knowledge to real-life scenarios. The continuous maturation of spatiotemporal data mining technology has led to the development of the technical support required to study the massive amounts of moving trajectory data [5, 6].

As a major component of public travel, taxis are often the focus of urban planning and construction [7–9]. Observations obtained from mining taxis' GPS data can be viewed from two perspectives. (1) From the perspective of taxi drivers, their accumulated driving experience throughout the years has granted them a deep understanding of a city's traffic conditions; when choosing a driving route, they often take into account the current traffic conditions, distance, trip time, and many other factors. Therefore, the driving trajectory information of taxi drivers can reflect their driving patterns and experience. (2) From the perspective of passengers, the large amount of taxi origin-destination (OD) data can reflect the distribution of people's travel demands regarding taxis, which will indirectly reveal the daily travel patterns of urban residents [10].

Taxi trajectory points have the characteristics of a large amount of data and local clustering. Thus, an area of study is divided into multiple regular grids based on the grid division method [11, 12], and the grids are used in place of the

original spatial data objects for analysis. This type of method is independent of the original data object and only relies on the number of grids; that is, it is insensitive to the input data. Therefore, this method can identify noisy data and is computationally fast. At the same time, the concept of information entropy is also introduced to quantify the distribution equilibrium degree of taxi pick-up points [13]. The distribution of taxi pick-up points can be analyzed according to the changes in information entropy and clustering degree. Mining that identifies taxi pick-up hotspots can serve as a reference for city management and planning, help improve the level of road traffic services, and alleviate the current "difficult to take a taxi" situation.

2. Related Research

The traditional density-based spatial clustering of applications with noise (DBSCAN) algorithm [14–16] clusters trajectory points; however, because this algorithm is data-driven and is sensitive to the input data when mining trajectory points, it is limited by the required computing space and has poor speed. Such shortcomings are manifested in terms of (1) different parameter combinations in the DBSCAN algorithm that will greatly affect the clustering results and the parameter values that are generally determined empirically and (2) the clustering quality that will drop if the taxi pick-up points have a nonuniform distribution and the clustering distances are largely different.

The ordering points to identify the clustering structure (OPTICS) algorithm [17, 18] has overcome the disadvantages of using global parameters in DBSCAN clustering analysis. This algorithm does not generate clustering results explicitly but instead extracts necessary clustering information by sorting output clusters. However, owing to its extremely high time complexity, this algorithm has low efficiency in a data-intensive computing environment. To solve the high time complexity problem of the OPTICS algorithm and resolve its inapplicability in data-intensive environments, An [19] proposed the CP-OPTICS algorithm, which is an improved OPTICS algorithm based on a grid and weighted information entropy strategy. By dividing the data set into a certain number of grid cells and introducing weighted information entropy, the minimum density threshold for each grid cell is calculated adaptively in this algorithm. The concept of a dense grid is defined for grid cells that meet a minimum density threshold, and the data points are compressed by replacing the grid data points with the centroid points; thus, the key purpose of this algorithm is to identify different centroid points.

Zhou [20] proposed the grid-based and information entropy-based clustering algorithm for multidensity (GICM). In this algorithm, the density threshold is automatically calculated from the information entropy carried by grids of different densities; then, the core grids of different density regions are separated, and the breadth-first search method and boundary processing technique are applied to perform clustering, thereby identifying different classes in the density data set. The algorithm performs clustering based on multidimensional data and identifies the core grid by calculating the information entropy. Starting from the core grid, all the density-reachable grids are finally classified as one class according to the breadth-first search method. The key to this algorithm is to identify the core grid and to apply the boundary processing technique constructively.

On related research topics, Georg et al. [21] estimated road direction and corresponding road boundaries using a grid-based route clustering method. Shen et al. [22] identified highway accident black spots using grid-based clustering and principal component clustering; An et al. [23] proposed a grid-based congestion detection method and measured the recurrent congestion areas using a custom clustering algorithm. Ma et al. [24] combined information entropy with principal component analysis to trace gridded taxi driving trajectories in space and extracted different patterns using a K-means clustering method. Dong et al. [25] determined the weights of features related to accidents based on the K-nearest neighbor algorithm and an information entropy index and built a retrieval database for road accident cases using a two-step clustering algorithm. Sun et al. [26] classified road segments based on the 24-h emission rates by using the temporal fuzzy C-means (FCM) clustering, while geographical detector and Moran's I were introduced to verify the impact of built environment on line source emissions and the similarity of emissions generated from the nearby road segments. Ke et al. [27] proposed an information entropy method using a histogram of optical flow to improve the accuracy and reliability of road congestion detection. Zhou and Zhou [28] classified tunnel images using an information entropy-based information clustering algorithm and then built a categorization quality evaluation model for urban tunnel traffic based on the images in each cluster. Hu and Thill [29] extracted empty taxi hotspots or hidden states using kernel density estimation (KDE). By utilizing both real-time and historical taxi data, Lu et al. [30] estimated the region-based taxi wait time and applied recurrent neural network (RNN) and deep learning algorithms to build a predictive model for the taxi service system and thus identify the taxi pick-up hotspots in a city.

Activity chain optimization (ACO) is the task of finding a minimum-cost tour that visits exactly one location for each required activity while respecting time window constraints. Esztergár-Kiss and Remeli [31] developed an exact algorithm that efficiently solves the ACO problem in all practical cases. Estrada et al. [32] present the optimization problem of three different on-demand transit systems operated by vehicles of different sizes. The problem is aimed at minimizing the total cost of the system. Pan [33] investigated college students' choice of train trips for homecoming during the Spring Festival travel rush. The estimation results answered the questions that which determinants and to what extent these determinants have an influence on college students' choice of homecoming train trips. Ali Shafabakhsh et al. [34] studied the frequency and severity of traffic accidents by combining geographic information systems with spatial analysis. Drawbacks of the traditional planar point pattern and the benefits of network analysis are offered in the paper. Wiley et al. [35] examined transit service intensity at the census tract level by assembling and analyzing a suitable GIS

database for the study area. The research results indicated that the core areas of municipalities were not necessarily well serviced by public transit. Sun and Ding [36] demonstrated significant positive associations between ride-sourcing demand and built environment factors, such as commercial/residential land use, public transport accessibility, as well as weather conditions. Chen et al. [37] analyzed the spatio-temporal characteristics of multimode travelers by combining the taxi floating car data, the metro smartcard data, and the GPS trajectories of Mobike, one of the most popular shared bicycles in China. Binomial logit models (BNL) were proposed to estimate mode choices for both peak and off-peak periods by incorporating socioeconomic, demographic, urban morphology, land use properties, and various trip-related variables.

In this paper, a grid information entropy clustering algorithm is proposed based on the grid spatial clustering of applications with noise (GSCAN) algorithm [38]. GSCAN is a type of grid-based density clustering algorithm that uses generated grids to replace the original data set, which is an effective data reduction method. In this algorithm, the original trajectory data are mapped to the grid cells through a mapping function, and the grid density is determined by counting the number of trajectory points in each grid; thus, it is a grid-based point density estimation algorithm. The proposed grid information entropy clustering algorithm adds the concept of information entropy to the GSCAN algorithm. This algorithm obtains the clustering degree of different grid cells by calculating the information entropy of the grid, then selects a hotspot grid cell according to a preset grid density threshold λ, and finally expands the hotspot grid cells according to their clustering degree. The critical part of the algorithm is to calculate the clustering degree of each grid cell. Both the CP-OPTICS algorithm and the GICM algorithm must traverse the grid cells to identify different centroid points and calculate the weight of information entropy and must then apply a breadth-first search method and boundary processing technique, which require additional computing time. In comparison, the grid information entropy clustering algorithm proposed in this study runs faster than these algorithms.

3. Data Introduction and Preprocessing

3.1. Study Areas. Nanjing and Beijing are two cities with relatively developed transportation in China. This paper selects these two cities as representative areas for research and verifies the effectiveness of the proposed clustering algorithm based on grid information entropy in mining taxi hotspots through comparative analysis. Among them, the geographic scope of Nanjing's research covers 11 districts in the city, as shown in Figure 1. The distribution of taxi passenger points in Beijing (Figure 2) is visualized through ArcGIS software. It is found that taxi passenger points are mainly distributed in 6 districts including Haidian District, Chaoyang District, Dongcheng District, Xicheng District, Shijingshan District, and Fengtai District. Therefore, these 6 districts are selected. The districts serve as the research geographic scope of Beijing, as shown in Figure 3.

3.2. Data Introduction. The taxi trajectory data used in this study is purchased from a third-party company (https://www.datatang.com). The data is uploaded every 15 to 30 seconds and stored in a table in the SQL (structured query language) server database. Each table contains 7 important fields, including taxi ID, time, longitude, latitude, speed, direction, and passenger-carrying status, which are all recorded in the track data (Table 1). Among them, Nanjing City contains daily trajectory data of more than 8,000 taxis, and Beijing city contains daily trajectory data of more than 7,000 taxis.

3.3. Data Preprocessing. This study mainly analyzes the passenger loading status field in the trajectory data to describe whether the taxi is carrying passengers. The passenger status field has two values: "0" and "1," which represent the no-load status and passenger status of the taxi, respectively. Before mining the taxi pick-up hotspots, the original trajectory data must be preprocessed, including data cleaning, map matching, and taxi pick-up point extraction.

Among them, data cleaning is mainly to eliminate errors in the original taxi GPS data due to equipment failures and human operations, including eliminating taxi trajectories that exceed the geographic scope of the study, eliminating duplicate recording data, and eliminating positioning device failure recording data. Map matching is aimed at the trajectory correction that must be carried out because of machine failure, data collection system coordinate deviation, or other reasons; the obtained GPS trajectory data does not match the corresponding road [39]. At present, the research of map matching algorithms has been relatively mature. This research uses the widely used point-to-line geometric analysis matching method to match the trajectory data [40, 41]. Figure 4 shows the flow chart of the map matching algorithm used in this study.

When the attribute value of the passenger loading status field is 1, the taxi is currently carrying passengers; when the attribute value is 0, the taxi is empty. As shown in Figure 5, eight trajectory points P1, P2, P3, ... P8 form a trajectory segment. The attribute O_s represents the passenger status of the taxi. When $O_s = 0$, the taxi has no passengers, which corresponds to the empty state; when $O_s = 1$, the taxi has passengers, which corresponds to the passenger-carrying state. Figure 5 shows that the passenger loading situation has changed at points P3 and P8. In P3, the operating system is changed from "0" to "1," which means that passengers get on the bus at this time, and P3 is defined as the taxi boarding point; similarly, in P8, the operating system is changed from "1" to "0," indicates that the passengers get off at this time, so P8 is defined as the taxi drop-off point. This research extracts taxi pick-up point data from taxi trajectory data for subsequent taxi pick-up hotspot mining.

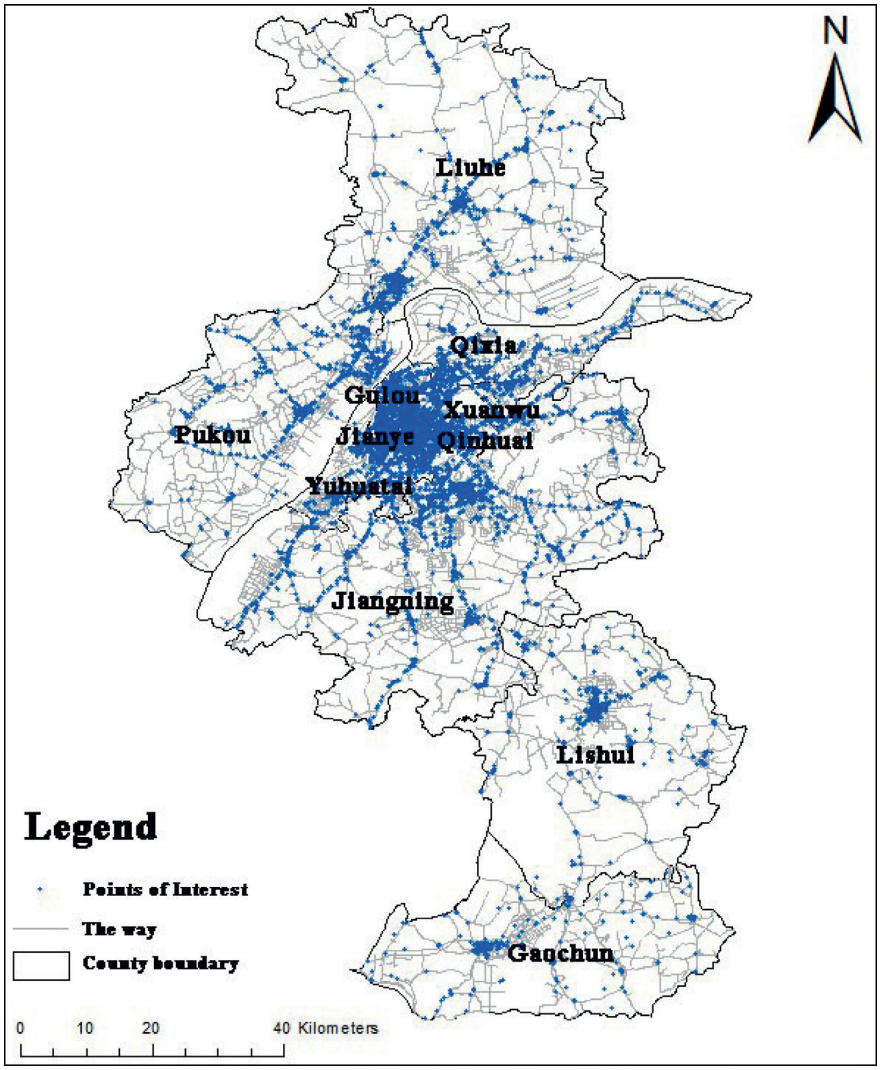


Figure 1: Schematic diagram of the study area (Nanjing).

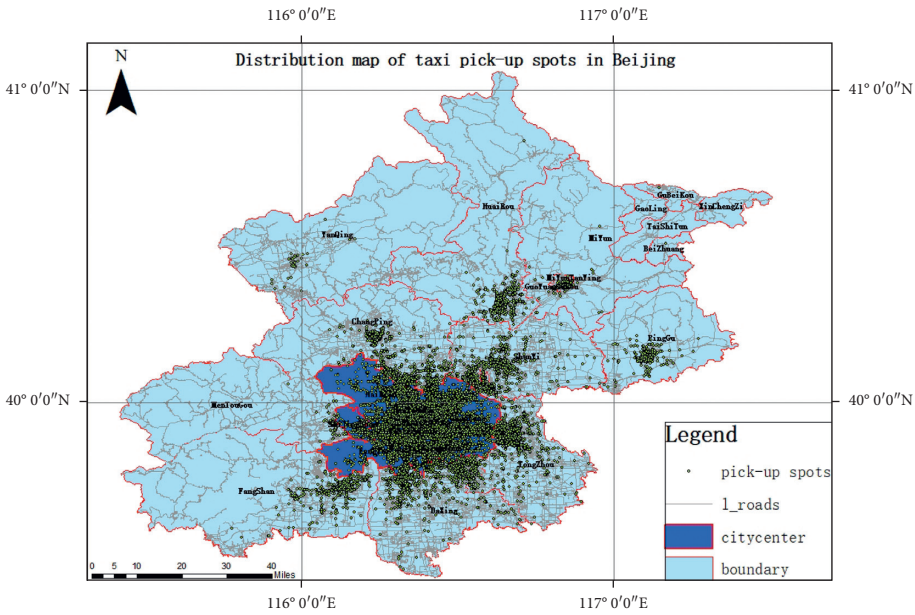


Figure 2: Distribution map of taxi pick-up spots in Beijing.

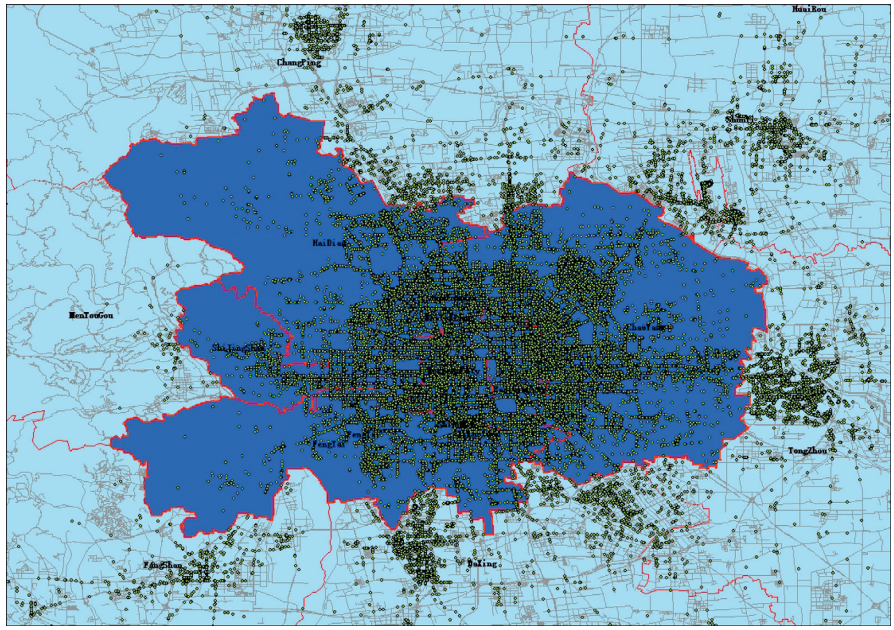


Figure 3: Distribution map of taxi pick-up spots in downtown Beijing.

Table 1: Examples of taxi trajectory data.

ID	Time	Longitude	Latitude	Speed	Direction	Passenger-carrying status
11051847361	07:00:08	118.797247	32.098116	0	250	0
11051847361	07:00:47	118.797212	32.09815	0	0	0
11051847361	07:01:24	118.797218	32.098146	0	160	1
11051847361	07:02:02	118.797186	32.098137	35	160	1
11051847361	07:02:39	118.797245	32.098119	40	160	1

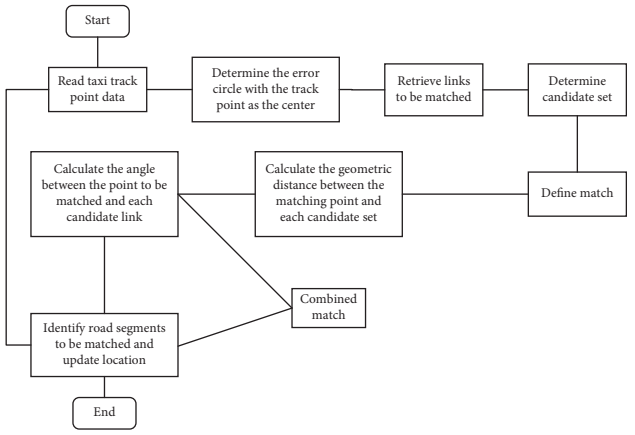


Figure 4: Map matching algorithm flowchart.

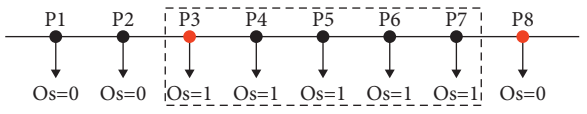


Figure 5: Schematic diagram of a trajectory segment.

4. Algorithm for Mining Taxi Pickup Hotspots

4.1. Related Definitions

Definition 1 (study area). The area containing the spatial data set $S = \{s_1, s_2, \ldots, s_n\}$, where $1 \le i \le n$ is defined as the study area in this paper, which is expressed in latitude and longitude coordinates as follows: $D = [118.357, 119.235] \times [31.232, 32.616]$.

Definition 2 (grid cell). Select the study area D and divide the area into grids based on dividing both the latitude and longitude by a unit length k; in this manner, the study area is divided into $k \times k$ nonoverlapping square grid cells, thereby obtaining $D = m = k \times k$ grids.

Definition 3 (grid mapping). S is a spatial data set, and the spatial coordinates of a point q are defined as [lat, lon]. The relationship between this point q and the grid cells can be represented by the following formula:

$$T(\text{lat}, \text{lon}) = \left(\arg\min_{1 \le i \le m1} \{\text{lat} \le \text{lat}_{\min} + i \times k\}, \cdot \arg\min_{1 \le i \le m2} \{\text{lon} \le \text{lon}_{\min} + j \times k\}\right). \quad (1)$$

The function of the spatial data point q mapped to the grid cell is shown in the following formula:

$$\begin{aligned} \mathrm{lat}_{\mathrm{ind}} &= \left\lfloor \frac{\mathrm{lat} - \mathrm{lat}_{\min}}{k} \right\rfloor, \\ \mathrm{lon}_{\mathrm{ind}} &= \left\lfloor \frac{\mathrm{lon} - \mathrm{lon}_{\min}}{k} \right\rfloor, \end{aligned} \tag{2}$$

where lon and lat represent the current longitude and latitude of q, respectively, and $\mathrm{lon}_{\mathrm{ind}}$ and $\mathrm{lat}_{\mathrm{ind}}$ represent the longitude and latitude of the grid cell to which q belongs, respectively. After grid mapping, all data points that fall in the same grid cell are represented by coordinates $(\mathrm{ind}_{\mathrm{lat}}, \mathrm{ind}_{\mathrm{lon}})$.

Definition 4 (hotspot grid cell). a grid cell G_i that satisfies the following formula is called a hotspot grid cell:

$$\mathrm{des}(G_i) \geq \lambda, \tag{3}$$

where λ is the density threshold of the hotspot grid cell.

Definition 5 (hotspot area). G is a set of grid cells after division, and there exists a nonempty subset K that satisfies the following conditions:

(1) $\forall G_i \in K$, $\forall G_j \in K$, G_i, and G_j have equal clustering degrees

(2) G_i and G_j satisfy $\mathrm{des}(G_i) \geq \lambda$ and $\mathrm{des}(G_j) \geq \lambda$, that is, G_i and G_j are hotspot grid cells

(3) G_i and G_j exhibit connectivity

An area that includes a subset K that satisfies the above conditions is called a hotspot area.

4.2. Algorithm Principles. The taxi trajectory points contain a large amount of data, local clustering, and other characteristics. If using a density-based algorithm for cluster analysis of trajectory points, the clustering results will be limited by the computing space requirements and will require extensive computation time, as this type of algorithm is data-driven and is sensitive to the input data when mining trajectory points. Inspired by the idea of grid division and information entropy, this paper improves the GSCAN algorithm presented in the literature [38] and proposes a clustering algorithm based on grid information entropy. This algorithm is used to mine the taxi pick-up hotspots in a city. The GSCAN clustering algorithm is a grid-based density clustering algorithm that maps the original trajectory data to grid cells through a mapping function and then determines the grid density by counting the number of trajectory points in each grid. On this basis, this study introduces the concept of information entropy and analyzes the distribution of taxi pick-up points by calculating the changes in information entropy and in the clustering degree of each grid.

The proposed grid information entropy clustering algorithm first divides the study area into grids with k as the unit, generating $k \times k$ nonoverlapping square grid cells; it then traverses the extracted data set of taxi pick-up points and maps the pick-up points to their corresponding grid cells through a mapping function; then, it extracts the hotspot grid cells by calculating the information entropy and clustering degree of the pick-up points in each grid; and finally, it traverses the hotspot grid cells and expands the hotspot areas according to the clustering degrees and the distances between corresponding grid cells. This algorithm greatly reduces the data size, improves the calculation speed, has good flexibility, and can be used to analyze massive data.

4.2.1. Grid Division Method. Grid division [11, 12] typically uses space-driven methods, which divide the study area into multiple regular grids to replace the original spatial data objects for analysis. This type of method is independent of the original data object and only relies on the number of grids. Thus, it is not sensitive to the input data. Therefore, this method can identify noisy data and is computationally fast.

4.2.2. Information Entropy of Taxi Pickup Point Distribution. Information entropy reflects an object's equilibrium degree and complexity using entropy [42]. The key of this study is to calculate the information entropy and clustering degree of each grid, which are then used to analyze the distribution of taxi pick-up hotspots in the entire study area.

In the analysis, a set of random pickup point variables $\{\lambda_0, \lambda_1, \ldots, \lambda_k\}$ is selected, and the probability for each pickup point variable to appear is $p(\lambda_i)$. The randomness of the pick-up point distribution can be measured by calculating the information entropy $H(\lambda)$. After grid mapping, each grid cell's information entropy can be calculated as follows [13]:

$$H(\lambda) = \sum_{i=0}^{r} p(g_i) I(g_i) = -\sum_{i=0}^{r} p(g_i) \log_b p(g_i), \tag{4}$$

where b is the logarithm base and usually takes a value of 2, 10, or the natural constant "e."

In general, there are two special cases. (1) All taxi pick-up points are concentrated in the same grid; in this case, the probability of this grid is 1, corresponding to the lowest information entropy of 0, exhibiting the minimum randomness. (2) All taxi pick-up points appear with an equal probability and fall into each grid at the same average value; in this case, the information entropy is the maximum, which is denoted as $H_{\max}$.

To simplify the calculation, the following formula was used to calculate the "normalized" information entropy and construct the equilibrium degree index J:

$$J = \frac{H}{H_{\max}}, \tag{5}$$

where J is the equilibrium degree of the taxi pick-up point distribution, H is the information entropy of the current grid, and $H_{\max}$ is the maximum information entropy. Because, then $0 \leq J \leq 1$.

The calculated information entropy was then "standardized," where a clustering degree index I of the taxi pick-

up points distribution was constructed according to the information entropy of the pick-up points in the grid cell and the maximum information entropy:

$$I = 1 - \frac{H_i}{H_{\max}}, \tag{6}$$

where H_i is the current grid's information entropy and $H_{\max}$ is the maximum information entropy. The clustering degree can be used to measure the cluster distribution degree in the grid.

Information entropy can be used to describe and evaluate the distribution of collection elements in a system, such as the clustering state and the dispersion degree, the order and disorder, as well as the disparity and equilibrium degree of the distribution. The following features apply to information entropy:

(1) The uncertainty of the occurrence probability of a variable can be expressed by information entropy. The more uncertain the occurrence probability of a variable is, the higher the corresponding information entropy will be.

(2) Information entropy can be used to indicate the information amount in an event. The greater the amount of information is in an event, the lower the corresponding information entropy will be, and the easier it is to predict the event.

(3) Information entropy can be used to measure the equilibrium status of the overall distribution. The higher the information entropy is, the more chaotic the distribution will be; the lower the information entropy is, the more balanced the distribution will be.

4.3. Algorithm Steps. To address the shortcomings of density-based clustering algorithms, such as sensitivity to input data and slow computational speed, this paper introduces the concept of information entropy. The changes in information entropy and clustering degree are utilized to analyze the distribution of taxi pick-up points. A clustering algorithm based on grid information entropy is proposed for mining taxi pick-up hotspots. The technical flowchart of this algorithm is shown in Figure 6, and the main algorithm steps are detailed as follows:

(1) The specific steps for mining taxi pick-up hotspots based on the grid information entropy clustering algorithm are as follows:

Step A: traverse the data set of taxi pick-up points within the range of the study area D and perform grid division on the study area.

Step B: map the taxi pick-up points data to the divided grid cells, calculate the information entropy and clustering degree of taxi pick-up points in each grid, extract the hotspot grids according to a preset density threshold λ, and sort the obtained hotspot grids from large to small.

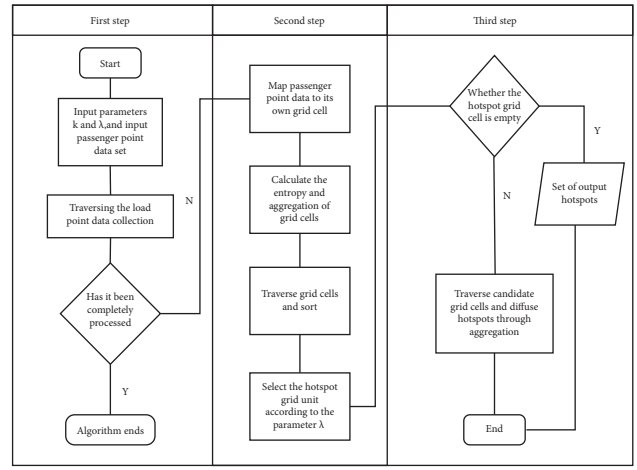


Figure 6: Flow chart of mining taxi pick-up hotspots.

Step C: determine whether the extracted hotspot grid cell set is empty; if not, traverse it and expand the hotspot area according to the clustering degree.

(2) In step A, the original trajectory data is preprocessed, and the taxi pick-up points are extracted, and finally, the taxi pick-up points are traversed to determine the range of the study area. The specific steps are as follows:

Step A-1: preprocess the original trajectory data, including data cleaning and map matching, and then extract the taxi pick-up points from the preprocessed trajectory data.

Step A-2: input the parameters k and λ, traverse the extracted taxi pick-up points data, determine the range of the study area, perform grid division on it, and then check whether all the taxi pick-up points data have been processed; if yes, then end the algorithm.

(3) In step B, the information entropy and clustering degree of the grid cells are calculated, and candidate grid cells are selected according to the parameter λ. The specific steps are as follows:

Step B-1: perform grid mapping, that is, map the extracted taxi pick-up points to the corresponding grid cells using the mapping function.

Step B-2: calculate the information entropy H and clustering degree I of each grid cell and sort from highest to lowest according to the grid clustering degree.

Step B-3: select hotspot grid cells according to the parameter λ; grids with values less than λ are not considered further.

(4) In step C, the candidate hotspot areas are expanded according to the clustering degree of the grid cell. The specific steps are as follows:

Step C-1: first, determine whether the hotspot grid cell is empty; if yes, then output the hotspot area set and end the algorithm.

Step C-2: traverse the hotspot grid cell set, divide the clustering degree into 5 classes, select a hotspot grid cell, and include its surrounding hotspot grid cells, having the same class and exhibiting connectivity, as one hotspot area; then, traverse the hotspot grid cell set and perform the same operation until all the hotspot grid cells are divided; finally, single and independent hotspot grid cells are grouped and divided into blocks of hotspots.

4.4. Algorithm Comparison Experiment and Analysis. To verify the effectiveness of the proposed grid information entropy clustering algorithm, the algorithm was compared with the DBSCAN and the TR-OPTICS algorithms [43]. Figure 7 shows a comparison of the running efficiencies of the three clustering algorithms, namely DBSCAN, TR-OPTICS, and the proposed grid information entropy clustering algorithm, under different data volumes. The figure indicates that when the amount of data is below 40,000, the running times of the DBSCAN and TR-OPTICS algorithms are slightly shorter than that of the grid information entropy clustering algorithm. This occurs because the grid information entropy clustering algorithm needs to perform data mapping during the calculation process, which will consume some time; thus, when the amount of data is low, its calculation time will be longer than the other two algorithms. However, as the amount of data continues to increase, the grid information entropy clustering algorithm requires less calculation time than the other two algorithms. The reason is that grids are used to replace a large amount of spatial data in this algorithm, thereby significantly reducing the data size and improving the calculation efficiency.

4.5. Analysis of Parameter Selection. Before we mine the taxi pick-up hotspots, two important parameters, namely the grid size k and the grid density threshold λ, need to be set in advance in the grid information entropy clustering algorithm. The nuances of different factors often have a great impact on clustering. Through a large number of experiments, Qiu and Zhang [44] proved that $k = \sqrt{N}$ (N is the number of data points in the data set) is an ideal input value for the grid division factor and is suitable for most clustering algorithms. Therefore, in this study, the formula $k = \sqrt{N}$ was used to first determine a reasonable grid size parameter k; on this basis, n numbers of k were selected by expanding the length of k to both the left and right in the area to be evaluated, which were then substituted into the grid information entropy clustering algorithm to calculate the final effect. Grid density threshold λ represents the density of taxi pick-up points that fall within each grid cell. The clustering results of the grid information entropy algorithm were analyzed under different grid sizes k, and n reasonable values for λ were selected. By experimenting and comparing the clustering effects using different values of λ and different corresponding grid size k, the most reasonable λ was selected for given grid size k according to the experiments. The parameter selection in this paper was mainly achieved by choosing the appropriate parameters through multiple experiments (empirical analyses). As shown in Figure 8, different parameter values will affect the algorithm's accuracy. Figure 8(a) indicates that for a given grid size, the smaller λ is, the more hotspots are generated; furthermore, when the grid size k is between 100 and 150, more taxi pick-up hotspots will be generated. In contrast, Figure 8(b) indicates that when λ is small, small areas of hotspots will be formed, but the amount of trajectory points that fall within these hotspot areas are relatively small and not representative of the actual data. Thus, the value of λ has to be set within a reasonable range, λ should not be too small and cause the generated hotspot areas to be nonrepresentative, nor should it be too large and generate too few hotspot areas. Combined with the results shown in Figure 8(a), the changing trends observed when $\lambda = 130$ and $\lambda = 150$ tend to be consistent, and the amplitude of fluctuation is relatively stable, exhibiting a relatively high reference value. Therefore, this paper selected $k = 140$ and $\lambda = 130$ for the experiment.

4.6. Algorithm Characteristics. Compared with existing methods, the solution adopted in this study has the following characteristics:

(1) Traditional density-based clustering algorithms are data-driven and sensitive to input parameters, resulting in large computing space requirements and poor computational speed. This paper proposes an algorithm to address these shortcomings when performing clustering analysis on large-scale trajectory data.

(2) Inspired by the idea of grid division and the information entropy method, this study introduces the concept of information entropy on the basis of the GSCAN algorithm and analyzes the distribution of taxi pick-up points using the changes of information entropy and clustering degree.

(3) The algorithm proposed in this paper uses grids to replace the original trajectory points data for calculation, which overcomes various shortcomings, computing space requirements and extensive computation time, of traditional density-based clustering algorithms, reduces the size of the data to be processed, and improves the calculation speed. The proposed algorithm also has good flexibility. Compared with the CP-OPTICS algorithm and GICM algorithm, it can process and analyze massive data more quickly.

(4) The efficiency of the proposed grid information entropy clustering algorithm is evaluated and analyzed. By comparing it with the DBSCAN algorithm and the TR-OPTICS algorithm, it is found that when the data volume is large, its computational efficiency is much higher than the DBSCAN and TR-OPTICS algorithms. The time complexity of the proposed grid information entropy clustering algorithm is calculated to be $O(n + m^2)$, where n is the number of original taxi pick-up points and m is the number of grids after grid division; this time complexity is

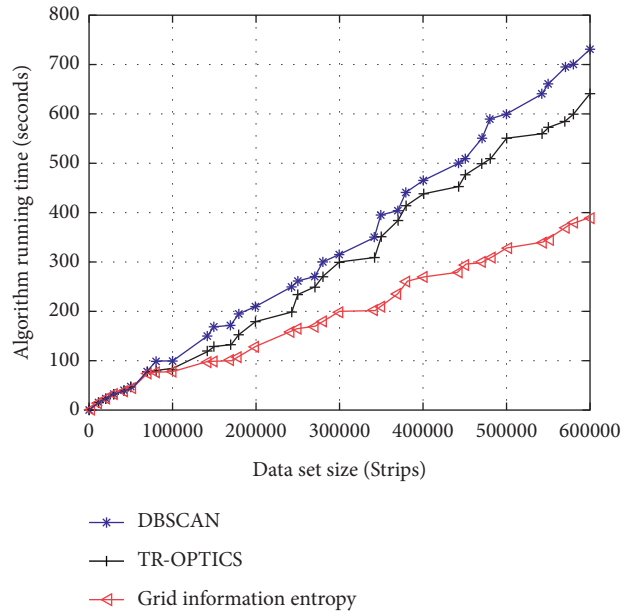


Figure 7: Performance comparison of three algorithms.

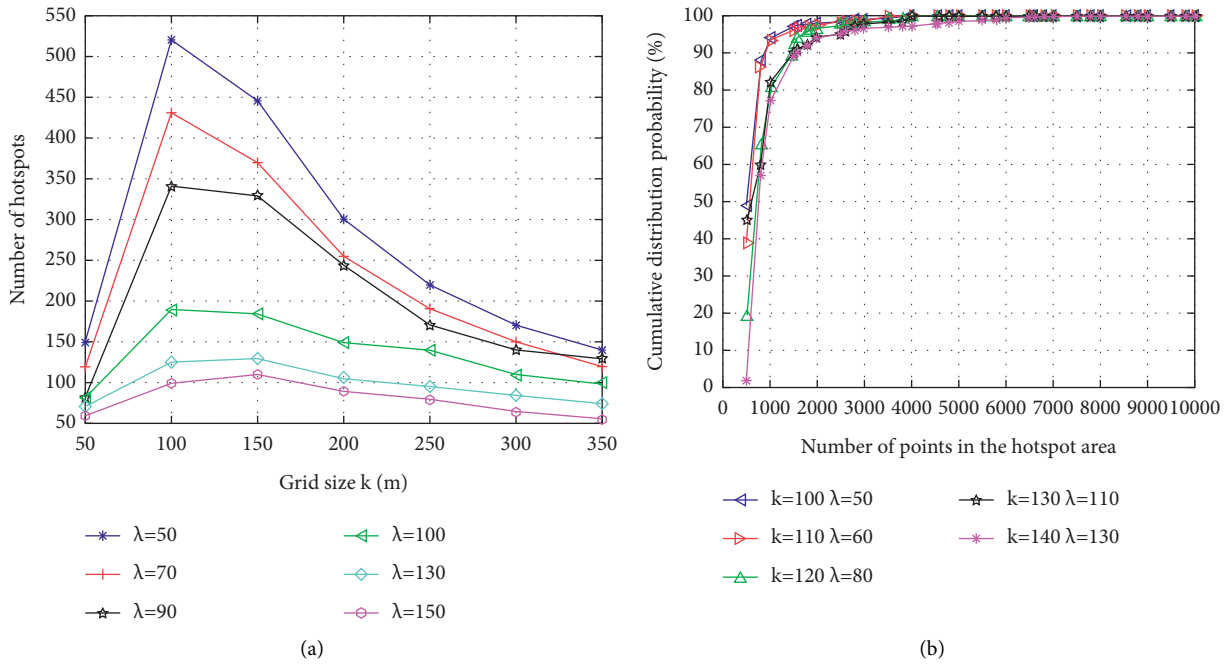
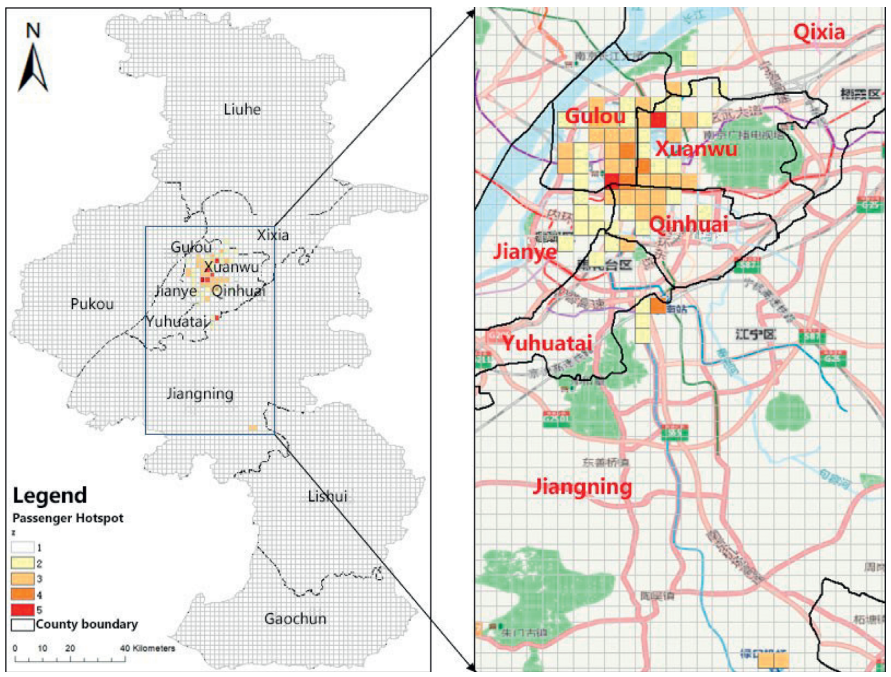


Figure 8: Parameter analysis.

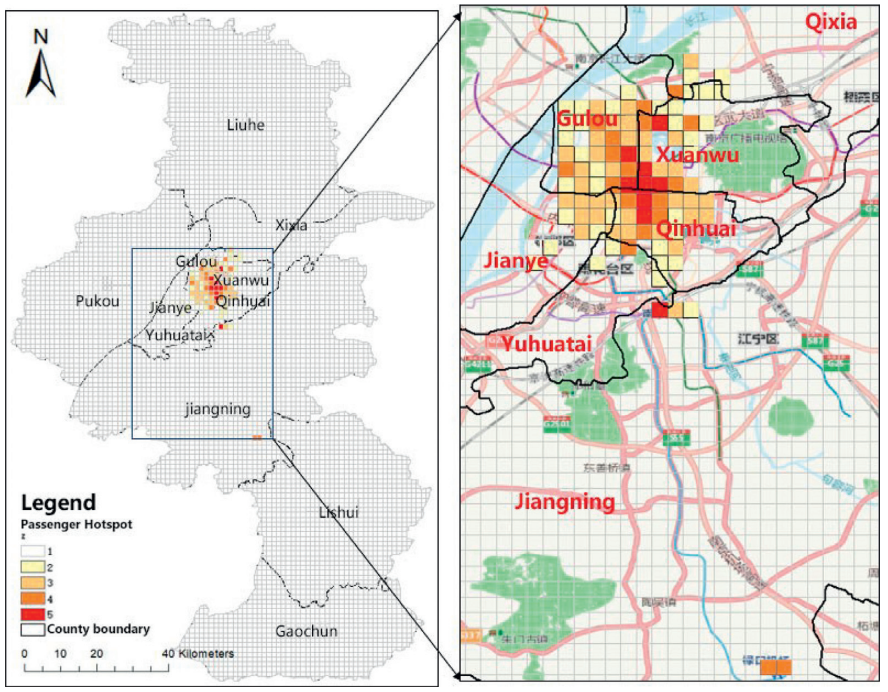
much lower than the $O(n^2)$ complexity of the DBSCAN algorithm under large data volume conditions.

5. Taxi Pickup Hotspots Mining Results and Interpretation in Nanjing

This study focuses on and analyzes taxi pick-up hotspots during four periods. The first period is from 8:00 to 10:00, which is the morning rush hour period, when the demand for rides is high; it is defined as the T1 period. The second period is from 12:00 to 14:00, which is the midday period, and trips are relatively evenly distributed; it is defined as the T2 period. The third period is from 18:00 to 20:00, which is the evening rush hour period, the complement of the morning rush hour period; it is defined as the T3 period. The fourth period is from 22:00 to 24:00, which is the period when people participate in night activities after work; it is defined as the T4 period. Figure 9 shows the clustering results. The figure on the left depicts the hotspot analysis result of all of Nanjing city, and the figure on the right is an enlarged view of the hotspot areas. It can be clearly seen from



(a)



(b)

Figure 9: Continued.

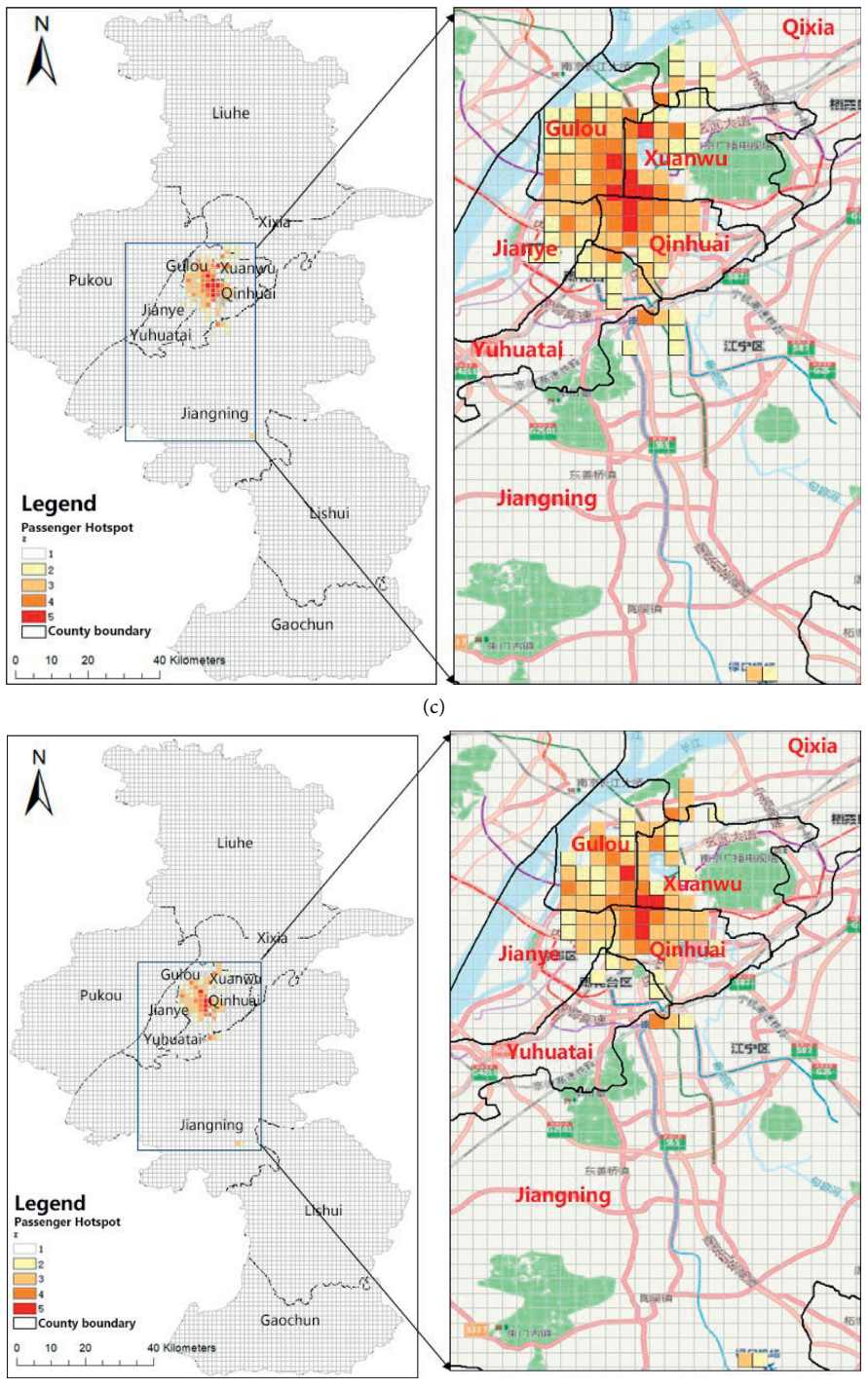


Figure 9: Taxi pick-up hotspot areas in Nanjing.

the figures that the taxi pick-up hotspots are mostly located in the Gulou, Xuanwu, Yuhuatai, Qixia, Qinhuai, and Jianye districts. Compared with the other three periods, the hotspots in the T3 period are of greater number and cover a wider range. The T1 and T2 periods are daytime hours, during which the residents are working, and their travel locations are often in regions consisting of concentrated residential areas and office sites. The T3 period covers the off-hours, during which residents have more freedom of movement, and the hotspot distribution is more dispersed. The T4 period is close to the early morning, and because many residents must go to work the next day, the number of trips begins to decrease, and the range of hotspot areas shrinks correspondingly. At the same time, it was found that the clustering degree of hotspot areas around train stations, high-speed rail stations, and airports remained at a high level

during all four periods, as these areas host a relatively large flow of people in Nanjing; furthermore, trains and planes arrive and depart 24 hours a day, so the demand for taxis to and from these sites is generally high.

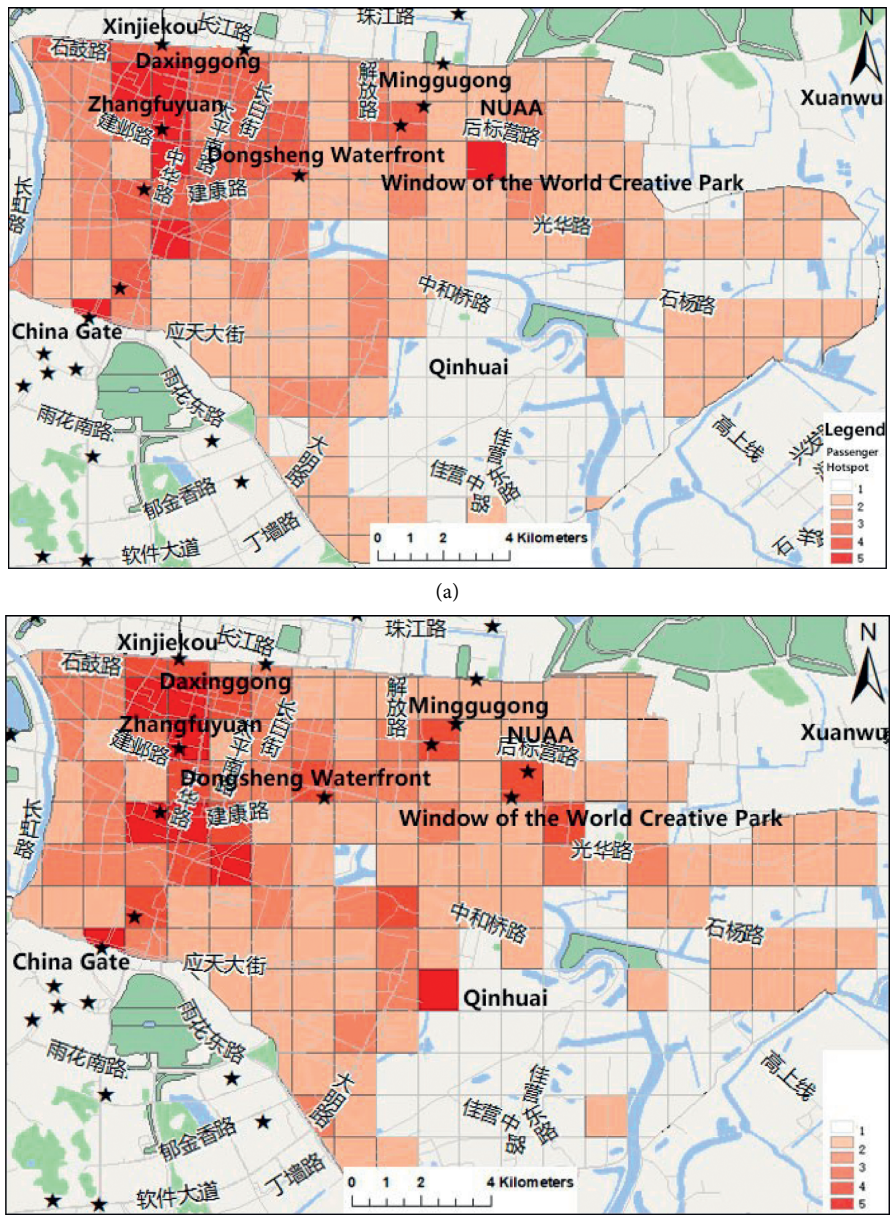
Further analysis of Figure 9 indicates that in the T1 period, which is the peak morning period regarding residents' trips during the day, the hotspot areas with the highest clustering degree are located in areas with large flows of people, such as train stations, high-speed rail stations, and airports, as well as around Mochou Lake and Hunan Road. These regions have dense residential areas, and the demand for taxi travel is high, so the clustering degree of taxi pick-up points is also high. The next highly clustered hotspot areas for taxi travel are mainly located around Xinjiekou, Daxinggong, Confucius Temple, Gulou, and Longjiang. These areas include large-scale commercial centers and office buildings, which are areas where people go to work in the morning, so the demand for taxi travel is high, as shown in Figure 9(a). In the T2 period, which is the lunch break, the hotspot areas are mainly concentrated in commercial and office centers such as Xinjiekou, Daxinggong, Confucius Temple, and Hunan Road. The clustering degree of these hotspot areas increases significantly compared to the T1 period. At the same time, cultural scenic areas such as Yuhuatai and Zijinshan also become hotspots for trips, as shown in Figure 9(b). In the T3 period, which is the evening peak for getting off work, the hotspot areas have further expanded, and the clustering degree of each hotspot has increased. The hotspot areas are mainly concentrated in places where commercial trade is concentrated and residential areas are widely distributed. At the same time, Zhongyang Shopping Mall, Grand Ocean Department Store, and other commercial centers at Xinjiekou, the New Century Plaza at Daxinggong and Confucius Temple, as well as places for leisure, entertainment, catering, and shopping around Zhujiang Road and Gulou, become the main areas for people's activities, as shown in Figure 9(c). In the T4 period, as the time approaches midnight, residents' travel activities begin to decline; most of the public transport is out of service; and taxis become the primary method for people to travel. Xinjiekou, Daxinggong, and other leisure and entertainment places become the main hotspots for trips, as shown in Figure 9(d).

Through the mining of the overall taxi pick-up hotspots in Nanjing, it is found that the taxi pick-up hotspots are mostly located in the six districts of Xuanwu, Gulou, Yuhuatai, Qinhuai, Qixia, and Jianye. Next, the grid information entropy clustering algorithm is used to perform cluster mining research on the distributions of taxi pick-up hotspots during the morning and evening rush hour periods in these six districts.

Figure 10 shows the taxi pick-up hotspots during the morning and evening rush hours in Qinhuai District. Qinhuai District, one of the central urban areas of Nanjing, is located in the southeast of Nanjing, with an area of 49.11 km^2 and a population of 1.026 million. It can be seen from the figure that owing to the large population in this area, the taxi pick-up hotspot areas are widely distributed. The most clustered areas for taxi pick-up hotspots are located in the neighborhood of Xinjiekou and Confucius Temple and the surrounding residential areas. Confucius Temple is one of the traditional commercial centers in Nanjing. It has a large number of antique markets and is characterized by catering and entertainment. The visitor flow is large here, and the demand for taxi travel is correspondingly high, so the distribution of taxi pick-up hotspot areas is also concentrated. The second relatively clustered area for taxi pick-up hotspots is the surrounding area of the Nanjing University of Aeronautics and Astronautics. Scenic tourist areas such as the Ming Palace, Yueyahu Park, as well as many residential areas can be found here. The daily visitor flow is large, as is the demand for taxi travel. Comparing the distributions of taxi pick-up hotspots during the morning and evening rush hours, the two distributions are similar, and only some regions have different clustering degrees of taxi pick-up hotspots. The distribution during evening rush hours is more extensive and clustered because residents have more freedom of movement after work, and some residents will congregate in areas where shopping, entertainment, and catering are concentrated. Therefore, the hotspot distribution range is more extensive, and clustering degrees are higher in some areas.

Figure 11 shows the distribution of taxi pick-up hotspots in the Gulou District. Gulou District is located in the northwest of Nanjing, with a total area of 54.18 km^2 and a permanent population of 1.293 million. From Figure 11, the taxi pick-up hotspots in Gulou District are also widely distributed, and the hotspot areas are mainly concentrated in areas extending northwest from Gulou to Xinjiekou. These areas include important commercial areas in Nanjing, such as Hunan Road and Gulou; they are the main leisure, catering, and shopping areas for residents. There are also offices and residential areas such as Huju Building and Sanpailou Unit. Thus, taxi pick-up hotspots are concentrated in these areas. Furthermore, commercial areas such as Longjiang and the Zhongyang Gate Overpass, as well as transportation hubs, have also formed taxi pick-up hotspot areas with relatively high clustering degrees. Comparing the distributions of taxi pick-up hotspots during the morning and evening rush hours, the hotspot areas during morning rush hours are more widely distributed, but the clustering degrees are relatively lower than those during the evening rush hours. This is because Gulou District is an old urban area of Nanjing, which has a large population and many residential areas. Thus, regions with relatively concentrated residential areas will generate some hotspots correspondingly. However, during the evening rush hours, most residents will gather near commercial, office, trade, and catering centers, so the population distribution is relatively concentrated. Therefore, taxi pick-up hotspots are also more concentrated, and clustering degrees are higher.

Figure 12 shows the distribution of taxi pick-up hotspots in the Qixia District. Qixia District is located in the northeast of Nanjing, with an area of 395.44 km^2 and a permanent population of 668,000. It has as many as 40 institutions for scientific research and higher education and is an important district that houses petrochemical, electronics, and building materials industries, as well as concentrated capital,



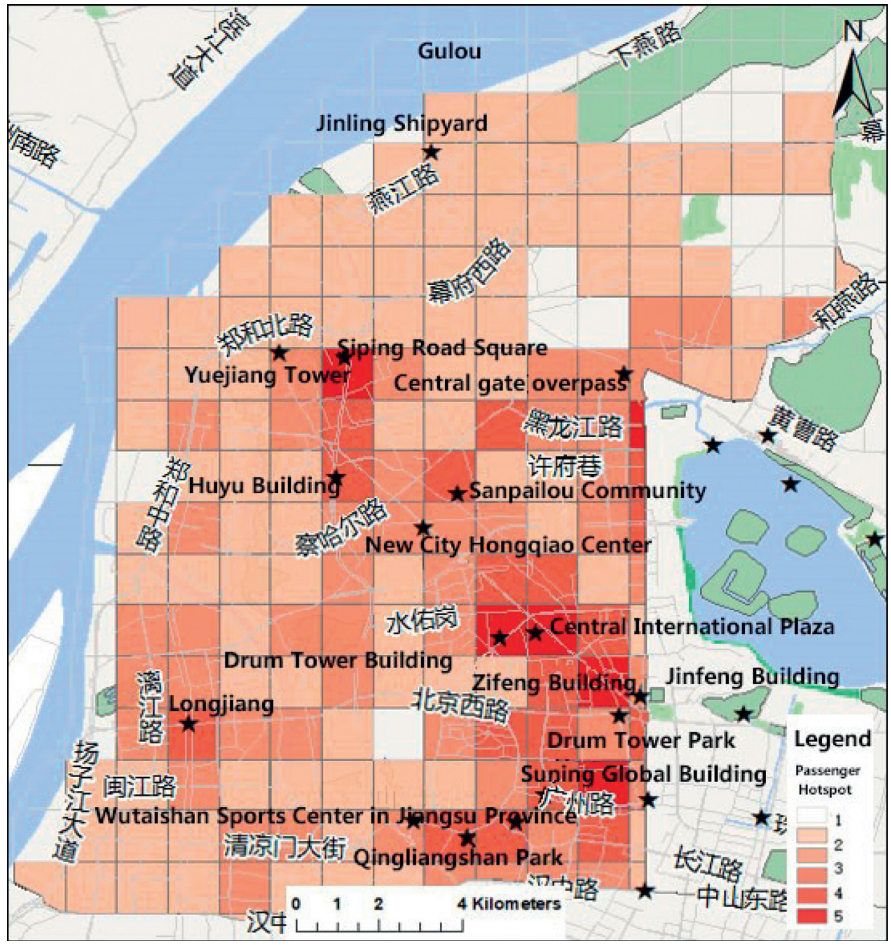
(a)

(b)

Figure 10: Taxi pick-up hotspots in Qinhuai District.

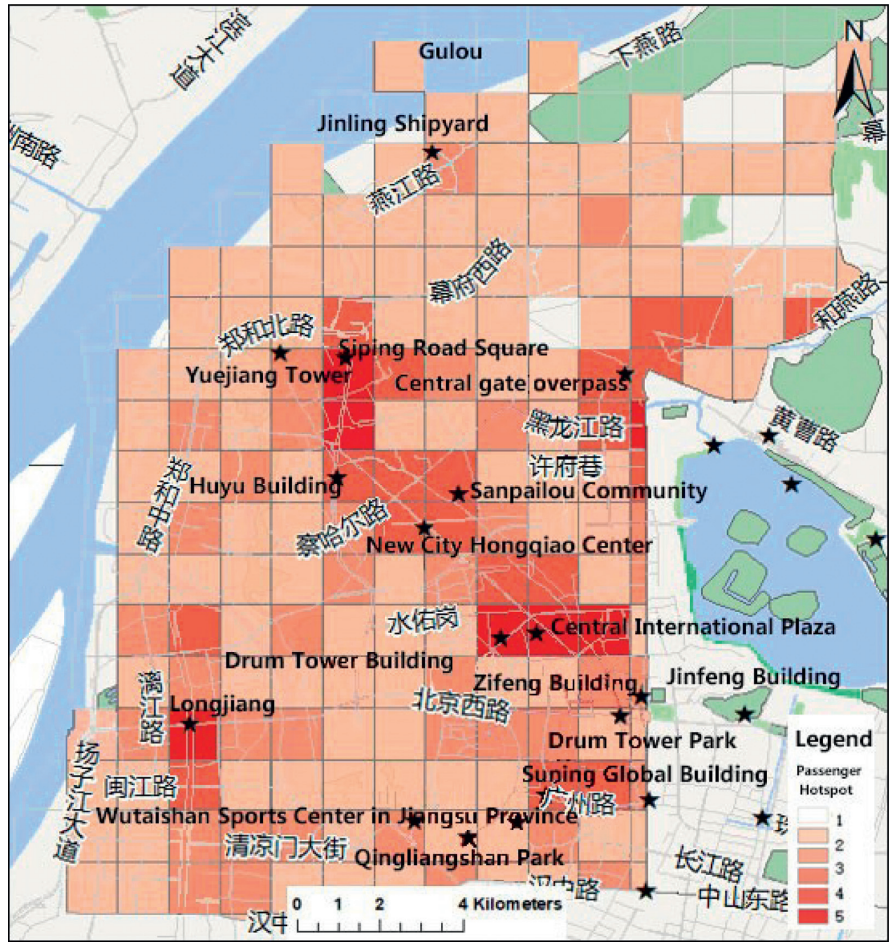
technology, and cultural centers. The figure indicates that the distribution of taxi pick-up hotspots in the Qixia District is relatively scattered. The hotspot areas are mainly distributed around the Zhongshan Scenic Area; for example, the clustering degrees of hotspot areas are relatively high near the Maigao Bridge and Heyan Road. This region is near the Qiaoqiao subway station, has concentrated residential areas around Heyan Road, and possesses a complete infrastructure. It is also close to important transportation hubs in Nanjing, such as Nanjing Station and East Coach Station. Thus, the clustering degrees of taxi pick-up hotspot areas are relatively high in this region. At the same time, there is another relatively highly clustered taxi pick-up hotspot area in the vicinity of Maqun. This region contains Maqun Science and Technology Park; the population is also relatively concentrated; and the demand for taxi travel is high. Comparing the distributions of taxi pick-up hotspots during the morning and evening rush hours, the distribution during the evening rush hours in Qixia District is more widespread than that during the morning rush hours; especially around Maigao Bridge, where the range of taxi pick-up hotspots has expanded further. There are many residential areas around Maigao Bridge, and there are large supermarkets and shopping malls in the neighborhood, which are the main activity areas for residents after work.

Figure 13 shows the distribution of taxi pick-up hotspots in the Xuanwu District. Xuanwu District, one of the central urban areas of Nanjing, is located northeast of Nanjing, with



(a)

Figure 11: Continued.



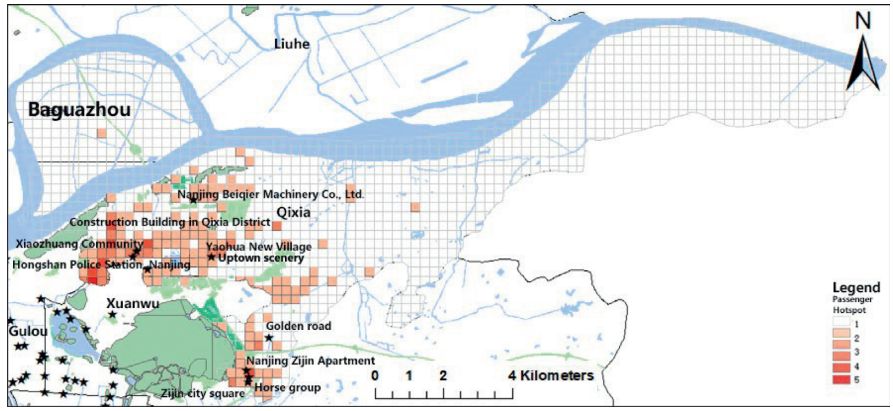
(b)

Figure 11: Taxi pick-up hotspots in Gulou District.

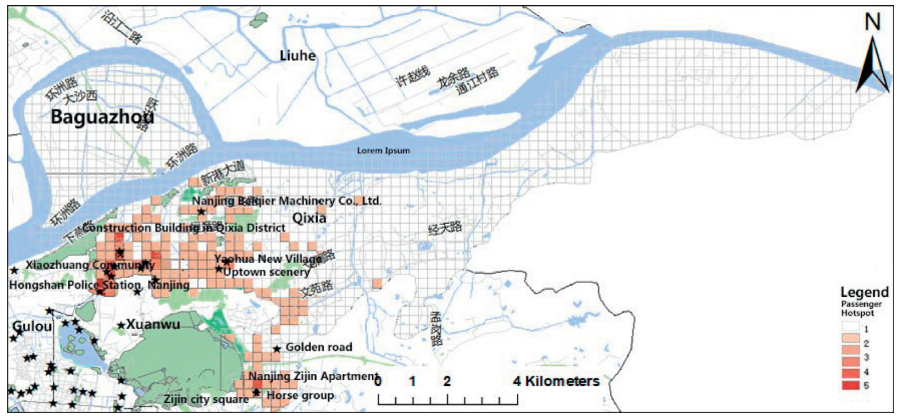
an area of 75.46 km^2 and a permanent population of 602,000. The figure indicates that the taxi pick-up hotspot areas during the morning and evening rush hours in this district can be divided into three parts, namely the areas around Zhujiang Road, Xinjiekou, Daxinggong, and Jiming Temple, areas near Nanjing Station, and areas close to the Nanjing Long-distance Bus East Station. Both Nanjing Station and Nanjing Long-distance Bus East Station are main transportation hubs in Nanjing, having large population flow, and therefore a large demand for taxis. The areas around Zhujiang Road, Xinjiekou, and Daxinggong consist of the central business district in Nanjing, involving numerous and constant economic trading and commercial activities. These areas have complete infrastructures and experience large flows of people, and the clustering degrees of taxi pick-up hotspot areas remain at relatively high levels.

Figure 14 shows the distribution of taxi pick-up hotspots in Jianye District. Jianye District, another central urban area of Nanjing, is located in the southwest of Nanjing, with an area of 83 km^2 and a total population of 600,000. In the figure, the taxi pick-up hotspot areas in Jianye District during the morning and evening rush hours are mainly distributed around Jiqingmen Street, which is mainly residential. However, the Jiangdongmen business district to its west is one of the five major business districts in Nanjing, and large shopping malls such as Hexi Wanda and Leji Plaza are located around Jiqingmen Street. This region is the main leisure, shopping, entertainment, and catering center in Jianye District. Secondly, Hexi CBD, the second-largest central business district in East China after Shanghai Lujiazui, has a relatively high clustering degree of taxi pick-up hotspots in its surrounding area. It was found that the clustering degree of the Olympic Sports Center increased significantly during the evening rush hours. The Nanjing Olympic Sports Center is a multifunctional national-level sports complex, which includes stadiums, gymnasiums, swimming pools, tennis courts, sports science and technology centers, and cultural and sports entrepreneurship centers. It often hosts various sports, science and technology, and cultural events and is the main activity center for residents of Jianye District after work.

Figure 15 shows the distribution of taxi pick-up hotspots during the morning and evening rush hours in Yuhuatai District. Yuhuatai District is located in the south of Nanjing, with an area of 134.6 km^2 and a permanent population of 413,000. It is China's largest research and development

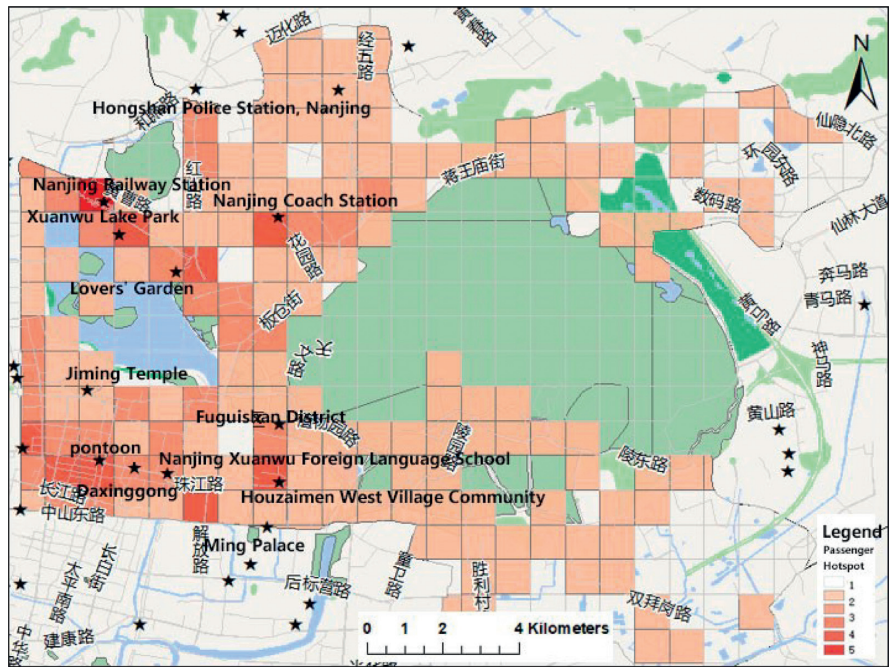


(a)



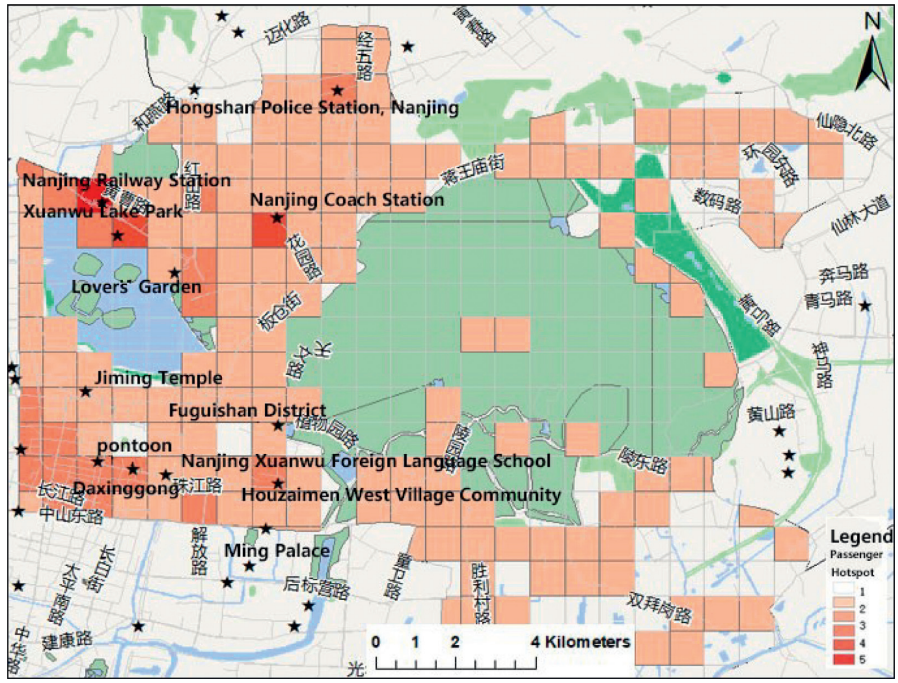
(b)

Figure 12: Taxi pick-up hotspots in Qixia District.



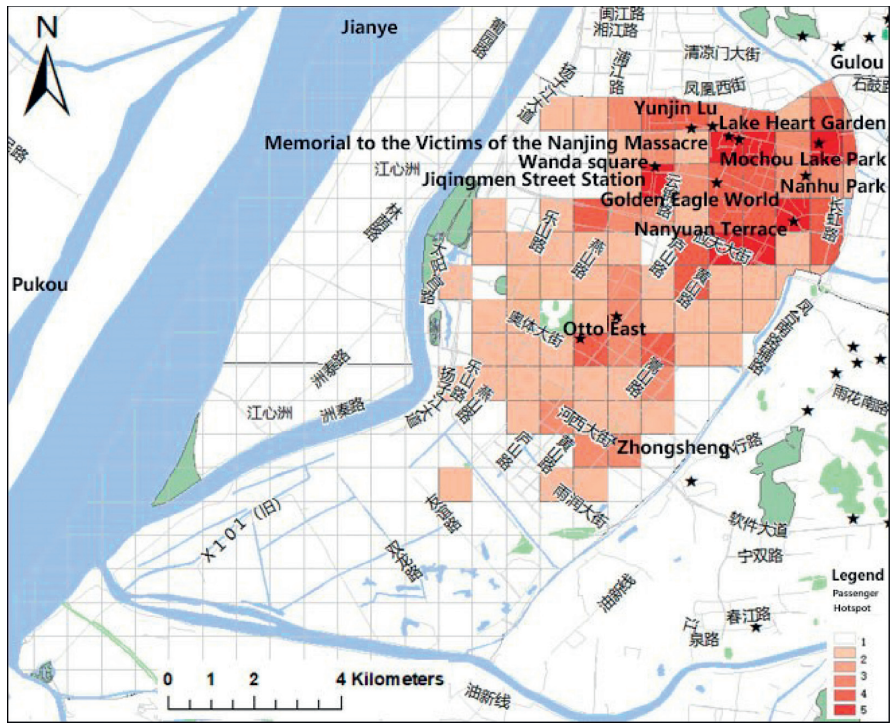
(a)

Figure 13: Continued.



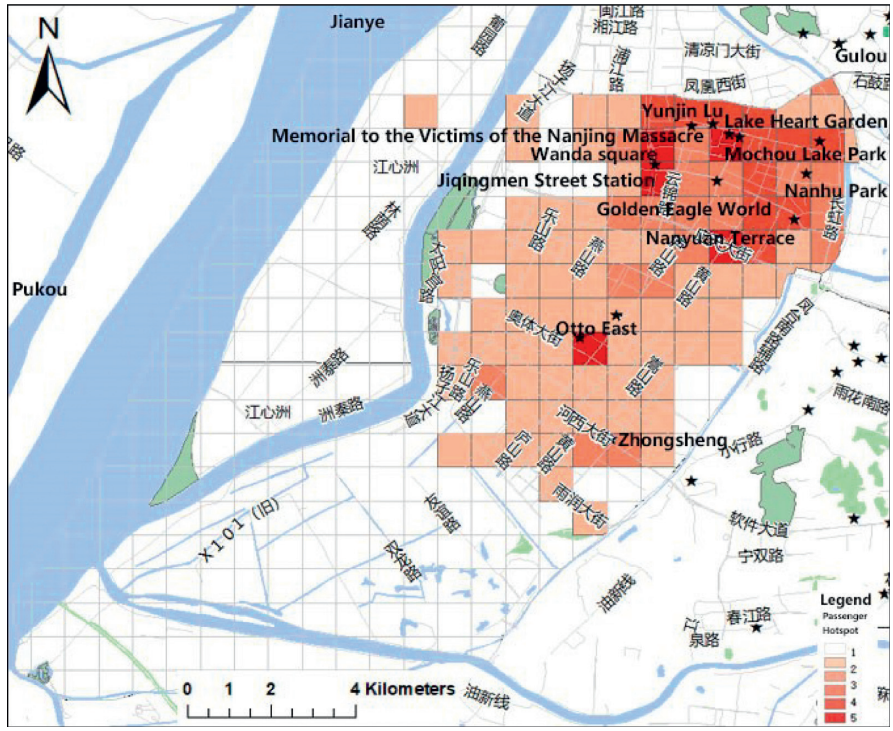
(b)

Figure 13: Taxi pick-up hotspots in Xuanwu District.



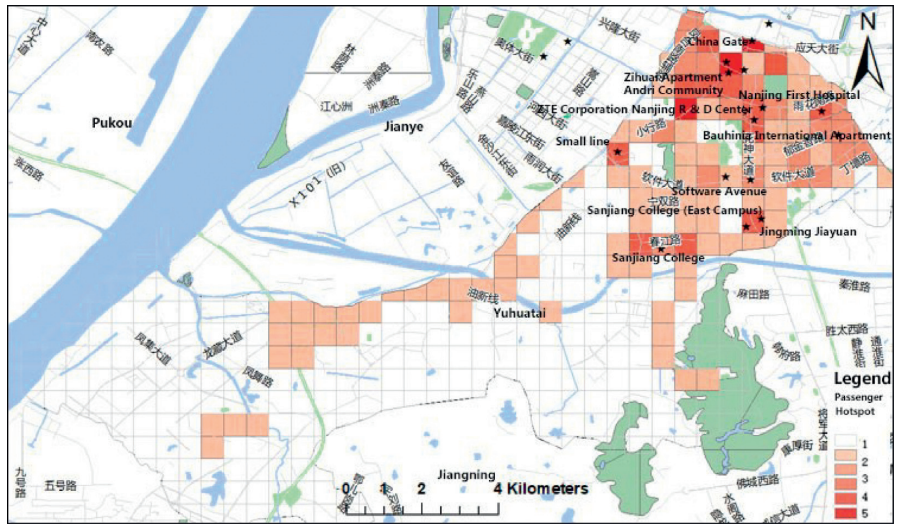
(a)

Figure 14: Continued.



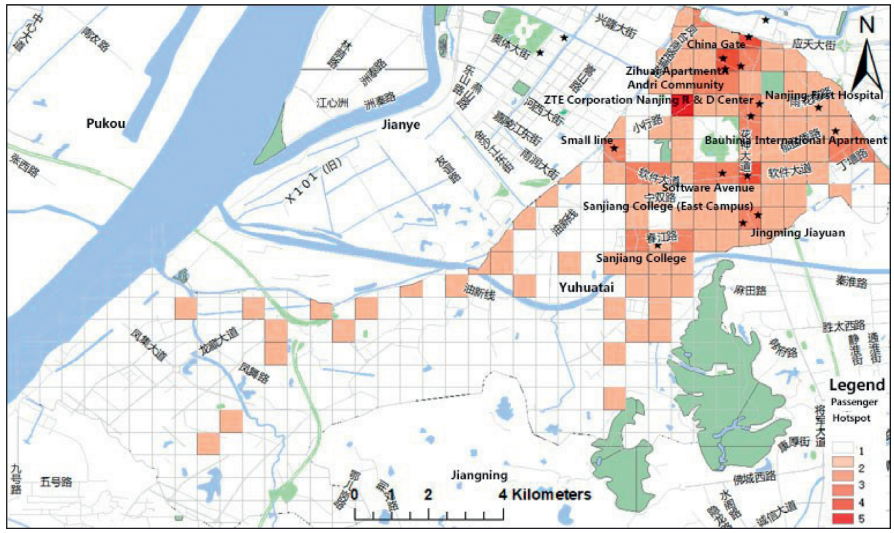
(b)

Figure 14: Taxi pick-up hotspots in Jianye District.



(a)

Figure 15: Continued.



(b)

Figure 15: Taxi pick-up hotspot in Yuhuatai District.

(R&D) base for communications software such that Nanjing is known as "China's famous software city." The taxi pick-up hotspots in this district are mainly located near subway stations, such as Zhonghua Gate, Xiaohang, and Software Avenue. At the same time, the clustering degrees of taxi pick-up hotspots in software R&D centers such as ZTE and Huawei remain at relatively high levels. These areas in Yuhuatai District are all areas with concentrated population distributions and large flows of people, where the demands for taxis remain at relatively high levels. Comparing the distributions of taxi pick-up hotspot areas during the morning and evening rush hours, the clustering degrees of taxi pick-up hotspots in these areas are higher during the morning rush hours, as residential areas are concentrated near some subway stations; during the evening rush hours, residents have more freedom of movement, so compared to the morning rush hours, the taxi pick-up hotspot areas are more widely distributed, but the clustering degrees are relatively lower.

6. Experimental Results and Analysis in Beijing

To verify the effectiveness of the grid information entropy algorithm, this paper conducted a series of experiments based on the working day data of March 1, 2017, in Beijing. As was done with the taxi trajectory data in downtown Nanjing, the day was divided into four time periods: 8:00 a.m. to 10:00 a.m. for the morning commute time period, 12:00 p.m. to 2:00 p.m. for the noontime period, 6:00 pm to 8:00 p.m. for the after-work time period, and 10:00 p.m. to 12:00 a.m. (midnight) for the nighttime period.

6.1. Morning Commute Period. The distribution of taxi passenger hotspots in the morning commuting time period (8:00 a.m. to 10:00 a.m.) in downtown Beijing is shown in Figure 16.

Hotspot areas in Haidian District included South St. of Software Park, junction of West Shangdi Road and Information Road, junction of Chengfu Road and Zhongguancun Road, Zhixin Road, Zhichun Road, West Road of North 3rd Ring Road, North Road of West 3rd Ring Road, and Wanshou Road. Hotspot areas in Chaoyang District included the junction of Xiaoyun Road and Tianze Road, the junction of the East 3rd Ring Road and Xinyuan South Road, San Lutun, Ritan North Road and Shenlu St., Zhaofeng St. and Jinghua St., the junction of Guanghua Road and Jintong East Road, and Dawang Bridge. Hotspot areas in Shijingshan District included North Stadium Road, South Yinhe St., and Lugu Road. Hotspot areas in Fengtai District included Wanfeng Road; the junction of Guang'an Road and Beijing West Railway Station South Road; the junction of Kaiyang Road and South Railway Station Happy Road; the junction of Beijing South Railway Station Road and Majiapu Road; and South Nanxiaojie Road. Hotspot areas in Xicheng District included the junction of Fuxingmen Inner St. and Xuanwumen St. Hotspot areas in Dongcheng District included the junction of Andingmen East St. and Hepingli West St., Jiaodaokou East St. and Dongzhimen Inner St., the junction of East Chang'an St. and Wangfujing St., and the junction of Chongwenmen Outer St. and Xihuashi St.

6.2. Noontime Period. The distribution of taxi passenger hotspots during the noontime period (12:00 p.m. to 2:00 p.m.) in downtown Beijing is shown in Figure 17.

Hotspot areas in Haidian District included the junction of Zhixin East Road and Huanyuan North Road, Astronautics Bridge Roundabout, Yuanda Road, Haidian South Road, the junction of Wangzhuang Road and Chengfu Road, Dahui Temple, Xizhimen St. hotspot areas in Chaoyang District included Jinfang North St., Huizhong Road, Xiaoying West Road, Wenxueguan Road, Jingmi Road, Jing'an East St., Chaoyangmen Outer St., the junction of

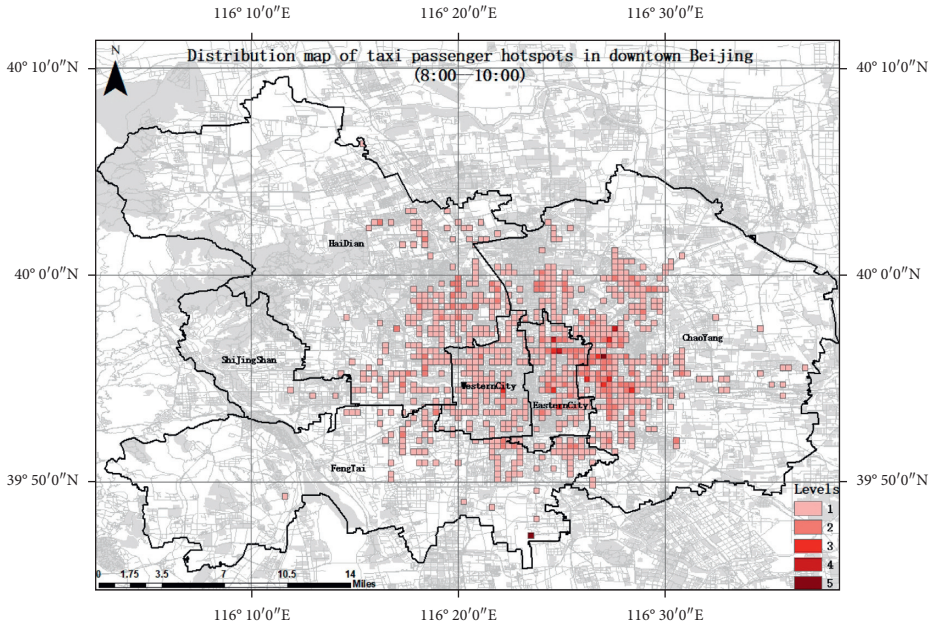


Figure 16: Distribution map of taxi passenger hotspots in downtown Beijing (8:00 a.m. to 10:00 a.m.).

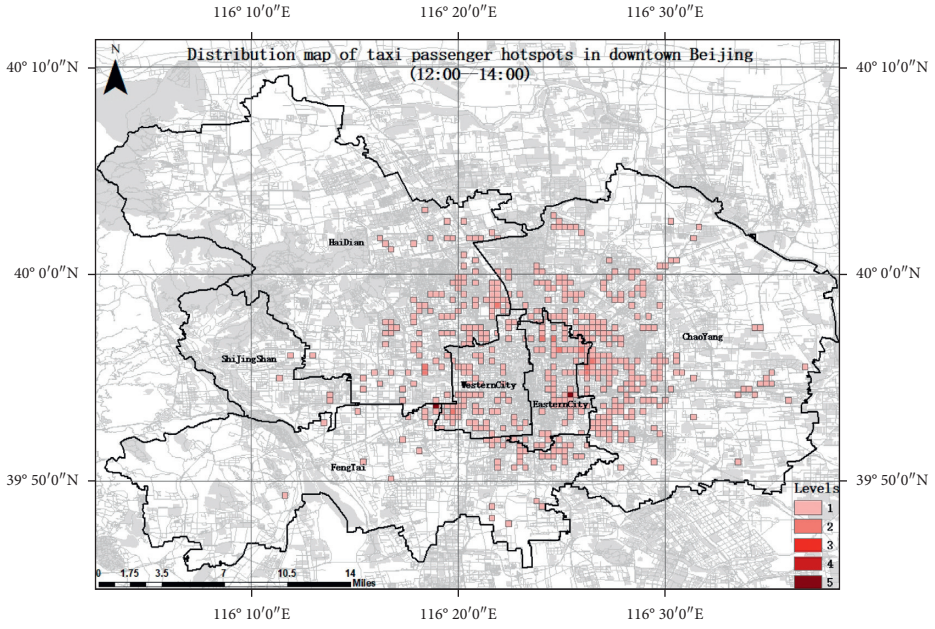


Figure 17: Distribution map of taxi passenger hotspots in downtown Beijing (12:00 p.m. to 2:00 p.m.).

Shilipu Road and Shifoying East Road, Panjiayuan, the junction of Chaoyang Road and Guanzhuang Road, and Tonghui River. Hotspot areas in Shijingshan District included Tiancun Road, Gucheng North Road, and the junction of Yongle East St., and Lugu East St. htspot areas in Fengtai District included South 3rd Ring Middle Road, Majiapu East Road, Shi'anmen, Shuikouzi St., and Lotus Pond East Road. Hotspot areas in Xicheng Districtincluded Guang'anmen, Zhenwu Temple Road, Baiwangzhuang St., Guoying Hutong, Xisi North St., Baisifang West St., Hufang Road, and Deshengmen. Hotspot areas in Dongcheng District included Andingmen West St., Gulou East St., Yonghegong St., Dongzhimen, Beijing Railway Station, Chaoyangmen Inner St., Chongwenmen, Guangqumen Inner St., and Yongdingmen.

6.3. After-Work Time Period. The distribution of taxi passenger hotspots in the after-work period (6:00 p.m. to 8:00 p.m.) in downtown Beijing is shown in Figure 18.

Hotspot areas in Haidian District included Xiyuan Hospital, Haidian St., Central Garden, Suzhou St., the junction of Zhongguancun East Road and Chengfu Road, Huayuan North Road, Zhichun Road, Dahui Temple,

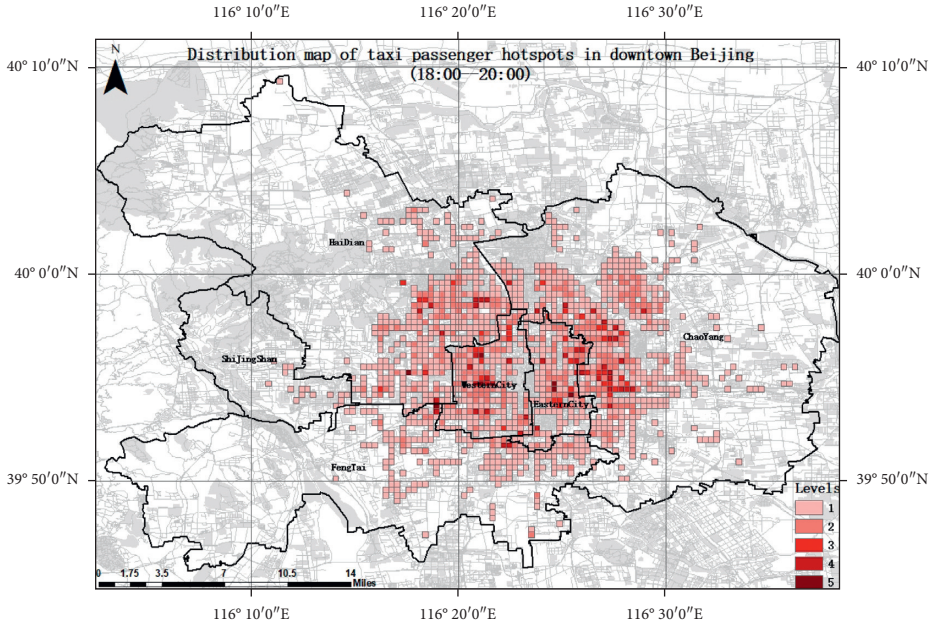


Figure 18: Distribution map of taxi passenger hotspots in downtown Beijing (6:00 p.m. to 8:00 p.m.).

Xizhimen, Wanshou Road, Fucheng Road, Fuxing Road, and Yangfangdian Road. Hotspot areas in Chaoyang District included Anxiang Road, Huizhong Road, Anzhen Road, Yinghuayuan East St., Sun Palace, Sanyuan Bridge, Xinyuan South Road and Liangmaqiao Road, Jiangtai Road, Chaoyang Road, East 3rd Ring Middle Road, Chaoyang Park South Road, and Jianguo Road. Hotspot areas in Shijingshan District included Fushi Road, Lugu Road, Lianshi East Road, and Sculpture Garden South St. Hotspot areas in Fengtai District included Fengtai North Road, West 4th Ring South Road, Science and Technology Park Ave., South 3rd Ring Road, Dacheng Road, Guang'an Road, Beijing West Railway Station South Road, Lize Road, Majiapu, Qunxing Road, and Fangzhuang Road. Hotspot areas in Xicheng District included Deshengmen, Xizhimen St., Fuchengmen Outer St., Guangningbo St., Fuxingmen St., Xuanwumen St., Guang'anmen St., Baisifang St., and Yong'an Road. Hotspot areas in Dongcheng District included Hepingli North St., Dongzhimen Inner St., Guangqumen, Chongwenmen, Beijing Railway Station, and Jianguo Road.

6.4. Nighttime Period. The distribution of taxi passenger hotspots in the nighttime period (10:00 p.m. to 12:00 a.m.) in downtown Beijing is shown in Figure 19.

Hotspot areas in Haidian District included the junction of Software Park St. and Shangdi St., Chengfu Road, Zhongguancun, Huayuan North Road, Zhichun Road, Dahui Temple, North 3rd Ring West Road, Xizhimen North St., Fucheng Road, Fuxing Road, Zizhuyuan Road, Wanshou Road, and Chegongzhuang West Road. Hotspot areas in Chaoyang District included North 4th Ring East Road, Datun Road, Beitucheng East Road, Jiangtai Road, Guangshun St., Wangjing, East Third Ring Road, Chaoyang Road, Jianguomen St., Guangqu Road, and Panjiayuan. Hotspot area in Shijingshan District included Lugu Road, Fushi Road, Yuquan Road, and Shijingshan Road. Hotspot areas in Fengtai District included Dacheng Road, Guang'an Road, Beijing West Railway Station, Fengtai North Road, Beijing Automobile Museum, Lize Road, Majiapu West Road, Xiluoyuan North Road, Nanyuan Road, Wanzhuang Road, and South 3rd Ring Middle Road. Hotspot areas in Xicheng District included Deshengmen Outer St., Xizhimen Outer St., Fuchengmen South St., Xizhimen Outer South Road, Caishikou St., Guang'anmen Outer St., Baisifang St., Fuxingmen Outer St., Di'anmen West St., and Zhushikou West St. Hotspot areas in Dongcheng District included Hepingli North St., Andingmen St., Dongzhimen St., Chongwenmen, Guangqumen St., Wangfujing, Beijing Railway Station, Jianguomen St., Chaoyangmen Inner St., Yongdingmen, Guangming Road, and Tiantan East Road.

6.5. Experimental Conclusions in Downtown Beijing. Beijing, as the capital of China, has a high population density and an obvious pattern of density distribution. By using the grid information entropy algorithm to visualize the clustering analysis of taxi pick-up spots in downtown Beijing, we found that the pick-up spots are mostly spatially distributed in the same areas, and there is a very high number of taxi pick-up spots on the streets along the 3rd Ring and 4th Ring in downtown Beijing, especially on the 3rd Ring streets. In each of these areas, there are hotspots for passenger pick-ups. The east and west city areas have a high number of passenger pick-ups around the Old Jiucheng in Beijing (Zhengyangmen, Chongwenmen, Xuanwumen, Andingmen, Deshengmen, Dongzhimen, Xizhimen, Chaoyangmen, and Fuchengmen). This is due to the many tourist attractions, as well as Beijing's urban architectural layout (with the palace city in the center and other areas arranged in circles around it). In other areas, there are many taxi pick-up spots at similar locations, such as Panjiayuan, Majiapu, Zhichun

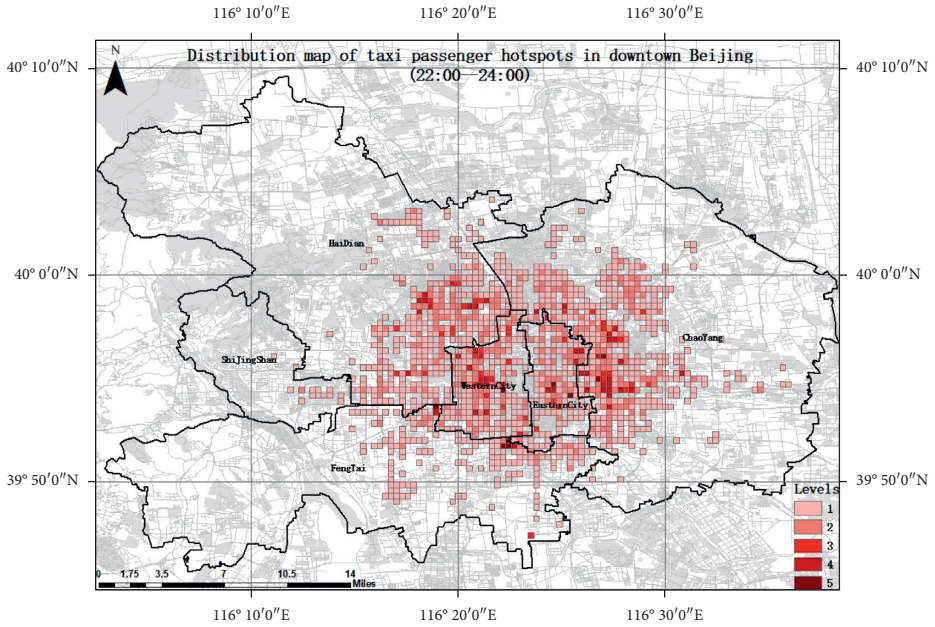


Figure 19: Distribution map of taxi passenger hotspots in downtown Beijing (10:00 p.m. to 12:00 a.m.).

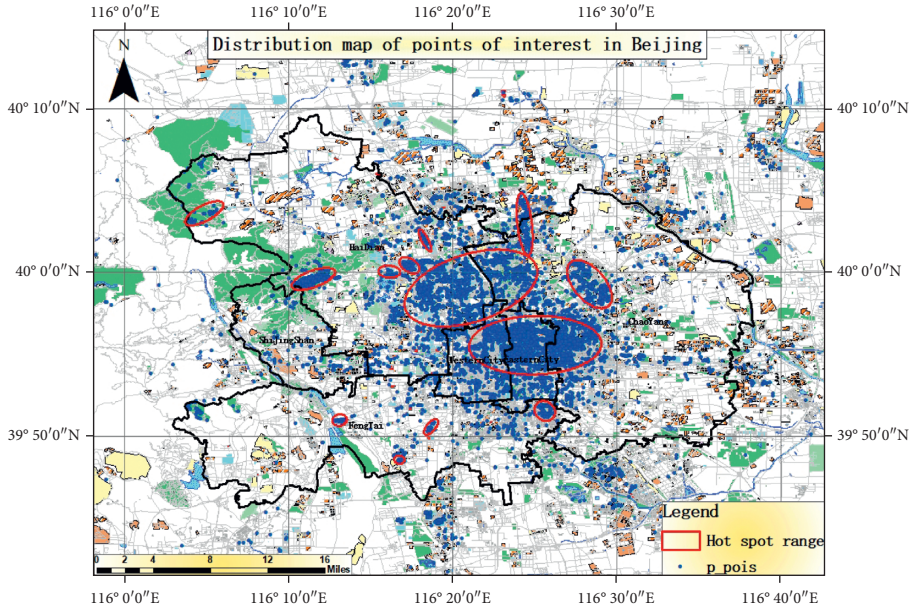


Figure 20: Distribution map of points of interest in downtown Beijing.

Table 2: Statistical indicators of the study areas.

	Nanjing	Beijing (downtown)
Study area (km^2)	6587.02	1368.93
Number of administrative districts	11	6
Number of passenger points	194918	142301
Number of grids	6671	8240
Grid size (m)	1000	140
Maximum entropy	2.311129	2.300098
Minimum entropy	1.562788	0.693147
Maximum aggregation	0.3238	0.6986
Minimum aggregation	0	0
Number of hotspots (average number in the study area)	56	72
Clustering time (s)	127	97

Road, Wangfujing, and Baipifang St. In terms of time, the experiment shows that the number of taxi pick-up spots peaks during the morning commute period, decreases significantly during the noontime period, increases sharply during the after-work period, and continues to increase throughout the nighttime period. This is consistent with the normal work patterns of residents in downtown Beijing.

By visualizing the spatial distribution of the points of interest of Beijing's ground objects, we found that the spatial distribution pattern of these points of interest is almost the same as that of the passenger pick-up spots, as shown in Figure 20. This further verifies the feasibility and effectiveness of the grid information entropy algorithm.

7. Comparison and Analysis of the Statistical Indexes

The statistical indicators in the two study areas are shown in Table 2 and described as follows:

(1) Aggregation degree = 1 − entropy value/maximum entropy value of a grid unit. According to this formula, it is certain that the minimum entropy obtained by the degree of aggregation is 0 because 0 = 1 − maximum entropy/maximum entropy. However, the maximum entropy in different study areas is not necessarily the same. At the same time, because the taxi distribution density in different regions is different, different grid cell sizes are selected. The distribution density of taxis in the six areas of the main urban area of Beijing is greater than that in Nanjing, so the size of the Beijing grid unit is correspondingly smaller than that in Nanjing.

(2) Although the maximum and minimum entropy values of Beijing are lower than those of Nanjing, the maximum aggregation degree and the number of hotspots (global average) in Beijing are greater than those in Nanjing, which is consistent with the fact that the economy and transportation of the capital Beijing are more developed than that of the ancient capital Nanjing.

(3) The number of hotspots (global average) is the average number of hotspots in all time periods of the study area. Because the hotspots in 2 regions and 4 time periods (8:00–10:00, 12:00–14:00, 18:00–20:00, and 22:00–24:00) are studied in this paper, the number of hotspots in each time period is different for each study area, so the average number of hotspots in 4 time periods and the average number of hotspots in each study area are statistically calculated. From the average number of hotspots, we can see that there are more in Beijing than in Nanjing.

Therefore, from the comparative analysis of the above experimental results, it can be seen that although the maximum and minimum entropy values of Beijing are smaller than that of Nanjing, its agglomeration and the number of hotspots are larger than that of Nanjing, indicating that Beijing's taxi traffic is busier and more developed than Nanjing. This result is consistent with the traffic conditions in Beijing and Nanjing in real life and also verifies the effectiveness of the algorithm in this paper.

8. Conclusions and Future Works

8.1. Conclusions. This paper proposed a clustering algorithm based on grid information entropy, which reduces the size of the data to be processed, improves the calculation speed, and exhibits greater flexibility for analyzing massive data. On the basis of the algorithm, the passenger pick-up hotspots for taxis in Nanjing were mined from GPS trajectory data, and it was found that the taxi pick-up hotspot areas are primarily located in six districts, namely Xuanwu District, Gulou District, Yuhuatai District, Qinhuai District, Qixia District, and Jianye District. The hotspots for taxi boarding in Beijing are mainly concentrated around Dongcheng District and Xicheng District. The grid information entropy clustering algorithm was used to perform data mining research on taxi pick-up hotspots in the study area during the morning and evening rush hours, and specific distributions of taxi pick-up hotspots in different districts were identified and analyzed. The taxi pick-up hotspot areas obtained through mining can be used to recommend current passenger pick-up hotspots for taxi drivers, thereby increasing the passenger-carrying rate of taxis, increasing income, and at the same time, reducing vehicle exhaust emissions and protecting the environment.

8.2. Innovations. This study introduced the idea of information entropy in physics to quantify the equilibrium degree of the distributions of taxi pick-up points and described the overall characteristics of the spatial distributions of taxi pick-up points from different perspectives. The study proposed a clustering algorithm based on grid information entropy by first dividing the study area into grids with k as the distance unit, traversing the taxi pick-up points data set, and mapping the taxi pick-up points to the corresponding grids using a mapping function. The algorithm then calculated the information entropy and clustering degree of each grid cell to extract the cells containing pick-up hotspots, and finally, the algorithm traversed the hotspot grid cells and expanded each hotspot area according to the clustering degrees and the distances between grid cells. This algorithm greatly reduces the size of the data, improves the calculation speed, exhibits good flexibility, and overcomes the shortcomings of the density-based method when processing massive data.

8.3. Future Study. In this study, the trajectory data of taxis in Nanjing and Beijing are used as examples to conduct mining and analysis and get the hotspots of taxi boarding. However, only taxi data from the trajectory data were used in this paper; at present, owing to the popularity of taxi booking apps such as Didi Taxi and Meituan Taxi, passengers are more willing to use these apps when going out. Therefore, subsequent studies should combine the trajectory data in the taxi booking apps to perform more comprehensive analyses; at the same time, the taxi floating car data, the metro smart

card data, and the GPS trajectories of Mobike can also be used for analyzing the spatiotemporal characteristics of multimode travelers [37]. As of now, this study focused on the hotspot areas for picking up passengers; subsequent studies can further recommend taxi waiting locations and passenger pick-up locations for passengers and drivers, respectively, as well as plan the shortest driving route for a particular trip.

Data Availability

All data, models, and codes that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they do not have any conflicts of interest.

Authors' Contributions

Bi Shuoben, Wan Lei, and Liu Aili conceived and designed the experiments. Wan Lei and Xu Ruizhuang performed the experiments. Bi Shuoben, Wan Lei, and Wang Luye wrote the Chinese paper. Bi Shuoben, Xu Ruizhuang, and Wang Luye translated the paper. All authors read and approved the final manuscript.

Acknowledgments

This work was financially supported by the National Natural Science Foundation of China (nos. 41971340 and 41271410).

References

[1] L. Wang, K. Hu, T. Ku, and J. W. Wu, "Mining urban moving trajectory patterns based on multi-scale space partition and road network modeling," *Acta Automatica Sinica*, vol. 41, no. 1, pp. 47–58, 2015.

[2] J. Huang, P. Zhang, H. Xuanjun, and S. Heli, "A trajectory prediction approach for mobile objects by combining semantic feature," *Journal of Computer Research and Development*, vol. 51, no. 1, pp. 76–87, 2014.

[3] Y. Wang, Q. Han, and H. Pan, "A clustering scheme for trajectories in road networks," in *Proceedings of the 2009 3rd International Conference on Teaching and Computational Science (WTCS 2009)*, pp. 11–18, Shenzhen, China, December 2012.

[4] Z. Deng, Y. Hu, M. Zhu, and X. Huang, "A scalable and fast OPTICS for clustering trajectory big data," *Cluster Computing*, vol. 18, no. 2, pp. 549–562, 2015.

[5] H. J. Miller and J. Han, *Geographic Data Mining and Knowledge Discovery*, pp. 352–366, Taylor & Francis, Oxfordshire, UK, 2001.

[6] J. Steenbruggen, M. T. Borzacchiello, P. Nijkamp, and H. Scholten, "Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities," *GeoJournal*, vol. 78, no. 2, pp. 1–21, 2013.

[7] H. Han, *Research of Intelligent Transport Platform Based on Big Data*, Chengdu University of Technology, Chengdu, China, 2014.

[8] L. Li, R. Dong, and B. He, "Development on TaxiReservation and dispatching system base on wireless network," *Journal of Lanzhou Jiaotong University*, vol. 32, no. 3, pp. 103–107, 2013.

[9] J. Weng, W. Liu, and Z. Chen, "Research on floating car data based taxi operation and management," *Journal of Beijing University of Technology*, vol. 36, no. 6, pp. 779–784, 2010.

[10] L. Tang, W. Zheng, and Z. Wang, "Space time analysis on the pick-up and drop-off of taxi passengers based on GPS big data," *Journal of Geo-Information Science*, vol. 17, no. 10, pp. 1179–1184, 2015.

[11] Y. Hu and G. Chen, "An effective cluster Analysis algorithm based on grid and intensity," *Journal of Computer Applications*, vol. 23, no. 12, pp. 64–67, 2003.

[12] H. Zhao, X. Liu, and H. Cui, "Grid -based clustering algorithm," *Computer Technology and Development*, vol. 20, no. 9, pp. 83–85, 2010.

[13] W. Qi, *Analysis of Spatial Characteristics of Taxi Passenger Points Based on GPS Data*, Jinlin University, Changchun, China, 2013.

[14] X. Li and D. Li, "DBSCAN spatial clustering algorithm and its application in urban planning," *Science Surveying and Mapping*, vol. 30, no. 3, pp. 51–53, 2005.

[15] H. Su Khaing and T. Thein, "An efficient clustering algorithm for moving object trajectories," in *Proceedings of the 3rd International Conference on Computational Techniques and Artificial Intelligence (ICCTAI 2014)*, pp. 74–78, Singapore, 2014.

[16] B. Yun, D. J. Sun, Y. Zhang, S. Deng, and J. Xiong, "A charging location choice model for plug-in hybrid electric vehicle users," *Sustainability*, vol. 11, no. 20, 2019.

[17] J. Yang, J. Gao, and J. Liang, "An improved DBSCAN clustering algorithm based on data field," *Journal of Frontiers of Computer Science & Technology*, vol. 6, no. 10, pp. 903–911, 2012.

[18] M. Duan and C. Tang, "Realization of clustering algorithm based on density," *Journal of Jishou University (Natural Sciences Edition*, vol. 34, no. 1, pp. 26-27, 2013.

[19] J. An, *Research on the Pattern Mining Algorithm of User's Moving Track Sequence based on MapReduce*, Shandong University of Technology, Zibo, China, 2016.

[20] Y. Zhou, *Clustering Algorithm Based on Grid and Information Entropy*, Hunan University, Changsha, China, 2011.

[21] T. Georg, W. Dirk, and B. Martin, "Grid-based multi-road-course estimation using motion planning," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 1924–1935, 2016.

[22] L. Shen, J. Lu, L. Man, and T. Chen, "Identification of accident blackspots on rural roads using grid clustering and principal component clustering," *Mathematical Problems in Engineering*, vol. 2019, no. 4, Article ID 2151284, 12 pages, 2019.

[23] A. Shi, H. Yang, and J. Wang, "Revealing recurrent urban congestion evolution patterns with taxi trajectories," *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, pp. 128–146, 2018.

[24] Q. Y. Ma, H. Yang, H. Zhang, K. Xie, and Z. Y. Wang, "Modeling and analysis of daily driving patterns of taxis in reshuffled ride-hailing service market," *Journal of Transportation Engineering Part A-Systems*, vol. 145, no. 10, Article ID 04019045, pp. 1–19, 2019.

[25] X. Dong and L. Mingjie, "Optimal road accident case retrieval algorithm based on k-nearest neighbor," *Advances in Mechanical Engineering*, vol. 11, no. 2, pp. 1–7, 2019.

[26] D. J. Sun, K. Zhang, and S. Shen, "Analyzing spatiotemporal traffic line source emissions based on massive didi online car-hailing service data," *Transportation Research Part D*, vol. 62, pp. 699–714, 2018.

[27] X. Ke, L. Shi, W. Guo, and D. Chen, "Multi-dimensional traffic congestion detection based on fusion of visual features and convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2157–2170, 2019.

[28] H. Zhou and S. Zhou, "Scene categorization towards urban tunnel traffic by image quality assessment," *Journal of Visual Communication and Image Representation*, vol. 65, Article ID 102655, 2019.

[29] C. Hu and T. Jean-Claude, "Predicting the upcoming services of vacant taxis near fixed locations using taxi trajectories," *ISPRS International Journal of Geo-Information*, vol. 8, no. 7, pp. 295–310, 2019.

[30] Y. Lu, Z. Zeng, H. Wu, G. G. Chua, and J. Zhang, "An intelligent system for taxi service: analysis, prediction and visualization," *AI Communications*, vol. 31, pp. 33–46, 2018.

[31] D. Esztergár-Kiss and V. Remeli, "Toward practical algorithms for activity chain optimization," *Transportation Letters*, vol. 13, no. 1, pp. 64–76, 2021.

[32] M. Estrada, J. Maria Salanova, M. Medina-Tapia, and F. Robusté, "Operational cost and user performance analysis of on-demand bus and taxi systems," *Transportation Letters*, vol. 13, no. 3, pp. 229–242, 2021.

[33] X. Pan, "Investigating college students' choice of train trips for homecoming during the Spring festival travel rush in China: results from a stated preference approach," *Transportation Letters*, vol. 13, no. 1, pp. 36–44, 2021.

[34] G. Ali Shafabakhsh, A. Famili, and M. Akbari, "Spatial analysis of data frequency and severity of rural accidents," *Transportation Letters*, vol. 2016, Article ID 1138605, 2016.

[35] K. Wiley, M. Hanna, and P. Kanaroglou, "Exploring and modeling the level of service of urban public transit: the case of the Greater Toronto and Hamilton Area, Canada," *Transportation Letters*, vol. 3, no. 2, pp. 77–89, 2011.

[36] D. J. Sun and X. Ding, "Spatiotemporal evolution of ride-sourcing markets under the new restriction policy: a case study in Shanghai," *Transportation Research Part A*, vol. 130, pp. 227–239, 2019.

[37] F. Chen, Z. Yin, and Y. Ye, "Taxi hailing choice behavior and economic benefit analysis of emission reduction based on multi-mode travel big data," *Transport Policy*, vol. 97, pp. 73–84, 2020.

[38] X. Zhao, *Urban Hot Spot Area and Hot Spot Path Mining Based on Space-Time Constraints*, Chongqiang University, Chongqing, China, 2017.

[39] S. Brakatsoulas, D. Pfoser, and R. Salas, "On map-matching vehicle tracking data," in *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 853–864, VLDB Endowment, Trondheim, Norwaya, September 2005.

[40] L. Chen, *Research on the Characteristics of Taxi Operation Based on the Data Collected by GPS Floating Car*, Tongji University, Shanghai, China, 2008.

[41] 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," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 5, pp. 312–328, 2007.

[42] Y. Tan and C. Wu, "The law of the information entropy values of land use composition," *Journal of Natural Resources*, vol. 18, no. 1, pp. 112–117, 2003.

[43] S. Yang, S. Bi, and A. Nkunzimana, "A spatial clustering method of taxi passenger track," *Computer Engineering and Applications*, vol. 54, no. 14, pp. 249–255, 2018.

[44] B. Qiu and X. Zhang, "Grid -based clustering algorithm with the parameter automatization," *Journal of Zhengzhou University*, vol. 27, no. 2, pp. 90–93, 2006.