J. Shanghai Jiao Tong Univ. (Sci.), 2017, 22(6): 682-687
https://doi.org/10.1007/s12204-017-1890-9

Analyzing Behavior Differences of Occupied and Non-Occupied Taxi Drivers Using Floating Car Data

NIAN Guangyue[1,2] (年光跃), *LI Zhe*[2,3] (李 喆), *ZHU Weiquan*[4] (朱伟权), *SUN Jian*[1,2*] (孙 健)
(1. State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
2. Center for UAV Applications and ITS Research, School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;
3. Shanghai Municipal Transportation Design Institute Co. Ltd., Shanghai 200030, China;
4. Shenzhen Transportation Design and Research Institute Co. Ltd., Shenzhen 518003, Guangdong, China)

Abstract: As the travel purpose of non-occupied taxies is to find new passengers rather than to arrive at the destination, large differences exist in the route choice behavior between the occupied and non-occupied taxies. With the assistance of geographic information system (GIS) and taxi-based floating car data (FCD), this paper investigates the behavior differences between occupied and non-occupied taxi drivers with the same origin and destination. Descriptive statistical indexes from the FCD in Shenzhen, China are explored to identify the route choice characteristics of occupied and non-occupied taxies. Then, a conditional logit model is proposed to model the quantitative relationship between drivers' route choice and the related significant variables. Attributes of the variables related to non-occupied taxies' observed routes are compared with the case of occupied ones. The results indicate that, compared with their counterparts, non-occupied taxi drivers generally pay more attention to choosing arterial roads and avoiding congested segments. Additionally, they are also found less sensitive to fewer traffic lights and shorter travel time. Findings from this research can assist to improve urban road network planning and traffic management.
Key words: non-occupied taxies, route choice behavior, floating car data (FCD), geographic information system (GIS)
CLC number: U 491.1 **Document code:** A

0 Introduction

An in-depth comprehension on drivers' route choice decision-making is essential to urban road network planning and traffic management. Especially, taxi is an important mode of urban transportation. For example, until 2010, the share of road resources by taxi is about 20.43% in Beijing, and the non-occupied rate of taxi usually exceeds 50%[1]. Unfortunately, the vehicle behavior of non-occupied taxies is not as regular free flow or car-following, which may increase the uncertainty of traffic flow. Consequently, analyses of non-occupied taxi drivers' travel behaviors are critical in generating an appropriate assignment of traffic volume, and thus to improve the performance of urban road network.

Theories on travel behavior provide a novelty perspective to explore the relationship between drivers' sense perception (route attributes) and their route choice behavior[2]. Within the framework, the process of route choosing is explained as a utility maximization process[3]. The route choice behaviors of non-occupied taxi drivers or the behavior differences between non-occupied taxi drivers and occupied taxi drivers are seldom considered in previous literature, and consequently are worth further analyzing.

The main motivation of this study is to investigate the differences between occupied and non-occupied taxies in route choice behavior characteristics, with assistance of geographic information system (GIS) technologies. Route choice behavior data of taxi drivers are collected from the taxi-based floating car data (FCD). A conditional logit model is proposed to explain the observed route choice behavior of occupied and non-occupied taxies. More specifically, the first sub-objective is to analyze the statistics of the route chosen by non-occupied taxi drivers and occupied taxi drivers, with the corresponding characteristics; the

Received date: 2016-12-05
Foundation item: the Major Project of National Social Science Foundation of China (No. 16ZDA048), the Shanghai Municipal Natural Science Foundation, China (No. 17ZR1445500), and the Humanities and Social Science Research Project of Ministry of Education, China (No. 15YJCZH148)
*E-mail: danielsun@sjtu.edu.cn

second sub-objective is to model the choice process of the two types of drivers and expose the possible impact degree of the attributes on the process; the third sub-objective is to explain the route choice decision by the observed preference characteristics.

As urban road network planning requires complete understanding of travel behavior, researchers in transportation area are rather interested in characterizing travel behavior of various drivers. In the early studies, the routes chosen were considered as either the shortest one or the fastest one[4-5]. However, the actual choices by drivers do not always comply with the traditional assumption. For studying the route choice behavior, scholars have endeavored to explain the route choice process with various experiments. Some studies[6-7] assumed that the travel time of alternative routes is independent of drivers' route choice behavior and is only related to the attributes of the alternative routes. Beyond transportation area, Barron and Erev[8-9] assumed that the route information provided by drivers is solely depended on a feedback mechanism on their previous driving experiences. Studies were carried out to investigate the accuracy of traveler perceptions, indicating that the travelers' perception usually differs with the environment[10]. However, Raghubir et al.[11] found that the travel time perceptions are also related to the route trip familiarity.

Various data were collected from multiple sources and used to analyze the travel behavior. Previous studies, in particular, were based on either stated preference surveys[12] or travel simulators[13]. Nonetheless, they were impossible to capture driver behavior in real-world traffic conditions without actual driving experiences. This indicates the importance of global position system (GPS) based and real-life experiments in the travel behavior studies[14-16]. Logit and probit specifications were introduced into the classical random utility models to simulate the process of route choice decision-making. Cascetta et al.[17] proposed the conditional logit model to improve the prediction accuracy of route choice, which is empirical in dealing with discrete choice data and supplying a set of alternative routes with little overlap. Additional studies considering aspects of overlapping paths and other modeling strategies to account for these correlations may be found from Refs. [18-19]. A significant amount of research has been performed to identify variables which may better explain drivers' route choice. Generally, travel time, travel time variability and the attributes of alternative routes are regarded as the important variables affecting route choice behavior[2-3]. Papinski et al.[14] demonstrated that a constrained choice set using potential path areas provides the best model fit results. Spissu et al.[20] developed a methodology to compare the observed and minimum cost routes based on several route choice information, while Li et al.[21] revealed that route choice behavior relates to the familiarity to origin-destination pairs. From the perspective of a taxi driver, Sun et al.[18] demonstrated that travel distance, travel time and road preference have high influence on taxi drivers' route choice. Zhang et al.[22] proposed a comprehensive taxi assessment index to quantify the quality of existing urban taxi system with the assistance of GIS technology. Additional studies considering aspects of overlapping paths and other modeling strategies to account for these correlations may be found from Refs. [23].

In summary, drivers' route choice behaviors have aroused extensive research effort, and GPS data have been a normal resource for travel behavior study in recent years[23]. Unfortunately, the route choice behavior of non-occupied taxi has not been studied adequately, which may be an uncertain factor influencing urban road traffic.

1 Data and Methods

1.1 Data Preparation

Data for this study are obtained from the taxi-based GPS devices in Shenzhen, including taxi position, occupancy state (with passengers or not), occurrence time and real-time operation speed. The detailed structure is shown in Table 1.

Table 1 The FCD structure

Field	Type	Comment
Time	Time	HH: MM: SS
Company	Character	—
Car ID	Integer	Each taxi with a unique ID
Longitude	Float	—
Latitude	Float	—
Speed	Float	—
Angle	Float	Degree to the north direction
Operation	Binary	1, with passenger; 0, without passenger

The external environment such as weather, tunnels and buildings usually produces deviations to the GPS data. Consequently, it is critical to adjust vehicle location for further analyses. ArcGIS 10 is used to adjust the floating car location to the nearest road segment with the assistance of "spatial join" tool.

As the interval between two continuous floating car points is 1 min, which is difficult to track the exact trajectory of cars, two additional steps are used to enhance the accuracy of positioning.

Firstly, spatial constraint is used to exclude unreasonable trips, such as U-turn. The occurrence of these is perhaps due to drivers' mistake, and should not be considered.

Secondly, the floating car location is adjusted to the nearest road by using the "spatial join" tool within the ArcGIS software.

As the research purpose is to explore the differences of route choice behavior between occupied and non-occupied taxies, 600 non-occupied and 600 occupied taxi trips from 954 drivers are chosen to form the basis of the study. For each category, 25 field data sets are chosen from each hour within a day. Trips from the same origin location O to the same destination location D are selected to eliminate the influence of travel purpose on route choice behavior. The junction of Nanyuan Road and Huafa South Road surrounded by office buildings, and the junction of Shangbu Road and Zhenhua Road near a small business center are selected as the origin O and the destination D, respectively. The reason is that roads in this area have different route attributes such as choice frequency. Locations of the origin O and the destination D are presented in Fig. 1.

Fig. 1 Four selected routes

Given the huge number of taxi drivers (greater than 15 000), only 126 drivers selected are identified with 2 or more trips. In order to remove the interference of abnormal travel behaviors and simplify the calculation procedure, a feasible set of alternative routes are identified as follows. The frequency of each alternative route is chosen as an assessment criterion, since routes with low frequency may have terrible traffic condition and are not suitable as an alternative route. Based on the statistics of travel trajectory, four routes with the highest frequencies are selected as the alternative routes of this study, as shown in Fig. 1.

1.2 Conditional Logit Model

During the route choice, a driver generally faces various alternative routes with different characteristics. One route has to be chosen according to the criterion of maximal utility. As a result, the choice among alternative routes can be modeled as a function, reflecting the unique characteristics during different phases. The conditional logit model proposed by McFadden[24] is appropriate to simulate route choice behaviors under this situation, and it is usually used to deal with discrete choice data. The advantage is that the conditional logit model can calculate condition likelihood function separately for each group.

The conditional logit model used in this study is based on several important assumptions. The first assumption considers that when a driver makes a route choice, the utility of each route choice needs to be estimated and the individual's utility maximization decision needs to be made. For example, an individual i chooses among J alternatives. Each j provides a different level of utility with the utility function U_{ij} consisting of an observable or deterministic part V_{ij} and a random component ε_{ij}:

$$U_{ij} = V_{ij} + \varepsilon_{ij}. \tag{1}$$

Assuming the random component ε_{ij} follows an extreme-value distribution of Type I, the probability that the individual i chooses the alternative j is

$$P_{ij} = P\{U_{ij} > U_{i|m}\}, \quad \forall m \neq j, \tag{2}$$

where $U_{i|m}$ represents the utility value of the individual i choosing any potential alternative m.

The second assumption is that the error terms are independent of each other, which implies the probability of choosing between two alternatives is independent of the attributes of other routes. The probability that the individual i chooses the alternative j can be calculated as

$$P_{ij} = \frac{\exp(V_{ij})}{\sum_{j=1}^{J} \exp(V_{ij})}, \quad \forall m \neq j, j \in J, \tag{3}$$

where J represents the entire choice set of alternatives.

The estimated log likelihood function is

$$l = \sum_{i=1}^{n} \sum_{j=1}^{J} d_{ij} \ln P\{y_i = j\}, \tag{4}$$

where d_{ij} equals 1 if the individual i chooses the alternative j and 0 otherwise, n represents the total number of taxi drviers, and y_i refers to the alternative chosen by the individual i.

1.3 Model Specification and Variable Selection

In order to understand the mechanism of taxi driver's route choice based on the GPS data, a set of explanatory variables have to be obtained[25]. Then, the conditional logit model is introduced to test if the route choice behavior can be explained by these explanatory variables. The explanatory variables in this study are selected as follows: the longest leg distance d_{ll}, traffic light count C, arterial road percentage R_{ar}, expected travel time $\bar{T}_{\mathrm{t}}$ and congested segment percentage R_{cs}.

The longest leg refers to a set of segments in which car can pass through directly and need not make turn. Disregarding the influence of travel time, drivers tend to select a route with longer D to improve driver's comfort level.

To decrease the time wasted on controlling delay of the intersection, drivers tend to avoid passing through the traffic lights.

Different road levels reflect different population densities and traffic volumes. Area adjacent to arterials is usually accompanied with robust transportation facilities and entertainments attracting more vehicles.

These explanatory variables are independent of time, and several descriptive statistical values of the explanatory variables of four routes are calculated, as shown in Table 2.

Table 2 Characteristics of the time-independent explanation variables

Route	$d_{\rm ll}$/m	C	$R_{\rm ar}$
1	738	3	0.465
2	719	3	0.813
3	478	2	0.507
4	420	4	0.276

The expected travel time is a significant factor influencing drivers' route choice behavior. Taxi drivers are usually veterans with rather sufficient understanding of road conditions. When a taxi driver faces alternative routes, he/she need estimate the travel time of each alternative route at the moment by driving experiences. In this study, the average travel time of each alternative route per hour is calculated on the basis of the historical data. By tracking taxi trajectories of 1 200 trips selected in this study, the expected travel time of four trips at each hour of a day is shown in Fig. 2, where t represents each hour within a day.

At a certain period of a day, some roads may undertake considerably congested traffic and result in a rather low vehicle speed. In this study, any segment with average vehicle speed under 20 km/h is considered as congested segment. Then, a taxi driver will deliberately avoid the congested roads according to the judgement. The percentage of congested segments at the moment can be presented as

$$R_{\rm cs} = L_{\rm leg}/L_{\rm total}, \tag{5}$$

where $L_{\rm total}$ represents the distance of the entire route, and $L_{\rm leg}$ represents the distance of the congested segments at the moment.

At different periods, the congestion level of each route is different. The hourly percentage of congested segments for each route is calculated, as shown in Fig. 3.

The historical data used in this study cover 1 200 trips in September 2014. By analyzing the data, trip characteristics of 600 occupied taxies and 600 non-occupied taxies, including $d_{\rm ll}$, C, $R_{\rm ar}$, $\bar{T}_{\rm t}$ and $R_{\rm cs}$, are obtained. The average value and standard deviation (s.d.) of each variable are calculated, as shown in Table 3. It can be found by analyzing the route choice characteristics that the preferences of non-occupied and occupied taxi drivers are noticeable. Many slight differences exist between the route choice behaviors of non-occupied and occupied taxies, as shown in Table 3. The route of non-occupied taxies has averagely longer $d_{\rm ll}$, indicating that drivers of non-occupied taxi are unlikely to switch to reduce the travel time during driving process. In the aspect of traffic light count, occupied taxi drivers try to avoid heavy traffic light deliberately. The statistical results about $R_{\rm ar}$ show that non-occupied taxi drivers tend to choose the route with higher traffic volume and road level to hunt more passenger flows. The comparison of the average travel time of the two types shows that non-occupied taxies are less sensitive to decrease travel time than the occupied taxies. It is difficult to ascertain as the congested road ratios of two types are similar.

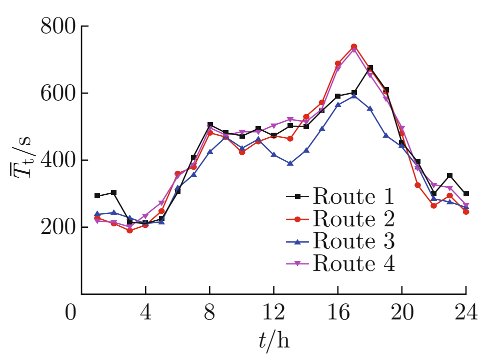

Fig. 2 Expected travel time for four proposed routes

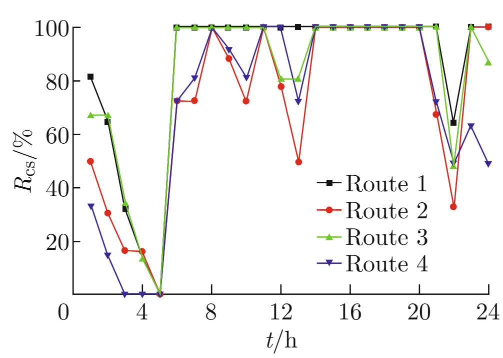

Fig. 3 Percentages of congested segments for four routes

Table 3 Route attribute statistics (mean ± s.d.) of occupied and non-occupied taxi trips

Taxi type	$d_{\rm ll}$/m	C	$R_{\rm ar}$	$\bar{T}_{\rm t}$	$R_{\rm cs}$
Occupied	562.1 ± 120.9	2.47 ± 0.61	0.58 ± 0.16	385.4 ± 124.5	0.75 ± 0.33
Non-occupied	601.1 ± 129.4	2.71 ± 0.63	0.61 ± 0.18	406.6 ± 133.2	0.77 ± 0.33

2 Analyses of Results

As the conditional logit model is introduced to test various variables, the problem of partial multicollinearity in some variables may result in the failure of significance test. Without the variable of d_{ll}, four factors including C, R_{ar}, $\bar{T}_{\mathrm{t}}$ and R_{cs} are selected. Tables 4 and 5 present the results of occupied and non-occupied taxies based on the conditional logit model, where B represents the coefficient of each attribute.

Table 4 Omni-bus test of model coefficients for occupied and non-occupied taxies

Taxi Type	−2 log likelihood	Overall score	
		Chi-square	Significance
Occupied	215.377	106.416	0
Non-occupied	258.6	66.279	0

According to likelihood ratio and Wald tests related to occupied taxies and non-occupied taxies, the estimated conditional logit model based on maximum likelihood estimation is significant at 99% significance level.

Wald chi-square for all factors is significant at 95% importance level.

According to Tables 4 and 5, among these factors, the percentage of arterial road has a positive effect on the route choice both for occupied and non-occupied taxi drivers. An increase of one unit of the arterial percentage increases the probability of choosing route 1.479 units for occupied taxi drivers and 2.194 units for non-occupied taxi drivers. Traffic light count, expected travel time and percentage of congested road are considered as the negative factors in route choice decision. An increase of one unit in them will decrease the probability of choosing route 0.962, 0.019, 2.765 units for occupied taxi drivers and 0.366, 0.008, 4.281 units for non-occupied taxi drivers. The results indicate that the routes with higher arterial road percentage, fewer traffic light count, shorter travel time and less congested road percentage are generally more preferable.

Behavior differences of occupied and non-occupied taxi drivers are further investigated. The percentages of arterials and congested segments generally have more influence on non-occupied taxi drivers, which means non-occupied taxi drivers pay more attention to choosing arterial roads and avoiding congested segments. The regression coefficients of traffic light count and expected travel time indicate that occupied-taxi drivers are more sensitive to fewer traffic light count and shorter travel time.

Table 5 Variables in the equation for occupied and non-occupied taxies

Route attribute	B		Error of s.d.		Wald test		Significance		exp (B)		95% confidence interval for exp (B)			
											Lower		Upper	
	OT	NOT	OT	NOT	OT	NOT	OT	NOT	OT	NOT	OT	NOT	OT	NOT
C	−0.962	−0.636	0.246	0.238	15.339	7.125	0	0.008	0.382	0.529	0.236	0.332	0.618	0.844
R_{ar}	1.479	2.194	0.653	0.552	5.133	15.804	0.023	0	4.389	8.967	1.221	3.041	15.776	26.445
$\bar{T}_{\mathrm{t}}$	−0.019	−0.008	0.005	0.004	14.982	4.738	0	0.029	0.982	0.992	0.972	0.984	0.991	0.999
R_{cs}	−2.765	−4.281	1.326	1.225	4.351	12.212	0.037	0	0.063	0.014	0.005	0.001	0.846	0.153

OT—Occupied taxi, NOT—Non-occupied taxi

3 Conclusion

In this study, taxi GPS data from Shenzhen, China are analyzed to investigate the difference of route choice behavior between occupied and non-occupied taxies. In order to obtain the trajectory of taxies, 1 200 trips are traced using GPS data with several factors chosen and compared to demonstrate the characteristics of drivers. After modeling the mechanism of drivers' route choice process, the results indicate that traffic light count, percentage of arterial road, expected travel time and percentage of congested road are the important factors affecting drivers' route choice. The regression coefficients of these factors are −0.965, 1.479, −0.019, −2.765 for occupied taxies and −0.636, 2.194, −0.008, −4.281 for non-occupied ones. This study demonstrates that non-occupied taxi drivers generally pay more attention to choosing arterial road and avoiding congested segments, rather than to fewer traffic light count and shorter expected travel time.

With the case study of Shenzhen, the non-occupied taxi drivers' route choice preference is revealed. The findings have a significant importance in congestion mitigation and road network planning. For example, rules may be stipulated to regulate the passenger hunting behavior, e.g. to restrain or even forbid them on certain roads, particular for the Uber or Di-Di drivers. As the route choice preference of non-occupied taxi drivers may induce urban congestion on the roads, location optimization for taxi stands is essential in improving the operational level of service (LOS) of traffic. While the results are promising, further studies need to be conducted to improve the performance of the model. For example, additional information on drivers and their associated demand profile for pick-ups are helpful, because the aspect of route choice is also based on available demand (passenger pick-ups). However, as the age, gender and education background of drivers are unknown from the GPS data in this study, the model can better explain the choice behavior by combining the intra-individual variabilities. Moreover, although the study only considers the effect of traffic condition, real-time external data increase the accuracy of results, and further context of land-use and destination choice is also important in driver route choice. To this end, the influence of weather and the condition of route facility need to be explored in the further studies.

References

[1] Beijing Transportation Research Center. 2015 Beijing transport annual report [EB/OL]. (2016-04-06). http://www.bjtrc.org.cn/InfoCenter/NewsAttach/aeb7c878-d31e-4f08-982f-3c17c717c87b.pdf.

[2] BOGERS E A I, VITI F, HOOGENDOORN S P. Joint modeling of ATIS, habit and learning impacts on route choice by laboratory simulator experiments [J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2004(1926): 189-197.

[3] AVINERI E, PRASHKER J N. Sensitivity to travel time variability: Travelers' learning perspective [J]. *Transportation Research Part C: Emerging Technologies*, 2005, **13**(2): 157-183.

[4] BOVY P H L, STERN E. Route choice: Way finding in transport networks [M]. Berlin: Springer, 1990.

[5] MIWA T, SAKAI T, MORIKAWA T. Route identification and travel time prediction using probe-car data [J]. *International Journal of ITS Research*, 2004, **2**(1): 21-28.

[6] BEN-ELIA E, EREV I, SHIFTAN Y. The combined effect of information and experience on drivers' route-choice behavior [J]. *Transportation*, 2008, **35**(2): 165-177.

[7] BEN-ELIA E, SHIFTAN Y. Which road do I take? A learning-based model of route-choice behavior with real-time information [J]. *Transportation Research Part A: Policy and Practice*, 2010, **44**(4): 249-264.

[8] BARRON G, EREV I. Small feedback-based decisions and their limited correspondence to description-based decisions [J]. *Journal of Behavioral Decision Making*, 2003, **16**(3): 215-233.

[9] EREV I, BARRON G. On adaptation, maximization, and reinforcement learning among cognitive strategies [J]. *Psychological Review*, 2005, **112**(4): 912-931.

[10] MOREAU A. Public transport waiting times as experienced by customers [J]. *Public Transport International*, 1992, **41**(3): 52-71.

[11] RAGHUBIR P, MORWITZ V G, CHAKRAVARTI A. Spatial categorization and time perception: Why does it take less time to get home? [J]. *Journal of Consumer Psychology*, 2011, **21**(2): 192-198.

[12] TILAHUN N Y, LEVINSON D M. A moment of time: Reliability in route choice using stated preference [J]. *Journal of Intelligent Transportation Systems*, 2010, **14**(3): 179-187.

[13] BOGERS E A I, BIERLAIRE M, HOOGENDOORN S P. Modeling learning in route choice [J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2007(2014): 1-8.

[14] PAPINSKI D, SCOTT D M, DOHERTY S T. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS [J]. *Transportation Research Part F: Traffic Psychology and Behavior*, 2009, **12**(4): 347-358.

[15] SUN D J, LIU Q, PENG Z R. Research and analysis on causality and spatial-temporal evolution of urban traffic congestions: A case study on Shenzhen of China [J]. *Journal of Transportation Systems Engineering and Information Technology*, 2011, **11**(5): 86-93.

[16] SUN D J, ZHANG C, ZHANG L H, et al. Urban travel behavior analyses and route prediction based on floating car data [J]. *Transportation Letters: The International Journal of Transportation Research*, 2014, **6**(3): 118-125.

[17] CASCETTA E, RUSSO F, VIOLA F A, et al. A model of route perception in urban road networks [J]. *Transportation Research Part B: Methodological*, 2002, **36**(7): 577-592.

[18] PRATO C G. Route choice modeling: Past, present and future research directions [J]. *Journal of Choice Model*, 2009, **2**(1): 65-100.

[19] HOOD J, SALL E, CHARLTON B. A GPS-based bicycle route choice model for San Francisco, California [J]. *Transportation Letters: The International Journal of Transportation Research*, 2011, **3**(1): 63-75.

[20] SPISSU E, MELONI I, SANJUST B. A behavioral analysis of daily route choice using GPS-based-data [J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2011(2230): 96-103.

[21] LI D, MIWA T, MORIKAWA T. Use of private probe data in route choice analysis to explore heterogeneity in drivers' familiarity with origin-destination pairs [J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2013(2338): 20-28.

[22] ZHANG D Z, SUN D J, PENG Z R. A comprehensive taxi assessment index using floating car data [J]. *Journal of Harbin Institute of Technology*, 2014, **21**(1): 7-16.

[23] RAHMANI M, JENELIUS E, KOUTSOPOULOS H N. Non-parametric estimation of route travel time distributions from low-frequency floating car data [J]. *Transportation Research Part C: Emerging Technologies*, 2015, **58**: 343-362.

[24] MCFADDEN D. Conditional logit analysis of qualitative choice behavior [C]//*Frontiers in Econometrics*. New York: Academic Press, 1974: 105-142.

[25] SUN D J, ELEFTERIADOU L. A driver behavior based lane-changing model for urban arterial streets [J]. *Transportation Science*, 2014, **48**(2): 184-205.