Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 8836511, 16 pages
https://doi.org/10.1155/2020/8836511

Research Article

Heterogenous Trip Distance-Based Route Choice Behavior Analysis Using Real-World Large-Scale Taxi Trajectory Data

Yajuan Deng,[1] Meiye Li,[2] Qing Tang,[3] Renjie He,[1] and Xianbiao Hu[3]

[1]*College of Transportation Engineering, Chang'an University, Xi'an 710064, China*
[2]*School of Transportation, Southeast University, Nanjing 211189, China*
[3]*Department of Civil, Architectural and Environmental Engineering, Missouri University of Science and Technology, Rolla, MO 65409, USA*

Correspondence should be addressed to Xianbiao Hu; xbhu@mst.edu

Received 4 May 2020; Revised 4 August 2020; Accepted 25 August 2020; Published 9 September 2020

Academic Editor: Kun Xie

Most early research on route choice behavior analysis relied on the data collected from the stated preference survey or through small-scale experiments. This manuscript focused on the understanding of commuters' route choice behavior based on the massive amount of trajectory data collected from occupied taxicabs. The underlying assumption was that travel behavior of occupied taxi drivers can be considered as no different than the well-experienced commuters. To this end, the DBSCAN algorithm and Akaike information criterion (AIC) were first used to classify trips into different categories based on the trip length. Next, a total of 9 explanatory variables were defined to describe the route choice behavior, and and the path size (PS) logit model was then built, which avoided the invalid assumption of independence of irrelevant alternatives (IIA) in the commonly seen multinomial logit (MNL) model. The taxi trajectory data from over 11,000 taxicabs in Xi'an, China, with 40 million trajectory records each day were used in the case study. The results confirmed that commuters' route choice behavior are heterogenous for trips with varying distances and that considering such heterogeneity in the modeling process would better explain commuters' route choice behaviors, when compared with the traditional MNL model.

1. Introduction

Analysis of the routing choice behavior provides theoretical support for route guidance and traffic assignment. Most early research studies on route choice behavior were based on the data collected from stated preference (SP) surveys or through small-scale experiments that were usually limited in data size or number of participants. In the modeling process, discrete choice models especially logit models were commonly used. The differences among these models were mainly reflected in the differences of data set, explanatory variables, or the model structure. For example, McFadden and Reid applied logit models to travel behavior analysis [1]. After that, based on the hypothesis that the random term of route utility function follows the Gumbel distribution, Dial constructed a discrete multinomial logit (MNL) model for multimode selection [2, 3]. In order to address the independence of irrelevant alternatives (IIA) issue of the MNL model, various modified models were proposed, such as the C-logit model and PS-logit model [4, 5], which were built by adding a modification term in the utility function to characterize the interactions among different routes. In addition, according to the generalized extreme value (GEV) theorem proposed by McFadden, some researchers proposed CNL and PCL models [6, 7] to avoid the IIA assumption of the MNL model. In general, these early research studies on route choice behavior lacked real-world data and were restricted by the algorithm complexity, and the numbers of explanatory variables used were usually limited as well.

With the rapid advancement of information and communication technologies (ICT), GPS technology has made significant progress, and the data collected by GPS devices have been widely used in various transportation research,

such as in travel time estimation [8–10], driving risk analysis [11, 12], departure time modeling [13, 14], and many others [15–17]. Such data have also been used to directly support the route choice behavior analysis, and the data-driven route choice models were qualitatively improved in terms of both effectiveness and accuracy. For example, route choice behaviors and network information in Chicago were studied using data collected using portable GPS devices, and path size (PS) logit models for different travel purposes in different time periods were proposed [18]. Based on the same method, Schussler and Axhausen collected travel data in Zurich area and calibrated C-logit model and PS-logit model [19]. Kim Mahmassani proposed a trajectory clustering algorithm to analyze the spatial and temporal travel patterns in a network [20], in which a framework for clustering and classifying vehicle trajectory data was built. Additionally, several medium-sized cities in Netherlands were selected as research objects and an MNL model based on GPS data was proposed to analyze the route choice behavior [21]. Li et al. collected the GPS data of private cars in Toyota City, explored the effect of travelers' heterogeneity on route choice, and concluded that the route choice behavior is affected by travelers' age, gender, vehicle displacement, and O-D's characteristics [22]. However, the analysis focus was on the traveler's heterogeneity, as opposed to the differences on the route characteristics. Bierlaire and Frejinger used the GPS data in Swiss to study the behavior characteristics of long-distance travel route selection and gave the estimation results of the PS-logit model and subnetwork model [23]. Miwa et al. used the taxi travel data of Nagoya City to analyze the characteristics of dynamic route choice behavior, an MNL model was built, and it was concluded that there are differences in the route choice behavior at different O-D distances [24]. Yamamoto et al. used the pedestrian GPS data from Nagoya to build a nested logit model [25], and Hu et al. used GPS data to analyze route choice behavior changes under preplanned road closures [26].

This manuscript focused on the analysis of route choice behavior of general traffic, based on the massive amount of trajectory data collected from the occupied taxicabs. Taxicabs, especially those work with the e-hailing platform such as Uber and Lyft, on the other hand, are mostly installed with the GPS devices for dispatching and safety purposes. However, most existing research studies based on the taxi GPS trajectory data focused on the routing behavior of the vacant taxi drivers, with the objective of minimizing the search time for the next customer [27, 28] or maximizing the profit [29–31], which was significantly different from regular drivers. Our underlying assumption was that when a taxi was occupied by customers, the taxi driver would seek to arrive at the destination in the least amount of time or distance as expected or required by the customer, similar to the objective of a commuter in his/her own car. Additionally, taxi drivers usually had good knowledge on the roadway network and traffic conditions, and thus their travel behavior can be considered as very similar to, and no different than, the well-experienced commuters.

Furthermore, this manuscript tested a hypothesis that trips with different lengths may exhibit different characteristics in driver's route choice behavior. As opposed to the common practice of developing and calibrating a unified model to describe the route choice behavior of all trips, the Akaike information criterion (AIC) was first used to classify trips into different categories based on the trip length. Next, a total of 9 explanatory variables were defined to describe the route choice behavior, and a PS-logit model was then built, which avoided the invalid assumption of IIA in the commonly seen multinomial logit model [24]. The taxi trajectory data from over 11,000 taxicabs in Xi'an, China, with 40 million trajectory records each day were used in the case study. The results confirmed the hypothesis that commuters' route choice behaviors are heterogenous for trips with varying distances and that considering such heterogeneity in the modeling process would better explain commuters' route choice behaviors.

The rest of this paper is organized as follows: Section 2 presents the data used in this research, including the GPS trajectory data and the traffic network. Section 3 discusses the analysis methodology in depth, and Section 4 presents the numerical analysis results. Section 5 concludes this research.

2. Data Preparation

2.1. GPS Data Set. The GPS trajectory data used in this research came from the historical database of the taxi dispatch system in Xi'an, China. The recording time was from 0 : 00 to 24 : 00, the recording interval was 30 s, and each record contained license plate number, timestamp, longitude, latitude, speed, driving direction, and loading state. The data set included data from over 11,000 taxicabs with 40 million trajectory records each day. Such a huge amount of data can meet the needs of this research. The following data cleaning and preprocessing were performed:

(1) Removed the flawed data with missing values.

(2) Only kept the data with loading state being "5 (passenger)."

(3) Removed the data with driving direction beyond 0°–360°.

(4) Removed the data with key attributes being "0 (invalid)."

2.2. Traffic Network. The OpenStreetMap (OSM) network of Xi'an was downloaded and utilized for this research. Post-processing efforts were made, including the removing the duplicate or redundant roads and adding the length of road segment and node information. Additionally, the road segments were classified into seven categories, including expressway, national highway, other highways, urban expressway, main road, secondary road, and neighborhood street. The research region is shown in Figure 1.

2.3. Hotspot OD Trips Extraction. Occupied trips between frequent origin–destination (OD) pairs were extracted from the database as the target data for analysis. We first identified

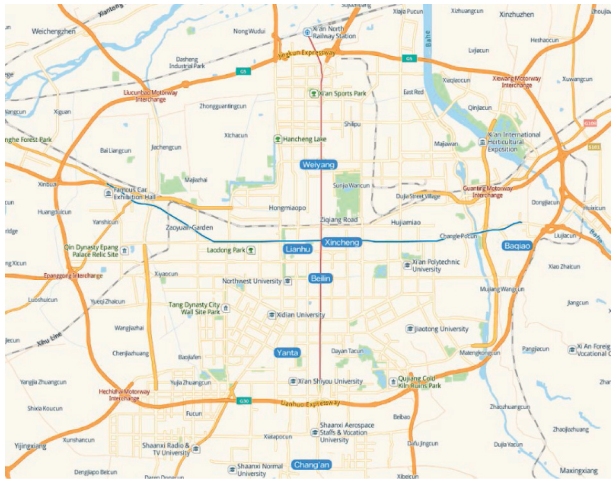


Figure 1: Traffic network in Xi'an city.

pick-up and drop-off hotspots and then extracted the frequent OD between these hotspots.

2.3.1. Identification of Drop-Off Hotspots. This step aimed to identify the areas with high density of drop-off events, with the goal of providing the basis for hotspot OD matching and ensuring that there was a sufficient number of passenger-carrying trips between the same OD pair (from pick-up to drop-off).

According to the change of loading state between two adjacent GPS data records, pick-up points and drop-off points can be identified. Taking GPS data of taxi on 19 April 2017 in Xi'an as an example, from 40 million trajectory data generated by 11,281 taxis, nearly 594 thousand drop-off points in the research region were obtained. The DBSCAN spatial clustering algorithm [32] was adopted to identify the drop-off hotspots. The algorithm contained two parameters: cluster neighborhood radius (Eps) and minimum density threshold (MinPts). In this paper, the K-distance method was used to determine the reasonable Eps. The method contained three steps:

Step 1: assuming that the drop-off points data set $D = \{P_i(x, y), \quad i = 1, 2, \ldots, n\}$ contained n points, we selected a drop-off point $P_i(x, y)$ and calculated the Euclidean distances between $P_i(x, y)$ and $P_1(x, y), P_2(x, y), \ldots, P_{i-1}(x, y), P_{i+1} \quad (x, y), \ldots, P_n$ (x, y), respectively. Then, they were sorted by Euclidean distances in ascending order as $d_i^1, d_i^2, \ldots, d_i^{k-1}, d_i^{k+1}, \ldots, d_i^n$ in which d_i^k indicated the K-distance of drop-off point $P_i(x, y)$.

Step 2: we calculated the K-distance of each drop-off point in the data set based on Step 1.

Step 3: we sorted the K-distances of all drop-off points in ascending order and plotted the K-distance figure. In the figure, the K-distance of the inflection point was defined as Eps of the data set.

Taking the drop-off points data set on Wednesday, April 19, 2017, as an example, we analyzed the data in different lengths of time. We found that when the length of the time period exceeds 8 minutes, the change of the K-distance figure tends to be stable, and the characteristics of the inflection point are more clear, which was shown in Figure 2. Finally, considering the limitations of computer performance, we took the drop-off points data set (5,000–5,400 points in total) of 10 : 00–10 : 10 am on Wednesday, April 19, 2017, as an example, and its K-distance figure is shown in Figure 2(d), which showed the K-distance changed significantly around 0.00211. Therefore, 0.00211 was selected as the Eps. This value will be used in the clustering of one day's drop-off points data set to identify the hotspot ODs.

MinPts indicated the density of drop-off points in each cluster. In this paper, with the given Eps and assuming MinPts, clustering results of drop-off points can be obtained. According to the clustering results under different MinPts, reasonable MinPts can be determined. Under different MinPts, the clustering results of the drop-off points are shown in Table 1.

To obtain as many clusters as possible and to ensure each cluster has a sufficient number of pick-up or drop-off points, the value of MinPts was set to be 800. The 594 thousand drop-off points were clustered into 11 clusters (Table 1). When the value of MinPts was set to be 800, spatial distribution of drop-off clusters and number of trips of each cluster were obtained as shown in Figure 3.

2.3.2. Identification of Hotspot OD. In order to ensure that the trip between the selected OD pairs is of sufficient quantity and effectiveness, a hotspot OD identification method was proposed in this step. It consisted of the following two steps: (1) For each drop-off point in drop-off clusters shown in Figure 3 (14283 points in total), search the corresponding pick-up point and trajectory data in between; (2) Re-cluster the pick-up points. The DBSCAN algorithm was used for the re-clustering of pick-up points. The pick-up points generated by the 11 drop-off clusters, as shown in Figure 3, were re-clustered. Eighteen pairs of hotspot ODs were obtained (Table 2). The results show that using the method above only needs to process one day's data to ensure that the number of passenger-carrying routes between ODs is sufficient.

In Table 2, CCluster means a pick-up hotspot that was re-clustered. "Cluster 1–CCluster : 245" means there are 245 single passenger-carrying trips between the pick-up point Cluster1 and the drop-off point CCluster.

3. Analysis Methodology

3.1. Trip Length Classification. To test the hypothesis of heterogenous route choice behavior for trips with different lengths, the Akaike information criterion (AIC) was first used to classify trips into different categories based on their length.

A few studies on the classification of trips by travel distance can be found in the literature. In the survey of urban residents' travel, the travel distance was subjectively divided into few distance segments, such as 0~3 km, 3~6 km,

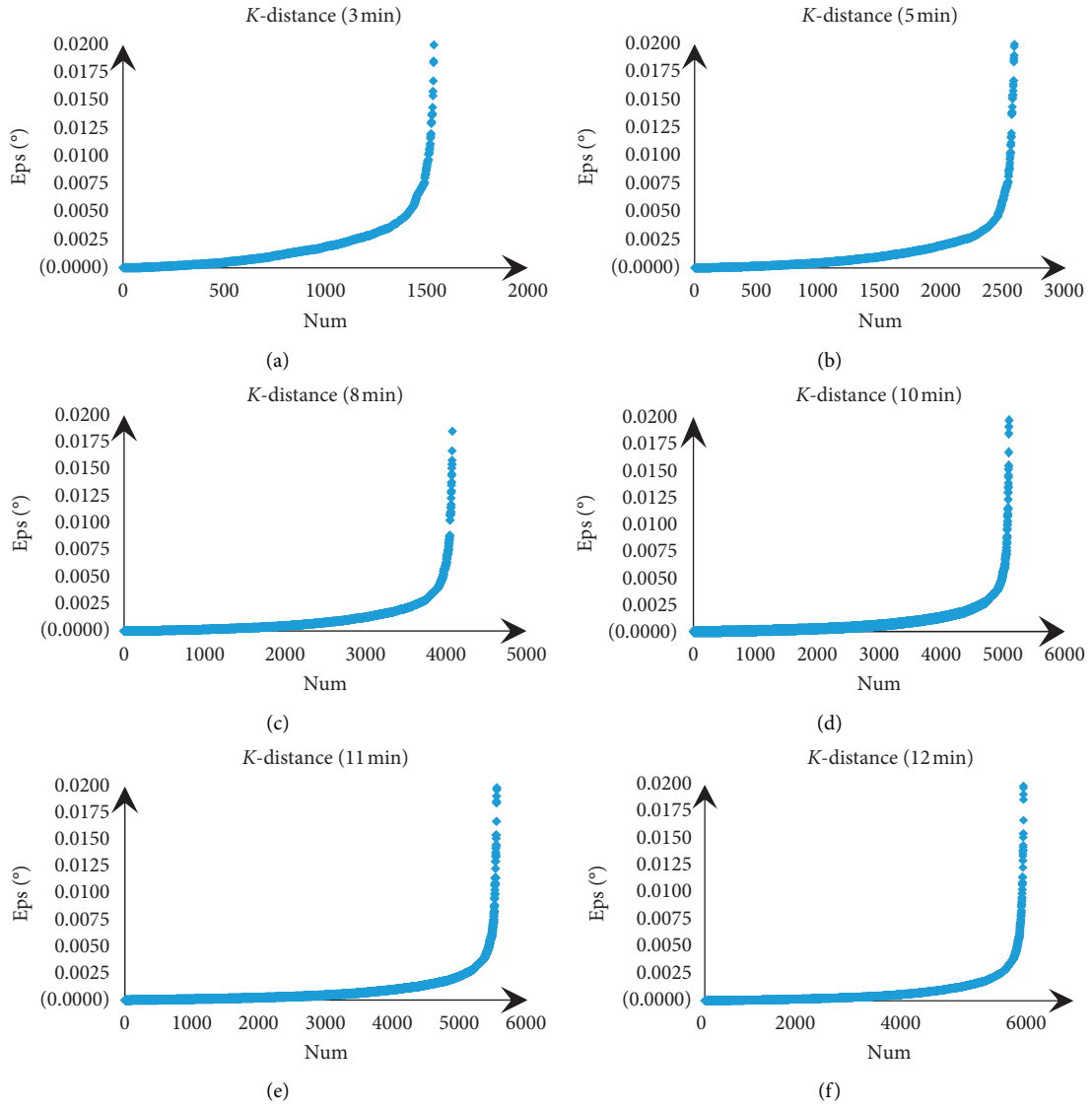


Figure 2: *K*-distance graph of GPS data in different time periods: (a) 3 min, (b) 5 min, (c) 8 min, (d) 10 min, (e) 11 min, and (f) 12 min.

Table 1: Clustering results of the pick-up (drop-off) points.

MinPts	Cluster of drop-off points	Noise data	Proportion of noise data	Cluster of pick-up points	Noise data	Proportion of noise data (%)
650	35	135277	68.03	34	135137	68.02
700	24	158980	79.95	26	158839	79.95
750	20	174013	87.51	17	173819	87.49
800	11	183478	92.27	11	183316	92.27
850	6	190179	95.64	6	190031	95.65
900	4	191671	96.39	3	191521	96.40

6~9 km, 9~12 km, and longer than 12 km [33, 34]. For mode split purpose, only qualitative classification of travel distance (short distance and long distance) was performed [35, 36]. In the route choice model, most studies used only one model to describe all the route choice behaviors [8, 21, 37]. For different types of passenger-carrying routes, the behavior of travelers was different. As such, currently a theoretical-sound method for classifying the travel routes is missing.

Based on the OD-Euclidean distance distribution of passenger-carrying routes, we sought for the eigenvalues with the travel volume changes significantly. These eigenvalues were used as the basis for the preliminary classification. The OD-Euclidean distance distribution of the 14,283 trips in 11 drop-off clusters mentioned in Section 2.3 is shown in Figure 4. In this section, we use this part of data for research.

Cluster	Cluster 1	Cluster 2	Cluster 3
Number of trips	914	2950	1004
Cluster	Cluster 4	Cluster 5	Cluster 6
Number of trips	977	1508	1338
Cluster	Cluster 7	Cluster 8	Cluster 9
Number of trips	979	855	1077
Cluster	Cluster 10	Cluster 11	Sum
Number of trips	1880	802	14283

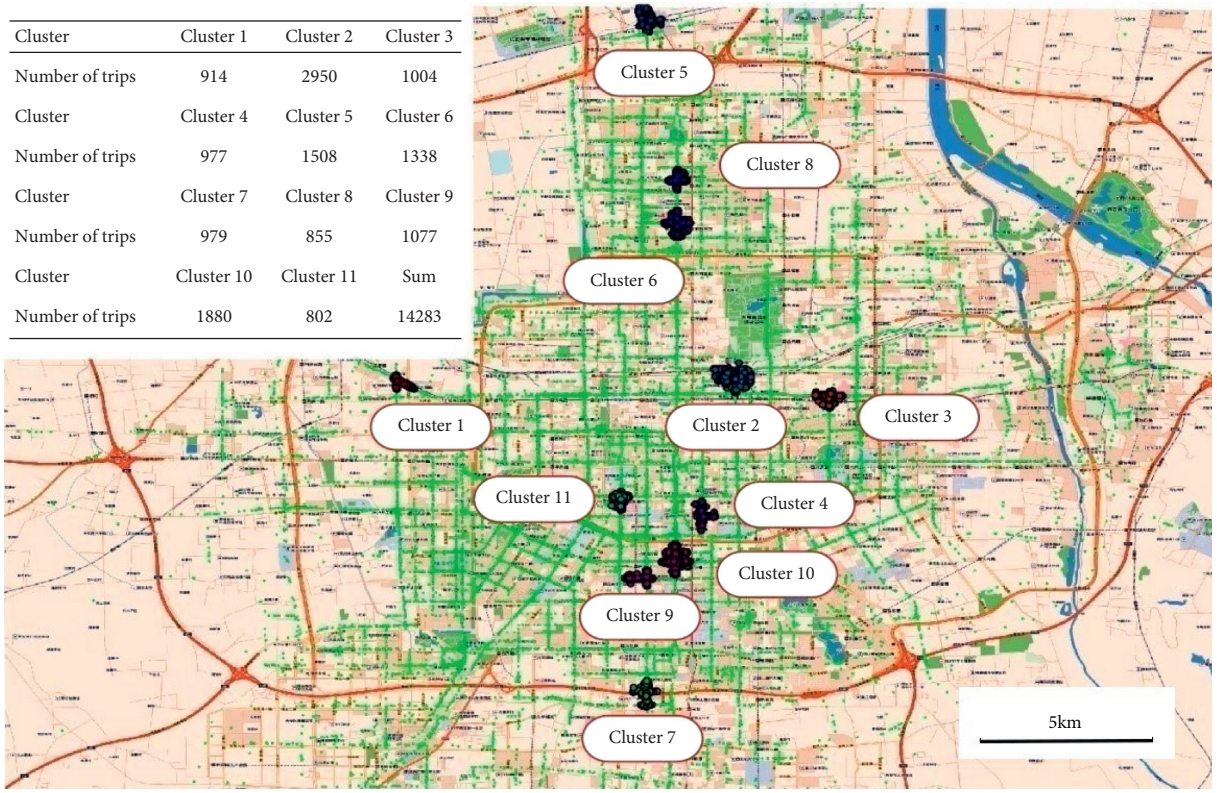
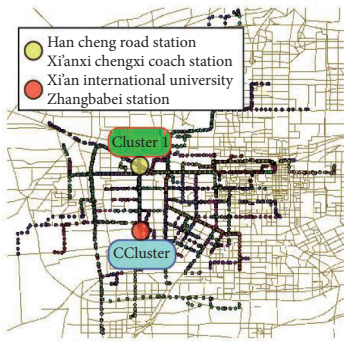
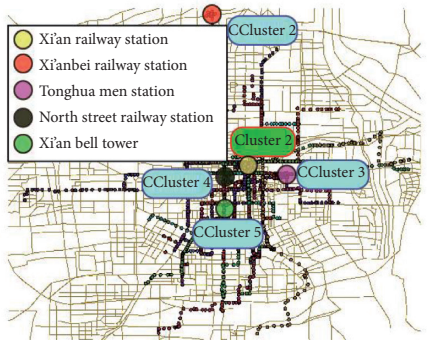


Figure 3: Spatial distribution of drop-off clusters.

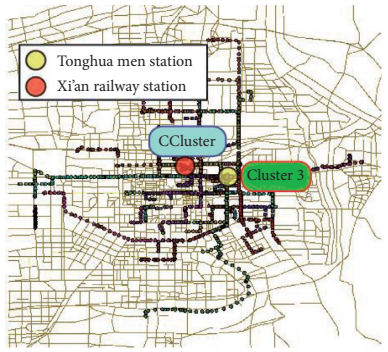
Table 2: Results of hotspot OD matching.



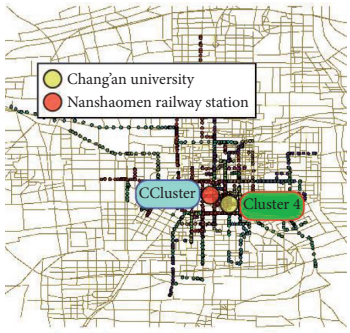
(1) Cluster 1–CCluster : 245



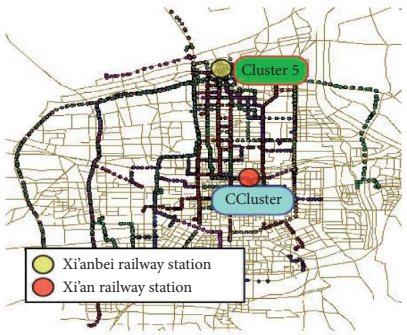
(2) Cluster 2–CCluster 1 : 262
(3) Cluster 2–CCluster 2 : 458
(4) Cluster 2–CCluster 3 : 289
(5) Cluster 2–CCluster 4 : 274



(6) Cluster 3–CCluster : 260



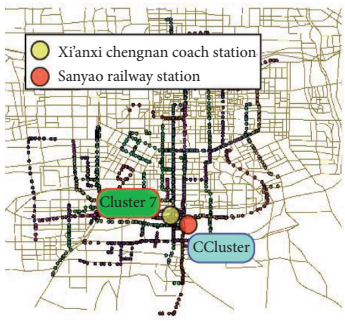
(7) Cluster 4–CCluster : 172



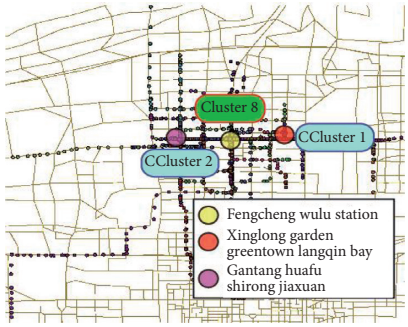
(8) Cluster 5–CCluster : 297

(9) Cluster 6–CCluster : 264

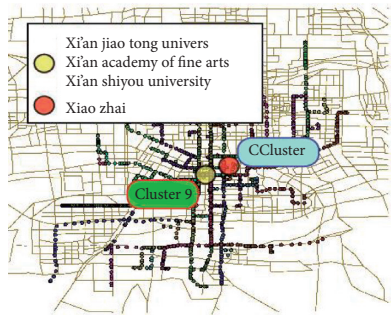
Table 2: Continued.



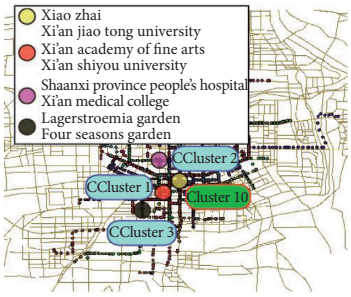
(10) Cluster 7–CCluster : 327



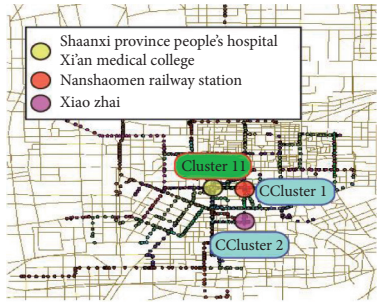
(11) Cluster 8–CCluster 1 : 137
(12) Cluster 8–CCluster 2 : 267



(13) Cluster 9–CCluster : 109



(14) Cluster 10–CCluster 1 : 306
(15) Cluster 10–CCluster 2 : 193
(16) Cluster 10–CCluster 3 : 176



(17) Cluster 11–CCluster 1 : 137
(18) Cluster 11–CCluster 2 : 204

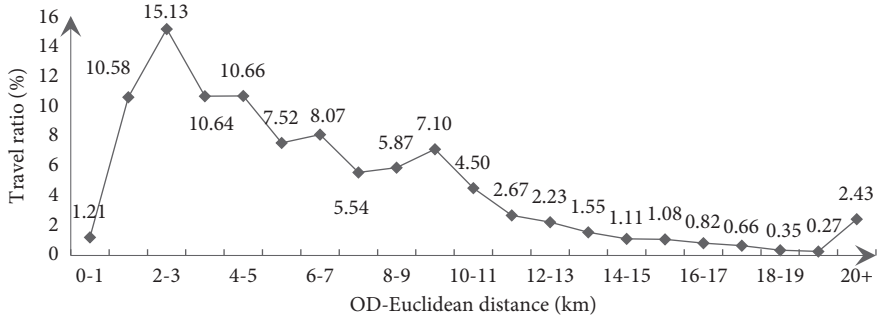


Figure 4: OD-Euclidean distance distribution of 11 drop-off clusters.

Figure 4 shows that at 3, 7, and 10 km, three peak values of travel volume can be observed. It is believed that these three peaks were consistent with the urban structure of Xi'an:

(1) 3 km radius: within 1–3 km of the Central Business District (CBD), there were many service facilities. These facilities can serve residents well and residents can fulfill their daily needs in this region, such as working, schooling, and shopping.

(2) 7 km radius: as a city with thousands of years of history, the CBD of Xi'an attracted a large number of trips. The CBD of Xi'an is located in the geometric center of the city, and CBD-centered 6-7 km covered major urban areas.

(3) 10 km radius: there are many passenger stations, airports, and tourist areas around the city, and these important points of interests also attracted a lot of travel. This phenomenon explains the occurrence of the third peak.

According to the above analysis, single passenger-carrying route of taxi can be divided into four categories: 0–3 km, 3–7, 7–10 km, and longer than 10 km. It should be noted that these were OD-Euclidean distances, which represented the linear distances between pick-up point and

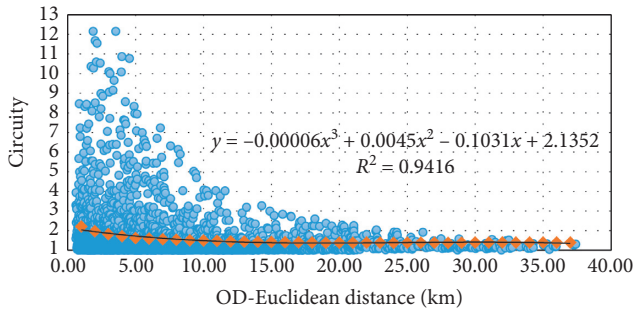


Figure 5: OD-Euclidean distances and circuity distribution.

drop-off point. It was difficult to reflect the actual length or travel time of the routes. In order to reflect the actual length of taxi passenger-carrying route, circuity was selected as another route classification index. We screened out the data of 14,283 trips, including the Euclidean distances and circuity of each OD as shown in Figure 5. The relationship between OD-Euclidean distances and average circuity for different types of passenger-carrying routes was fitted as follows. This is a typical regression curve fitting done using Microsoft Excel, and the results show an R square value of 0.9416, which indicates satisfactory results:

$$\mathrm{cir}_{rs} = -0.00006 \cdot \mathrm{eucl}_{rs}^3 + 0.0045 \cdot \mathrm{eucl}_{re}^2 - 0.1031 \cdot \mathrm{eucl}_{rs} + 2.1352, \tag{1}$$

where cir_{rs} is the circuity of passenger-carrying route from pick-up point r to drop-off point s and was calculated by the ratio of OD-Euclidean distance to the actual travel distance. eucl_{rs} is the OD-Euclidean distance of passenger-carrying route from pick-up point r to drop-off point s (unit: kilometer).

The mean values of 0–3 km, 3–7 km, and 7–10 were 1.5 km, 5 km, and 8.5 km, respectively. Considering that only 13.17% of OD-Euclidean distances were over 10 km, and 80% of them were distributed in 10–15 km, and 12.5 km was selected as the representative value. By introducing 1.5 km, 5 km, 8.5 km, and 12.5 km into equation (1), the initial clustering centers of five schemes can be calculated (1.9905, 1.7247, 1.5471, and 1.4324). In addition, there have been studies that divide the travel distance of travelers into 3 categories and above [35, 36]. Therefore, we decided to set the number of clusters to 3 or 4. If the number of clusters is 3, depending on the cluster center, there are 4 optional clustering schemes; if the number of clusters is 4, there is 1 optional clustering scheme. Five clustering schemes are shown in Table 3.

In order to compare the effect of the five clustering schemes, the AIC criterion, proposed by H. Akaike in information theory, was introduced to identify the best scheme.

$$\mathrm{AIC} = -2 \cdot \ln L(\widehat{\theta}) + 2 \cdot k, \tag{2}$$

where $L(\widehat{\theta})$ is the maximum likelihood estimation of the model, and with the increase in the difference among clusters, the value becomes larger. k is the number of parameters in the model, and the more classifications the model consists, the greater the value will be. The value of AIC depends on $L(\widehat{\theta})$ and k. The smaller the k is, the more concise the model becomes, and the larger the $L(\widehat{\theta})$ is, the more accurate the model will be. The AIC therefore considered both complexity and precision in identifying the best scheme.

Table 3: Initial clustering centers of five clustering schemes.

Scheme	Number of clusters	Initial clustering centers			
1	3	1.9905	1.7247	1.5471	—
2		1.9905	1.7247	1.4324	
3		1.9905	1.5471	1.4324	
4		1.7247	1.5471	1.4324	
5	4	1.9905	1.7247	1.5471	1.4324

For circuity data sets $\{X_i \mid i = 1, 2, \ldots, K\}$, which contained K circuitries of passenger-carrying routes. The number of clusters was N, the final cluster center of each cluster was $\{C_m \mid m = 1, 2, \ldots, N\}$, sample size of each cluster was $\{Q_m \mid m = 1, 2, \ldots, N\}$, and internal deviation of each cluster was $\{D_m \mid m = 1, 2, \ldots, N\}$.

$$D_m = \frac{1}{Q_m} \cdot \sum_{j=1}^{Q_m} \mathrm{dist}\left(X_j, C_m\right), \tag{3}$$

where $\mathrm{dist}(X_j, C_m)$ is the Euclidean distance between X_j and C_m, X_j is the circuity of a passenger-carrying route in cluster m, and C_m is the center of cluster m.

The density distribution of deviations in each cluster is shown in equation (4).

$$f(D_m) = \frac{Q_m/K}{(d_{\max} - d_{\min})/N} = \frac{N}{K} \cdot \frac{Q_m}{d_{\max} - d_{\min}}, \tag{4}$$

where $d_{\max} = \mathrm{Max}(D_m)$ and $d_{\min} = \mathrm{Min}(D_m)$.

According to the principle of logarithmic maximum likelihood estimation, the logarithmic maximum likelihood estimation functions of the internal deviations of each cluster ($\{D_m \mid m = 1, 2, \ldots, N\}$) can be obtained as follows:

$$\ln L\left(D_m \mid m = 1, 2, \ldots, N\right) = -N \cdot \ln \frac{K}{N} - \sum_{m=1}^{N} \ln \frac{d_{\max} - d_{\min}}{Q_m}. \tag{5}$$

Plug equation (5) into equation (2), the AIC, which was the basis of passenger-carrying route classification, can be calculated as follows:

$$\mathrm{AIC}(N) = 2 \sum_{m=1}^{N} \ln \frac{d_{\max} - d_{\min}}{Q_m} + 2N\left(1 + \ln \frac{K}{N}\right). \tag{6}$$

The clustering scheme with minimum AIC were selected as the optimal scheme. Five K-means clustering schemes were implemented by SPSS, which is a statistical analysis software package developed by IBM, and the AIC values of the five schemes, which as shown in Table 3, were 2.885, 2.6137, 2.8041, 3.5233, and 3.0231, respectively. The AIC value of scheme 2 was the smallest, which means that this scheme had the best balance in complexity and precision.

Accordingly, scheme 2 was considered as the optimal scheme.

In clustering scheme 2, the boundaries of cluster 1 were 1 and 1.489, which corresponded to the passenger-carrying routes with OD-Euclidean distance longer than 10 km. The boundaries of cluster 2 were 1.489 and 1.826, which corresponded to the passenger-carrying routes with OD-Euclidean distance between 3 km and 10 km. The boundaries of cluster 3 were 1.826 and 2.544, which corresponded to the passenger-carrying routes with OD-Euclidean distance between 0 km and 3 km. Accordingly, the classification results of taxi passenger-carrying routes were $0\,\text{km} \le D \le 3\,\text{km}$ (short distance), $3\,\text{km} \le D \le 10\,\text{km}$ (medium distance), and 10 km D (long distance), where D indicated the OD-Euclidean distance.

With such thresholds for trip lengths clarification, the Euclidean distance distribution of 18 pairs of hotspot ODs is shown in Figure 6.

The hotspot OD from Xiaozhai (Cluster18, pick-up cluster) to Shaanxi Province People's Hospital and Xi'an Medical College (Cluster11, drop-off cluster) was selected as the research object of short-distance taxi passenger-carrying route. The hotspot OD from Lagerstroemia Garden and Four Seasons Garden (Cluster16, pick-up cluster) to Xiaozhai (Cluster10, drop-off cluster) was selected as the research object of medium-distance taxi passenger-carrying route. The hotspot OD from Xi'an Bei Railway Station (Cluster2, pick-up cluster) to Xi'an Railway Station (Cluster2, drop-off cluster) was selected as the research object of long-distance taxi passenger-carrying route. These three OD pairs are illustrated in Figure 7.

3.2. Route Choice Probability Distribution Analysis. Figure 8 illustrates the actual probability distribution of route choice for different passenger-carrying route categories shown in Figure 7. The formula for calculating the fluctuation value of the path choice probability is as follow:

$$\Delta P = \max_k P_{ik}^{rs} - \min_k P_{ik}^{rs}, \tag{7}$$

where P_{ik}^{rs} stands for the probability of driver i choosing route k taxi from r to s.

It can be observed that the fluctuation of route choice probability can be summarized as follow: 0.2010 (short distance) <0.239 (long distance) <0.305 (medium distance). The following can be found:

(1) *Short-distance passenger-carrying routes* had the smallest fluctuation. A most likely explanation was that due to the limited scale of the network between short-distance hotspot OD pair, drivers did not have enough options to make a detour and utility values of difference routes were similar.

(2) *Medium-distance passenger-carrying routes* had the highest fluctuation. The scale of network between medium-distance hotspot OD pair was moderate, as drivers had more options to make a detour in acceptable travel time.

(3) The fluctuation of *long-distance passenger-carrying routes* was higher than short-distance routes but lower than medium-distance routes. It was probably because that the scale of network between long-distance hotspot OD pair was large and drivers had enough options to make a detour. However, the drivers' acceptable circuity or delays were small for long-distance passenger-carrying routes.

3.3. Explanatory Variables. In this study, route choice behavior modeling explanatory variables were selected from three aspects: path factor, road factor, and PS correction term. We defined the coefficients corresponding to the explanatory variables in the model as shown in Table 4 below.

In Table 4, the travel time (TT) equals to the difference between the origin and destination GPS timestamps of a single passenger-carrying trip, K represents the length of path, D represents the OD-Euclidean distance, N_p is the number of intersections, K_m stands for the length of main road, K_s represents the length of secondary, K_b represents the length of branch road, and K_{co} is the length of congested road, which is judged by the average travel speed of the road section from GPS data.

3.4. Path Size Logit Model. The traditional multinomial logit model was a discrete choice model based on the theory of random utility, which can be used to describe the individual's choice behavior. The model was simple and easy to understand. However, the IID assumption of utility random item led to the result that there were IIA characteristics in the model. The probability that two routes were selected was only related to the utility of them and not to other routes. However, according to Figure 6, we knew that there were many common roadway segments among different taxi passenger-carrying routes.

The path size logit model reflected this issue by introducing a correction term into the utility function. Therefore, the PS-logit model was adopted to analyze the taxi passenger-carrying route choice behavior in this paper. The utility function of PS-logit is shown in equation (8).

$$U_{ik}^{rs} = V_{ik}^{rs} + \beta_{PS}^{rs} \ln\left(\text{PS}_k^{rs}\right) + \varepsilon_k^{rs}. \tag{8}$$

U_{ik}^{rs}: utility of traveler i choosing route k, from pick-up point r to drop-off point s.

V_{ik}^{rs}: fixed utility of traveler i choosing route k, from pick-up point r to drop-off point s.

β_{PS}^{rs}: parameters to be calibrated.

PS_k^{rs}: path-size value of route k, from pick-up point r to drop-off point s.

$$\text{PS}_k^{rs} = \sum_{a \in \Gamma_k} \frac{l_a}{L_k^{rs}} \cdot \frac{1}{\sum_{j \in K_{rs}} \mu_{aj}^{rs}}, \quad \forall k \in K_{rs},\ rs \in \text{RS}. \tag{9}$$

Γ_k: roads set in route k.

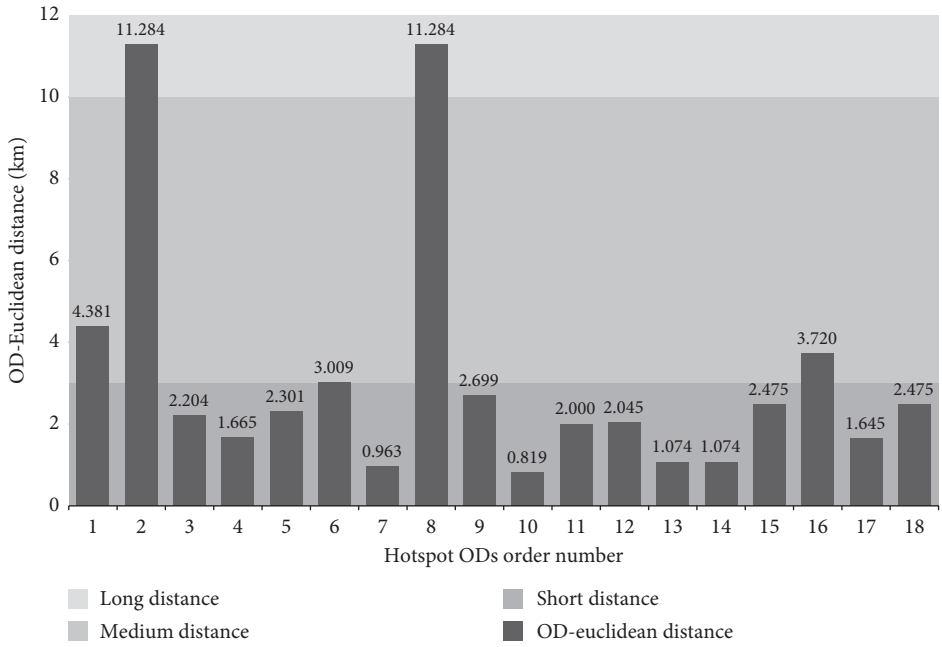
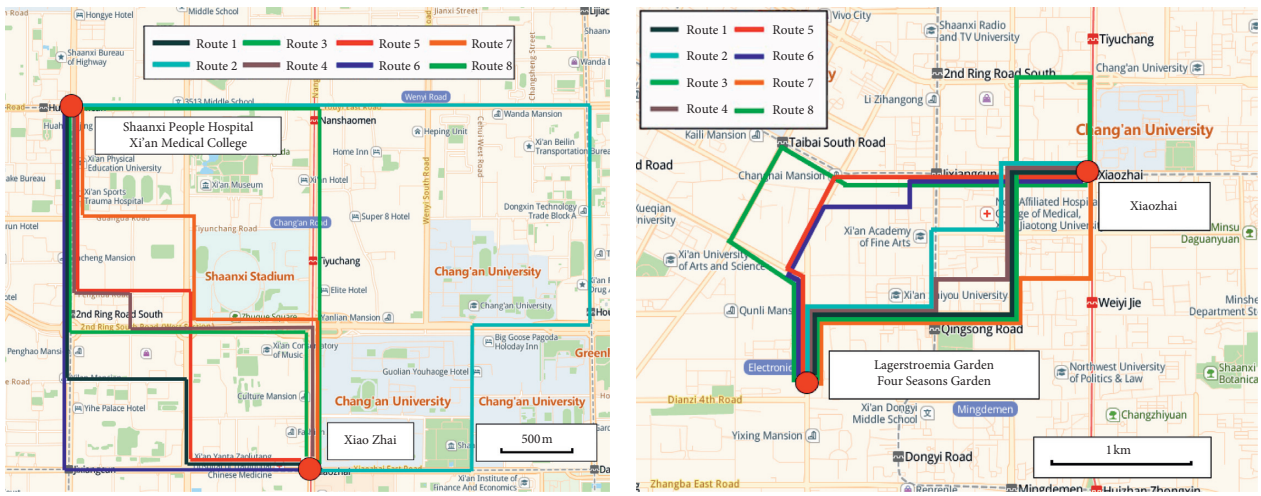
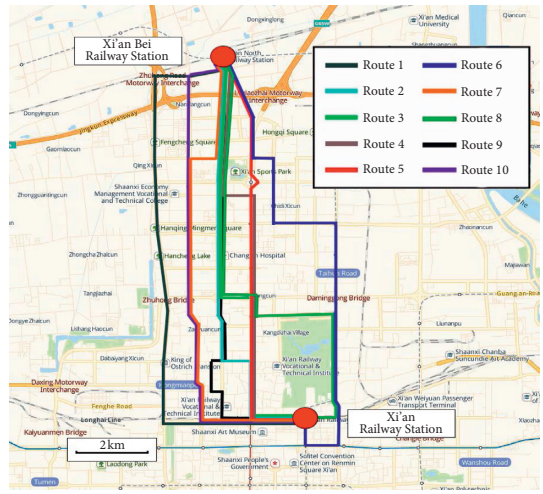


Figure 6: Euclidean distance distribution of hotspot ODs.



(a) (b)



(c)

Figure 7: Routes of different passenger-carrying route categories: (a) short-distance passenger-carrying routes, (b) medium-distance passenger-carrying routes, and (c) long-distance passenger-carrying routes.

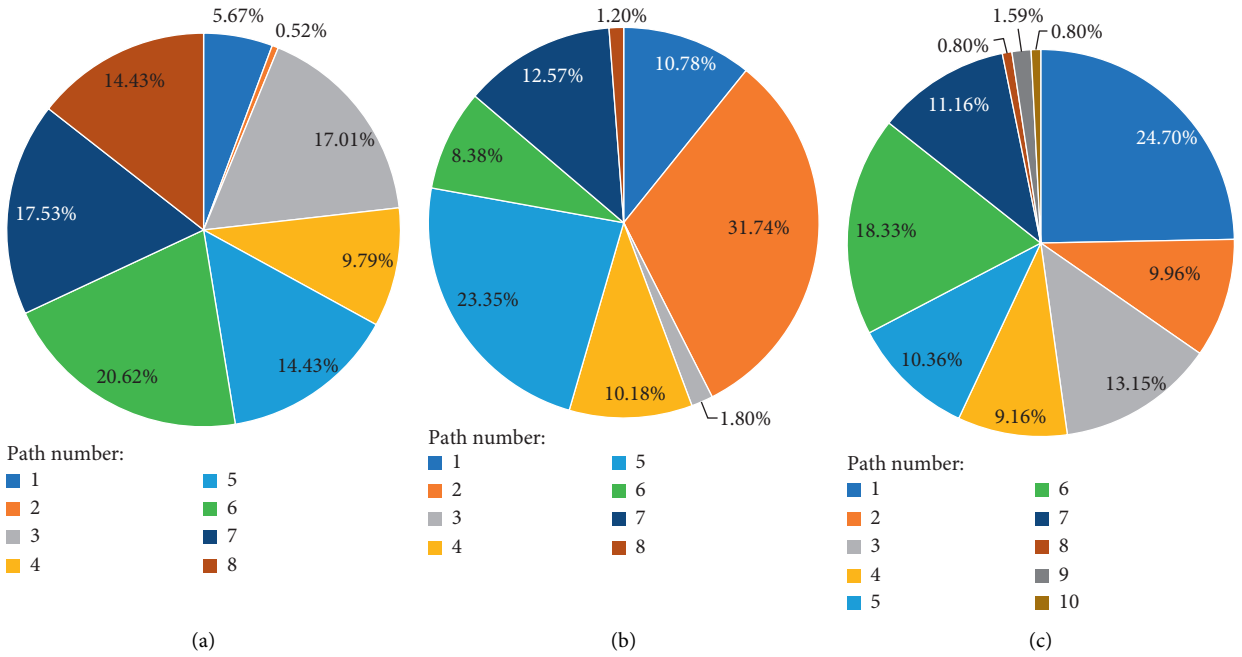


Figure 8: Probability distribution of route choice for different route categories: (a) short distance, (b) medium distance, and (c) long distance.

Table 4: Explanatory variables of taxi passenger-carrying route choice behavior.

Variables		Abbreviation	Coefficient	Unit	Equation
Route factors	Travel time	TT	β_{TT}	Minute	—
	Circuity	CI	β_{CI}	—	K/D
Road factors	Frequency of intersections	PT	β_{PT}	—/minute	N_p/TT
	Proportion of main road	TR	β_{TR}	%	K_m/K
	Proportion of secondary road	SR	β_{SR}	%	K_s/K
	Proportion of branch road	BR	β_{BR}	%	K_b/K
	Left turns	L	β_L	—	—
	Right turns	R	β_R	—	—
	Proportion of congestion	CO	β_{CO}	—	K_{co}/K
Path size		Ln(PS)	Ln(PS)	—	—

K_{rs}: routes set between r (pick-up point) and s (drop-off point).

RS: OD set.

l_a: length of road a.

L_k^{rs}: length of route k.

μ_{aj}^{rs}: if road a belongs to route j, μ_{aj}^{rs} equals to 1, otherwise μ_{aj}^{rs} equals to 0.

The PSL model for this study is constructed as follows:

$$P_{ik}^{rs} = \frac{\exp\left[\sum_{m=1}^{M}\left(\beta_m \cdot c_{m,ik}^{rs}\right) + \beta_{PS}^{rs} \cdot \ln\left(PS_k^{rs}\right)\right]}{\sum_{l \in K_{rs}} \exp\left[\sum_{m=1}^{M}\left(\beta_m \cdot c_{m,il}^{rs}\right) + \beta_{PS}^{rs} \cdot \ln\left(PS_l^{rs}\right)\right]}. \quad (10)$$

P_{ik}^{rs}: for the taxi passenger-carrying route from r to s, the probability that driver i chooses route k.

β_m: coefficient of explanatory variable m.

$c_{m,ik}^{rs}$: for the taxi passenger-carrying route from r to s, when the driver chooses route k, the value of explanatory variable m.

M: the number of explanatory variables.

4. Results and Discussion

4.1. Model Calibration Results. With the help of Biogeme software package, the parameters of MNL model and PS-Logit model with different types of passenger-carrying routes were calibrated, respectively. In addition, we aggregate all routes together as a control group. The results are shown in Table 5.

According to Table 5, for different route types, the t-statistics of explanatory parameters of the two models were statistically valid. The coefficient of PS correction term was positive, which was consistent with the basic principle of the

Table 5: Calibration results of taxi passenger-carrying route choice model.

Variable	Short distance				Medium distance				Long distance				All trips combined			
	PSL		MNL		PSL		MNL		PSL		MNL		PSL		MNL	
	Coef	*t*-stat	Coef	*t*-stat	Coef	*t*-stat	Coef	*t*-stat	Coef	*t*−stat	Coef	*t*−stat	Coef	*t*-stat	Coef	*t*-stat
β_{TT}	−0.41	−4.21	−0.61	−5.84	−0.46	−3.34	−0.62	−3.89	−0.55	−2.81	−0.83	−3.28	−0.74	−2.19	−0.86	−3.91
β_{CI}	−2.35	−2.62	−3.3	−3.42	−2.45	−2.88	−3.7	−3.92	−3.45	−2.2	−4.76	−3.77	−4.72	−2.71	−4.32	−2.66
β_{PT}	0.02	2.48	0.5	3.41	0.02	2.05	0.55	2.86	0.04	1.97	0.79	2.71	0.13	2.34	−0.12	3.26
β_{TR}	0.02	1.78	0.53	1.91	0.01	1.76	0.52	1.84	0.03	1.96	0.7	2.03	0.03	1.80	0.67	2.04
β_{SR}	−0.01	−1.69	−0.02	−1.83	−0.02	−1.88	−0.02	−1.9	−0.02	−1.78	−0.02	−1.7	−0.02	−1.72	−0.08	−2.49
β_{BR}	−3.87	−2.01	−5.43	−2.78	−8.29	−1.87	−7.83	−2.91	−22.35	−2.31	−24.37	−3.21	−10.10	−2.77	−7.54	−2.26
β_{L}	−0.42	−3.36	−0.53	−4.54	−0.45	−3.7	−0.54	−5.41	−0.58	−5.66	−0.82	−7.16	−0.47	−3.35	−0.19	−1.77
β_{R}	−0.17	−2.54	−0.43	−2.69	−0.17	−2.56	−0.45	−2.73	−0.21	−1.76	−0.48	−2.52	−0.38	−1.73	−0.43	−3.21
β_{CO}	−1.59	−3.07	−1.61	−3.56	−1.85	−2.29	−1.88	−2.58	−2.38	−3.56	−2.41	−3.17	−2.68	−2.06	−3.22	−2.01
Ln (PS)	3.47	3.2	—	—	4.02	3.69	—	—	4.71	5.11	—	—	3.37	2.12	—	—
Num of observations	204				176				262				642			
Likelihood ratio ρ^2	0.272		0.225		0.271		0.223		0.316		0.197		0.219		0.182	

Table 6: MRS of explanatory variables.

Factors	Symbol	MRS		
		Short distance	Medium distance	Long distance
Circuity	$MRS(\beta_{CI}, \beta_{TT})$	5.73170	5.32609	6.27272
Frequency of intersections	$MRS(\beta_{PT}, \beta_{TT})$	−0.04878	−0.04348	−0.07273
Proportion of main road	$MRS(\beta_{TR}, \beta_{TT})$	−0.04878	−0.02174	−0.05455
Proportion of secondary road	$MRS(\beta_{SR}, \beta_{TT})$	0.02439	0.04348	0.03636
Proportion of branch road	$MRS(\beta_{BR}, \beta_{TT})$	9.43902	18.02174	40.63636
Left turns	$MRS(\beta_{L}, \beta_{TT})$	1.02439	0.97826	1.05454
Right turns	$MRS(\beta_{R}, \beta_{TT})$	0.41463	0.36957	0.38182
Proportion of congestion	$MRS(\beta_{CO}, \beta_{TT})$	3.87805	4.02174	4.32727
Path size	$MRS(Ln(PS), \beta_{TT})$	−8.46341	−8.73913	−8.56364

PS-logit model. In addition, adjusted likelihood ratio of PS-Logit model was better than that of the MNL model, which meant that PS-logit model described drivers' passenger-carrying route choice behavior more accurately than the traditional MNL model. Finally, the adjusted likelihood of the control group was significantly lower than the other three groups, which showed that dividing the passenger-carrying route by distance can optimize the model. According to Table 5, the following conclusions can be drawn:

(1) The coefficients with positive values included β_{PT}, β_{TR}, and Ln(PS). The coefficients with negative values included β_{TT}, β_{CI}, β_{SR}, β_{BR}, β_{L}, β_{R}, and β_{CO}. This showed that when drivers chose routes, they tended to choose roads with high proportion of main roads, lower circuity, shorter travel time, and less congestion, regardless of the length of travel distance.

(2) With the increase of travel distance, the absolute value of β_{CI}, β_{PT}, β_{TR}, β_{SR}, β_{BR}, and β_{CO} increased obviously. This indicated that as travel distance increases, the impacts of circuity, path structure, and the congestion proportion of the choice of the driver will also increase.

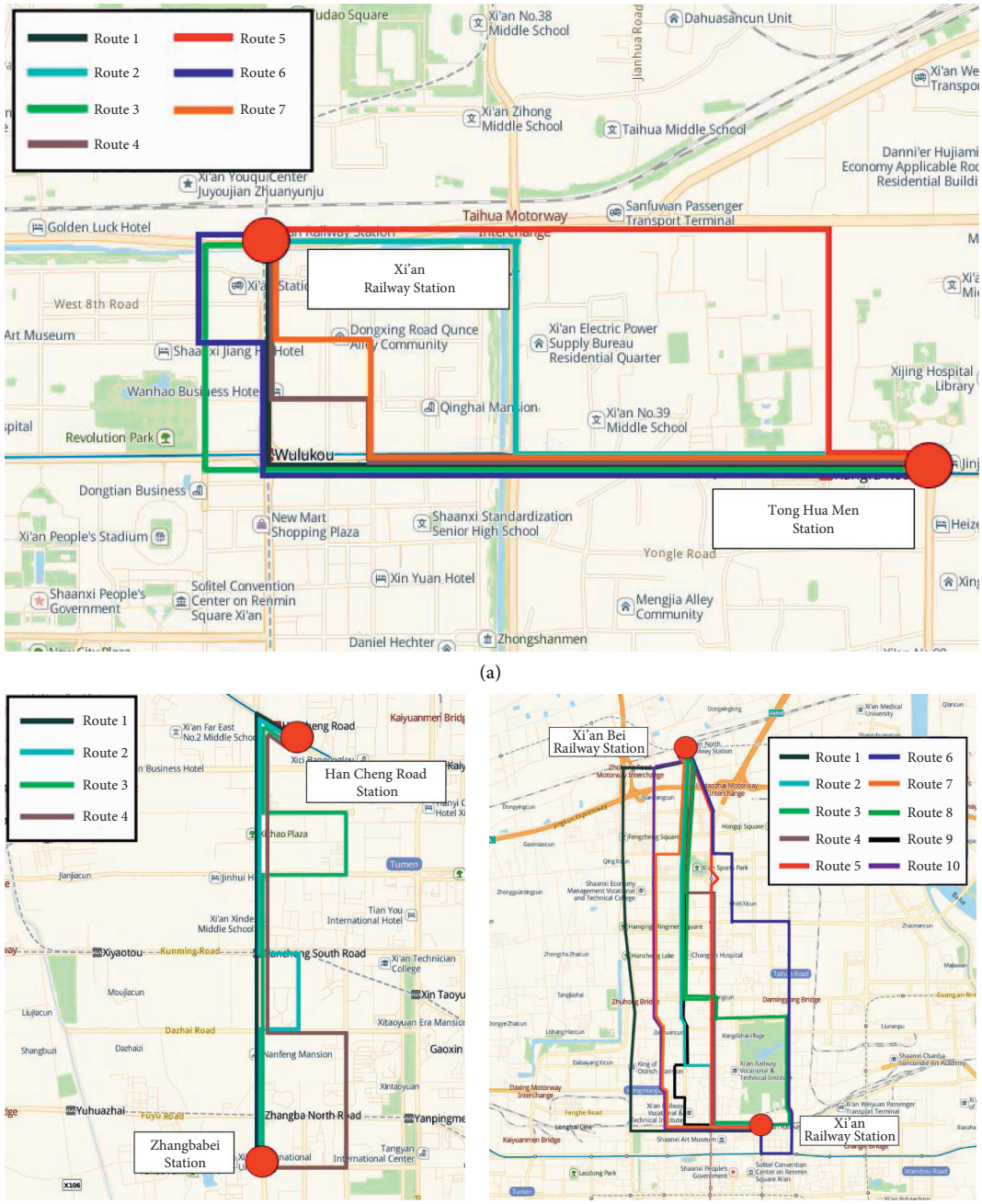
4.2. Route Choice Preference Analysis. With the level of consumer satisfaction unchanged, the marginal rate of substitution (MRS) referred to the scenario that when consumers increased one unit of a product and needed to abandon certain number of another product. Many existing research studies use MRS in the analysis of the calibration results of the choice model [38, 39]. In this paper, with the utility of passenger-carrying route kept unchanged, MRS was defined as the change of basic variable when the other explanatory variables increased by one unit. It can be calculated as follows:

$$\mathrm{MRS}\left(\alpha_i, \alpha_j\right) = \frac{\partial U/\partial \alpha_i}{\partial U/\partial \alpha_j}. \tag{11}$$

In this study, the PS-Logit model with a better adjust likelihood ratio was selected as the analysis object. Travel time was selected as the basic variable, the MRS between travel time and other explanatory variables are shown in Table 6.

According to Table 6, the following conclusions can be drawn:

(1) The relationship among the MRS of explanatory variables was found to be $MRS(\beta_{BR}, \beta_{TT}) > MRS(Ln(PS), \beta_{TT}) > MRS(\beta_{CI}, \beta_{TT}) > MRS(\beta_{CO}, \beta_{TT}) >> MRS(\beta_{L}, \beta_{TT}) > MRS(\beta_{R}, \beta_{TT}) > MRS(\beta_{PT}, \beta_{TT}) > MRS$



(a)

(b)

(c)

Figure 9: Routes of different passenger-carrying route categories for verification: (a) short-distance passenger-carrying routes, (b) medium-distance passenger-carrying routes, and (c) long-distance passenger-carrying routes.

Table 7: Short-distance passenger-carrying route choice model verification results.

Route number	Actual choice times	True choice ratio	Calculated choice times	Calculated choice ratio	Number of errors
Route 1	233	0.524	213	0.479	20
Route 2	79	0.178	90	0.203	11
Route 3	66	0.148	47	0.105	19
Route 4	23	0.052	41	0.092	18
Route 5	16	0.036	23	0.052	7
Route 6	11	0.025	8	0.018	3
Route 7	17	0.038	23	0.052	6
Total	445	1.000	445	1.000	84
Hit ratio			0.81421		

Table 8: Medium-distance passenger-carrying route choice model verification results.

Route number	Actual choice times	True choice ratio	Calculated choice times	Calculated choice ratio	Number of errors
Route 1	154	0.815	136	0.720	18
Route 2	12	0.063	19	0.101	7
Route 3	9	0.048	5	0.026	4
Route 4	14	0.074	29	0.153	15
Total	189	1.000	189	1.000	44
Hit ratio			0.76720		

Table 9: Long-distance passenger-carrying route choice model verification results.

Route number	Actual choice times	True choice ratio	Calculated choice times	Calculated choice ratio	Number of errors
Route 1	77	0.266	71	0.247	6
Route 2	32	0.111	29	0.100	3
Route 3	35	0.121	38	0.131	3
Route 4	22	0.076	26	0.092	4
Route 5	25	0.087	31	0.104	5
Route 6	57	0.197	53	0.183	4
Route 7	27	0.093	32	0.112	5
Route 8	3	0.010	2	0.008	1
Route 9	7	0.024	5	0.016	2
Route 10	4	0.014	2	0.008	2
Total	289	1.000	289	1.000	35
Hit ratio			0.87889		

$(\beta_{TR}, \beta_{TT}) > \mathrm{MRS}(\beta_{SR}, \beta_{TT})$. If the goal was to reduce travel time, the first and foremost factors to be considered should be proportion of branch road, path-size value, circuity, and proportion of congestion. The minor factors to be considered should be the number of left turn, right turn, number of nodes per minute, and the proportion of main road and secondary road.

(2) As the distance of passenger-carrying route increased, the MRS of circuity and proportion of branch road and congestion also increased. On the contrary, the MRS of frequency of intersections decreased. When the distance of passenger-carrying route was long, drivers usually avoided routes with high circuity and proportion of congestion and preferred to choose the routes with high proportion of freeway or highway segments.

(3) To maintain the utility of passenger-carrying route unchanged, if the number of left turn increased by one, for short-distance, medium-distance, and long-distance passenger-carrying routes, the travel time needed to be reduced by 1.02, 0.98, and 1.05 min, respectively. If the number of right turns increased by one, for short-distance, medium-distance, and long-distance passenger-carrying routes, the travel time needed to be reduced by 0.41, 0.37, and 0.38 min, respectively. Time cost of left turn was about 2.6 times as high as that of right turn.

4.3. Verification of Route Choice Model. The verification of the path selection model is mainly achieved by comparing the trial calculation results of the path selection model with the actual path selection results, and finally the model's hit ratio to evaluate the effectiveness of the model is obtained. The calculation steps of hit ratio are as follows:

Step 1: assuming that the total number of samples is N, the total number of alternatives is M, there are K parameters in the final calibration result of the model, and the parameter calibration value β_k and the corresponding parameter value C_k are brought into the calibration model to obtain the selection probability $\widehat{P}_m$ of the corresponding program.

Step 2: assuming that traveler n has the greatest probability of choosing the route m, then $\widehat{\delta}_{mn} = 1$, otherwise $\widehat{\delta}_{mn} = 0$.

Step 3: when the actual selection result δ_{mn} of the traveler is consistent with the predicted result of the calibration model, set $S_m^n = 1$, otherwise $S_m^n = 0$. Then, the hit rate can be calculated as follows:

$$\mathrm{HR} = \frac{1}{N} \sum_{n=1}^{N} S_m^n. \tag{12}$$

In this paper, three different types of OD, as shown in Table 2, were selected to verify the model: The hotspot OD from Tong Hua Men Station (CCluster3, pick-up cluster) to Xi'an Railway Station (Cluster2, drop-off cluster) was selected as the verification object of short-distance taxi passenger-carrying routes; the hotspot OD from Han Cheng Road Station (CCluster, pick-up cluster) to Zhangbabei Station (Cluster1, drop-off cluster) was selected as the verification object of medium-distance taxi passenger-carrying routes; and the hotspot OD from Xi'an Railway Station (Cluster2, pick-up cluster) to Xi'an Bei

Railway Station (CCluster2, drop-off cluster) was selected as the verification object of long-distance taxi passenger-carrying routes. After removing abnormal data, these three ODs have 445, 189, and 289 valid trips and 7, 4, and 10 valid routes, respectively. The routes between these three ODs are shown in Figure 9.

According to the route choice model constructed in Section 4.1, the route choice results of each hotspot OD are calculated and compared with the actual choice situation. The results are shown in Tables 7–9.

The Tables 7–9 show that the hit ratios of the short-distance, medium-distance, and long-distance passenger-carrying route choice models are 0.81421, 0.76720, and 0.87889, respectively, indicating that the three types of route choice models constructed are effective and can be explained reasonably the behavior of passenger-carrying route choice. The analysis of extra OD pairs requires the substantial amount of manual work.

5. Conclusion and Future Work

This manuscript, for the first time, focused on the analysis of route choice behavior based on the massive amount of real-world GPS trajectory data collected from the occupied taxi cabs. Our analysis based on the trajectory data from Xi'an, China, found that for trips with different lengths, the characteristics of route choice behavior could be very different. As such, according to the distribution of Euclidean distance and volume, five route classification schemes for taxi passenger-carrying routes were proposed based on the circuity K-means clustering method. The Akaike information criterion (AIC) principle was adopted to identify the best route classification scheme. After that, taxi passenger-carrying routes were divided into three categories: short distance, medium distance, and long distance. Based on the MNL model, three PS-Logit models were proposed to analyze the route choice behaviors. The numerical analysis validated our hypothesis and revealed heterogenous activity patterns and influencing factors for trips with different lengths.

According to the study, the following conclusions can be drawn: (1) taxi passenger-carrying routes can be classified based on the distribution of Euclidean distance and K-means clustering of circuities; (2) for different taxi passenger-carrying routes, the fluctuation of route choice probability can be summarized as follows: short distance < long distance < medium distance; (3) for different taxi passenger-carrying routes, the first and foremost factors to be considered were proportion of branch road, path-size value, circuity, and proportion of congestion. The minor factors to be considered were the number of left turns, right turns, the number of nodes per minute, and the proportion of main road and secondary road; (4) with the increase of travel distance, drivers usually avoided routes with high circuity and intersection density but preferred to choose the routes with high proportion of freeway or highway; and (5) the effects of circuity, frequency of intersections, path structure, and congestion degree on utility function were significantly different among different taxi passenger-carrying route categories.

Finally, we have selected another OD pair for each category for validation purpose, and the analysis shows consistent conclusions. Future research could be focused on using the data set from other cities to validate the model. The works to be improved are as follows: On the one hand, the variables considered in the model in this paper were easy to be defined, while some other factors that were difficult to be defined or computed were not taken into account such as trip purpose, preference, network familiarity, and influence of weather and environment. On the other hand, in this manuscript, only Euclidean distance, travel volume, and circuity were considered in the taxi passenger-carrying route classification. If more data types become available, more factors could be considered such as the network structure among the hotspot OD. How to identify and select sufficient factors to improve the route classification results may need further discussion.

Data Availability

The GPS trajectory data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The research is supported by the National Key Research and Development Program of China (grant no. 2018YFB1600900), the Shaanxi Provincial Science and Technological Project (grant nos. 2020JM-244), and the Science and Technological Project of Shaanxi Provincial Transport Department (grant no. 19-24X).

References

[1] D. Mcfadden and F. Reid, "Aggregate travel demand forecasting from disaggregate behavioral models," *Transportation Research Record*, vol. 534, pp. 24–37, 1975.

[2] R. B. Dail, "A probabilistic multipath traffic assignment model, which obviates path enumeration," *Transportation Research*, vol. 5, no. 2, pp. 83–111, 1971.

[3] M. Ben-Akiva and S. R. Lerman, "Disaggregate travel and mobility choice models and measures of accessibility," in *Proceedings of the 3rd International Conference on Behavioral Travel Modelling*, Tanunda, Australia, 1977.

[4] E. Cascetta, A. Nuzzolo, F. Russo, and F. Vitetta, "A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks," in *Proceedings of the Thirteenth International Symposium on Transportation and Traffic Theory*, pp. 697–711, Lyon, France, 1996.

[5] M. S. Ramming, *Network Knowledge and Route Choice*, Massachusetts Institute of Technology, Cambridge, MA, USA, 2002.

[6] C.-H. Wen and F. S. Koppelman, "The generalized nested logit model," *Transportation Research Part B: Methodological*, vol. 35, no. 7, pp. 627–641, 2001.

[7] F. S. Koppelman and C.-H. Wen, "The paired combinatorial logit model: properties, estimation and application," *Transportation Research Part B: Methodological*, vol. 34, no. 2, pp. 75–89, 2000.

[8] Q. Tang and X. Hu, "Modeling individual travel time with back propagation neural network approach for advanced traveler information systems," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 6, Article ID 04020039, 2020.

[9] Z. Li, R. Kluger, X. Hu, Y.-J. Wu, and X. Zhu, "Reconstructing vehicle trajectories to support travel time estimation," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2672, no. 42, pp. 148–158, 2018.

[10] H. Noh, A. Sun, X. Hu, and A. Rehan, "Development of a regional network performance measurement model for planning application based on high-frequency GPS data," in *Proceedings of the Transportation Research Board 96th Annual Meeting*, Washington, DC, USA, January 2017.

[11] X. Zhu, Y. Yuan, X. Hu, Y.-C. Chiu, and Y.-L. Ma, "A Bayesian network model for contextual versus non-contextual driving behavior assessment," *Transportation Research Part C: Emerging Technologies*, vol. 81, pp. 172–187, 2017.

[12] X. Hu, Y.-C. Chiu, Y.-L. Ma, and L. Zhu, "Studying driving risk factors using multi-source mobile computing data," *International Journal of Transportation Science and Technology*, vol. 4, no. 3, pp. 295–312, 2015.

[13] X. Hu, X. Zhu, Y.-C. Chiu, and Q. Tang, "Will information and incentive affect traveler's day-to-day departure time decisions?-An empirical study of decision making evolution process," *International Journal of Sustainable Transportation*, vol. 14, no. 6, pp. 403–412, 2020.

[14] H. Xianbiao, C. Yi-Chang, and Z. Lei, "Behavior insights for an incentive-based active demand management platform," *International Journal of Transportation Science and Technology*, vol. 4, no. 2, pp. 119–134, 2015.

[15] H. Qi and X. Hu, "Real-time headway state identification and saturation flow rate estimation: a hidden Markov chain model," *Transportmetrica A: Transport Science*, vol. 16, no. 3, pp. 840–864, 2020.

[16] Y. Deng, X. Luo, X. Hu, Y. Ma, and R. Ma, "Modeling and prediction of bus operation states for bunching analysis," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 9, Article ID 04020106, 2020.

[17] Y.-J. Deng, X.-H. Liu, X. Hu, and M. Zhang, "Reduce bus bunching with a real-time speed control algorithm considering heterogeneous roadway conditions and intersection delays," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 7, Article ID 04020048, 2020.

[18] N. S. Dhakar and S. Srinivasan, "Route choice modeling using GPS-based travel surveys," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2413, no. 1, pp. 65–73, 2014.

[19] N. Schussler and K. W. Axhausen, "Accounting for route overlap in urban and sub-urban route choice decisions derived from GPS observations," in *Proceedings of the 12th International Conference on Travel Behavior Research*, Jaipur, India, 2009.

[20] J. Kim and H. S. Mahmassani, "Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories," *Transportation Research Part C: Emerging Technologies*, vol. 59, pp. 375–390, 2015.

[21] T. Thomas and B. Tutert, "Route choice behavior in a radial structured urban network: do people choose the orbital or the route through the city center?" *Journal of Transport Geography*, vol. 48, pp. 85–95, 2015.

[22] D. Li, T. Miwa, T. Morikawa, and P. Liu, "Incorporating observed and unobserved heterogeneity in route choice analysis with sampled choice sets," *Transportation Research Part C: Emerging Technologies*, vol. 67, pp. 31–46, 2016.

[23] M. Bierlaire and E. Frejinger, "Route choice modeling with network-free data," *Transportation Research Part C: Emerging Technologies*, vol. 16, no. 2, pp. 187–198, 2008.

[24] T. Miwa, T. Morikawa, and S. Kurauchi, "Preliminary analysis on dynamic route choice behavior using probe-vehicle data," *Infrastructure Planning Review*, vol. 22, pp. 467–476, 2005.

[25] T. Yamamoto, S. Takamura, and T. Morikawa, "Structured random walk parameter for heterogeneity in trip distance on modeling pedestrian route choice behavior at downtown area," *Travel Behaviour and Society*, vol. 11, pp. 93–100, 2018.

[26] X. Hu, Y. Yuan, X. Zhu, H. Yang, and K. Xie, "Behavioral responses to pre-planned road capacity reduction based on smartphone GPS trajectory data: a functional data analysis approach," *Journal of Intelligent Transportation Systems*, vol. 23, no. 2, pp. 133–143, 2019.

[27] X. Hu, S. Gao, Y. Chiu, and D. Lin, "Modeling routing behavior for vacant taxicabs in urban traffic networks," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2284, no. 1, pp. 81–88, 2012.

[28] Q. Tang, X. Hu, and H. Qi, "Modeling routing behavior learning process for vacant taxis in a congested urban traffic network," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 6, Article ID 04020043, 2020.

[29] X. Yu, S. Gao, X. Hu, and H. Park, "A Markov decision process approach to vacant taxi routing with e-hailing," *Transportation Research Part B: Methodological*, vol. 121, pp. 114–134, 2019.

[30] X. Yu, S. Gao, X. Hu, and H. Park, "Multi-cycle optimal taxi routing with e-hailing," in *Proceedings of the Transportation Research Board 97th Annual Meeting*, Washington, DC, USA, 2018.

[31] X. Yu, S. Gao, and X. Hu, "Optimizing vacant taxis' routing decisions: model-based and model-free approaches," in *Proceedings of the Transportation Research Board Annual Conference*, Washington, DC, USA, 2019.

[32] T. N. Tran, K. Drab, and M. Daszykowski, "Revised DBSCAN algorithm to cluster data with dense adjacent clusters," *Chemometrics and Intelligent Laboratory Systems*, vol. 120, pp. 92–96, 2013.

[33] G. Wallner, S. Kriglstein, E. Chung, and S. A. Kashfi, "Visualisation of trip chaining behaviour and mode choice using household travel survey data," *Public Transport*, vol. 10, no. 3, pp. 427–453, 2018.

[34] Y. Sun, J. Shi, and P. M. Schonfeld, "Identifying passenger flow characteristics and evaluating travel time reliability by visualizing AFC data: a case study of Shanghai Metro," *Public Transport*, vol. 8, no. 3, pp. 341–363, 2016.

[35] W. Zhou, W. Wang, and Z. Guo, "Travel distance of different traffic modes based on different grades of roads," *Journal of Wuhan University of Technology (Transportation Science & Engineering)*, vol. 33, no. 5, pp. 976–979, 2009.

[36] M. Li, G. Song, and C. Ying, "Research on excessive short distance car trips in urban area," *Journal of Beijing Jiaotong University*, vol. 38, no. 3, pp. 15–21, 2014.

[37] Y. Yang, E. Yao, and L. Pan, "Taxi route choice behavior modeling based on GPS data," *Journal of Transportation*

Systems Engineering and Information Technology, vol. 15, no. 1, pp. 81–86, 2015.

[38] R. Sebastián, J. C. MunOz, and L. D. Grange, "A topological route choice model for metro," *Transportation Research Part A: Policy and Practice*, vol. 45, no. 2, pp. 138–147, 2011.

[39] H. Jeffrey, S. Elizabeth, and C. Billy, "A GPS-based bicycle route choice model for San Francisco, California," *Transportation Letters*, vol. 3, no. 1, pp. 63–75, 2011.