International Journal of Geo-Information

MDPI

Article

Urban Hotspot Area Detection Using Nearest-Neighborhood-Related Quality Clustering on Taxi Trajectory Data

Qingying Yu [1,2], Chuanming Chen [1,2,*], Liping Sun [1,2] and Xiaoyao Zheng [1,2]

1 School of Computer and Information, Anhui Normal University, Wuhu 241002, China; ahnuyuq@ahnu.edu.cn (Q.Y.); slp620@ahnu.edu.cn (L.S.); zxiaoyao@ahnu.edu.cn (X.Z.)
2 Anhui Provincial Key Laboratory of Network and Information Security, Anhui Normal University, Wuhu 241002, China
* Correspondence: ccm1981@ahnu.edu.cn

Abstract: Urban hotspot area detection is an important issue that needs to be explored for urban planning and traffic management. It is of great significance to mine hotspots from taxi trajectory data, which reflect residents' travel characteristics and the operational status of urban traffic. The existing clustering methods mainly concentrate on the number of objects contained in an area within a specified size, neglecting the impact of the local density and the tightness between objects. Hence, a novel algorithm is proposed for detecting urban hotspots from taxi trajectory data based on nearest neighborhood-related quality clustering techniques. The proposed spatial clustering algorithm not only considers the maximum clustering in a limited range but also considers the relationship between each cluster center and its nearest neighborhood, effectively addressing the clustering issue of unevenly distributed datasets. As a result, the proposed algorithm obtains high-quality clustering results. The visual representation and simulated experimental results on a real-life cab trajectory dataset show that the proposed algorithm is suitable for inferring urban hotspot areas, and that it obtains better accuracy than traditional density-based methods.

Keywords: passenger travel trajectory; neighborhood association; urban hotspot area detection; nearest neighborhood-related quality clustering

Citation: Yu, Q.; Chen, C.; Sun, L.; Zheng, X. Urban Hotspot Area Detection Using Nearest-Neighborhood-Related Quality Clustering on Taxi Trajectory Data. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 473. https://doi.org/10.3390/ijgi10070473

Academic Editor: Wolfgang Kainz

Received: 25 May 2021
Accepted: 9 July 2021
Published: 10 July 2021

1. Introduction

Urban hotspots are the embodiment of the frequent activities of urban residents. The locations and routes frequently traveled by residents can intuitively reflect the city's traffic conditions and user movement patterns. Urban hotspot area detection is an important issue that needs to be explored in urban planning and traffic management. Detected hotspot areas can be used as effective reference information for traffic guidance and the layout of urban public facilities. For example, hotspot detection results at the same time on different days can effectively guide where to place public service advertisements. The advertisements placed in hotspot areas are more likely to be noticed. By publishing hotspots that occur at certain times, drivers can be guided to avoid these hotspot areas, thereby alleviating traffic congestion. As it is well known, urban traffic conditions are influenced by the density of vehicles on the roads. Taxis are one of the most convenient means of public transportation for city dwellers, providing personalized travel services. As such, taxi trajectory data are spatio-temporal big data containing the travel behavior of residents. Information about residents' travel time, routes, and distance traveled is closely related to residents' travel activities. Hence, acquisition of the taxi location density can be used to analyze urban traffic conditions. Consequently, it is important to mine hotspots from taxi trajectory data, which reflect the travel characteristics of urban residents and the operational status of urban traffic [1].

Many scholars have used big data to conduct research on the detection of urban spatial hotspot areas, achieving rich results. For example, Ashbrook et al. [2] proposed a two-step method to infer hotspot locations and constructed a Markov model to predict future locations. Zhou et al. [3] proposed a clustering algorithm (DJ-Cluster) based on density and connection to infer hotspot access locations. Cao et al. [4] proposed a semantically enhanced clustering algorithm (SEM-CLS) to extract semantically meaningful locations and rank the locations through a unified probability model. Xia et al. [5] divided the original trajectory into a series of sub-trajectories using stay points and road intersections as feature points, clustered the sub-trajectories, and then analyzed the weights of the sub-trajectories to obtain the hot paths. Gui et al. [6] proposed a distributed parallel algorithm for extracting traffic hotspots from taxi trajectories. First, the information on taxi stops was extracted, and then a representative density-based clustering (DBSCAN, density-based spatial clustering of applications with noise) method was used to cluster the block data to discover hotspot areas in different time periods. Different from partitioning and hierarchical clustering methods, DBSCAN defines clusters as the largest sets of points connected by density. It can divide regions with a sufficiently high density into clusters and can find clusters of arbitrary shapes in a noisy spatial database. Ma et al. [7] applied an agglomerated hierarchical clustering algorithm and GIS (geographical information system) analysis method based on taxi trajectory data to mine the hotspot areas and spatio-temporal characteristics of residents who travel using taxis. Savage et al. [8] proposed a grid-based trajectory clustering algorithm to find hotspots, which can distinguish the direction of the route and analyze the sub-parts of the route. However, the abnormal data points contained in the dataset cannot be removed, which affects the clustering effect. Ferreira et al. [9] proposed a probabilistic model-based hotspot identification method based on simulated intersection data. This method is easy to apply, but it needs to consider risk factors to determine hotspots and is only suitable for intersection data. Scholz [10] analyzed the GPS trajectory data of 536 taxis in San Francisco over 22 days and proposed a method of modeling collective activity patterns to determine the location and time of activity hotspots in the metropolis of San Francisco. However, the time interval of the dataset used was one hour, which is not universal.

The spatial clustering method has mainly been adopted in hotspot mining based on trajectory data [11–14]. From the perspective of clustering objects, the existing research is mainly divided into three categories:

(1) Studies that directly perform density-based clustering on the locations of trajectory data [15,16]. This method is suitable for clustering noisy spatial data. It can effectively deal with abnormal data and can find clusters with arbitrary shapes by connecting adjacent regions with sufficient density. However, when the density of spatial clustering is uneven and the cluster spacing is very different, the clustering quality is poor.
(2) Studies that convert the sequence of locations into a sequence of trajectory segments and find hot paths and regions by clustering the trajectory segments [17]. This method can find the local similarities in complex spatio-temporal trajectories; the extracted feature points are concise and effective, but the clustering results mainly depend on the division quality of trajectory segments.
(3) Studies that convert the trajectory to a certain sequence of grids and then cluster on the grid sequence to find hotspots [18,19]. The advantage of this method is that it can accurately identify the complex coupling phenomenon of urban hotspots, but the algorithm relies too much on experimental parameters.

In order to improve adaptability, some studies have also utilized the aforementioned methods [20], the clustering results of which can be used in the analysis of hotspot areas. During the specified time period, the top k areas detected with the highest density of locations are hotspots where traffic congestion or parking dilemmas may occur. Therefore, analyzing and detecting hotspots can serve the management of urban planning or traffic control departments.

Most existing research performed density-based clustering directly on the locations in the trajectories. This type of method can find clusters with arbitrary shapes in a spatial database containing noise and can connect adjacent regions with sufficient density to effectively process abnormal data, but the cluster quality is poor when the density of spatial clusters is uneven and the inter-cluster distances are large [15,16]. Zheng et al. [21] proposed a grid density-based clustering algorithm to discover residents' preferred travel areas during different periods of the day. Liu et al. [22] identified highly congestion-prone areas using the DBSCAN clustering method.

The aforementioned studies are primarily applicable to data spaces with a uniform density; they ignore the effect of the local density and the tightness between objects. However, in an actual traffic network, the locations in the taxi trajectories are not evenly distributed. In order to solve this problem, this paper proposes an improved quality threshold clustering algorithm based on neighborhood association—denoted as QTNA—which is used to detect urban residents' travel hotspots from taxi trajectory data. The algorithm considers the relationship between each cluster center and its neighborhood to obtain high-quality clustering results, which are significant for the analysis of hotspot areas. The visual representation of and simulation experiments on a real taxi trajectory dataset in Beijing show that the proposed algorithm is suitable for the detection of urban residents' travel hotspots and has higher accuracy than traditional density-based methods.

The remainder of this paper is organized as follows. Section 2 introduces the preliminary concepts and problem definition. The novel spatial clustering method for urban hotspot area detection is also presented. The experimental results and analysis are discussed in Sections 3 and 4. Section 5 presents the conclusions, as well as the limitations and implications of this research.

2. Methods

2.1. Nearest Neighborhood Model

Definition 1. *(nearest neighborhood): Consider a dataset DS with n objects. For any object $O_i \in DS(i = 1, \ldots, n)$, the nearest neighborhood of O_i refers to a set consisting of any object (excluding O_i itself) with a distance from O_i that is less than θ_r , denoted by $NN(O_i)$. It is defined as follows:*

$$NN(O_i) = \{p | dist(p, O_i) \leq \theta_r,\ p \in DS \backslash \{O_i\}\}, \tag{1}$$

where θ_r is a raduis threshold, each p in $NN(O_i)$ is called a θ_r-neighbor of the object O_i, and $dist(x, y)$ represents the distance between objects x and y. That is, for any $q \in DS - NN(O_i)$, $dist(q, O_i) > \theta_r$, $q \neq O_i$. Each object within the nearest neighborhood of O_i is called a θ_r-neighbor of O_i.

Definition 2. *(nearest neighborhood distance): For any object $O_i \in DS(i = 1, \ldots, n)$, the nearest neighborhood distance of O_i refers to the average distance between O_i and all objects in $NN(O_i)$, denoted by $NNdist(O_i)$. It is defined as*

$$NNdist(O_i) = \frac{\sum_{p \in NN(O_i)} dist(p, O_i)}{|NN(O_i)|}. \tag{2}$$

In Equation (2), $|NN(O_i)|$ is the size of $NN(O_i)$, which needs to be compared with the size threshold θ_n.

In the proposed nearest neighborhood-related quality clustering algorithm, the locations of taxi trajectory data are the basic research objects.

2.2. Passenger Travel Trajectory Model

According to the description of empty and heavy vehicles in the original taxi trajectory dataset (0 is empty, others are heavy), the positions of passengers in all taxi trajectories can

be extracted. We used 0 and 1 to represent the empty and heavy states, respectively. The trajectory data of a certain taxi on a certain day can be converted into a passenger-carrying state sequence: 001111110001111 . . . 000. The state segment sequence corresponding to the trajectory can then be obtained, as shown in Figure 1. The continuous "1" sequence represents the passenger travel segment, and the continuous "0" sequence represents the no-load travel segment.

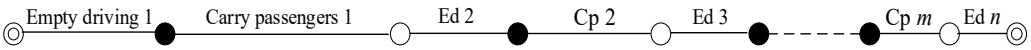


Figure 1. Sequence of a taxi's one-day trajectory state segments. There are m passenger travel segments.

Definition 3. *(pick-up location): A location whose status transfers from zero to nonzero in the taxi trajectory dataset is defined as a pick-up location. The set of all pick-up locations is denoted as PL.*

Definition 4. *(drop-off location): A location whose status transfers from nonzero to zero in the taxi trajectory dataset is defined as a drop-off location. The set of all drop-off locations is denoted as DL.*

All the locations in $PL \cup DL$ are the original and destination locations.

Definition 5. *(passenger travel trajectory): A trajectory consisting of a set of time-ordered locations with the pick-up location as the starting point and the drop-off location as the ending point is called a passenger travel trajectory.*

As shown in Figure 1, the double circle points indicate the starting and ending locations of the taxi on that day, the black solid circle points indicate the pick-up locations, and the hollow circle points indicate the drop-off locations. The empty segment of the taxi is represented by Ed (empty driving), and the passenger travel segment is represented by Cp (carry passengers). The trajectory shown in Figure 1 contains *m* passenger travel segments.

2.3. Urban Hotspot Area Detection Algorithm

Based on spatio-temporal analysis methods, mining the movement patterns of taxi trajectories and the regional distribution of pick-up and drop-off locations is helpful for in-depth analysis of urban residents' travel behavior characteristics and movements and can provide a powerful data reference for urban transportation planning departments.

2.3.1. Algorithm Framework

Based on the quality threshold (QT) clustering method [23], in this paper, we propose a nearest neighborhood-related quality clustering algorithm to detect urban hotspot areas from taxi trajectory data. The QT algorithm was originally proposed for gene clustering and later used in the clustering of time series. It ensures the clustering quality by finding dense clusters whose diameters do not exceed a given user-defined diameter threshold [23]. Considering the neighborhood relationship and clustering quality, the proposed algorithm can obtain more accurate clustering results and detect high-density hotspot areas.

Urban hotspots refer to areas where passengers are highly concentrated. Taking the taxi trajectory data as the research object, the largest cluster of pick-up and drop-off locations represents the densest area. The purpose of our algorithm is to find an optimal cluster that considers the nearest neighborhood feature each time. The algorithm contains two main phases. First, each location point and its nearest neighbors are grouped into one cluster. Second, the largest cluster that satisfies the size requirement is selected as a candidate cluster, and its neighborhood feature is analyzed to obtain the current optimal cluster. The framework of the proposed algorithm is shown in Figure 2.

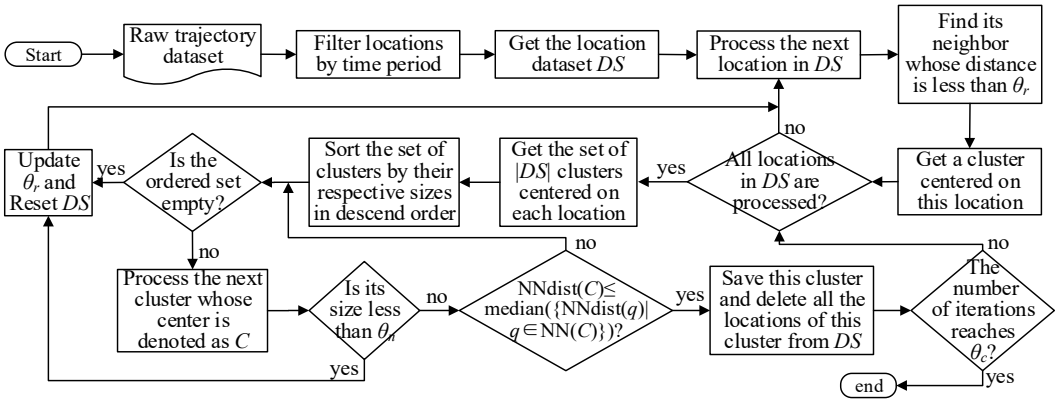


Figure 2. Schematic diagram of the proposed approach.

2.3.2. Algorithm Description

Based on the *PL* and *DL* datasets, the nearest neighborhood-related quality clustering method was used to mine the urban hotspots of residents in each period, including hot pick-up and drop-off areas.

The input of the algorithm was the set of locations to be analyzed, which was extracted from the raw trajectory dataset and represented by *DS*. Each iteration of the proposed algorithm was designed to include three specific parts, as follows:

(i) Find the θ_r-neighbors for each location to form $|DS|$ clusters;
(ii) Choose the maximal cluster that meets the size requirements;
(iii) Analyze the neighborhood feature of the candidate cluster to filter out the current optimal cluster.

The steps of the proposed algorithm are as follows:

Step 1. For each location $O_i \in DS$, the distance $dist_{ij}$ between O_i and O_j ($O_j \in DS \backslash \{O_i\}$) is calculated.

Step 2. For any location O_i and O_j, if $dist_{ij} \leq \theta_r$, put O_j into the cluster centered on O_i.

Step 3. Sort the set of all $|DS|$ clusters by their respective sizes in descending order and find the maximal cluster *Clus*.

Step 4. Let *C* be the center of the candidate cluster, *Clus*. If the number of objects in *Clus* is greater than or equal to the cluster size threshold θ_n, calculate the intra-cluster distance *NNdist*(C) of *Clus* (i.e., the average distance between the center point *C* and all other points in *Clus*) and calculate the intra-cluster distance *NNdist*(*q*) centered on all other points except point *C* in *Clus*, where $q \in NN(C)$; otherwise, update θ_r, reset *DS*, and go to Step 2.

Step 5. Calculate the value of *NNdist*(C)- *median*({*NNdist*(*q*) | *q*∈*NN*(C)}). If it is less than or equal to 0, select *Clus* as the maximal cluster and delete all objects in *Clus* from *DS*, and go to Step 1; otherwise, select the next maximal cluster (if it exists) from the ordered set as *Clus*, and go to Step 4.

Step 6. Repeat steps 1 to 5 until *DS* is empty or the number of iterations reaches θ_c.

The trajectory location clustering algorithm was denoted as QTNA; it was used to detect urban hotspot areas effectively. In this paper, the QTNA algorithm is also called the nearest neighborhood-related quality clustering algorithm. The pseudocode of the algorithm is given as follows:

The algorithm ends when the location dataset is empty, or the number of iterations reaches the threshold θ_c. The updated method of θ_r is as follows:

$$\theta_r = \theta_r \times (1 + \propto) \tag{3}$$

where $\propto$ is an adjustment parameter for θ_r.

The time complexity of each iteration of Algorithm 1 depends on the following: (a) the time to compute the distance $dist_{ij}$ between any two locations O_i and O_j, whose time complexity is O(n^2) (Lines 5–16); (b) the time to sort the set of n clusters according to the size of each cluster, whose time complexity is O(n^2) (Line 17); (c) the time to scan each cluster from large to small to determine whether the nearest neighborhood distance of each center meets the requirements, whose time complexity is O(n) (Lines 19–33). The maximum number of iterations is θ_c, where $\theta_c \ll n^2$. Therefore, the time complexity of Algorithm 1 is O(n^2).

Algorithm 1: QTNA (Quality threshold clustering based on neighborhood association)

Input: DS($= \{O_1, \ldots, O_n\}$, the set of pick-up/drop-off locations in taxi trajectory dataset), θ_r (the radius threshold), θ_n (the cluster size threshold), θ_c (the iteration number threshold)
Output: *CS* (the clustering results)
1: $itercnt \leftarrow 0$; $k \leftarrow 0$; $oldDS \leftarrow DS$;
2: $CS \leftarrow \varnothing$; $Centers \leftarrow \varnothing$;
3: **repeat:**
4: $itercnt \leftarrow itercnt + 1$;
5: **for** $i \leftarrow 1$ to $|DS|$ **do**
6: $tempCS_i \leftarrow \{O_i\}$;
7: **for** $j \leftarrow 1$ to $|DS|$ **do**
8: **if** (j==i) **then**
9: continue;
10: **end if**
11: Compute the distance $dist_{ij}$ between O_i and O_j;
12: **if** $dist_{ij} \leq \theta_r$ **then**
13: $tempCS_i \leftarrow tempCS_i \cup \left\{O_j\right\}$;
14: **end if**
15: **end for**
16: **end for**
17: Sort *tempCS* by $|tempCS_i|$ in descend order;
18: $Maxi \leftarrow 1$;
19: **while** ($Maxi \leq |tempCS|$) **do**
20: **if** $|tempCS_{Maxi}| \geq \theta_n$ **then**
21: $C \leftarrow$ the center of $tempCS_{Maxi}$;
22: **if** $NNdist(C) \leq median(\{NNdist(q) | q \in NN(C)\})$ **then**
23: $k \leftarrow k + 1$;
24: $CS_k \leftarrow tempCS_{Maxi}$;
25: $Centers_k \leftarrow C$;
26: $DS \leftarrow DS - tempCS_{Maxi}$;
27: break;
28: **else**
29: $Maxi \leftarrow Maxi + 1$;
30: **end if**
31: **else** break;
32: **end if**
33: **end while**
34: **until** isempty(*DS*) or $itercnt == \theta_c$ or isequal(*oldDS*, *DS*);
35: **return** *CS*;

3. Results

In this section, we present a case study based on the city of Beijing in China. Beijing is a provincial-level administrative region, a municipality directly under the control of the central government, and the political, economic, cultural, and transportation center of China. As the capital of China, the construction and development of urban road networks in Beijing are representative of its status. Beijing has a well-developed road network system, abundant taxi routes, diversified user travel modes, and extensive sources of trajectory data. Therefore, Beijing was the first choice for this study. The taxi trajectory data in Beijing

have obvious characteristics of large sample data and typical representative significance. They are suitable for the development of urban hotspot area detection based on trajectory data mining. Hotspot detection results can provide data support for traffic management departments, guide users to travel reasonably, save travel time, and alleviate traffic congestion in big cities. We performed a set of experiments to evaluate the performance of the proposed algorithm. We first present the experimental settings, including the introduction of the experimental environment and several parameters selected in the experiments. Then, the datasets and evaluation metrics used in the experiments are introduced. Finally, the visualization results of the experiments are shown, and the accuracy of the proposed approach is evaluated using the silhouette metric.

Our experimental process was specifically arranged as follows: (1) Preprocess the experimental datasets based on the proposed passenger travel trajectory model. (2) Detect the urban hotspot areas from the taxi trajectory data based on the QTNA algorithm. (3) Obtain experimental results and related analysis.

3.1. Experimental Environment and Parameter Selection

The experiments were conducted with MATLAB 8.3 on a PC with an Intel Core 2 Duo CPU 3.60 GHz and RAM of 32 GB. The operating system was Microsoft Windows 10.

In order to evaluate the effect and accuracy of hotspot area recognition, two sets of experiments are conducted in this section (one set is based on the data of different time periods on the same day, and the other is based on the data of the same time period on different days). The proposed QTNA algorithm is compared with DBSCAN and QT clustering methods in terms of its effect and accuracy. The reasons for choosing these two algorithms for comparative experiments are as follows. First, QTNA is a density-based clustering method. DBSCAN is the most classic density-based algorithm, and it is also the most widely used density-based clustering algorithm for detecting hotspots. Second, the idea of choosing the optimal cluster for each iteration contained in the QT algorithm is the basis of our proposed algorithm. The parameters used in Algorithm 1 include the radius threshold θ_r, the size threshold θ_n, and the iteration number threshold θ_c. Specifically, based on realistic scale requirements and our preprocessing experimental results, θ_r is assigned as 0.005, θ_n is set as 30, and θ_c is set as 100.

3.2. Dataset

As it was mentioned earlier, we used Beijing as the case study location. The experimental GPS trajectory data of approximately 20,000 taxis in Beijing in March 2017 came from Datatang (Beijing, China) Intelligent Technology Co., Ltd. (Zhongguancun Street, Haidian District, Beijing, China) which included the taxis' original equipment manufacturer (OEM) identification code, terminal phone number encryption, Universal Time Coordinated (UTC) time, message length, latitude, longitude, driving angle, driving speed, mileage, positioning description, empty and loaded vehicle description, status, status description, and other information. The daily data were stored in the txt file format—such as "20170301.txt", which recorded the GPS location data of all taxis in Beijing on 1 March 2017. The average daily trajectory data contained approximately 25.05 million locations. Owing to equipment or communication failures and other reasons, some sampling data will inevitably be wrong. Therefore, the dataset had to be cleaned to remove records with missing or obviously abnormal data. The taxi trajectory data used in the following were all preprocessed. This dataset is denoted as TaxiDS.

In the following experiments, we extracted a total of 12 sub-datasets from TaxiDS, which specifically included the data of the pick-up and drop-off locations of passenger travel trajectories during three identical periods—8:00–9:00, 13:00–14:00, and 18:00–19:00—on 1 March 2017 and 4 March 2017. In the three time periods on 1 March, the number of pick-up points is 17,675, 18,835, and 15,403, respectively, and the number of drop-off points is 16,517, 17,573, and 16,431, respectively. In the three time periods on 4 March, the number of pick-up points is 11,466, 14,011, and 11,381, respectively, and the number of drop-off

points is 10,490, 13,710, and 12,422, respectively. The total size of the 12 sub-datasets is 3.38 MB.

3.3. Evaluation Metrics

Let *Tclusters* be a set consisting of trajectory clusters and num_c be the number of trajectory clusters in *Tclusters*. The silhouette index [24] value of trajectory T_x can be used to measure the degree of cohesion between T_x and the trajectory cluster C_i to which it belongs, and the degree of separation between T_x and other trajectory clusters (C_j, $j \neq i$). It is denoted as $S(T_x)$, and its equation is as follows:

$$S(T_x) = \frac{b(T_x) - a(T_x)}{max\{a(T_x),\ b(T_x)\}}, \tag{4}$$

where $a(T_x)$ is the average distance of T_x to all T_y ($T_x, T_y \in C_i, T_y \neq T_x$), and $b(T_x)$ is the minimum distance over all clusters C_j ($j \neq i$), of the average distances to $T_y \in C_j$. $a(T_x)$ and $b(T_x)$ can be calculated as follows:

$$a(T_x) = \frac{1}{|C_i|-1} \sum_{T_x, T_y \in C_i,\ T_x \neq T_y} dist(T_x, T_y), \tag{5}$$

$$b(T_x) = min_{C_i, C_j \in TClusers,\ j \neq i} \left\{ \frac{1}{|C_j|} \sum_{T_y \in C_j} dist(T_x, T_y) \right\}. \tag{6}$$

The silhouette value $S(T_x)$ ranges from -1 to 1, where a high value indicates that the trajectory T_x is well matched to its own cluster and poorly matched to neighboring clusters. If most trajectories have high values, then the trajectory clustering result is appropriate.

We can then quantify the validity of the trajectory clustering using the silhouette index (*SI*), which is defined as follows:

$$SI = \frac{1}{num_c} \sum_{i=1}^{num_c} \{ \frac{1}{|C_i|} \sum_{T_x \in C_i} S(T_x) \}. \tag{7}$$

The *SI* is suitable for evaluating the performance of the clustering algorithm, and its result is representative in evaluating the clustering effect.

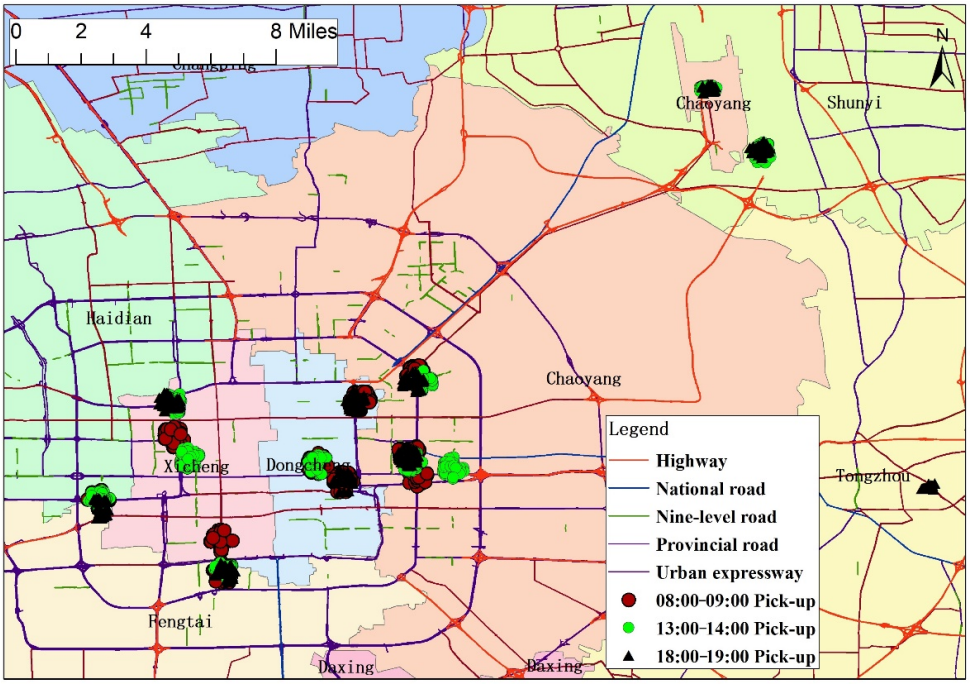
3.4. Case Study Results

In this section, both the experimental results and related analysis of this case study are presented. A set of comparative experiments is first conducted to evaluate the performance of the proposed approach. Then, the detected urban hotspot areas are visually displayed.

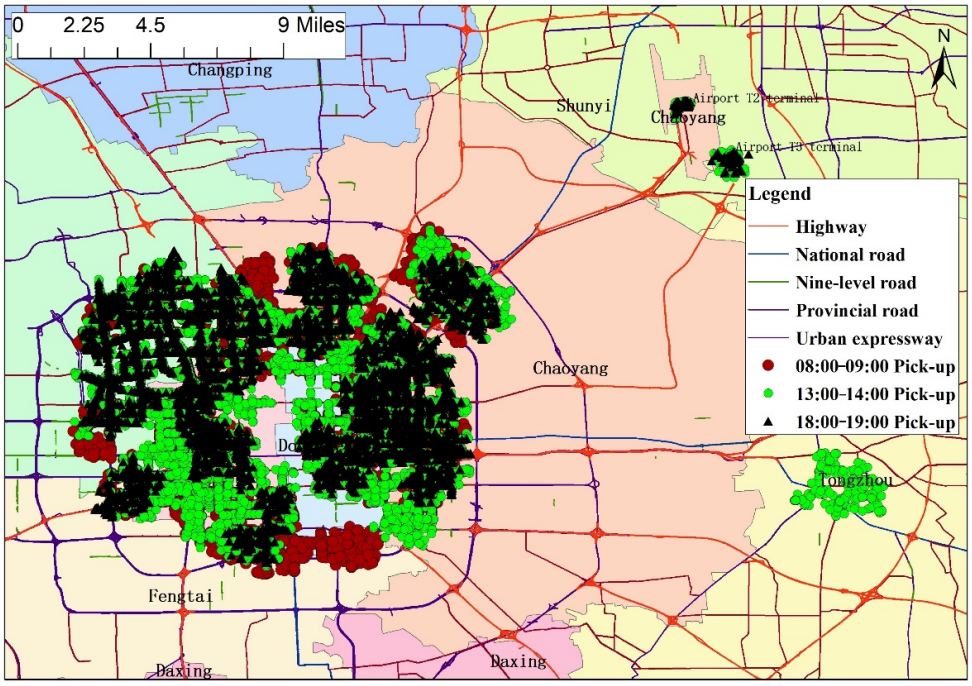
3.4.1. Different Time Periods on the Same Day

First, a set of experiments for inferring hotspot areas at different time periods on the same day was conducted. The dataset contained the pick-up and drop-off locations of passenger travel trajectories during three typical periods—8:00–9:00, 13:00–14:00, and 18:00–19:00—on 1 March 2017.

As it is shown in Figure 3, the top 10 hot pick-up areas detected by the QTNA and DBSCAN algorithms in the three time periods of the day are marked with three different symbols. Figure 3a shows the detection results of QTNA, and Figure 3b shows the detection results of DBSCAN. The hotspot area detected by DBSCAN covers almost half of the city center, which is meaningless for traffic management or urban planning. The detected results indicate that the proposed QTNA algorithm is more suitable for hotspot detection.



(**a**) QTNA



(**b**) DBSCAN

Figure 3. Top 10 hot pick-up areas during different periods on the same day. Some overlapping areas marked by different symbols are hot pick-up areas at different periods. (**a**) shows the detection results of the QTNA algorithm. (**b**) shows the detection results of the DBSCAN algorithm.

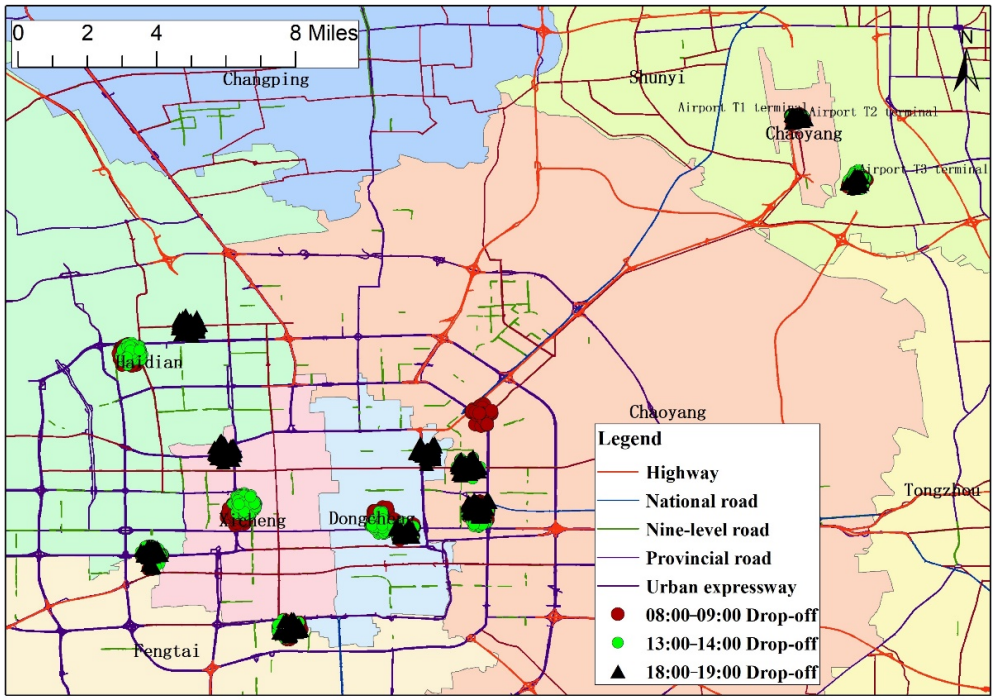
It can be seen from Figure 3a that the detected hot pick-up areas were mainly distributed in Chaoyang, Dongcheng, Xicheng, Fengtai, Shunyi, Haidian, and Tongzhou Districts, and some areas are hot pick-up areas at different periods. In other words, there are overlapping areas marked by different symbols, such as the area near the intersection of Guanghua Road and Jinghua South Street in Chaoyang District, which happens to be

the exit of the Jintaixizhao Subway Station with a large passenger flow. Some areas are only hot pick-up areas during a specific time period. For example, the area near the exit of the Taoranting Subway Station on Baizhifang East Street in Xicheng District is a hot pick-up area between 8:00 and 9:00. The area near the intersection of the Jingtong Expressway and Xidawang Road in Chaoyang District is a hot pick-up area between 13:00 and 14:00, and the area near the intersection of Binhe Middle Road and Yudaihe East Street in Tongzhou District is a hot pick-up area between 18:00 and 19:00. Such hotspot discovery is valuable for many applications. It can provide guidance for traffic management decisions during specific time periods.

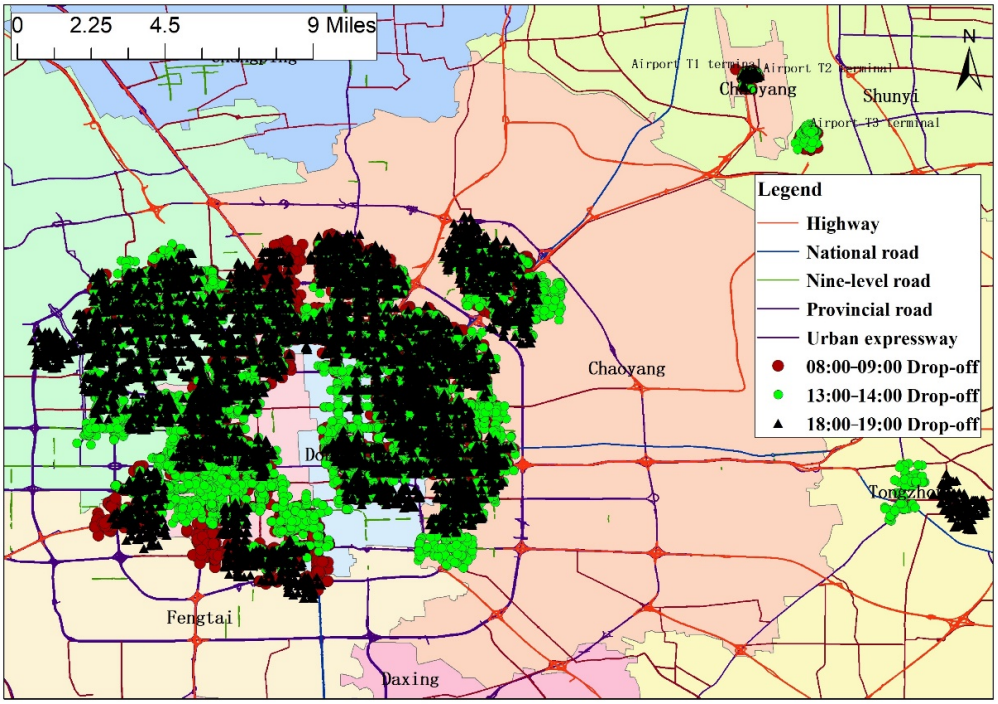
As it is shown in Figure 4, the top 10 hot drop-off areas detected by the QTNA and DBSCAN algorithms in the three time periods of the day are marked with three different symbols. Figure 4a shows the detection results of the QTNA algorithm, and Figure 4b shows the detection results of the DBSCAN algorithm. Similar to the result in Figure 3b, the hotspot area detected in Figure 4b also covers almost half of the city center, which indicates the DBSCAN algorithm is not appropriate for inferring hotspots. The detected hot drop-off areas were mainly distributed in Chaoyang, Dongcheng, Haidian, Xicheng, Fengtai, and Shunyi Districts. It can also be seen that some areas are hot drop-off areas at different time periods, such as the T1, T2, and T3 terminals of Beijing Capital International Airport located in Shunyi District and its inner Chaoyang District enclave. Some areas are only hot drop-off areas within a specific time period; for example, the area near the intersection of Xiaoyun Road and Dongsanhuan North Road in Chaoyang District is a hot drop-off area in the time period 8:00–9:00. The area near Zhongguancun East Road, Chengfu Road, and Heqing Road in Haidian District is a hot drop-off area during the time period 18:00–19:00.

The DBSCAN algorithm is essentially a process of finding core samples of high density and expanding clusters from them. For any point p, if it is the core point, a cluster C can be formed, with p as the center and r as the radius. The expansion process is conducted to traverse the points in the cluster. If the point q belonging to the r-neighborhood of p is the core point, the points in the r-neighborhood of q are also classified into cluster C. The process is executed recursively until C can no longer be expanded. DBSCAN is good for data which contain clusters of similar density. However, in an actual traffic network, the locations in the taxi trajectories are not evenly distributed. In contrast, the QTNA algorithm can detect high-density areas centered on each point, and it considers the relationship between each cluster center and its neighborhood to obtain high-quality clustering results, which are significant for the analysis of urban hotspots.

Based on the detection results of the top 10 hot pick-up and drop-off areas during three different time periods on the same day, it was found that (i) the hot pick-up and drop-off areas were unevenly distributed, concentrated in the capital functional core area and urban functional expansion area; (ii) some hot pick-up areas were also hot drop-off areas, generally concentrated in road sections with high passenger flows, such as subway entrances and bus stops; and (iii) compared with hot pick-up areas, the distribution of hot drop-off areas at different time periods on the same day was more focused. The hotspots within a certain time period reflect the travel aggregation characteristics of urban residents, and the discovery of hotspots is helpful for scientific traffic management and urban planning.



(**a**) QTNA

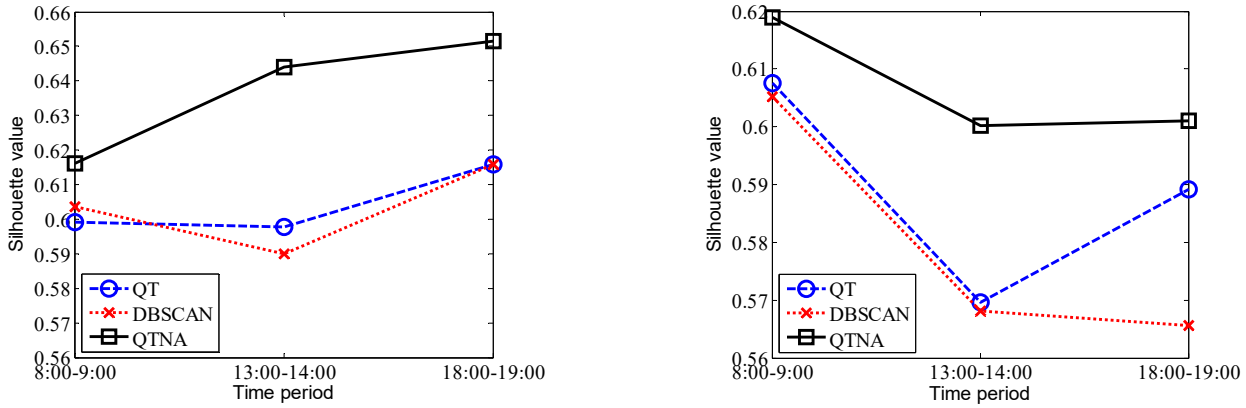


(**b**) DBSCAN

Figure 4. Top 10 hot drop-off areas during different time periods on the same day. Some areas are hot drop-off areas at different time periods. (**a**) shows the detection results of the QTNA algorithm. (**b**) shows the detection results of the DBSCAN algorithm.

In order to verify the detection accuracy of the proposed algorithm, the SI introduced in Section 3.3 was selected for quantitative evaluation. The pick-up and drop-off locations during the three time periods were clustered to find the hot pick-up and drop-off areas. Figure 5 shows the comparison of silhouette values of the QTNA, DBSCAN, and QT algorithms on the three datasets with different time periods. Figure 5a shows the clustering results based on the pick-up location dataset for different time periods, and Figure 5b shows

the clustering results based on the drop-off location dataset for different time periods. As it can be seen from Figure 5, the accuracy of the proposed QTNA algorithm was superior to that of the other two algorithms. Judging from the results of visualization and accuracy evaluation, the proposed algorithm is more suitable for hotspot detection.



(**a**) Pick-up location (**b**) Drop-off location

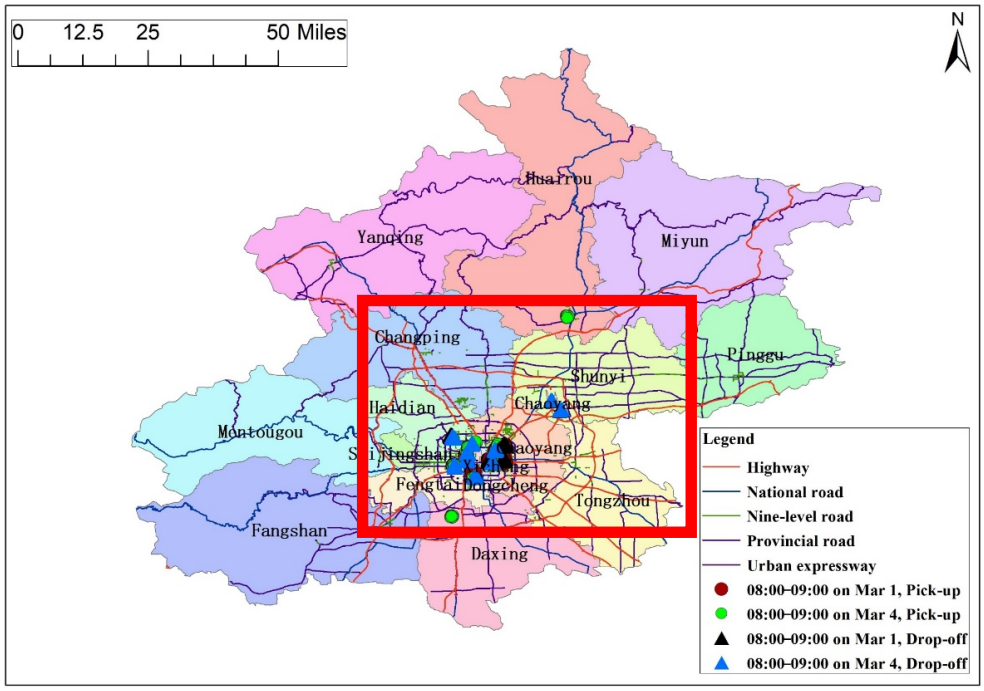
Figure 5. Performance comparison of clustering algorithms. (**a**) shows the clustering results based on the pick-up locations during different time periods on the same day. (**b**) shows the clustering results based on the drop-off locations during different time periods on the same day.

3.4.2. Same Time Period on Different Days

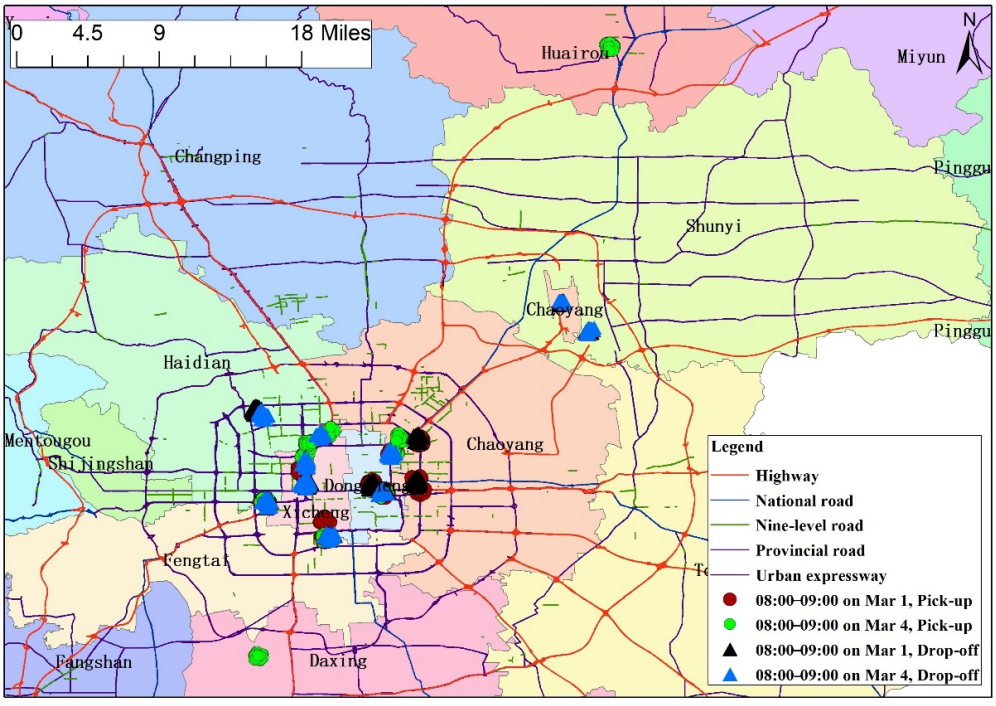
Second, a set of experiments for hotspot area inference during the same time period on different days was conducted. The dataset contained the pick-up and drop-off locations of passenger travel trajectories during three identical periods—8:00–9:00, 13:00–14:00, and 18:00–19:00—on 1 March 2017 and 4 March 2017. Figure 6 shows the top 10 hot pick-up and drop-off areas detected by the QTNA algorithm on the two days between 8:00 and 9:00, marked with symbols of four different colors and shapes. Figure 6a shows the distribution of hotspots on a global map of Beijing, and Figure 6b is an enlarged view of the area framed by the red box in Figure 6a. It can be seen that some high-density hotspot areas are marked with different symbols, indicating that these areas are hotspots during this time period on the two days.

From the detection results of the top 10 hot pick-up and drop-off areas during the same time period on different dates, we found the following: (i) In the time period between 8:00 and 9:00 on the working day (1 March) and the rest day (4 March), the hot pick-up areas were mainly distributed in Chaoyang, Dongcheng, Xicheng, Fengtai, and Huairou Districts. The hot drop-off areas were mainly distributed in Chaoyang, Dongcheng, Xicheng, Haidian, Fengtai, and Shunyi Districts. (ii) There were certain differences between the hot pick-up areas on rest days and those on weekdays. For example, Huairou District is an ecological conservation development zone, with a permanent population density of only 185 persons per km^2. The top 10 hot pick-up and drop-off areas detected during multiple time periods on working days were not distributed in Huairou District, and hot pick-up areas near Nanhua Street on the west side of the Yingbin Roundabout in the southern part of Huairou District were detected during rest days. Due to the diversification of travel purposes and travel distances on rest days, there were certain changes in hot pick-up areas. (iii) Among the top 10 hot drop-off areas detected in the 8:00–9:00 time period of the two days, there was a hot drop-off area located in the enclave of Shunyi District in Chaoyang District, as shown in Figure 6—the specific location was at the Beijing Capital International Airport (terminals T1 and T2). Another hot drop-off area was the T3 terminal located in Shunyi District, indicating that hot drop-off areas for passenger travel trajectories are often popular travel destinations for urban residents. (iv) In the 8:00–9:00 period of the two days, there

were areas that were both hot pick-up locations and hot drop-off locations, such as the areas located at the exits of the Beijing Railway Station subway station, as shown in Figure 7.



(**a**) Distribution of hotspots on a global map of Beijing



(**b**) Enlarged view of the area framed by the red box in (**a**)

Figure 6. Top 10 hot pick-up and drop-off areas during the same time period on different days. (**a**) shows the distribution of hotspots on a global map of Beijing. (**b**) shows the enlarged view of the area framed by the red box in (**a**).

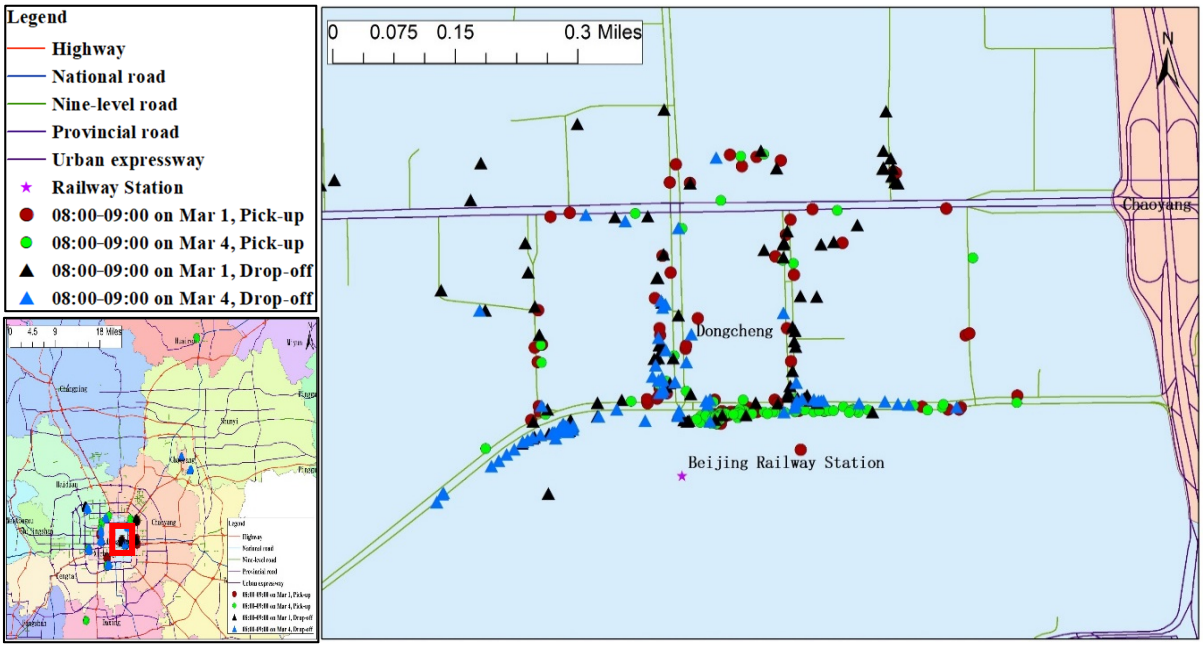


Figure 7. Example of a hot pick-up and drop-off area during the same time period on different days.

To verify the detection accuracy of the proposed algorithm, in addition to the data in the 8:00–9:00 time period, the data in the 13:00–14:00 and 18:00–19:00 time periods were also selected for comparison experiments. Table 1 shows the silhouette value comparison of the clustering results of the QTNA, DBSCAN, and QT algorithms based on pick-up locations during the same time period on different dates. Table 2 shows the silhouette value comparison of the clustering results of these three algorithms based on drop-off locations during the same time period on different dates.

Table 1. Performance comparison of clustering algorithms based on pick-up locations during the same time period on different dates.

Date and Time / Algorithm	8:00–9:00		13:00–14:00		18:00–19:00	
	1 Mar	4 Mar	1 Mar	4 Mar	1 Mar	4 Mar
QT	0.5993	0.6505	0.5977	0.6493	0.6157	0.7125
DBSCAN	0.6035	0.5656	0.5898	0.5985	0.6159	0.6189
QTNA	0.6162	0.7118	0.6440	0.6673	0.6515	0.7204

Table 2. Performance comparison of clustering algorithms based on drop-off locations during the same time period on different dates.

Date and Time / Algorithm	8:00–9:00		13:00–14:00		18:00–19:00	
	1 Mar	4 Mar	1 Mar	4 Mar	1 Mar	4 Mar
QT	0.6075	0.7033	0.5696	0.6592	0.5892	0.6711
DBSCAN	0.6052	0.6301	0.5681	0.5987	0.5656	0.6078
QTNA	0.6189	0.7574	0.6002	0.6910	0.6011	0.7072

In summary, as shown in Figure 5, based on the taxi pick-up location and drop-off location datasets during different time periods of the same day, the silhouette values of the clustering results obtained based on the QTNA algorithm were greater than those of the

DBSCAN and QT algorithms. As it is shown in Tables 1 and 2, based on the data during the same time period on different dates, the same results were obtained. Therefore, according to the clustering results on the taxi pick-up and drop-off location datasets, the density of taxi pick-up and drop-off locations can be clearly distinguished, and the corresponding hotspot areas can be inferred. The relative density based on the neighborhood association ensures the accuracy of the clustering results. In addition, the radius and size thresholds can be assigned different values to adapt to various actual situations. Experimental comparison results indicate that the proposed algorithm outperforms the traditional DBSCAN and QT clustering algorithms in terms of applicability and accuracy.

4. Discussion

Exploring hotspots of interest from taxi trajectory data is beneficial to urban traffic management, road planning, and location-based services. The hotspot areas hidden in trajectory data are the information that must be mastered when studying the travel characteristics of multiple users and which can be used to establish a predictive model of future user behavior. Hotspot area detection is an important scientific issue in trajectory data analysis. The locations and routes by which residents frequently travel can intuitively reflect urban traffic conditions and user movement patterns. Hot pick-up and drop-off locations detected during different periods of the day can effectively guide user travel and avoid traffic congestion. Hotspots detected during the same time period on different days can provide reasonable data support for the layout of urban public facilities [25]. As the capital of China, Beijing has a well-developed road network system, abundant taxi routes, and extensive sources of trajectory data. This paper studied GPS trajectory data of about 20,000 taxis in Beijing in March 2017 and analyzed the distribution of urban hotspot areas with respect to time. First, the pick-up and drop-off locations were extracted from taxi trajectory data based on the constructed passenger travel trajectory model. A nearest neighborhood-related quality clustering algorithm was then proposed to cluster the pick-up and drop-off locations. On the one hand, we detected and analyzed hotspot areas during different time periods on the same day; on the other hand, we focused on the detection results during the same time period on different days.

In summary, compared with previous research, the differences and advantages of this study are as follows:

(1) Nearest neighborhood model construction. By learning from our previous work [26], we proposed a nearest neighborhood model, which was adopted in location clustering and could help detect optimal clusters.
(2) Urban hotspot area detection using an improved quality threshold clustering algorithm based on neighborhood association. An improved quality threshold clustering algorithm was proposed that considers the neighborhood association in order to improve the accuracy of spatial clustering. The proposed algorithm was used to detect urban hotspot areas based on taxi trajectory data. Analysis of the relative density is important for spatial clustering of taxi locations.
(3) Case study. The proposed algorithm was tested on a real-life trajectory dataset of taxis in Beijing. The visual presentation and experimental results show that our algorithm detected urban hotspot areas with high accuracy, effectively providing data support for traffic guidance.

The results of this study are helpful for traffic guidance and urban facility planning, with the following practical implications:

(1) The publication of the detected hotspots during a certain time period can help alleviate traffic congestion and improve the quality of residents' travel experience.
(2) Hotspot detection results during the same time period on different days can effectively guide the location of urban public facilities and reduce wastage of resources.
(3) Hotspot areas detected from the taxi trajectory dataset can provide urban planning departments with guidance for setting up taxi parking spots.

5. Conclusions

In-depth mining results of taxi trajectory data are helpful for the analysis of the spatial characteristics of urban residents' travel. This paper addressed the issue of urban hotspot area detection using an improved quality threshold clustering algorithm based on neighborhood association. Each iteration of the proposed algorithm has three specific steps. First, the θ_r-neighbors for each location are found to form several clusters. Then, the maximal cluster that meets the size requirement is chosen. Finally, the neighborhood features of the candidate cluster are analyzed to filter out the current optimal cluster. The proposed method not only considers the maximum clustering in a limited range but also considers the local density of clusters and the tightness between objects through neighborhood analysis, effectively addressing the clustering issue of unevenly distributed datasets. The visual representation and simulated experimental results show that the proposed algorithm could obtain effective and reasonable urban hotspots from taxi trajectory data and provide valuable information for traffic management systems. This achievement is practical enough to be applied in travel recommendation and road planning and may also be used in urban hotspot area analysis based on other moving objects in the city. This paper studied clustering based on the pick-up and drop-off locations, without clustering the trajectory segments. In the future, we plan to further infer popular routes during different time periods and conduct personalized route recommendations.

Author Contributions: Conceptualization, Qingying Yu; data curation, Chuanming Chen; funding acquisition, Qingying Yu, Liping Sun and Xiaoyao Zheng; methodology, Qingying Yu; supervision, Chuanming Chen; validation, Chuanming Chen; visualization, Qingying Yu; writing—original draft, Qingying Yu; writing—review and editing, Liping Sun and Xiaoyao Zheng. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 61702010, 61972439, and 61672039, the Anhui Provincial Natural Science Foundation, grant number 1808085MF172, and the Key Program in the Youth Elite Support Plan in Universities of Anhui Province (gxyqZD2020004).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors would like to thank the reviewers for their useful comments and suggestions for this paper. This work was supported by the National Natural Science Foundation of China (Grant Nos. 61702010, 61972439, and 61672039), the Anhui Provincial Natural Science Foundation (Grant No. 1808085MF172), and the Key Program in the Youth Elite Support Plan in Universities of Anhui Province (Grant No. gxyqZD2020004).

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

References

1. Gong, S.; Cartlidge, J.; Ruibin, B.; Yue, Y.; Li, Q.; Qiu, G. Geographical and temporal huff model calibration using taxi trajectory data. *GeoInformatica* **2020**, *4*, 1–28. [CrossRef]
2. Ashbrook, D.; Starner, T. Using GPS to learn significant locations and predict movement across multiple users. *Pers. Ubiquitous Comput.* **2003**, *7*, 275–286. [CrossRef]
3. Zhou, C.; Frankowski, D.; Ludford, P.; Shekhar, S.; Terveen, L. Discovering personal gazetteers: An interactive clustering approach. In Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems (GIS), Arlington, VA, USA, 12–13 November 2004; pp. 266–273.
4. Cao, X.; Cong, G.; Jensen, C.S. Mining significant semantic locations from GPS data. *Proc. VLDB Endow.* **2010**, *3*, 1009–1020. [CrossRef]
5. Xia, Y.; Wen, H.; Zhang, X. Hot route analysis method based on trajectory clustering. *J. Chongqing Univ. Posts Telecommun. (Nat. Sci.)* **2011**, *23*, 602–606.
6. Gui, Z.; Xiang, Y.; Li, Y. Parallel discovering of city hot spot based on taxi trajectories. *J. Huazhong Univ. Sci. Technol. (Nat. Sci.)* **2012**, *40*, 187–190.
7. Ma, Y. *Research on Residents' Behavior of Attractive Areas and Spatio-Temporal Feature Based on Taxi Trajectory Data*; Nanjing Normal University: Nanjing, China, 2014.

8. Savage, N.S.; Nishimura, S.; Chavez, N.E.; Yan, X. Frequent trajectory mining on GPS data. In Proceedings of the 3rd International Workshop on Location and the Web, Tokyo, Japan, 29 November 2010; Volume 2010, pp. 8–11.
9. Ferreira, S.; Couto, A. Hot-spot identification: Categorical binary model approach. *Transp. Res. Rec.* **2013**, *2386*, 1–6. [CrossRef]
10. Scholz, R.W. *Space-Time Modeling of Urban Population Daily Travel-Activity Patterns Using GPS Trajectory Data*; Texas State University: San Marcos, TX, USA, 2018.
11. Hong, Z.; Chen, Y.; Mahmassani, H.S. Recognizing network trip patterns using a spatio-temporal vehicle trajectory clustering algorithm. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 2548–2557. [CrossRef]
12. Zhao, P.; Qin, K.; Ye, X.; Wang, Y.; Chen, Y. A trajectory clustering approach based on decision graph and data field for detecting hotspots. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1101–1127. [CrossRef]
13. Li, F.; Shi, W.; Zhang, H. A two-phase clustering approach for urban hotspot detection with spatiotemporal and network constraints. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3695–3705. [CrossRef]
14. Wu, B.; Wilamowski, B.M. A fast density and grid based clustering method for data with arbitrary shapes and noise. *IEEE Trans. Ind. Inform.* **2017**, *13*, 1620–1628. [CrossRef]
15. Jiang, Z.; Wang, M.; Chen, Y. Path recommendation based on geographic coordinates and trajectory data. *J. Commun.* **2017**, *38*, 165–171.
16. Qiao, S.; Han, N.; Ding, Z.; Jin, C.; Sun, W.; Shu, H. A Multiple-motion-pattern trajectory prediction model for uncertain moving objects. *Acta Autom. Sin.* **2018**, *44*, 608–618.
17. Zhang, D.; Lee, K.; Lee, I. Hierarchical trajectory clustering for spatio-temporal periodic pattern mining. *Expert Syst. Appl.* **2018**, *92*, 1–11. [CrossRef]
18. Yuan, G.; Sun, P.; Zhao, J.; Li, D.; Wang, C. A review of moving object trajectory clustering algorithms. *Artif. Intell. Rev.* **2017**, *47*, 123–144. [CrossRef]
19. Dong, S.; Liu, J.; Liu, Y.; Zeng, L.; Xu, C.; Zhou, T. Clustering based on grid and local density with priority-based expansion for multi-density data. *Inf. Sci. (NY)* **2018**, *468*, 103–116. [CrossRef]
20. Mao, Y.; Zhong, H.; Qi, H.; Ping, P.; Li, X. An adaptive trajectory clustering method based on grid and density in mobile pattern analysis. *Sensors* **2017**, *17*, 2013. [CrossRef]
21. Zheng, L.; Xia, D.; Zhao, X.; Tan, L.; Li, H. Spatial-temporal travel pattern mining using massive taxi trajectory data. *Phys. A Stat. Mech. Appl.* **2018**, *501*, 24–41. [CrossRef]
22. Liu, C.; Qin, K.; Kang, C. Exploring time-dependent traffic congestion patterns from taxi trajectory data. In Proceedings of the 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM), Fuzhou, China, 8–10 July 2015; pp. 39–44.
23. Pravilovic, S.; Appice, A.; Lanza, A.; Malerba, D. Wind power forecasting using time series cluster analysis. In *Discovery Science*; Springer International Publishing: Cham, Switzerland, 2014; pp. 276–287.
24. De Amorim, R.C.; Hennig, C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Inf. Sci. (NY)* **2015**, *324*, 126–145. [CrossRef]
25. Qiao, Y.; Cheng, Y.; Yang, J.; Liu, J.; Kato, N. A mobility analytical framework for big mobile data in densely populated area. *IEEE Trans. Veh. Technol.* **2017**, *66*, 1443–1455. [CrossRef]
26. Yu, Q.; Luo, Y.; Chen, C.; Bian, W. Neighborhood relevant outlier detection approach based on information entropy. *Intell. Data Anal.* **2016**, *20*, 1247–1265. [CrossRef]