International Journal of Geo-Information

MDPI

Article

Spatial–Temporal Analysis of Vehicle Routing Problem from Online Car-Hailing Trajectories

Xuyu Feng [1], Jianhua Yu [1], Zihan Kan [2], Lin Zhou [1], Luliang Tang [3] and Xue Yang [1,*]

[1] School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China; 20201003052@cug.edu.cn (X.F.); 1202021145@cug.edu.cn (J.Y.); zhoulin@cug.edu.cn (L.Z.)
[2] Department of Geography and Resource Management, Institute of Space and Earth Information Science, The Chinese University of Hong Kong, Shatin, Hong Kong, China; zihankan@cuhk.edu.hk
[3] State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; tll@whu.edu.cn
* Correspondence: yangxue@cug.edu.cn

Abstract: With the advent of the information age and rapid population growth, the urban transportation environment is deteriorating. Travel-route planning is a key issue in modern sustainable transportation systems. When conducting route planning, identifying the spatiotemporal disparities between planned routes and the routes chosen by actual drivers, as well as their underlying reasons, is an important method for optimizing route planning. In this study, we explore the spatial–temporal differences between planned routes and actual routes by studying the popular roads which are avoided by drivers (denoted as: PRAD) from car-hailing trajectories. By applying an improved Hidden Markov Model (HMM) map-matching algorithm to the original trajectories, we obtain the Origin-Destination (OD) matrix of vehicle travel and its corresponding actual routes, as well as the planned routes generated by the A* routing algorithm. We utilize the Jaccard index to quantify the similarity between actual and planned routes for the same OD pairs. The causes of PRADs are detected and further analyzed from the perspective of traffic conditions. By analyzing ride-hailing trajectories provided by DiDi, we examine the route behavior of drivers in Wuhan city on weekdays and weekends and discuss the relationship between traffic conditions and PRADs. The results indicate that the average accuracy of GNSS trajectory point-to-road map-matching reaches 88.83%, which is approximately 12% higher than the accuracy achieved by the HMM map-matching method proposed by Hu et al. Furthermore, the analysis of PRAD causes reveals that PRADs occurring on weekdays account for approximately 65% and are significantly associated with traffic congestion and accidents during that time. The findings of this study provide insights for future research on sustainable transportation systems and contribute to the development of improved route-planning strategies.

Keywords: car-hailing trajectories; vehicle-routing problem; popular roads avoided by drivers; map matching; spatial–temporal analysis

check for updates

Citation: Feng, X.; Yu, J.; Kan, Z.; Zhou, L.; Tang, L.; Yang, X. Spatial–Temporal Analysis of Vehicle Routing Problem from Online Car-Hailing Trajectories. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 319. https://doi.org/10.3390/ijgi12080319

Academic Editors: Wolfgang Kainz and Hartwig H. Hochmair

Received: 26 June 2023
Revised: 22 July 2023
Accepted: 28 July 2023
Published: 1 August 2023

1. Introduction

The transportation system serves as a crucial catalyst for urbanization and infrastructure development, playing a significant role in a country's economic growth and sustainability [1]. With the continued increase in the level of urban motorization, urban transportation issues have become increasingly severe. According to statistical data, as of the end of 2021 China's motor-vehicle ownership had reached 395 million, with 35 cities having more than two million motor vehicles (http://www.gov.cn/xinwen/2022-01/12/content_5667715.htm, accessed on 25 February 2022), leading to numerous traffic problems, particularly congestion on certain key transportation routes during specific periods. The development of public transportation systems is paramount for achieving environmental sustainability goals. In recent years, promoting green transportation in global cities has

become a focal point of attention [2,3].To alleviate the increasing congestion and improve traffic efficiency, the topic of analyzing drivers' route-choice behavior in both the spatial and temporal dimensions is crucial [4–6]. With the advancements in mobile technologies and location-based services (LBS), GNSS-enabled navigation systems play an important role in people's route-choice behaviors. A key function of such systems is route search/planning to the destination, based on the assumption that the planned route is optimal. The A* algorithm [7], as a classical path-planning algorithm, has been widely studied and received considerable attention. In recent years, new path-planning algorithms have also been proposed, such as a hybrid approach that combines Tabu search and the Artificial Bee Colony algorithm [8], a hybrid beetle swarm optimization algorithm (HBSO) [9], and the Colony Search Optimization Algorithm (CSOA) [10]. These algorithms exhibit strong scalability in the search space, but they also require significant computational resources. Sturtevan [11] provided a series of standardized test datasets for path-planning problems based on grid maps, which are used to evaluate the performance of different path-planning algorithms. Researchers have compared the efficiency and accuracy of the A* algorithm with other path-planning algorithms and found that the A* algorithm performs optimally in most cases.

However, the routes searched by algorithms do not always match people's actual choices of routes. Therefore, it is important to understand how well the planned routes match the actual routes and the factors that may contribute to the differences between them. Detecting the popular roads avoided by drivers by comparing the planned route and the actual route between the same OD pairs becomes a vital step towards grasping the characteristics of drivers' routing behaviors in both the spatial and temporal dimensions. Meanwhile, it can also be used to optimize vehicle routing algorithms to further reduce drivers' travel costs and alleviate traffic congestion.

In this study, we apply car-hailing trajectories to explore drivers' routing behaviors and further explore why drivers tend to avoid using some roads that are suggested by routing algorithms (e.g., the A* method) by identifying popular dodging roads. To acquire the popular dodging roads, we first match the raw car-hailing trajectories to the motor vehicle road network based on an improved Hidden Markov (HMM) map-matching algorithm. Then, the OD matrix is extracted according to the matched trajectories and the corresponding actual route that vehicles traveled is further obtained. By using the A* searching algorithm, we generate the planned route between each OD pair. We use the Jaccard index to quantify the similarity between the actual route and the planned route of the same OD pair and visualize the similarity. Next, we employ the Network-aware Trajectory Clustering (NEAT) method [12] to detect high-frequency PRADs by using clustering techniques, and their causes are further analyzed in relation to traffic jams and accidents. By using the massive car-hailing trajectories provided by the DiDi company in the city of Wuhan, we find that about 65% of PRADs correspond to serious traffic jams on workdays. The main contributions of this study include:

(1) We improved the original version of the HMM algorithm by optimizing the computations of angle feature in observation probability and velocity in transition probability to increase the accuracy of map matching and provide accurate results for the subsequent analysis of PRADs.
(2) We utilized the trajectory data of DiDi ride-hailing vehicles in Wuhan City as our dataset and employed the NEAT algorithm to identify PRADs. Through our analysis we made an interesting discovery; existing route-planning algorithms do not avoid heavily congested or accident-prone road segments during route planning, despite the fact that these segments lead to poor travel experiences in real-world trips. This finding sheds light on the reasons behind the existence of PRADs. Our research results provide new insights and evidence for the study of optimizing route-planning strategies.

2. Literature Review

2.1. Map Matching

The widespread adoption of Global Navigation Satellite System (GNSS) devices and the continuous development of positioning technology have generated a significant amount of GNSS trajectory data, which serve as an important data source for research on Location-Based Services (LBS) and Intelligent Transportation Systems (ITS). Currently, in-vehicle GNSS devices have overcome the issues of low coverage and high costs associated with dedicated sensors. However, these devices still exhibit problems such as deviation errors and low sampling rates during data collection, leading to a mismatch or substantial differences between GNSS trajectory data and road network data. Consequently, research applications such as route planning, traffic flow analysis, geographic social network analysis, autonomous vehicles, and trajectory anomaly detection are hindered, necessitating the use of map matching to associate trajectory points with road network locations and obtain matched trajectory point positions and information, thus constituting the map-matching process. Therefore, map matching is an essential data processing step in various Location-Based Services (LBS) and Intelligent Transportation Systems. Consequently, numerous researchers employ map-matching algorithms to establish correspondences between the original GNSS trajectories and road networks [13,14]. Quddus et al. [15] categorized map-matching methods into four types: (1) geometry-based algorithms primarily focus on the geometric characteristics of road networks and trajectories [16], but they overlook topological information, resulting in confusion or errors in matching for complex road network scenarios such as overpasses; (2) topology-based algorithms emphasize the topological relationships between trajectory data and road networks [17] (these first two methods establish a graph structure based on the road network and integrate topological information, but they are not suitable for low-sampling-rate trajectory data); (3) probability-based algorithms treat GNSS positions and trajectories as random variables and stochastic processes, with Hidden Markov Models (HMMs) being commonly used [18] (these algorithms consider both geometric and topological information and do not require training data, yielding excellent application results; however, due to the need for computing the shortest path, they incur certain computational costs); (4) algorithms based on advanced mathematical theories, such as fuzzy logic [19], neural networks [20], Kalman filters [21], and Dempster–Shafer theory [22]. These algorithm-based methods require large training datasets for point-by-point matching, making their application challenging. Additionally, treating trajectory data simply as collections of independent and identically distributed random variables in a stochastic process is highly unreasonable.

In summary, the Hidden Markov Model (HMM) is a preferable choice for map matching due to its ability to consider geometric and topological information, treat trajectories as complete stochastic processes, and not require training datasets. The HMM has been widely applied in map matching in recent years. Hu et al. [23] considered trajectory direction factors based on the HMM algorithm by incorporating direction angles into the calculation of observation probabilities, which improved the accuracy of trajectory-point matching. However, they did not consider trajectory velocity factors, and their calculation of direction-angle probabilities was complex. Paul Newson was the first to use the Hidden Markov Model for map matching [18], using the normal distribution to calculate observation probabilities based on GNSS error distances and the exponential probability distribution to calculate transition probabilities based on road network distances. Anders Hansson et al. [24] considered distance, velocity, heading, lane-deviation rate, and lane markings to calculate observation and transition probabilities, applying this method to lane-level road network map matching, which heavily relies on the quality of trajectory and road network data. Therefore, to address these issues, this study improves upon Hu's work by first optimizing the calculation of direction angles in the observation-probability model and then considering the trajectory velocity factor and incorporating it into the calculation of the transition-probability model. The improved HMM algorithm shows enhanced matching accuracy compared to Hu's algorithm, particularly for trajectory points collected at complex

intersections, and it can effectively and accurately map-match large-scale, low-frequency GNSS trajectory data.

2.2. The Issue of Vehicle Route-Choice Behavior

At present, most of the research on drivers' routing behavior mainly depends on the stated-preference (SP) surveys or data collected by small-scale experiments, which are limited by the number of participants involved [25,26]. Approaches to mining routing behavior focus on discrete choice models, e.g., the multinomial logit (MNL) model [27],the CNL (Cross-nested logit) model [28], and the GEV (Generalized extreme value) model [29]. The differences between these models are manifested in the characteristics of the datasets, explained variables, and model structures [30]. For example, Dial et al. [27] proposed a discrete multinomial logit (MNL) model for multimode selection. To address the independence of irrelevant alternatives (IIA) problem of the MNL model, many studies have developed some new models based on the MNL model by adding a modification section to represent the interactions between different routes, such as the C-logit model [31] and the PS-logit model [32]. Apart from this, some studies have demonstrated the use of the developed CNL (Cross-nested logit) model [28] and PCL (the paired combinatorial logit) model [33] to avoid the IIA issue of the MNL model based on the GEV principle (McFadden, 1978) [29]. However, these previous studies mainly analyzed drivers' routing behaviors depending on a small amount of survey data, which is time consuming and biased due to the limited data collected.

With the rapid development of information and communication technologies, positioning technologies and the collection or storage capabilities of massive data, have advanced the application of GNSS trajectories in the field of transportation, such as in travel-time estimation [34], risk assessment of driving behaviors [35,36], departure time modeling [30], route-choice behavior analysis [30,37–40], etc. Among them, vehicle routing driven by large tracking datasets has been improved in both effectiveness and accuracy. For instance, Kim et al. [41] established a framework for clustering and categorizing vehicle trajectories to analyze vehicles' travel patterns in space and time. Lu et al. [37] developed a visualization system to help users deal with the massive trajectories and discover the causes of route selections. Based on their study, the contributing factors of routing problems included route-related elements (e.g., route length, traffic light number, route importance, time-cost distribution) and trajectory-related elements (e.g., departure time in a day, departure day, trajectory length). Li et al. [42] discovered the effect of heterogeneity of route selection in relation to drivers' ages and genders, engine capacity, and the characteristics of the OD matrix, by using the vehicle trajectories collected by private cars in Toyota City. Deng et al. [30] applied taxi trajectories to explore route selection behavior based on heterogeneous travel distances. They first used the DBSCAN (Density-based spatial clustering of application with noise) algorithm and AIC (Akaike information criterion) to categorize travel distances into several types, and built the PS-logit model by defining nine explanatory variables to analyze the heterogeneity of trips of varying distance. In summary, most studies which have analyzed drivers' routing behavior by using trajectories have focused on exploring the causes of route selection, while lacking an understanding of the differences between planned routes and actual routes.

3. Methodology

The methodological framework of the spatial–temporal analysis for vehicle routing from the aspect of PRAD detection is constructed by using the car-hailing trajectories, as shown in Figure 1.

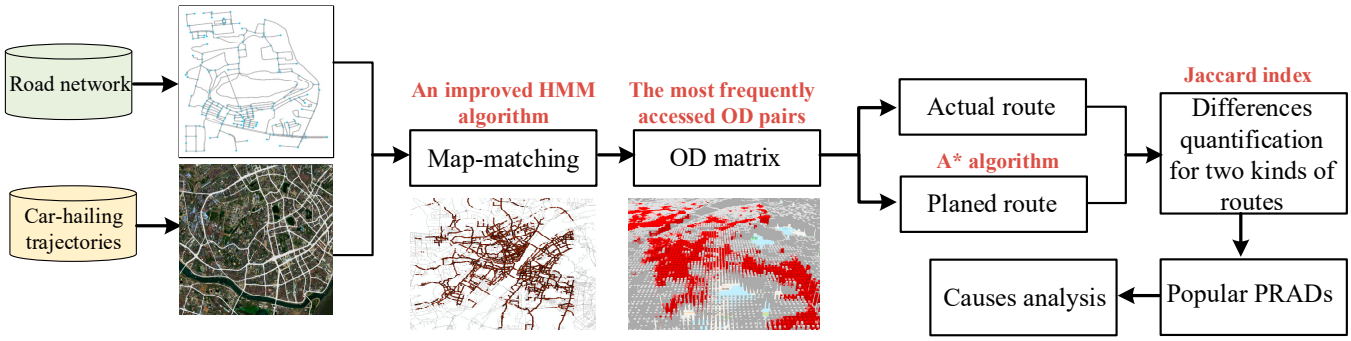


Figure 1. Methodological framework of spatial–temporal analysis for vehicle routing problem from online car-hailing trajectories.

In this study, two kinds of spatial data were used to analyze the PRADs. The first was car-hailing trajectory data collected by residents who worked as part-time drivers. Specifically, a trajectory comprises a set of corresponding tracking points and can be denoted as $Tra = (p_1, \ldots, p_n)$, where n is the number of tracking points belonging to the trajectory. Each tracking point records the location (e.g., longitude and latitude), time, speed, and heading direction of the moving objects, denoted as p_i $(x_i, y_i, t_i, s_i, a_i)$, $i = 1, 2, \ldots, n$. The second type of spatial data used in this study is the road network of the study area, including road segments, nodes, topology information, and the direction of traffic flow.

3.1. Map Matching Based on the Improved HMM Algorithm

Map matching is the first step for exploring driving behavior from car-hailing trajectories. Its detailed mechanisms have significant effects on the results of drivers' routing behaviors. Quddus et al. [15] summarized the existing map-matching algorithm into four categories, including geometry-related methods [16], topology-based methods [17], probability-based methods [18], and mathematical methods [19–21,23]. Among these kinds of methods, HMM-based map matching has been widely applied because it does not need to train data and considers the features of trajectories and road networks both in the geometry and topology. Hu et al. [23] added the driving direction to the computation of observation probability based on the original model of HMM to improve the accuracy of map matching. However, their method was limited by the complexity of the direction-angle probability calculation and poor performance for low-frequency tracking-point matching. To address these issues, we improved the computational model of observation probability and transition probability, based on the work conducted by Hu et al. (2019) [23], to enhance the accuracy of map matching, especially for tracking points collected at complex road intersections.

For a trajectory $Tra = (p_1, \ldots, p_{i-1}, p_i, p_{i+1}, \ldots, p_n)$, assuming p_{i-1}, p_i, and p_{i+1} are candidates for map matching and s_{i-1}^k, s_i^k, s_i^k correspond to their state points. The key idea for HMM-based map matching is to compute the observation probability and transition probability of tracking points based on their corresponding state points. The computational process of observation probability and transition probability of tracking points are conducted in two layers, including the observation layer and state layer. Specifically, the observation probability quantifies the possibility of tracking points matched with the state of candidates. For most of HMM-based map-matching algorithms, the observation probability was computed based on the distance from the candidate tracking point to the road network [23,24]. In this study, we improved the computational method of observation probability for the angle feature by adding the angle between tracking points and directed road segments (see Equation (1)). So, the observation probability from the aspect of the angle feature can be calculated based on Equation (2).

$$a = \begin{cases} |\beta - \gamma| & |\beta - \gamma| < 180^\circ \\ 360^\circ - |\beta - \gamma| & |\beta - \gamma| \geq 180^\circ \end{cases} \quad (1)$$

$$P_{angle}\left(o_i|s_i^k\right) = {\cos(a)+1}\Big/{2} \tag{2}$$

where β indicates the heading angle of tracking point p_i, γ represents the angle of the candidate matching road segment with the direction of north, and a is the difference between β and γ. The parameter $P_{angle}(o_i \mid s_i^k)$ represents the observation probability of the observation point o_i and its corresponding candidate state point s_i^r from the perspective of the angle.

The computation method for observation probability in the aspect of distance is the same as the original version of the HMM-based map-matching algorithm (Paul and John, 2009) [18]. The comprehensive observation probability (denoted as $P_{o\_dis\_ang}$ $(o_i \mid s_i^k)$) both in the aspects of angle and distance is calculated based on Equation (3), where P_{dis} $(o_i \mid s_i^k)$ represents the observation probability of the observation point o_i and its corresponding candidate state point s_i^k from the perspective of distance, ω_d and ω_a are their weight, respectively, and $\omega_d + \omega_a = 1$.

$$P_{o\_dis\_ang}\left(o_i\middle|s_i^k\right) = \omega_d P_{dis}\left(o_i\middle|s_i^k\right) * \omega_a P_{angle}\left(o_i\middle|s_i^k\right) \tag{3}$$

The transition probability quantifies the possibility of the state point of the previous tracking point changing to the state of the current tracking point. The existing research on the HMM-based map-matching algorithm mainly considers the distance feature of tracking points [18,43,44]. In this study, we added the speed of the tracking point to the computational method of transition probability based on the observation that the speed restrictions of different kinds of roads are different. For example, the driving speed of a ramp in China is limited to 40 km/h which is lower than its adjacent main road's 60 km/h. The transition probability in the perspective of speed can be calculated according to Equation (4), where v_{i-1} and v_i denote the speed of the previous tracking point p_{i-1} and the current tracking point p_i, respectively. The parameter in Equation (4) represents the average speed from the candidate state point s_{i-1}^t of the tracking point p_{i-1} to the candidate state point s_i^r of the tracking point p_i. Here, the distance from s_{i-1}^t to s_i^r is the network distance which is obtained based on the shortest routing algorithm A* [45]. The transition probability in the perspective of distance is the same as with the original version of the HMM-based map-matching algorithm [18]. Furthermore, the comprehensive transition probability, both in the aspects of speed and distance, can be calculated based on Equation (5), where P_{t_dis} $(s_{i-1}^t \mid s_i^r)$ denotes the transition probability of the candidate state point s_{i-1}^t of tracking point p_{i-1} and the candidate state point s_i^r of tracking point p_i, ω_{t_dis}, and ω_{t_speed} are their weight, respectively, and $\omega_{t\_dis} + \omega_{t\_speed} = 1$.

$$P_{t\_v}\left(s_{i-1}^t \to s_i^r\right) = \frac{(v_{i-1}+v_i)/2}{\overline{v}_{(i-1,t)\to(i,r)}} \tag{4}$$

$$P_{t\_dis\_speed}\left(s_{i-1}^t \to s_i^r\right) = \omega_{t\_dis} P_{t\_dis}\left(s_{i-1}^t \to s_i^r\right) * \omega_{t\_speed} P_{t\_speed}\left(s_{i-1}^t \to s_i^r\right) \tag{5}$$

3.2. Identification for PRAD

The trajectory data used in this study were collected by car hailing company DiDi. Each trajectory records the actual route of a vehicle with passengers between an OD pair. Based on the improved map-matching algorithm, we can obtain the OD pairs of trajectories that are matched to the road network. We calculate an OD matrix of travel time and distance based on the OD pairs. Then, the actual routes of all OD pairs are extracted based on the map-matching results. The corresponding planned routes between the OD pairs are obtained based on the A* routing algorithm because of its performance [45]. Meanwhile, the travel cost of an OD pair during routing planning using the A* routing algorithm is decided by travel distance and time. It should be noted that the travel distance and time are obtained based on the network distance and speed restrictions of different roads.

For analyzing the differences between actual routes and planned routes of OD pairs, we apply the Jaccard index to quantify their similarity. The Jaccard index (JI), also known as the Jaccard similarity coefficient, is mainly used for comparing the differences or similarities of two finite sample sets [46] and calculated based on Equation (5), where A and B represent two finite sample sets, respectively. The larger the JI is, the higher the sample similarity is, and the smaller the JI is, the lower the sample similarity is. In this study, the actual routes and their corresponding planned routes are regarded as the sample sets A and B. The spatial distribution and OD clustering are further visualized by dividing the intervals of the JI values.

$$\mathrm{JI}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A}| + |\mathrm{B}| - |\mathrm{A} \cap \mathrm{B}|} \tag{6}$$

We define the PRADs as road segments which are included in the planned route but do not appear in the corresponding actual route. Thus, a PRAD is obtained by comparing the actual route and the planned route of a same OD pair. As shown in Figure 2, 'road_1' and 'road_2' are two-way roads, and 'road_3' is a one-way road. Based on the map-matching results, the actual route of trajectory *trai* = $(\ldots, p_i, p_{i+1}, \ldots)$ includes 'road_1' and 'road_2' (see Figure 2). However, the planned route of trajectory *trai* contains 'road_1' and 'road_3'. Thus, for trai, 'road_3' is its dodging road. For different trajectories with different OD pairs, their PRADs may also be different. To explore the spatiotemporal pattern of drivers' traveling, we need to identify the PRADs first.

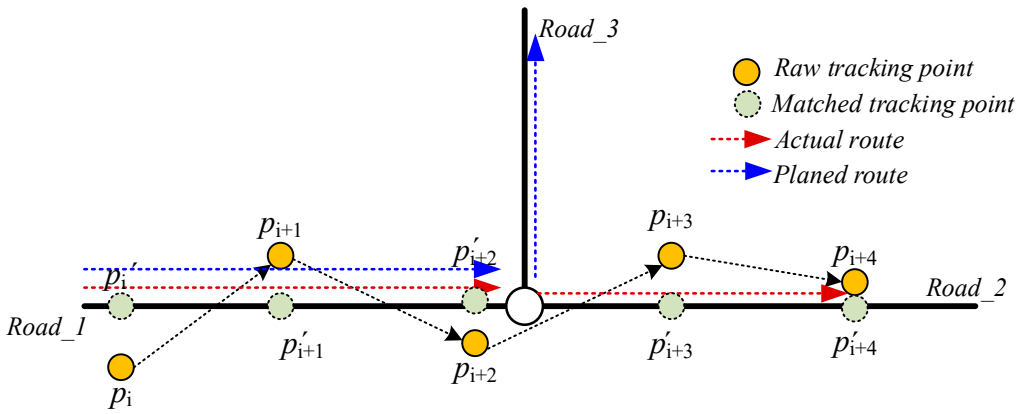


Figure 2. Definition of a PRAD.

The NEAT (road NEtwork Aware Trajectory Clustering) algorithm is a trajectory-based clustering algorithm that incorporates road network information into the clustering process to better identify trajectories with similar movement patterns. The detailed steps of the NEAT algorithm are described as follows: (1) Data preprocessing: In the NEAT algorithm, the trajectory data needs to be preprocessed initially. This preprocessing step involves operations such as noise removal and data cleaning of trajectory points to extract relevant trajectory features effectively. (2) Feature extraction: For each preprocessed trajectory, a set of feature vectors needs to be extracted. The NEAT algorithm employs two feature extraction methods: (1) Point-based feature extraction: The coordinates of trajectory points are used as feature vectors, where each point's latitude and longitude serve as feature dimensions. (2) Segment-based feature extraction: The direction and distance between trajectory points are used as feature vectors, where the direction and distance of each segment of movement serve as feature dimensions. (3) Clustering method: The NEAT algorithm utilizes a density-based clustering method called DBSCAN (Density-Based Spatial Clustering of Applications with Noise). DBSCAN classifies data points into core points, boundary points, and noise points, and it can handle clusters of arbitrary shapes. (4) Road network constraint. During the clustering process, the NEAT algorithm also takes into account the information of the road network. Specifically, the NEAT algorithm treats the road network as a graph structure and maps trajectory points to the nearest road.

During the clustering process, it merges trajectory points that are distant from each other but located on the same road. This approach avoids erroneously separating trajectories of travel on the same road into different clusters.

The NEAT algorithm incorporates road network information into the clustering process, enabling better identification of trajectories with similar movement patterns. It ensures road continuity by assigning trajectory points on the same road to the same cluster. Therefore, to identify popular detour segments, we applied a clustering method based on the NEAT algorithm instead of directly calculating the frequency of driver detours and estimating popular segments based on descending order. This is because the occurrence of detour segments several times does not necessarily imply that they are popular segments avoided by drivers, as chance factors may be involved. In addition, this study faces the challenge of determining an appropriate threshold to define popular detour segments based on their occurrence frequency. Furthermore, some detour segments may be overlooked due to their similar traffic directions.

To address these issues, we first group the PRADs into a set of clusters based on their location. As shown in Figure 3, assuming RSi represents the road network segment, $i = 1, 2, \ldots, 5$, among them RS5 is a two-way lane, and RS1 to RS4 are one-way lanes. The parameter DS_i represents a set of clusters of PRADs on the corresponding RSi segment. That is, for two-way lanes, PRADs in two directions on the same road are grouped into the same cluster, and PRADs in the same direction on the same section of the one-way lane are grouped into another cluster. For example, RS1 and RS3, RS2 and RS4, are all connected with the same intersections. In this study, we group DS_i in this case into the same cluster. Then, we improve the clustering method in the third stage of NEAT (road NEtwork Aware Trajectory clustering) proposed by Han et al. [12] to cluster all DS groups and detect the road segments avoided by drivers. Specifically, the improvement mainly includes: (1) the clustering unit is a road segment with the same direction of traffic flow; (2) the distance between two clustering units is calculated by using the Hausdorff distance [47]; (3) the threshold of clustering is adaptively acquired based on the input dataset by using the method proposed by Lee et al. [48]. Based on the clustering results, the clusters of DS_i shown in Figure 3 will be identified as PRADs if they satisfy the clustering threshold and vice versa.

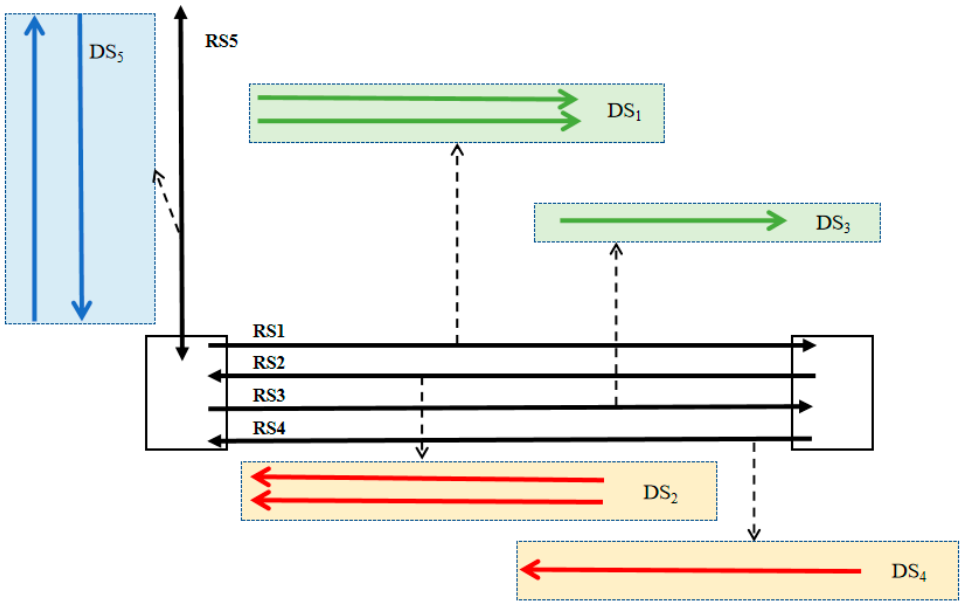


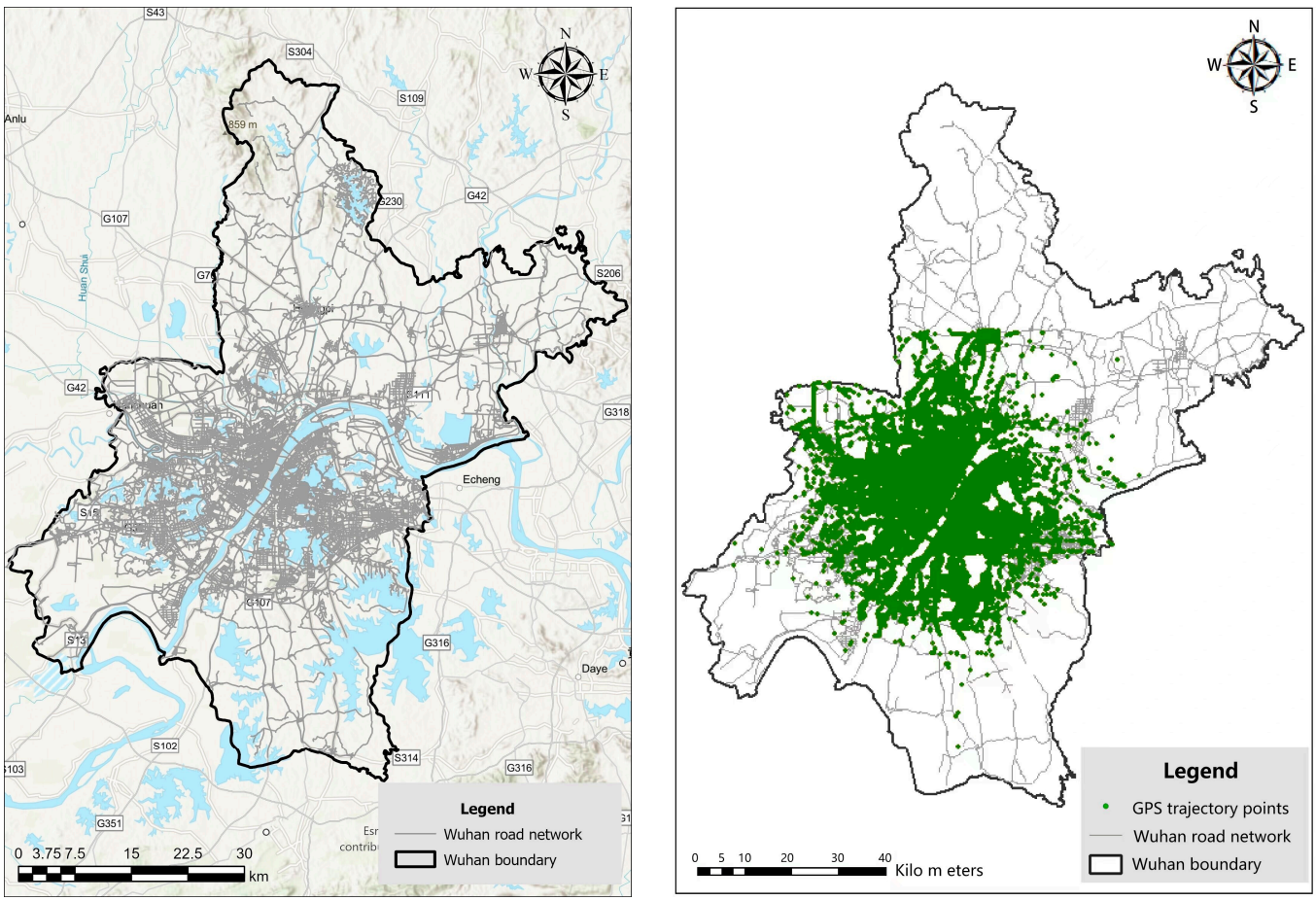
Figure 3. PRAD clustering.

4. Case Study: Vehicle-Routing Behavior Analysis in the City of Wuhan

4.1. Data Collection and Map Matching

4.1.1. Car-Hailing Trajectories Collection and Preprocessing

Taking Wuhan as the experimental area, trajectories collected by 300,000 car-hailing taxis belonging to the DiDi company, from 8–16 August 2017, were used to analyze the spatiotemporal pattern of vehicle-routing problems. The sampling interval and positional accuracy of these trajectories range from 30 s to 120 s and from 5 m to 20 m, respectively. Each tracking point records the information of the current moving object, including latitude, longitude, time, speed, and heading angle. The average amount of tracking records every day was about 600,000 in the experimental region and each trajectory is composed of about 200 tracking points. These car-hailing trajectories cover the main districts of the city of Wuhan, as shown in Figure 4. In addition, there are outliers in the trajectory data caused by signal drift or irregular driving behavior. Here, we applied the method proposed by Yang et al. (2018) [49] to remove outliers from the raw trajectories. The experimental results show that about 35.95% of tracking points were removed. Apart from the trajectory data, road networks of motor vehicles used in this study were acquired from the platform of OSM (OpenStreetMap). Based on the statistics, there are about 77,086 road segments located in the experimental region. Road networks obtained from the OSM platform were provided by volunteers, so there also existed issues such as topological errors. This study applied the method proposed by Hu et al. (2019) [23] to revise these topological errors. Based on the statistics, about 315 road segments had topological errors which were corrected.



(**a**) (**b**)

Figure 4. Study region and data diagram. (**a**) Study area and road network; (**b**) GPS trajectory points.

4.1.2. Map Matching Based on the Improved HMM Algorithm

The processed trajectories were matched to the road network by using the improved HMM algorithm. Based on the principle of the map-matching algorithm proposed in this study, we need to set the value of weights of ω_d, ω_a, ω_{t_dis} and ω_{t_speed}. To obtain the optimum value of these weights, we randomly selected 200 road segments and manually estimated the matching accuracy of tracking points by tuning the value of them from zero to one, as shown in Table 1. In Table 1, we find that the values of ω_d, ω_a, ω_{t_dis} and ω_{t_speed} were set, respectively, at 0.7, 0.3, 0.7, and 0.3 with maximum accuracy of map matching. The accuracy results in Table 1 illustrate that the accuracy of map matching for vehicle trajectories with a low sampling rate was closely related to the distance from the observation point to the candidate state point.

Table 1. Matching accuracy of the improved HMM algorithm with different values of parameters.

ω_d	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ω_a	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
ω_{t_dis}	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ω_{t_speed}	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Matching accuracy (%)	69.8	70.6	75.2	77.7	80.3	83.5	88.3	83.8	79.5

To verify the effectiveness of the improved HMM algorithm. We randomly selected 35 trajectories and compared the map-matching results with the method proposed by Hu et al. (2019) [23]. Specifically, we converted the original trajectories into several kinds of trajectories with fixed sampling rates such as 30 s, 30–60 s, and more than 60 s by using cubic spline interpolation and the Douglas–Peuker compress method. Then, these processed trajectories were matched to the road network and we manually estimated the average value, variance, and standard deviation of the accuracy of map matching. As shown in Figure 5, the left panel of box plots indicate the results by using the method proposed by Hu et al. (2019) [23] and the right panel of box plots demonstrate the matching results using the improved HMM method which we optimized based on the work conducted by Hu et al. (2019) [23]. The experimental results show that the average value, variance, and standard deviation of the accuracy of map matching of three kinds of sampling rate trajectories by using the method proposed by Hu et al. (2019) [23] are lower than those based on the improved HMM algorithm in this study. This also means that the distribution of correct matching results based on the improved HMM algorithm is more concentrated. The experimental results mentioned above indicate that the average accuracy of GNSS trajectory-point matching with road network maps reaches 88.83%, demonstrating an improvement in accuracy of approximately 12% compared to the HMM-based map-matching method proposed by Hu et al. (2019) [23].

Beyond that, experimental results illustrate that the matching results by using the algorithm proposed in this paper are more suitable for road segments located on complex road intersections. As we can see in Figure 6, trajectories collected at a roundabout were matched based on the improved HMM method in this study (Figure 6b) and the HMM algorithm proposed by Hu et al. (2019) [23], (Figure 6a). Based on the manual inspection, the raw tracking points were collected in Luoyu Road, Lumo Road, and the Guanggu Roundabout which is connected to these two roads. After map matching, these tracking points should be matched to these road segments. However, only three tracking points were correctly matched to the road where they were collected by using the method proposed by Hu et al. (2019) [23], as shown in Figure 6a. The other tracking points of the raw trajectory shown in Figure 6a were regarded as matching failures and abandoned. By comparing with Hu's method, our algorithm can correctly match all raw tracking points to the correct places (Figure 6b).

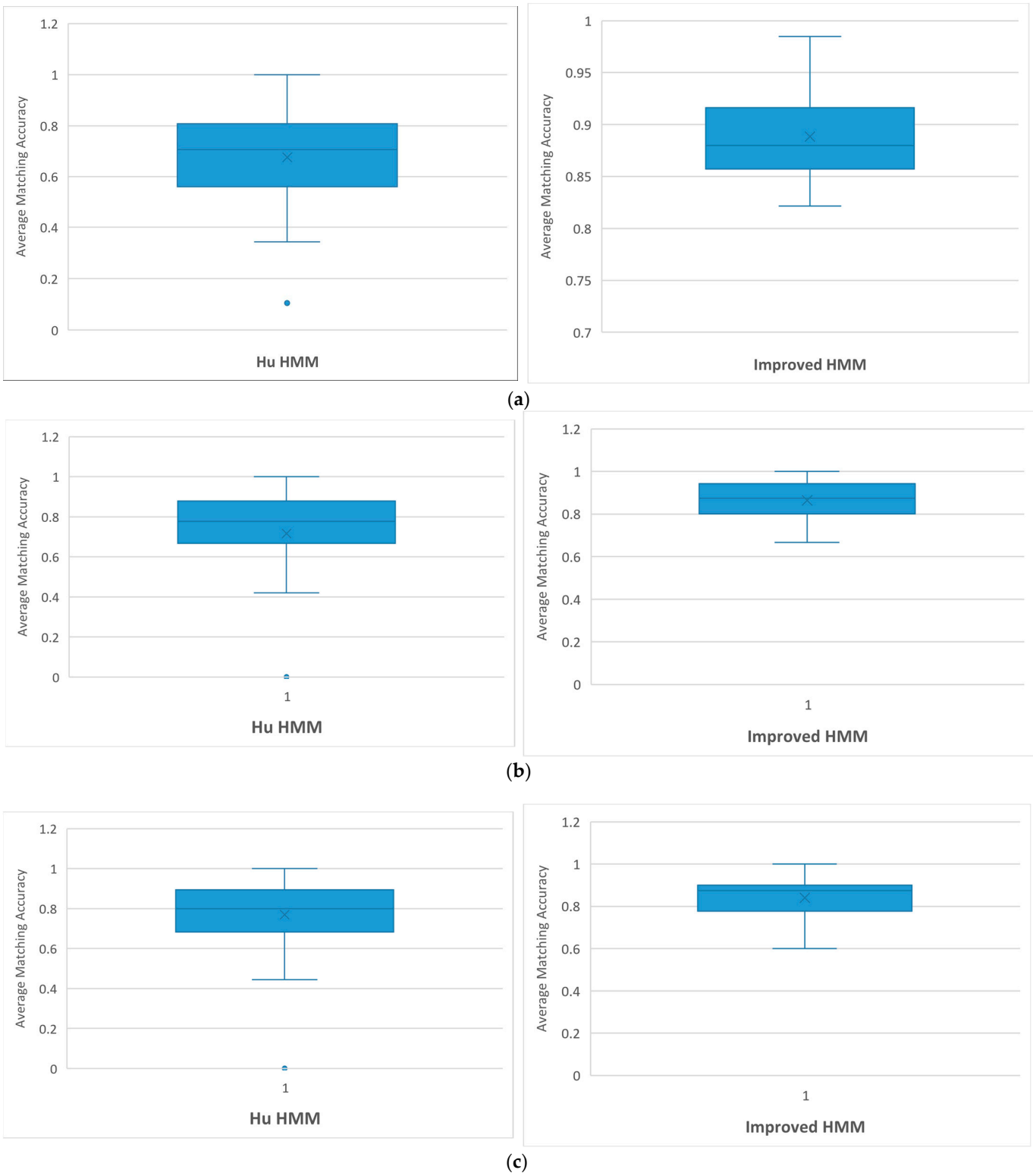


Figure 5. Comparison of map-matching results based on the HMM algorithm proposed by Hu et al. (2019) (**left**) and the improved HMM algorithm (**right**) proposed in this study. (**a**) 30 s sample interval (left mean: 0.676; var: 0.045; std: 0.212, right: mean: 0.888; var: 0.002; std: 0.043); (**b**) 30–60 S sample interval (left mean: 0.717; var: 0.064; std: 0.253, right: mean: 0.864; var: 0.011; std: 0.103); (**c**) >60 S sample interval (left mean: 0.770; var: 0.048; std: 0.218, right: mean: 0.84; var: 0.013; std: 0.112).

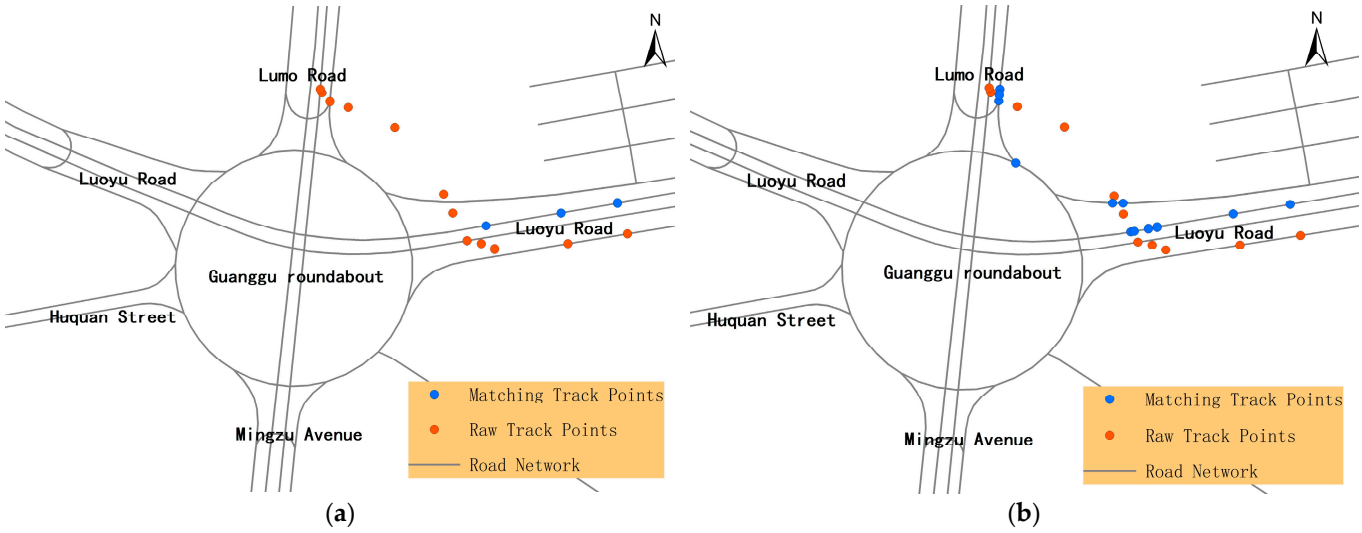


Figure 6. Comparison of results at complex road intersections. (**a**) Hu HMM; (**b**) Our HMM.

4.2. PRAD Visualization and Analysis

4.2.1. JI Value Categorization and Visualization

Based on the map-matching results, we computed the JI value of an OD pair to estimate the similarity between its planned routes and the corresponding actual routes. Since massive trajectories will bring the problem of poor expression in the visualization of the JI value, we applied kernel density statistics to divide the JI value into several classes to facilitate the visualization. Then, we visualized the JI value of all OD pairs according to their categorization. Figure 7 shows the kernel density distribution of the JI value of car-hailing trajectories which were collected, respectively, on 12 and 15 August 2017.

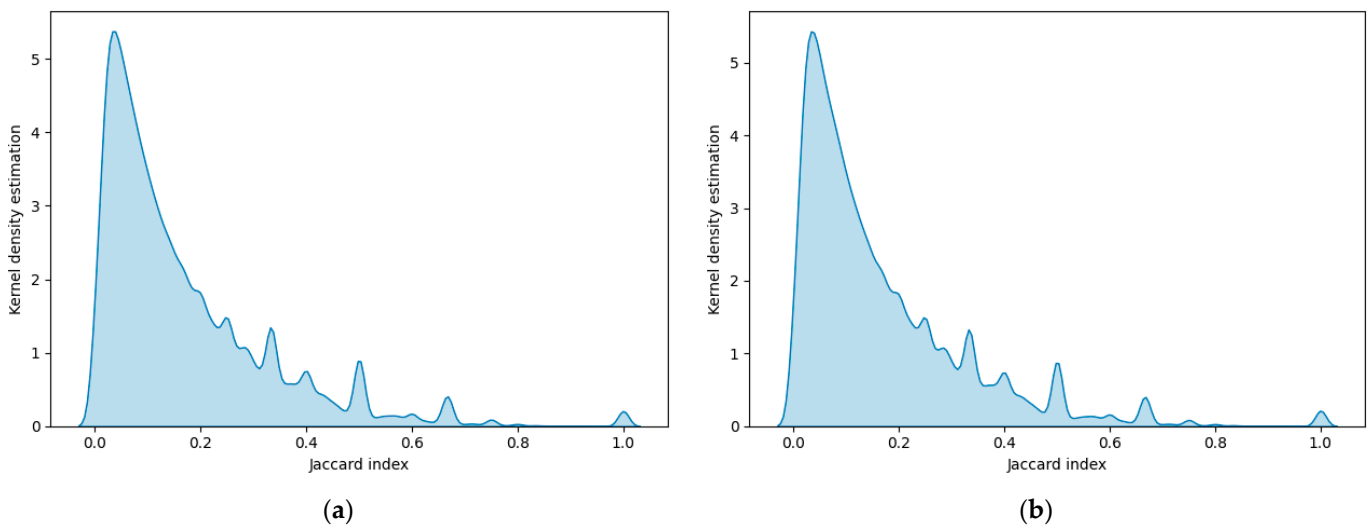


Figure 7. Kernel density distribution of JI values for GNSS trajectories. (**a**) 12 August 2017 Jaccard index value distribution; (**b**) 15 August 2017 Jaccard index value distribution.

In Figure 7, we can find that the trends of the JI value of car-hailing trajectories obtained on two different dates are roughly the same, and most of the JI values concentrate around 0.1. Further, there is almost no JI value in the range between 0.8 and 0.9, which means that most actual routes are significantly different from the corresponding planned routes. Meanwhile, we can see several peaks in the distribution of the JI values at 0.1, 0.2, 0.4, 0.6, and 1.0. The number of the JI values distributed within the interval of [0, 0.1] is highest, and the densities of other ranges including (0.1, 0.2], (0.2, 0.4], (0.4, 0.6], and (0.6,1.0)

show a descending order. Based on the kernel distribution pattern of JI values of all OD pairs, we divided the JI values into five intervals of [0, 0.1], (0.1, 0.2], (0.2, 0.4], (0.4, 0.6], and (0.6, 1.0), respectively. Based on this, we define the similarity between the actual routes and planned routes of five kinds of JI value as 'totally different', 'different', 'slightly different', 'similar', and 'no difference'. These different categories of JI values were visualized by the platform of Kepler. Figure 8 shows the visualization results of the JI values of trajectories collected on 12 August 2017. The statistics show that the proportion of JI values within the interval of [0, 0.1], (0.1, 0.2], (0.2, 0.4], (0.4, 0.6], and (0.6, 1.0) of Figure 8 is 45.54%, 24.85%, 21.00%, 6.21%, and 2.40%, respectively. This result indicates that the proportion of actual routes which are totally the same as the planned routes is very small.

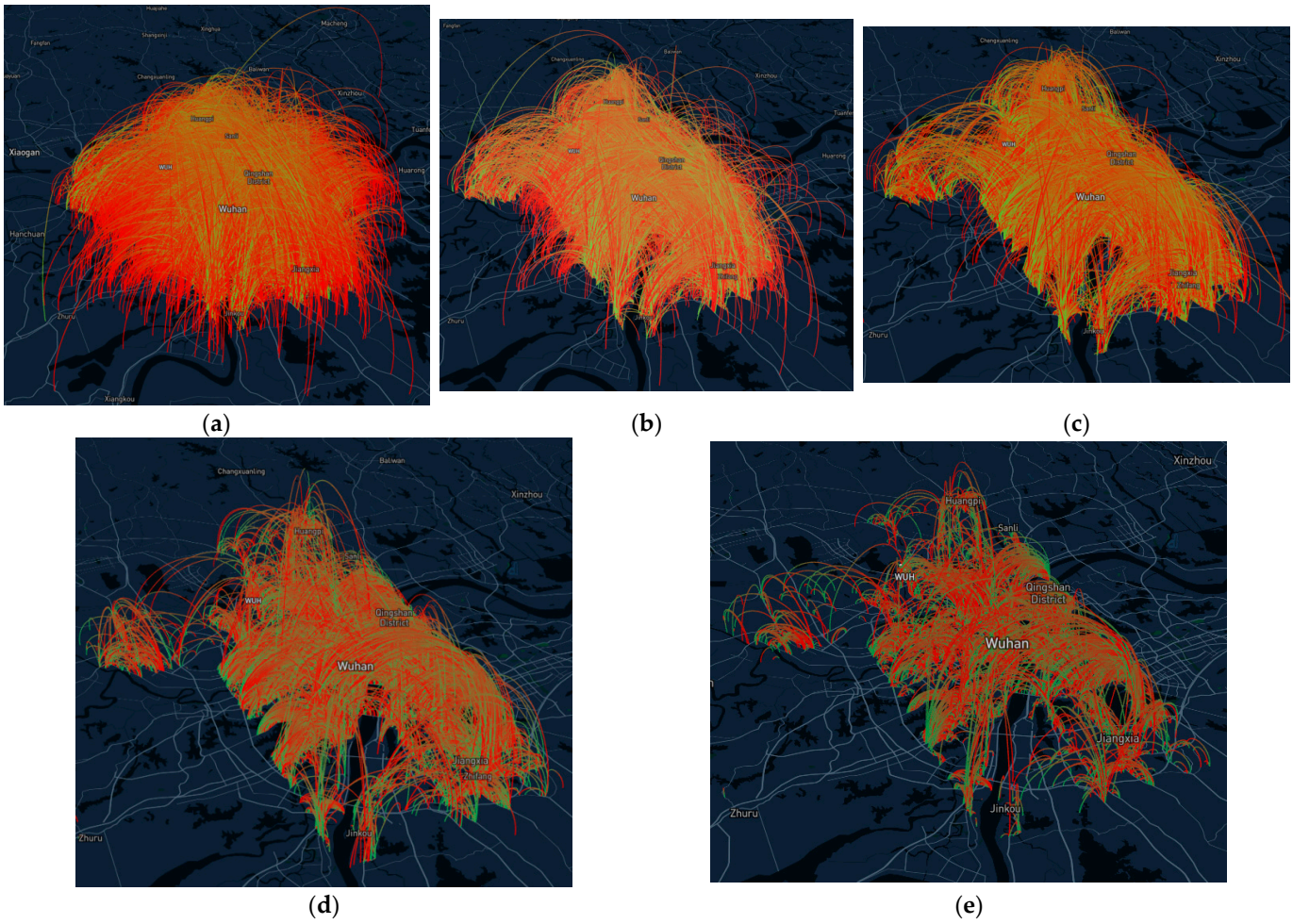


Figure 8. Spatial distribution patterns of JI values of trajectories collected on 12 August 2017. In the figure, each line is divided into red and green halves, representing the direction from the starting point (red) to the destination point (green): (**a**) JI values with the type of 'totally different', (**b**) JI values with the type of 'different', (**c**) JI values with the type of 'slightly different', (**d**) JI values with the type of 'similar', (**e**) JI values with the type of 'no difference'.

To further analyze the pattern of each type of OD pair of JI values, we clustered their own OD pairs by using the FlowmapBlue (https://flowmap.blue/, accessed on 25 March 2022) method which is a free tool for illustrating aggregated numbers of movements between geographic locations as flow maps. Figure 9 illustrates the clustering results of OD pairs that belong to each category of JI values. Specifically, the size of the dot shown in Figure 9 represents the number of OD pairs connected to it. The larger the dot, the greater number of OD pairs connected to it. Furthermore, the arrows indicate the direction from the origin point to the destination. Similarly, the size and brightness of these arrows

also indicate the volume of traffic flow from the origin point to the destination. The volume of traffic flow is proportionate to the size or brightness of the arrows. As we can see from Figure 9, the distribution of OD pairs becomes gradually dispersed as the JI values decrease, although most of OD pairs still gather at the central area of the experimental region (see Figure 9a). This indicates that the larger the travel distance, the greater the difference between the actual travel route and the planned route. When the travel distance is large, there are more alternative routes to avoid certain risks (such as traffic congestion), thus, the possibility of inconsistency with the planned route is higher. So, the route-searching algorithm should consider or improve the accuracy of long-distance trips. In contrast, when the driving distance is smaller, avoiding certain risks may lead to higher costs of travel distance or increased travel time, thus it is less likely to be inconsistent with the planned route. Figure A1 in Appendix A presents the visualization results of JI values for trajectories collected on 15 August 2017. Figure A2 displays the OD clustering results for the same date. Likewise, Figure A3 depicts the relationship between the JI values and the distances between OD pairs on the same day. The features of the JI values and OD pairs clustering results shown in the Appendix A are similar to the results of Figures 8 and 9. That means that the traveling patterns of car-hailing drivers every day are similar.

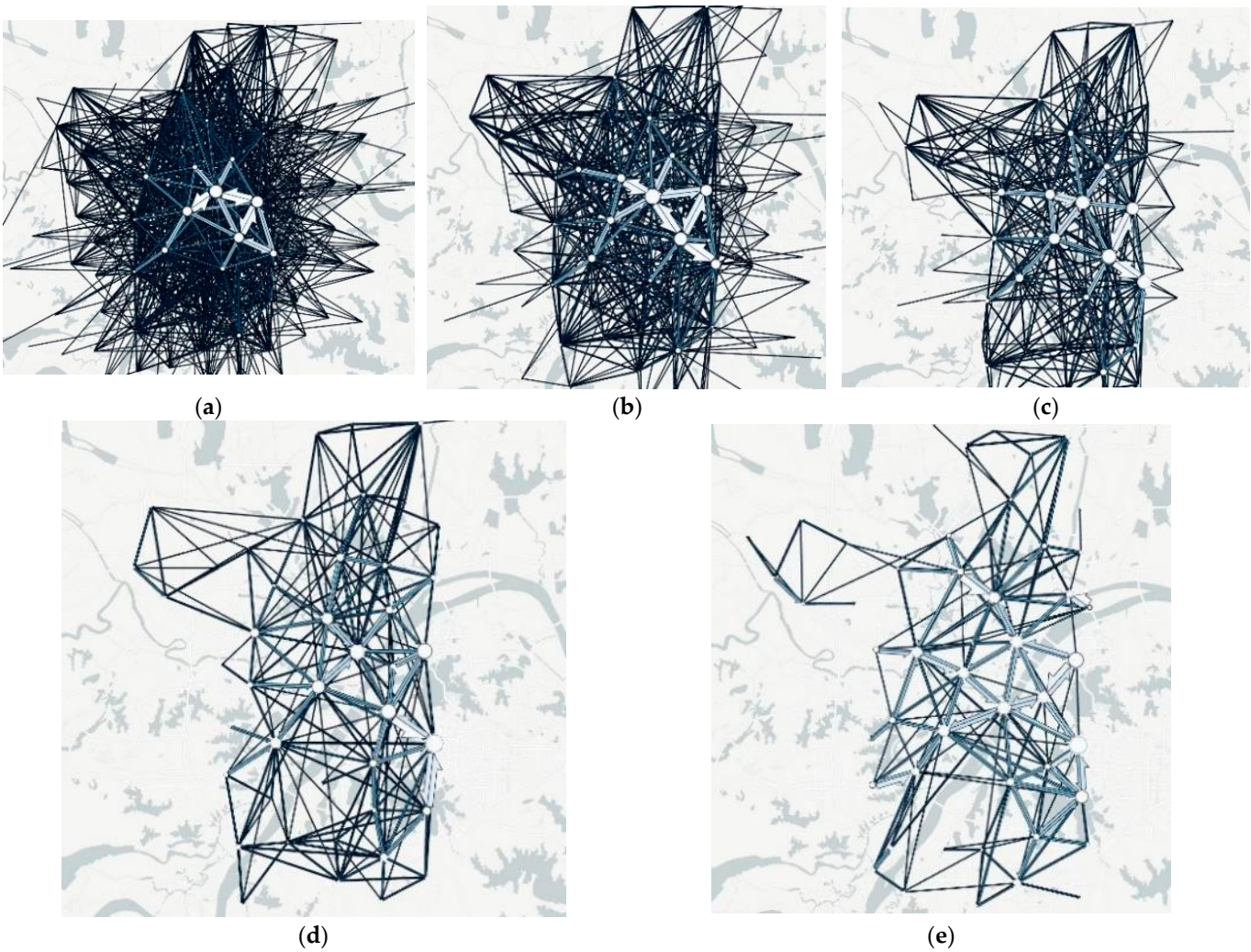


Figure 9. OD clustering results on 12 August 2017, (**a**) OD pairs of JI values within [0, 0.1], (**b**) OD pairs of JI values within (0.1, 0.2], (**c**) OD pairs of JI values within (0.2, 0.4], (**d**) OD pairs of JI values within (0.4, 0.6], (**e**) OD pairs of JI values within (0.6,1.0).

Figure 10 shows the relationship between the JI values and the distance of each OD pair. In Figure 10, we can find that the JI values of OD pairs decrease as the traveling distance increases. This means that traveling patterns with a higher JI value mainly exist in short-distance trips. For long-distance trips, car-hailing drivers tend to select a route which is totally different or different from the planned route.

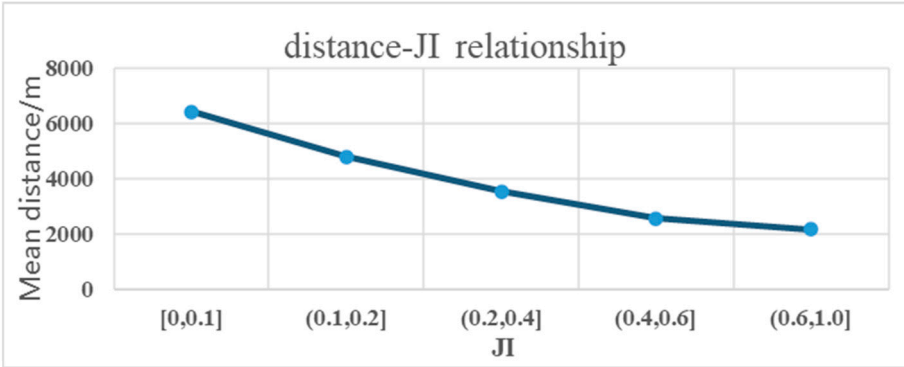
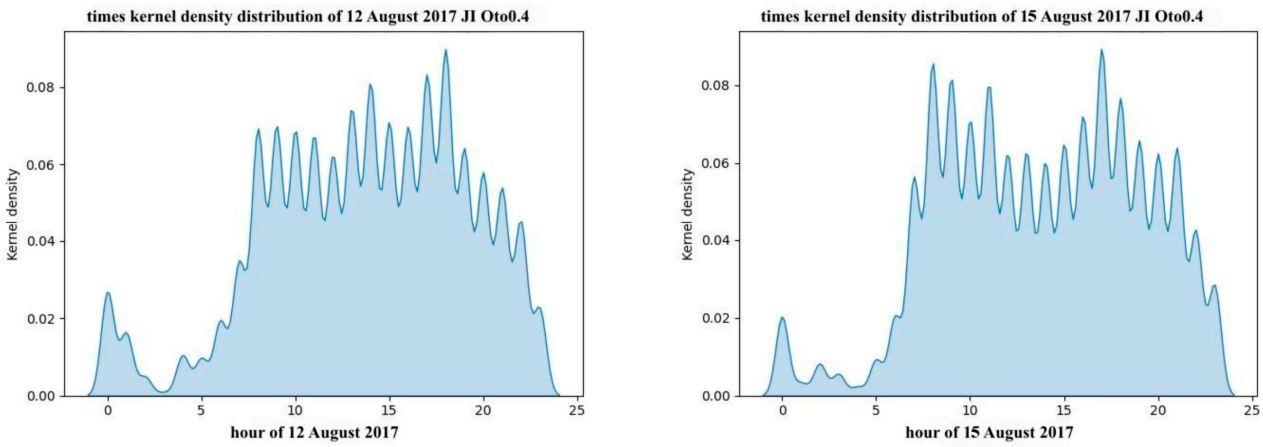


Figure 10. Relationship between JI values and the distances between OD pairs on 12 August 2017.

To further explore the distribution of JI values on all routes, we analyze the time distribution of OD pairs on workdays and weekends, respectively, as shown in Figure 11. Based on the experimental results, travel activities mainly concentrated on the period of 8:00 am–10:00 pm, regardless of weekend or workday. However, on the weekend, residents' travel activities usually occurred during three time periods: 8:00 am–9:00 am, 1:00 pm–2:00 pm, and 5:00 pm–6:00 pm (see Figure 11a). During workdays, the traveling activities mainly occurred in two time periods, 8:00 am– 9:00 am and 5:00 pm–6:00 pm, as shown in Figure 11b. That is, the traveling activities are concentrated on the morning peak and evening peak, especially in the workday. That means, during the workday, drivers tend to select driving routes from the original point to the destination based on their experience or the real-time situation. These routes may not be the shortest in time or distance for drivers.



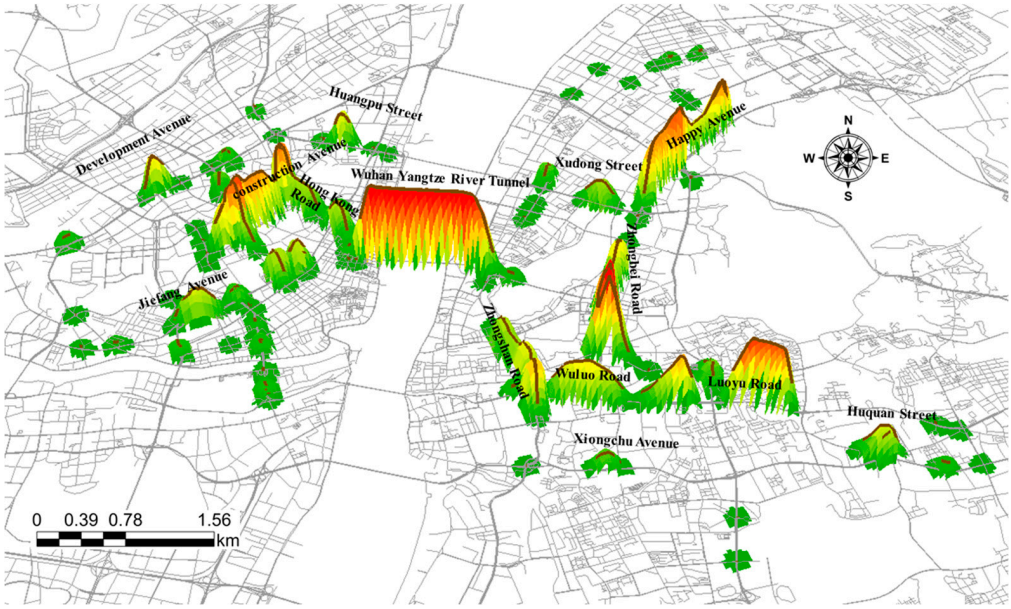
(**a**) (**b**)

Figure 11. The distribution of JI value of all OD pairs in workday and weekday, (**a**) traveling activities in weekend, (**b**) traveling activities in workday.

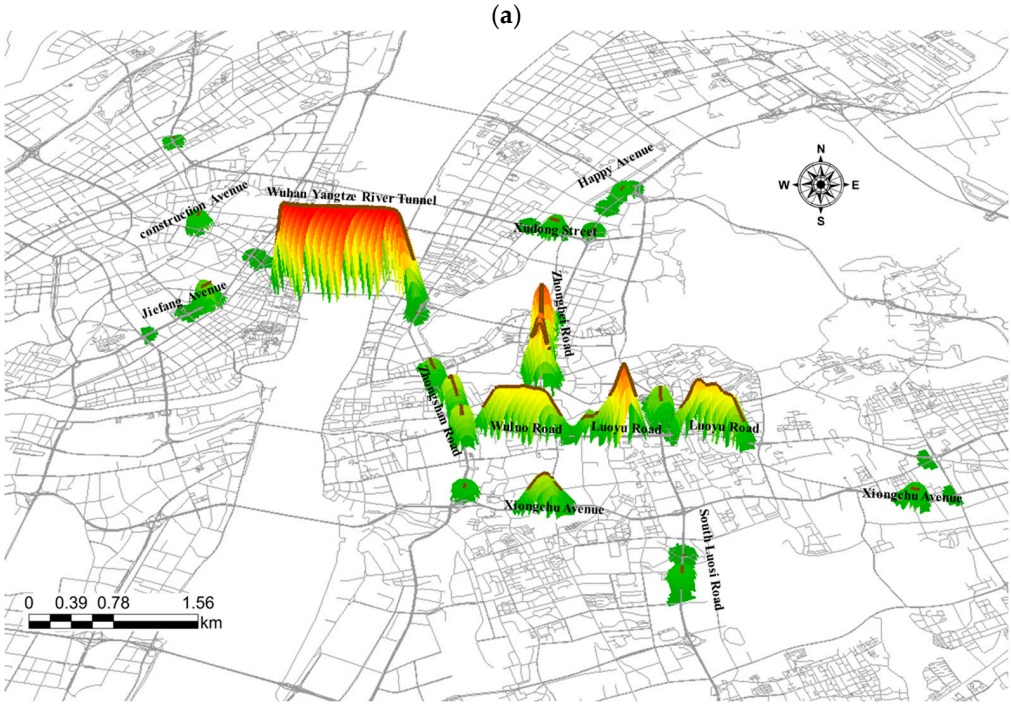
4.2.2. Detection of Hottest Road Segments Avoided by Drivers

Through analyzing the JI values of all OD pairs, we found that most of the travel routes of online car-hailing drivers were entirely or partly different with the corresponding planned routes. In this study, the road segments in planned routes but not in actual routes

are defined as the PRAD. To investigate the possible reasons why the drivers did not select these routes, we detected and analyzed the hottest PRADs in a temporal and spatial context. Since these PRADs are more likely to occur during the morning and evening peak hours when the traffic is congested, we detected the hottest ones from traveling activities which occurred in morning and evening peak hours on workdays and weekends, respectively, based on the NEAT clustering method. As we can see from Figure 12, the hottest PRADs are shown based on the heat map and grey lines represent the road network. Based on the distribution of the hottest PRADs on workdays, we can find that some road segments have always been the PRAD, no matter whether in morning peak hours or evening peak hours, such as Wuluo road, Zhongbei road, etc. (See Figure 12a,b).

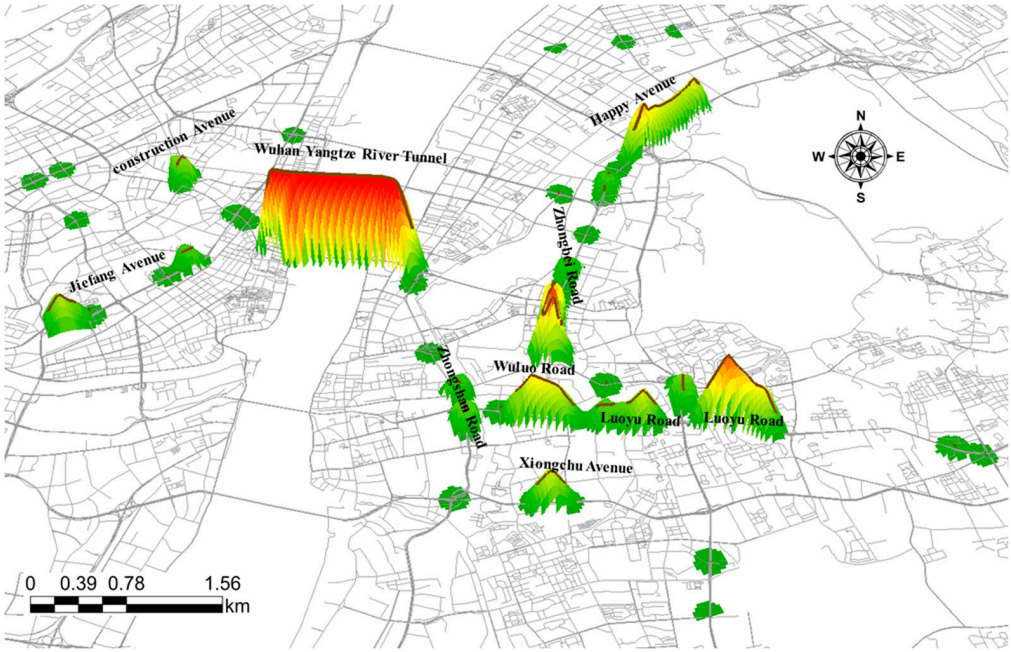


(**a**)

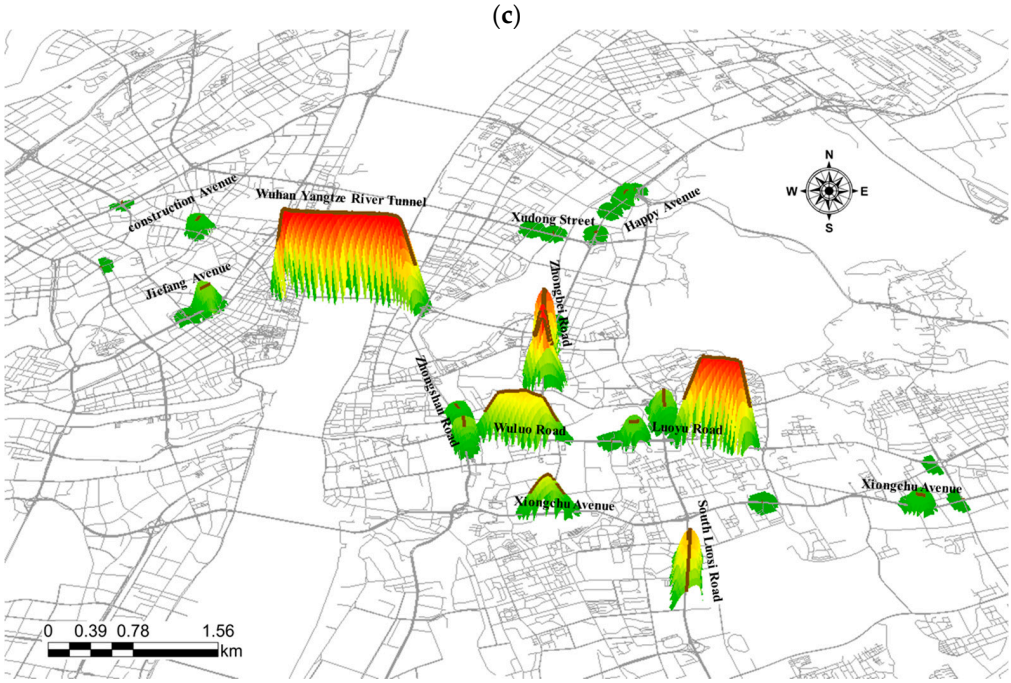


(**b**)

Figure 12. *Cont.*



(**c**)



(**d**)

Figure 12. The distribution of hottest PRADs on a workday (15 August 2017) and a weekend day (12 August 2017): (**a**) the hottest PRAD occurred during morning peak hours on a workday, (**b**) the hottest PRAD occurred during evening peak hours on a workday, (**c**) the hottest PRAD occurred during morning peak hours on weekend day, (**d**) the hottest PRAD occurred during evening peak hours on a weekend day.

The number of road segments which are avoided by drivers (also named PRADs) during morning peak hours on 15 August 2017 (on a workday), was about 46. The total length of these PRADs was about 42.03 km. On a workday, the number of PRADs in the evening peak was about 21, with 15.4 km total length. Compared with that on a workday, the number of PRADs in the morning peak and evening peak on a weekend day (12 August 2017), was 22 and 17, respectively. The total length of these PRADs was 14.9 km and 16.47 km, respectively. These statistics indicate the number of the hottest PRADs on

weekend days is obviously less than on workdays in the morning peak. Furthermore, some of them are very similar to PRADs which appeared on a workday, such as the Wuhan Yangtse River tunnel (see Figure 12c,d). Table 2 summarizes the road names and types of all the hottest PRADs which appeared on both workdays and weekends. In Table 2, we can find that most of the PRADs are main roads.

Table 2. The information on the hottest PRADs shown in Figure 12.

Road Name	Road Type
Wuluo road	trunk
Luoyu road	trunk
Zhongbei road	primary
Wuhan Yangtse river tunnel	primary
Huanle avenue	primary
Jianshe avenue	primary
Qingnian road	primary

Drivers possibly avoided these road segments for two reasons: Firstly, the traffic congestion on these roads is serious. Most drivers avoided these roads during the operating time to save time and earn a higher income. To validate this assumption, we obtained the traffic monitoring data from the Gaode Map traffic monitoring platform (https://report.amap.com/detail.do?city=420100, accessed on 1 April 2022). Based on the data derived from this platform, we found that about 65% of PRADs with a serious traffic jam coincided on a workday. On weekends, traffic congestion happened in about 30% of PRADs. In addition, based on the public information provided by the website of WPCOM (https://www.sanyefengji.cn/qichezatan/414998.html, accessed on 1 April 2022), about 20 road segments often experienced traffic jams in the city of Wuhan. Among them, about 13 road segments are identified as being on a congested road with a cyclical pattern and nine of them are considered as PRADs, including 'Air Road interchange', 'Fazhan Avenue', 'Gusaoshu road', 'Zhongshan road', 'Qingtai road', 'Wutaizha road', 'Wuluo road', 'Zhongbei road', 'Guanggu roundabout on Luoyu Road', 'Wuhan Yangtze River bridge and tunnel'. The main reason that these road segments often experience traffic jams is because road networks around these road segments are inadequate, which means they cannot adequately alleviate the enormous transportation pressure coming from neighboring commercial places, e.g., large shopping malls, restaurants, and other amenities. Apart from this, about seven road segments belong to the congested roads because of construction work to the road surface, and five of them are identified as PRADs, including 'Zhongnan Road', 'Jiefang Road', 'the starting part of Development Avenue', and 'Hanzheng road' and its surrounding roads.

We also investigated the traffic accidents that happened in the experimental area, based on the information collected from the paper published by Fan et al. (2018) [50]. The results showed that the incidence of traffic accidents that occurred in PRADs ranged from 0 to 8.52%. To further quantify the relationship between PRADs, traffic jams and accidents, we estimated their correlation by using Spearman's correlation. In Appendix A Table A1, we counted the rate of avoidance of all PRADs identified based on the trajectories collected on a workday (15 August 2017) and a weekend day (12 August 2017). Here, the rate of avoidance of the PRADs was computed based on its occurrence frequency. That is the rate of avoidance of one PRAD is equal to dividing its occurrence frequency by the total of the occurrence frequency of all PRADs. Meanwhile, we also obtained the traffic jam index and traffic accident rate of these PRADs through the Gaode Map traffic monitoring platform and the data provided by Fan et al. (2018) [50]. The correlation between these PRADs, traffic jams and accidents, both on workdays and weekends was computed(see Table 3).

Table 3. Spearman's correlation between PRADs, traffic jams and traffic accidents.

	Traffic Jam	Traffic Accident
	Coefficient (*p*-Value)	
Morning peak of workday	0.472 ** (0.001)	0.779 ** (0.000)
Evening peak of workday	0.801 ** (0.000)	0.737 ** (0.000)
Morning peak of weekend	0.878 ** (0.000)	0.772 ** (0.000)
Evening peak of weekend	0.604 * (0.010)	0.596 * (0.012)

Note: ** denotes significance at 0.01 level; * denotes significance at 0.05 level.

In Table 3, we find that traffic jams and accidents are associated with PRADs no matter whether on workdays or weekends. Specifically, traffic jams and traffic accidents occurring in the peaks of workdays are significantly associated with PRADs. During the peaks of weekends, traffic jams and traffic accidents are also associated with PRADs, but this is not significant in the evening peaks. This result indicates that the occurrences of traffic jams and accidents are dynamic in the evening peaks of weekends. In general, most drivers avoid these roads during the operating times to save time and obtain a higher income.

4.2.3. Results and Analysis

In this chapter, we employ a methodological framework that involves pre-processing through map matching and adaptive detection of PRADS using methods such as NEAT clustering. We apply this framework to detect popular route avoidance segments in DiDi travel data for Wuhan City on 12 and 15 August 2017. Furthermore, we conduct statistical analysis on the results for both weekdays and weekends. The results revealed that during the morning peak period of 15 August 2017 (weekday), there were approximately 46 PRADs identified, with a total length of approximately 42.03 km. On the weekday, during the evening peak period, around 21 PRADs were detected, with a total length of 15.4 km. In comparison, on the non-working day (12 August 2017), there were 22 PRADs during the morning peak and 17 PRADs during the evening peak, with total lengths of 14.9 km and 16.47 km, respectively. These statistical data indicate that the number of PRADs during non-working days is significantly lower than during weekday morning peaks. According to the data obtained on traffic congestion in Wuhan, we found that about 65% of the popular avoided roads with serious traffic jams occur on weekdays.

Furthermore, the Spearman correlation coefficient analysis revealed a significant relationship between traffic congestion, traffic accidents, and PRADs during both weekday and non-working day morning and evening peaks. Specifically, there was a significant correlation between traffic congestion, traffic accidents, and PRADs during weekday peak periods. During non-working day morning and evening peaks, traffic congestion and traffic accidents were also correlated with PRADs, but the correlation was not significant during the evening peak. Based on these findings, it can be concluded that there is an association between PRADs and traffic congestion as well as traffic accidents during weekday and non-working day morning and evening peak periods in Wuhan city.

From a data perspective, this study utilizes the extensive travel data provided by DiDi for analysis. Compared to the data collected through stated preference (SP) surveys or small-scale experiments, as used by Kroes E.P. and Hensher D.A. and others, the experimental data in this study are larger in scale, more representative, and have smaller biases. From an analytical perspective, in contrast to the majority of the literature exploring the reasons for route choice, we have made a breakthrough by analyzing the differences between planned and actual routes. This approach has successfully revealed that drivers tend to avoid certain planned routes primarily due to considerations related to traffic congestion and accidents.

5. Discussion

The purpose of this study is to investigate the factors influencing driver's route choice by analyzing the phenomenon of disparities between planned routes and actual routes. To the best of our knowledge, this may be one of the few empirical research cases that explore

drivers' route-choice behavior through the analysis of Planned Route Actual Deviation (PRAD), which is of significant importance for optimizing route-planning strategy research.

5.1. Inspiration for Reality

In this paper, we explore the hidden reasons behind PRADs in terms of traffic conditions on weekdays and weekends, and, analyzing the results, we obtain an interesting finding. Generally, the choice of a planned path is based on several principles, including the minimum travel distance, the shortest travel time, or a comprehensive optimal solution from the consideration of travel time, distance, number of traffic lights, and road speed limit. Although the path-planning algorithm has considered many factors to obtain an optimal path for the driver, the actual traffic situation is very complex. There are two main reasons why drivers avoid the planned path and choose other roads to reach their destination. First, some of the roads on the planned path may not be the best due to severe traffic congestion. Statistics obtained from the real-time city traffic detail platform of GODE Maps show that about 65% of people avoid roads with severe traffic jams on weekdays. The results of this study can ease urban traffic congestion and provide a basis for decision making by traffic management authorities. Another major reason is that drivers want to choose a safe way to reach their destinations. However, some of the planned roads have very high accident rates. Correlation analysis based on the obtained data shows that there is a significant relationship between the accident rate of the popular avoidance roads and the roads identified as popular avoidance roads. The experimental results show that the popular avoidance road sections on weekend days are significantly fewer than those on weekdays. By analyzing the popular avoidance road sections, these findings can be used to optimize the path-planning strategy or path-search algorithm, which means that users can avoid these road sections at specific times (e.g., peak hours on weekdays) to reduce their travel time and cost.

5.2. Limitation and Future Prospects

Although this paper has achieved some research results, there are still certain shortcomings and limitations that require further in-depth research. It is necessary to apply the research findings to path-planning strategies and effectively reduce the travel time and cost for urban residents while providing optimization strategies for urban transportation planning.

(1) Although our proposed map-matching algorithm, based on an improved Hidden Markov Model (HMM), and clustering algorithm based on NeuroEvolution of Augmenting Topologies (NEAT), have shown excellent performance in theory, their practical applications are limited by runtime constraints, preventing real-time processing and detection of popular detour sections on urban roads. Therefore, further research is needed to optimize the algorithms and achieve the goal of online real-time detection of popular detour sections. This is one of the key focuses and directions of our future research.
(2) When exploring the reasons for the formation of popular detour sections in the path selection behavior of ride-hailing services, we have only conducted cause analysis and verification from the perspectives of traffic congestion and traffic accidents. In future research, it is important to explore more relevant influencing factors that contribute to the formation of popular detour sections, such as land use, road construction, and driver personalized preferences.

6. Conclusions

Addressing vehicle-routing problems needs to consider many factors, including travel distance, road conditions (e.g., traffic jams and accidents), personalized preference, etc. The basis of weighing these factors during routing planning is to figure out the differences between planned routes based on these factors with the actual routes selected by drivers. In this study, we answer this question by studying the popular dodging roads from a large

number of car-hailing trajectories. Specifically, we optimized the HMM map-matching algorithm by improving the computations of angle feature in observation probability and velocity in transition probability to increase the accuracy of map matching and provide accurate matched results for the subsequent analysis of PRADs. The actual routes of the OD matrix were generated based on the map-matching results. The planned routes of the corresponding OD pairs were generated by the A* routing algorithm. By using the Jaccard index, we quantified and visualized the similarity between the actual routes and the planned routes between the same OD pairs. The most popular road segments avoided by drivers were detected based on the clustering method of NEAT and its causes were further analyzed in relation to traffic conditions, including traffic jams and accidents.

Taking online car-hailing in Wuhan as a case study, we explored the spatiotemporal patterns of PRADs. The experimental results showed that the hottest PRADs on a weekend day are significantly fewer than that those on a workday. In general, a planned route is selected based on several principles including minimum travel distance, shortest travel time, or the comprehensive optimum scheme incorporating aspects of travel time, distance, number of traffic lights, road speed limits, etc. Although a route-planning algorithm has considered many factors to obtain an optimal route for drivers, the actual traffic conditions are very complicated. There are two main reasons why drivers avoided the planned routes and selected other roads to arrive at their destination. First, some roads of the planned routes may not have been optimal due to serious traffic congestion. The statistics obtained from Gaode Map validated that about 65% of dodging routes with serious traffic jams coincided with a workday. Apart from that, drivers also want to select a safe way to arrive at their destinations. However, the traffic accident rate on some planned roads is very high. Based on the correlation analysis, the accident rate of PRADs was significantly associated with the road segments which were identified as the PRADs. These findings from analyzing PRADs can be used for optimizing route-planning strategies, which means users can avoid these road segments at a specific time, such as the peak of a workday, to reduce their travel time and cost.

Our research findings can provide insights into modern route navigation. Large-scale ride-hailing companies such as DiDi can expediently detect PRADS by regularly analyzing the collected data. Consequently, they can de-prioritize these road segments during actual route navigation and plan them as low-traffic routes. However, traffic conditions often change rapidly, and the proposed framework in this paper cannot detect PRADs in real time due to operational time constraints. Therefore, occasional errors in the provided PRADs may result in inadequate route planning. Nevertheless, in this study we conducted data analysis for two days and found that the traveling patterns of car-hailing drivers were highly consistent daily. Thus, with prolonged observation and analysis, if certain road segments are consistently detected as PRADs over an extended period, our research results can also provide insights to government transportation departments. They can optimize and improve these segments based on PRADs results, thereby enhancing traffic flow efficiency.

Author Contributions: Conceptualization, Xue Yang and Zihan Kan; Methodology, Xue Yang, Jianhua Yu, Zihan Kan and Xuyu Feng; Formal analysis, Jianhua Yu and Xuyu Feng; Writing—original draft, Xue Yang and Jianhua Yu; Writing—review and editing, Zihan Kan, Xuyu Feng, Lin Zhou and Luliang Tang. All authors have read and agreed to the published version of the manuscript.

Funding: This work was founded by the National Natural Science Foundation of China (No. 42271449) and the national college students' innovation and entrepreneurship training program (S202310491067).

Data Availability Statement: Due to the nature of this research, participants in this study did not agree for their data to be shared publicly, so supporting data is not available.

Acknowledgments: The authors would like to sincerely thank the anonymous reviewers for their constructive comments and valuable suggestions to improve the quality of this article. All authors have read and agreed to the published version of the manuscript.

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

Appendix A

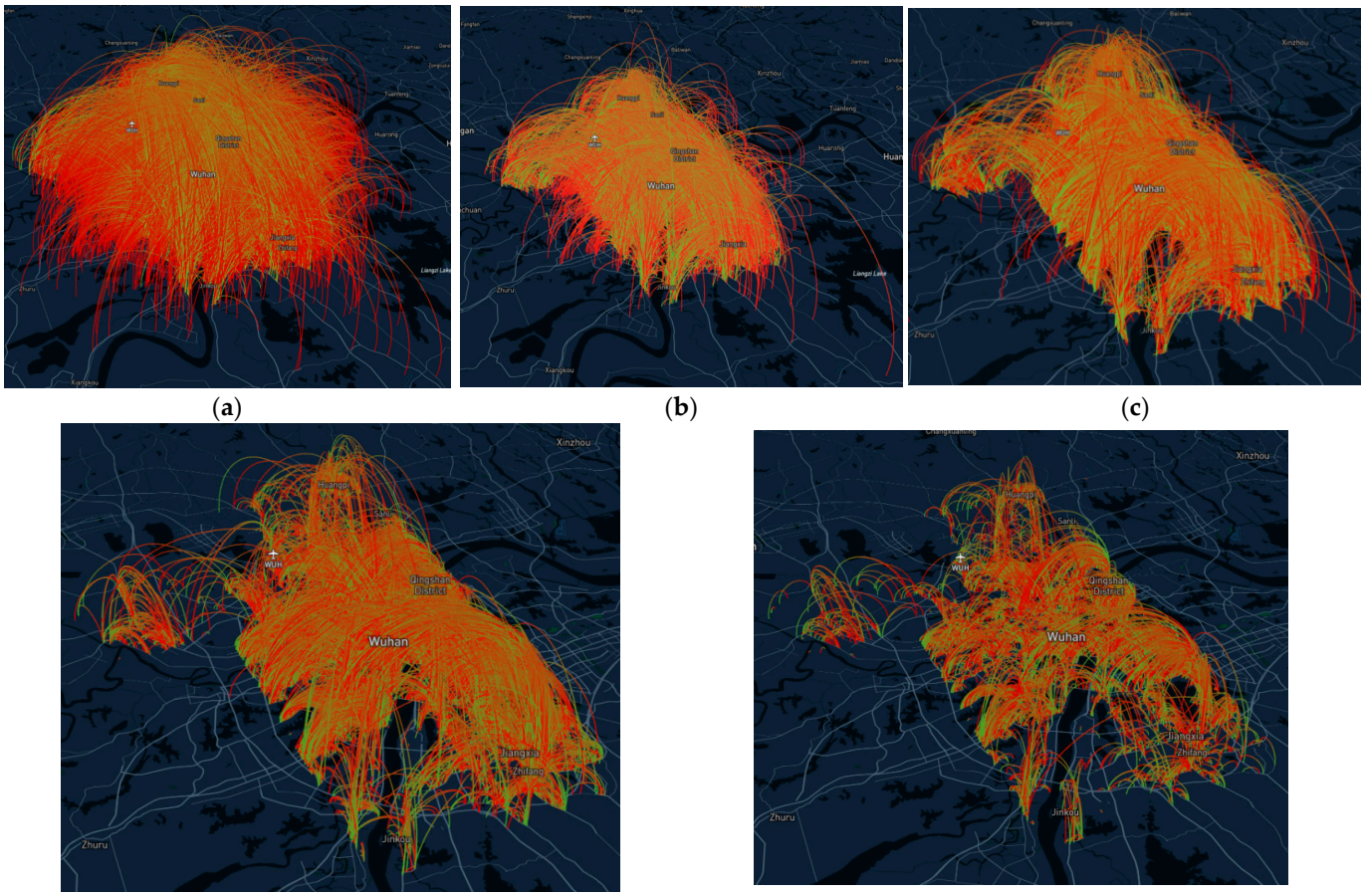


Figure A1. Spatial distribution pattern of JI value of trajectories collected on 15 August 2017. In the figure, each line is divided into red and green halves, representing the direction from the starting point (red) to the destination point (green): (**a**) JI values with the type of 'totally different', (**b**) JI values with the type of 'different', (**c**) JI values with the type of 'slightly different', (**d**) JI values with the type of 'similar', (**e**) JI values with the type of 'no difference'.

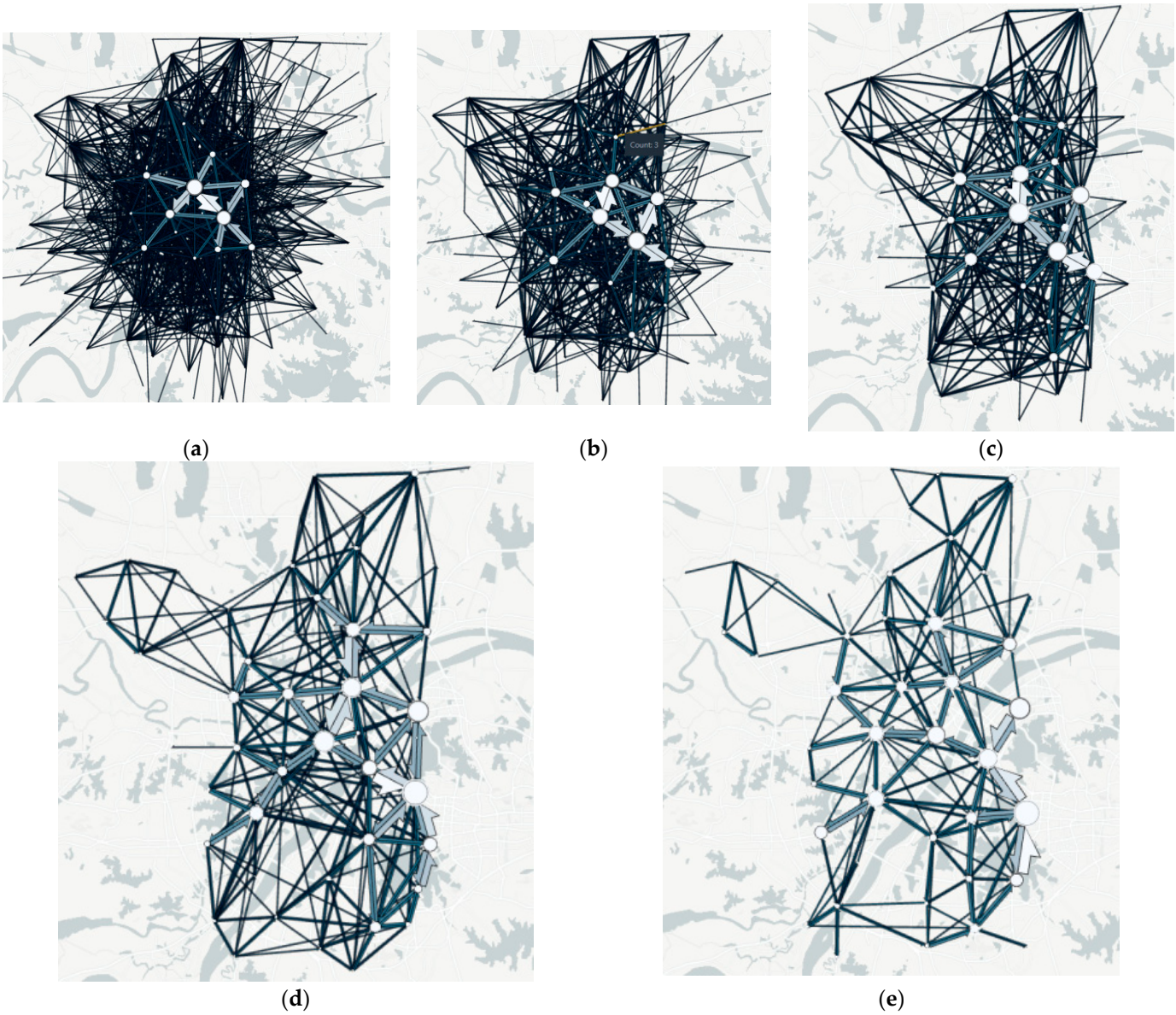


Figure A2. OD clustering results on 15 August 2017: (**a**) OD pairs of JI values within [0, 0.1], (**b**) OD pairs of JI values within (0.1, 0.2], (**c**) OD pairs of JI values within (0.2, 0.4], (**d**) OD pairs of JI values within (0.4, 0.6], (**e**) OD pairs of JI values within (0.6, 1.0).

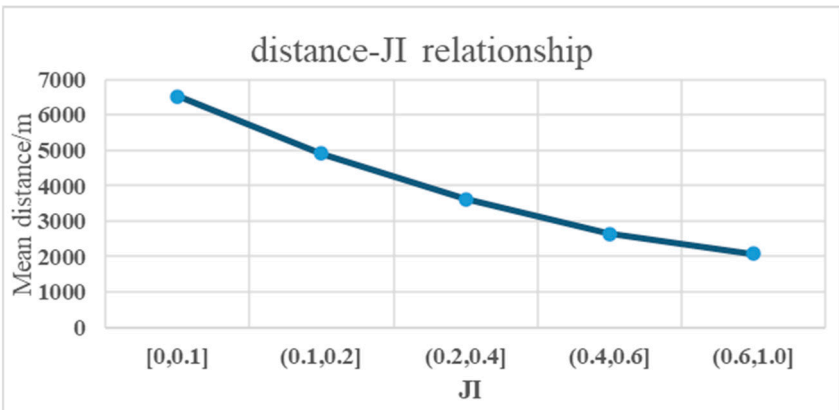


Figure A3. Relationship between JI values and the distances between OD pairs on 12 August 2017.

Table A1. Details of PRADs including their name, the rate of avoidance, traffic jam index, and traffic accident rate.

Road Name of PRADs	The Rate of Avoidance on the PRAD	Traffic Jam Index	Traffic Accident Rate
Luoyu Road	2.27	3.6	1.35
Jianshe Avenue	2.17	2.3	1.13
Zhongbei Road	4.17	4.2	2.46
Huanle Avenue	1.17	2.1	0.52
Jiefang Avenue	3.17	3.6	2.32
Youyi Avenue	2.03	2.2	1.08
Zhongshan Road	2.38	3.3	0.28
Xinhua Road	4.32	2	2.27
Fazhan Avenue	4.2	3.9	2.26
Wuluo Road	3.59	4.1	2.09
Wusheng Road	2.95	1.6	0.25
Hongkong Road	2.87	2.3	0.54
Huangpu Avenue	2.49	2.5	0.37
Air Road interchange	2.14	4.3	0.52
Youyi Road	1.82	1.3	1.05
Xiongchu Avenue	1.64	1	0.54
Luoshi Road	1.58	1.2	0.5
Xudong Avenue	1.54	1.9	0.11
Qingnian Road	1.35	1.5	0.08
Qiuchang Street	1.14	1.3	0.07
Dazhi Road	1.13	1.6	0.05
South Luoshi Road	0.94	1.4	0.4
Yinwu Avenue	0.86	0.9	0.54
Wuhan Yangtze River tunnel	0.79	3.6	0
Bayi Road	0.78	0	1.04
Heping Avenue	0.6	1.1	0
Qiaokou Road	0.54	1.3	0.28
Zhongshan Avenue	0.54	3.5	0
Jinqiao Avenue	0.53	0	0
Renhe Road	0.52	0	0.28
Huquan Street	0.52	0	0
Tangjiadun Road	0.51	3.8	0
Sanyanqiao Road	0.5	1.8	0.52
Guanggu roundabout	0.37	2.1	0.31
Jinghan Avenue	0.31	0	0.95
Hanxi Road	0.29	1.6	0

Table A1. *Cont.*

Road Name of PRADs	The Rate of Avoidance on the PRAD	Traffic Jam Index	Traffic Accident Rate
Yangtze River tunnel	0.28	3.6	0
Shengli Street	0.27	0	0
Nanjing Road	0.26	0	0
Zhongnan Road	0.26	3.5	0.52
Jinghan Road	0.25	0	0
Mingzu Avenue	0.25	2.4	0.94
Xingye Road	0.25	0.3	0
Second ring road	0.24	0.6	0

References

1. Almatar, K.M. Transit-Oriented Development in Saudi Arabia: Riyadh as a Case Study. *Sustainability* **2022**, *14*, 16129. [CrossRef]
2. Almatar, K.M. Towards sustainable green mobility in the future of Saudi Arabia cities: Implication for reducing carbon emissions and increasing renewable energy capacity. *Heliyon* **2023**, *9*, e13977. [CrossRef]
3. Almatar, K.M. Traffic congestion patterns in the urban road network: (Dammam metropolitan area). *Ain Shams Eng. J.* **2023**, *14*, 101886. [CrossRef]
4. Jing, P.; Zhao, M.; He, M.; Chen, L. Travel Mode and Travel Route Choice Behavior Based on Random Regret Minimization: A Systematic Review. *Sustainability* **2018**, *10*, 1185. [CrossRef]
5. Li, L.; Wang, S.; Wang, F.-Y. An Analysis of Taxi Driver's Route Choice Behavior Using the Trace Records. *IEEE Trans. Comput. Soc. Syst.* **2018**, *5*, 576–582. [CrossRef]
6. Xu, Q.; Ji, X. User Equilibrium Analysis Considering Travelers' Context-Dependent Route Choice Behavior on the Risky Traffic Network. *Sustainability* **2020**, *12*, 6706. [CrossRef]
7. Hart, P.E.; Nilsson, N.J.; Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 100–107. [CrossRef]
8. Zhang, D.; Cai, S.; Ye, F.; Si, Y.-W.; Nguyen, T.T. A hybrid algorithm for a vehicle routing problem with realistic constraints. *Inf. Sci.* **2017**, *394–395*, 167–182. [CrossRef]
9. Lu, F.; Chen, W.; Feng, W.; Bi, H. 4PL routing problem using hybrid beetle swarm optimization. *Soft Comput.* **2023**, 1–14. [CrossRef]
10. Wen, H.; Wang, S.X.; Lu, F.Q.; Feng, M.; Wang, L.Z.; Xiong, J.K.; Si, M.C. Colony search optimization algorithm using global optimization. *J. Supercomput.* **2021**, *78*, 6567–6611. [CrossRef]
11. Sturtevant, N.R. Benchmarks for grid-based pathfinding. *IEEE Trans. Comput. Intell. AI Games* **2012**, *4*, 144–148. [CrossRef]
12. Han, B.; Liu, L.; Omiecinski, E. Road-Network Aware Trajectory Clustering: Integrating Locality, Flow, and Density. *IEEE Trans. Mob. Comput.* **2013**, *14*, 416–429. [CrossRef]
13. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous Methods for Deep Reinforcement Learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1928–1937.
14. Chen, F.U.; Shengke, H.U.; Yan, T.A.; Hangbin, W.U.; Chun, L.I.; Lianbi, Y.A.; Wei, H.U. A real-time map matching method for road network using driving scenario classification. *Acta Geod. Cartogr. Sin.* **2021**, *50*, 1617.
15. Quddus, M.A.; Ochieng, W.Y.; Noland, R.B. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transp. Res. Part C Emerg. Technol.* **2007**, *15*, 312–328. [CrossRef]
16. Phuyal, B.P. Method and Use of Aggregated Dead Reckoning Sensor and GPS Data For Map Matching. In Proceedings of the 15th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2002), Portland, OR, USA, 24–27 September 2002.
17. Yu, M. Improved Positioning of Land Vehicle in ITS Using Digital Map and Other Accessory Information. Ph.D. Thesis, Hong Kong Polytechnic University, Hong Kong, 2006.
18. Newson, P.; Krumm, J. Hidden Markov Map Matching through Noise and Sparseness. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 4–6 November 2009.
19. Syed, S.; Cannon, M.E. Fuzzy Logic Based-Map Matching Algorithm for Vehicle Navigation System in Urban Canyons. In Proceedings of the 2004 National Technical Meeting of the Institute of Navigation, San Diego, CA, USA, 26–28 January 2004.
20. Li, H.; Wu, G. Map Matching for Taxi GPS Data with Extreme Learning Machine. In *Advanced Data Mining and Applications: 10th International Conference, ADMA 2014, Guilin, China, 19–21 December 2014*; Springer International Publishing: Cham, Germany, 2014; pp. 447–460. [CrossRef]

21. Dai, P.; Li, Z.; Wang, J. Research on map-matching algorithm using kaman filter to improve localization accuracy from Baidu map based on android. In Proceedings of the 2016 6th International Conference on Information Technology for Manufacturing Systems (ITMS 2016), Prague, Czech Republic, 10–11 May 2016; pp. 249–253.
22. Zhao, X.; Cheng, X.; Zhou, J.; Xu, Z.; Dey, N.; Ashour, A.S.; Satapathy, S.C. Advanced Topological Map Matching Algorithm Based on D–S Theory. *Arab. J. Sci. Eng.* **2017**, *43*, 3863–3874. [CrossRef]
23. Hu, Y.; Lu, B. A Hidden Markov Model-Based Map Matching Algorithm for Low Sampling Rate Trajectory Data. *IEEE Access* **2019**, *7*, 178235–178245. [CrossRef]
24. Hansson, A.; Korsberg, E.; Maghsood, R.; Norden, E.; Selpi, S. Lane-Level Map Matching Based on HMM. *IEEE Trans. Intell. Veh.* **2020**, *6*, 430–439. [CrossRef]
25. Kroes, E.P.; Sheldon, R.J. Stated preference methods. An introduction. *J. Transp. Econ. Policy* **1988**, *22*, 11–25.
26. Hensher, D.A. Stated preference analysis of travel choices: The state of practice. *Transportation* **1994**, *21*, 107–133. [CrossRef]
27. Dial, R.B. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transp. Res.* **1971**, *5*, 83–111. [CrossRef]
28. Wen, C.-H.; Koppelman, F.S. The generalized nested logit model. *Transp. Res. Part B Methodol.* **2001**, *35*, 627–641. [CrossRef]
29. McFadden, D. Econometric models for probabilistic choice among products. *J. Bus.* **1980**, *53*, S13–S29. [CrossRef]
30. Deng, Y.; Li, M.; Tang, Q.; He, R.; Hu, X. Heterogenous Trip Distance-Based Route Choice Behavior Analysis Using Real-World Large-Scale Taxi Trajectory Data. *J. Adv. Transp.* **2020**, *2020*, 8836511. [CrossRef]
31. Cascetta, E.; Nuzzolo, A.; Russo, F.; Vitetta, A. A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and Some Calibration Results for Interurban Networks. In Transportation and Traffic Theory. Proceedings of the 13th International Symposium On Transportation And Traffic Theory, Lyon, France, 24–26 July 1996; IEEE: Piscataway, NJ, USA, 1996.
32. Ramming, M. *Network Knowledge and Route Choice*; MIT Press: Cambridge, MA, USA, 2009.
33. Koppelman, F.S.; Wen, C.-H. The paired combinatorial logit model: Properties, estimation and application. *Transp. Res. Part B Methodol.* **2000**, *34*, 75–89. [CrossRef]
34. Tang, Q.; Hu, X. Modeling individual travel time with back propagation neural network approach for advanced traveler information systems. *J. Transp. Eng. Part A Syst.* **2020**, *146*, 04020039. [CrossRef]
35. Hu, X.; Chiu, Y.-C.; Ma, Y.-L.; Zhu, L. Studying Driving Risk Factors using Multi-Source Mobile Computing Data. *Int. J. Transp. Sci. Technol.* **2015**, *4*, 295–312. [CrossRef]
36. Zhu, X.; Yuan, Y.; Hu, X.; Chiu, Y.-C.; Ma, Y.-L. A Bayesian Network model for contextual versus non-contextual driving behavior assessment. *Transp. Res. Part C Emerg. Technol.* **2017**, *81*, 172–187. [CrossRef]
37. Lu, M.; Lai, C.; Ye, T.; Liang, J.; Yuan, X. Visual Analysis of Multiple Route Choices Based on General GPS Trajectories. *IEEE Trans. Big Data* **2017**, *3*, 234–247. [CrossRef]
38. Deng, Y.; Luo, X.; Hu, X.; Ma, Y.; Ma, R. Modeling and Prediction of Bus Operation States for Bunching Analysis. *J. Transp. Eng. Part A Syst.* **2020**, *146*, 04020106. [CrossRef]
39. Deng, Y.-J.; Liu, X.-H.; Hu, X.; Zhang, M. Reduce Bus Bunching with a Real-Time Speed Control Algorithm Considering Heterogeneous Roadway Conditions and Intersection Delays. *J. Transp. Eng. Part A Syst.* **2020**, *146*, 04020048. [CrossRef]
40. Qi, H.; Hu, X. Real-time headway state identification and saturation flow rate estimation: A hidden Markov Chain model. *Transp. A Transp. Sci.* **2020**, *16*, 840–864. [CrossRef]
41. Kim, J.; Mahmassani, H.S. Spatial and Temporal Characterization of Travel Patterns in a Traffic Network Using Vehicle Trajectories. *Transp. Res. Procedia* **2015**, *9*, 164–184. [CrossRef]
42. Li, D.; Miwa, T.; Morikawa, T.; Liu, P. Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets. *Transp. Res. Part C Emerg. Technol.* **2016**, *67*, 31–46. [CrossRef]
43. Goh, C.; Dauwels, J.; Mitrovic, N.; Asif, M.T.; Oran, A.; Jaillet, P. Online Map-Matching Based on Hidden Markov Model for Real-Time Traffic Sensing Applications. In Proceedings of the 2012 15th International IEEE Conference on Intelligent Transportation Systems, Anchorage, AK, USA, 16–19 September 2012; pp. 776–781. [CrossRef]
44. Jagadeesh, G.R.; Srikanthan, T. Online Map-Matching of Noisy and Sparse Location Data With Hidden Markov and Route Choice Models. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 2423–2434. [CrossRef]
45. Candra, A.; Budiman, M.A.; Hartanto, K. Dijkstra's and a-Star in Finding the Shortest Path: A Tutorial. In Proceedings of the 2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, 16–17 July 2020; pp. 28–32.
46. Verma, V.; Aggarwal, R.K. A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: Empirical and theoretical perspective. *Soc. Netw. Anal. Min.* **2020**, *10*, 43. [CrossRef]
47. Rogers, C.A. *Hausdorff Measures*; Cambridge University Press: Cambridge, UK, 1998.
48. Lee, J.-G.; Han, J.; Whang, K.-Y. Trajectory clustering: A partition-and-group framework. In Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data—SIGMOD'07, Beijing, China, 11–14 June 2007; pp. 593–604.

49. Yang, X.; Tang, L.; Niu, L.; Zhang, X.; Li, Q. Generating lane-based intersection maps from crowdsourcing big trace data. *Transp. Res. Part C Emerg. Technol.* **2018**, *89*, 168–187. [CrossRef]
50. Fan, X. Spatial and Temporal Analysis of Urban Road Traffic Accidents and Optimization of Multi-Constrained Spatial Zoning. Ph.D. Thesis, Wuhan University, Wuhan, China, 2019.