Short Paper

A Two-Layer Model for Taxi Customer Searching Behaviors Using GPS Trajectory Data

Jinjun Tang, Han Jiang, Zhibin Li, Meng Li, Fang Liu, and Yinhai Wang

***Abstract*—This paper proposes a two-layer decision framework to model taxi drivers' customer-search behaviors within urban areas. The first layer models taxi drivers' pickup location choice decisions, and a Huff model is used to describe the attractiveness of pickup locations. Then, a path size logit (PSL) model is used in the second layer to analyze route choice behaviors considering information such as path size, path distance, travel time, and intersection delay. Global Positioning System data are collected from more than 36 000 taxis in Beijing, China, at the interval of 30 s during six months. The Xidan district with a large shopping center is selected to validate the proposed model. Path travel time is estimated based on probe taxi vehicles on the network. The validation results show that the proposed Huff model achieved high accuracy to estimate drivers' pickup location choices. The PSL outperforms traditional multinomial logit in modeling drivers' route choice behaviors. The findings of this paper can help understand taxi drivers' customer searching decisions and provide strategies to improve the system services.**

***Index Terms*—Customer-search behavior, Huff model, path overlapping, path size logit model, taxi GPS traces.**

I. Introduction

Taxi is an important travel mode in urban transportation system. It plays a critical role to meet travel needs within urban areas. As compared to other travel modes, taxi is very convenient, comfortable, and fast and is considered a competitive service to satisfy the medium-to-long distance trips. However, in urban areas the taxi service has been in a dilemma. On one hand, it is difficult for customers to take a taxi during peak hours. On the other hand, more cruising behaviors of taxi drivers for searching passengers increase the traffic congestion and result in unnecessary pollution.

Previously, researchers have put forward a lot of effective and practical methods to solve the above problems in taxi service. Those methods can be classified into four categories: (1) System optimization. In this category, studies mainly focus on how to improve customers' satisfaction towards taxi service [1], enhance operational effectiveness of taxi business [2], [3], and manage taxi fleets size [4]; (2) Economic modeling. Studies in this category implement various regulatory policies to improve the efficiency of taxi system [5], [6] but ignoring the spatial structure of the market [7]; (3) Network modeling. The models in this category are mainly proposed by Yang and Wong [8]. They constructed various network models by considering customer OD patterns, congestions, model calibration efficiency, demand and supply equilibrium while taking into account the influence of spatial structure in the market; and (4) Route choice behaviors. A number of studies applied the logit and probability based models to analyze taxi drivers' behaviors in occupied situations [9] and customer searching behaviors in vacant situations [7]; [10], [11].

After a review of the relevant literature, two limitations were found in the previous work: (1) Previous zone or cell based approach cannot accurately describe true customer searching behavior, since those approaches simply assume the route choice is in zones or cells. In our study, actual routes in the road network are identified to solve the problems caused by the zone or cell based approaches; (2) The Multinomial Logit (MNL) model is widely used in previous studies and it assumes that the unobserved utilities for different alternatives are identically and independently distributed, which may not be valid in the context of route choice particularly due to path overlapping during the customer searching process [12].

The main contribution of this paper includes following several aspects: (1) We use the DBSCAN algorithm to cluster extracted pick-up and drop-off records (PDR) and recognize some hot spots for travel, and we also discuss approach to find reasonable parameters; (2) Customer searching model includes two internal layers. The first layer models the probability of the decision of a driver which zone or area to go to pick up a passenger. A Huff model is used to describe the attractiveness of a pick-up area, considering the historical number of pick-up passengers and distance between drop-off and pick-up areas as explanatory variables. The second layer models driver's route choice decision in the road network. A PSL model is developed to explore route choice behaviors by considering path size, intersection delay, path travel time, and path distance. (3) As travel time express obvious variability, we present a statistical method to estimate path travel time in different time periods based on instantaneous speed of taxi vehicles. (4) We finally compare the calibration and validation results between PSL model and traditional Multinomial Logit (MNL) model based on the actual routes choice of vacant taxi drivers. The higher log-likelihood, adjusted R-squared and lower fitting errors of PSL proves it is superior to MNL model.

II. Methodology

A. Model Structure

Fig. 1 provides calculation procedure of this work. We firstly collect raw taxi GPS data and extract pick-up or drop-off locations using state

Manuscript received September 20, 2015; revised January 2, 2016 and March 3, 2016; accepted March 12, 2016. Date of publication April 21, 2016; date of current version October 28, 2016. This work was supported in part by the National Natural Science Foundation of China under Grants 51138003 and 51329801. The Associate Editor for this paper was F.-Y. Wang.

J. Tang is with the School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: jinjuntang@163.com).

H. Jiang is with the Department of Automation, Tsinghua National Laboratory for Information Science and Technology (TNlist), Tsinghua University, Beijing 100084, China (e-mail: jianghan0521@sina.cn).

Z. Li and Y. Wang are with the Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: lizhibin@uw.edu; yinhai@uw.edu).

M. Li is with the Department of Civil Engineering Tsinghua University, Beijing 100084, China (e-mail: limengall@gmail.com).

F. Liu is with the School of Energy and Traffic Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China (e-mail: rcliufang@163.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2544140

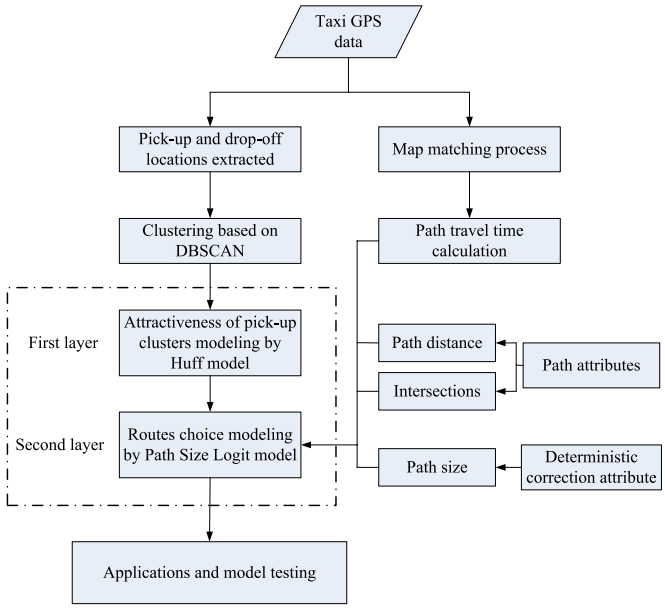

Fig. 1. Structure of the two-layer decision model for customer search.

information about whether loading passengers. Secondly, a Density-based Spatial Clustering of Applications with Noise (DBSCAN) method is used to cluster pick-up and drop-off locations, in which two parameters, radius of neighborhood (Eps) and minimum points (its means the minimum GPS locations a cluster, MinPts), are optimized in application. Then, in the first layer of decision, a Huff model is implemented to present attractiveness distribution of pick-up clusters. Furthermore, in the second layer, a Path Size Logit (PSL) model is used to analyze routes choice behavior in actual road networks. In the PSL model, we consider four main factors: path travel time, path distance, intersections delay and path size. Before calculate path travel time, a map matching process should be completed to project raw GPS locations onto the roads. Finally, an application in Xidan district of Beijing city is used to show the effectiveness of proposed model.

B. First Layer: Attractiveness Modeling

1) Clustering Based on DBSCAN: In this section, we use DBSCAN to cluster pick-up and drop-off locations and explore their spatial characters. The DBSCAN algorithm [14] is widely used in density-based clustering from large scale data for its simple calculation structure and low computing cost. It directly divides all point densities reachable from different points into clusters.

Thus, from the introduction mentioned above, pick-up and drop-off locations can be accurately clustered based on proper Eps and MinPts. For each cluster, we compute the cluster centers under the condition of minimizing an objective function J, which is defined as:

$$J = \sum_{i=1}^{n} d(\text{location}_i, \text{center}) = \sum_{i=1}^{n} |\text{location}_i - \text{center}| \tag{1}$$

where, $|\cdot|$ represents the general Euclidean distance, location is the location of pick-up or drop-off, n is the number of locations in different clusters, center means the cluster centers.

2) Huff Model: A classical Huff model [15] is used to analyze drivers' choice behavior, the Huff model is a variant on the gravity and spatial interaction models and it measures the percentage of demand in each origin zone that will visit various destinations.

$$P_{ij} = \frac{T_{ij}}{\sum_{k=1}^{m} T_{ik}} = \frac{W_j^{\alpha} \text{Cost}_{ij}^{-\beta}}{\sum_{k=1}^{m} W_k^{\alpha} \text{Cost}_{ik}^{-\beta}} \tag{2}$$

where P_{ij} is the probability of a taxi driver located at drop-off cluster center i choosing pick-up cluster center j; T_{ij} means the sum of trips from i to j; W_j is the attractiveness of pick-up cluster j, the number of pick-up locations in cluster is used to represent attractiveness; Cost_{ij} is the cost from i to j, the distance is used to estimate the cost; α and β are the sensitivity parameters; m is the number of pick-up cluster centers corresponding to drop-off center i. In [16], four types of cost and attraction function combinations were compared to model attractiveness:

① $T_{ij} = \exp(\alpha \ln W_j - \beta \ln C_{ij}) = W_j^{\alpha} C_{ij}^{-\beta}$
② $T_{ij} = \exp(\alpha W_j - \beta C_{ij})$
③ $T_{ij} = \exp(\alpha W_j - \beta \ln C_{ij}) = C_{ij}^{-\beta} \exp(\alpha W)$
④ $T_{ij} = \exp(\alpha \ln W_j - \beta C_{ij}) = W_j^{\alpha} \exp) - \beta C_{ij})$.

Here, we also compare the accuracy of four models. In order to calibrate the parameters, we construct an error function as follows:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(P_{ij}^{\text{real}} - P_{ij}\right)^2 \tag{3}$$

where P_{ij}^{real} means the observed choose probability; n is the number of drop-off cluster centers; m is the number of pick-up cluster centers corresponding to drop-off center i. As E is a nonlinear objective function, the Levenberg–Marquardt (LM) method [17] is used to solve this non-linear least square problem. The LM method is a widely used optimization algorithm in solving least square curve fitting and nonlinear programming problems.

C. Second Layer: Routes Choice Behavior Modeling

In the road networks, adding a deterministic correction term to the path utility function is a practical way to address the path overlapping issue in route choice model [12]. This correction term is defined as path size (PS), which was introduced in PSL model proposed by Ben-Akiva and Ramming [13]. The correction attribute PS is defined as:

$$\text{PS}_{\text{in}} = \sum_{a \in \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{aj}} \tag{4}$$

where PS_{in} is the path size of path i in choice set C_n, l_a is the length of link a, Γ_i is the set of all links along path i, L_i is the total length of path i, δ_{aj} equals to 1 if link a is on path j and 0 otherwise. If a path is completely independent with other paths, its PS value equals to 1. For a path, more links are shared with others, and then its PS value will become smaller. Although various extended PS formulations proposed in [18], [19], the main ideas of these improved method are based on Equation (4).

Delay caused by intersections is another important factor should be considered in routes choice model. Here, we define a penalty function to evaluate the influence of intersections, and its structure is similar to the path cost estimation in Rahmani and Koutsopoulos [20]:

$$\text{PV}_{\text{in}} = c_1 \cdot N_{\text{intersection}} + c_2 \cdot N_{\text{straight}} + c_3 \cdot N_{\text{leftturn}} \tag{5}$$

where PV_{in} is the penalty value of path i in choice set C_n, $N_{\text{intersection}}$ is the number of intersections on path i, N_{straight} is the number of straight when across intersections, N_{leftturn} is the number of left

turning, c_1, c_2 and c_3 are the penalty factors, we set $c_1 = 1.0$, $c_2 = 0.5$, $c_3 = 1.0$. The calculation process of path travel time will be introduced in the next section.

Accordingly, the utility of path i in choice n is defined as:

$$U_{\text{in}}=\gamma_1 \cdot TT_{\text{in}}+\gamma_2 \cdot \ln(\text{PS}_{\text{in}})+\gamma_3 \cdot \text{PV}_{\text{in}}+\gamma_4 \cdot Dis_{\text{in}}+\varepsilon_{\text{in}} \quad (6)$$

where U_i is the utility for path i in choice set n, TT represents the path travel time, PS means the path size value, PV indicates penalty value for delays in intersections, Dis is the distance of path, γ_1, γ_2, γ_3 and γ_4 are the coefficients of four factors, ε_{in} is defined as the unobserved error components and followed standard Gumbel distribution. It should be noted that the distance is a static factor and travel time is a dynamic factor which varies with time periods in a day. For a single route, the values of travel time will not consistently increase as the values of distance increase. Thus, these two important factors are considered in the utility.

The probability that path i will be chosen in the PSL model is calculated then:

$$p\left(i|C_n\right) = \frac{e^{U_{\text{in}}}}{\sum_{j\in C_n} e^{U_{jn}}} \quad (7)$$

where C_n means the choice set for OD pair n.

III. Data

A. Collection of GPS Data

The taxi GPS data we used in this study were collected from about 36,000 drivers in Beijing city. The data starts from September 2014 to February 2015. It is recorded at a rate of 30 seconds. The duration of collection stats from 5:00 AM to 24:00 PM. We divide data set into two parts: training dataset and testing dataset, dataset in first four months is used to calibrate models and dataset in last two months is used to validate models. As the training and testing data are collected from same drivers, we randomly extract 20% of testing dataset to validate the model. In order to reduce the bias associated with random selection samples, five experiments with different testing sets were conducted for model validation. As the travel time varies according to the different time period in a day, and travel time is the significant factor to affect driving behavior of taxi drivers. In order to analyze driving behavior characteristics in different periods, we divide daytime into five periods: period 1, from 5:00 to 9:00; period 2, from 9:00 to 13:00; period 3, from 13:00 to 17:00; period 4, from 17:00 to 21:00; period 5, from 21:00 to 24:00. In each time period, we calculate path travel time and collect routes choice observations to calibrate PSL model.

B. Calculation of Path Travel Time

In this paper, path travel time is estimated by using taxis as a probe vehicle in the road networks. Thus, in the implementation process, we have to match the taxi GPS locations onto the road on which drivers travel. Meanwhile, in order to obtain higher calculation accuracy of the travel time, the whole path should be divided into segments. Through collecting instantaneous speed of taxis driving on the segments, we are able to estimate the travel time of each segment. Eventually, the path travel time can be estimated as the sum of all segments contained in this path.

The path travel time can be calculated by following equation:

$$TT_{kn} = \sum_{i=1}^{N_{\text{seg}}} \left[\left(\frac{l_i^k}{\left(\frac{\sum_{j=1}^{N_{\text{taxi}}} v_{ij}}{N_{\text{taxi}}} \right)} \right) \right] \quad (8)$$

where TT_{kn} is the travel time of path k in choice set $n, l^k i$ represents the length of the ith segment, v_{ij} means the instantaneous speed of the jth taxi travelling on the ith segment, N_{taxi} is the total number of taxis and N_{seg} is the number of divided segments. In the division of segments, the locations where road line shape change (the direction of road line changes) or intersections are generally treated as the end of segments. The average length of a segment is about 100 meters. We set the data collection interval as 5 min, that is, in every 5 min we check whether there are taxis appeared on each segment. If there are no data samples, we should use the average values of upstream and downstream to impute missing data. As the collection interval is relatively long, the data samples can be obtained in almost all the segments.

IV. Results and Discussion

In the model application, we select Xidan district as case study, which is one of the most important shopping center in Beijing city. The longitude of application area ranges from 116.342466° to 116.382135°, and latitude ranges from 39.906959° to 39.924189°.

A. Clustering Results Based on DBSCAN

Fig. 2(a) shows the changes of cluster number with different parameters (Eps and MinPts) based on drop-off locations collecting during six months. For a given MinPts, the cluster number rises gradually as Eps increases at first stage. Its value reaches to a peak and then decreases when Eps increases at second stage. The reason is that a small Eps value can cause the clusters to be separated, and if Eps is too large, the clusters will merge into a larger one. Another important parameter MinPts can also affect the clustering results to a given Eps. As the MinPts increases, the number of clusters decreases. A small MinPts value can extract more clusters and a large value means the number of points clustered into the region with radius of Eps increases. Fig. 2(b) shows the cumulative percentage distribution for drop-off locations based on different parameter pairs when the cluster numbers reach maximum. As we can see, when MinPts is 4 and Eps is 6, the number of clusters with less than 50 locations takes up more than 90% of all clusters. This means large amount of small clusters are generated. This clustering result is not able to actually reflect clustering properties in some large area, for example shopping centers or railway stations. When MinPts is 12 or 14 and Eps is 10, the proportion of clusters with less than 200 locations reaches to 90%. This indicates that most locations are classified into large clusters to estimate main attractive spots in urban city. Furthermore, the curves of cluster number when MinPts equal to 12 and 14 have similar distribution patterns, this means the cluster number become stable. So, in this study, a proper parameters are set as MinPts = 12 and Eps = 10, and 154 clusters are generated from drop-off locations. By using the similar method, we can obtain the proper parameters (MinPts = 12, Eps = 12) and 147 clusters for pick-up locations shown in Fig. 2(c) and (d).

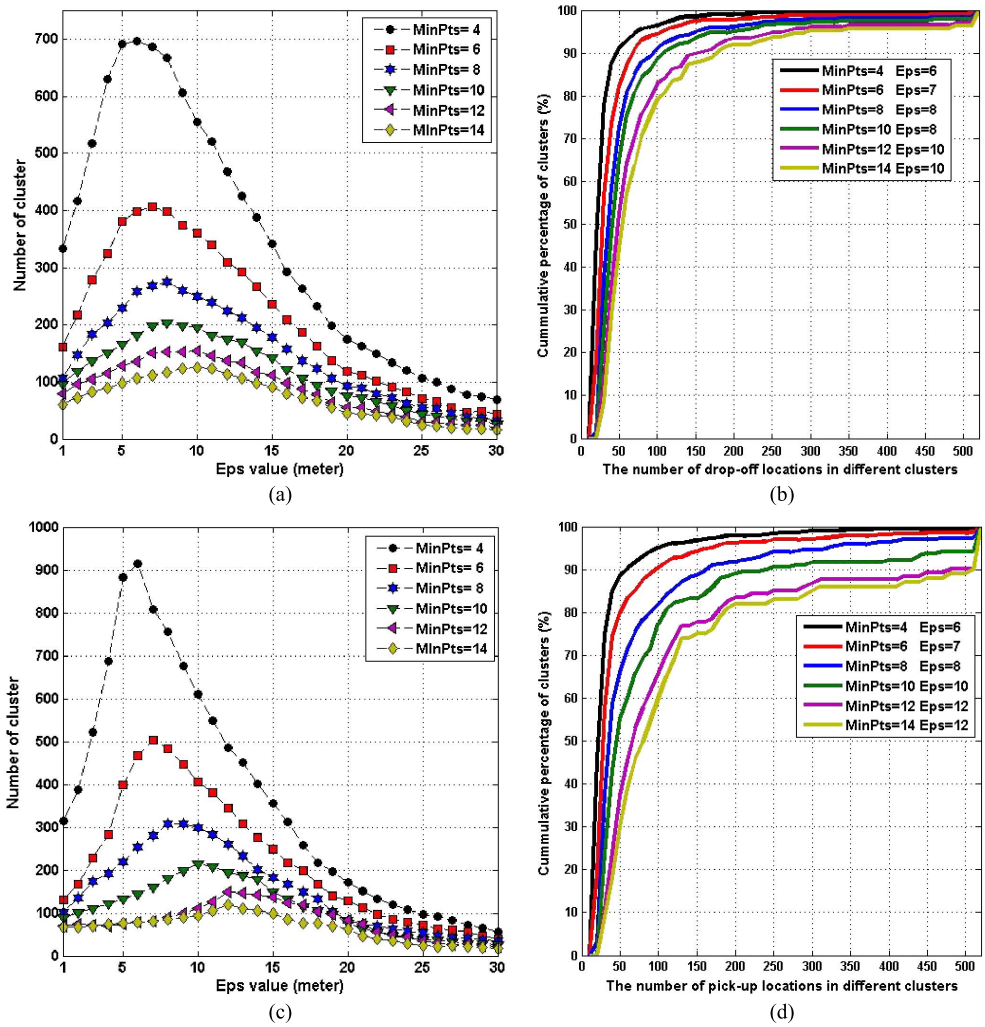

Fig. 2. Clustering results for different parameters in DBSCAN. (a) Clustering result of drop-off locations. (b) Cumulative proportion distribution of drop-off clusters. (c) Clustering result of pickup locations. (d) Cumulative proportion distribution of pickup clusters.

TABLE I
PARAMETER ESTIMATION RESULTS BASED ON THE LM METHOD IN THE HUFF MODEL

Models	α	β	squared sum of the residual
1	0.8467	-0.1316	0.9336
2	0.0771	-0.0324	1.0468
3	0.0743	-0.1505	1.1791
4	0.8724	-0.0299	1.0271

B. Model Calibration

For Huff model, we obtain 154 drop-off clusters and 147 pick-up clusters in clustering process. However, we find that there are no trips between several cluster centers, such as from drop-off center in the southeast corner to the pick-up center in the northwest corner. In order to save cruising time for searching customers, taxi drivers seldom travel for long distance to find next customer. Thus, we eliminate some cluster with few trips. We compare the calibration accuracy of four different attractive models introduced before. Table I provides the calibration results based on LM algorithm in all time periods. The results show that the classic Huff model has the best fitting performance with parameters $\alpha = 0.8467$ and $\beta = -0.1316$.

Secondly, for the PSL model, we select 4 drop-off clusters shown in Fig. 3 as origins and 31 corresponding pick-up clusters as destinations. There are total 134 paths taxi driver frequently choose in the observations. Fig. 3 shows the distribution of drop-off and pick-up clusters in the Xidan district, in which we regard drop-off clusters and pick-up clusters as origins and destinations respectively. There are total 31 OD pairs in the case. The travel time of each path can be estimated from equation (8).

Fig. 4 shows the travel time of one path estimated from probe taxi vehicles in five periods. Fig. 4(a) only expresses about 3000 travel time samples collected at interval of 5 min. Fig. 4(b) represents the fitting results based on lognormal distribution. As we can see, the mean travel time in period 1 is relatively low, and then it increases in period 2 and 3, finally decreases in period 4 and 5. This change is consistent with the variation of travel time in a day. In the study, we use the mean values of fitting functions as travel time for each path.

We compare the calibration results between PSL model and MNL model based on the actual routes choice of vacant taxi drivers in Xidan district during four months. The estimation results are presented in Tables II–VI. All estimated values of factors are statistically significant at the 95% level. As we expected, the path travel time has high negative coefficient value at different time periods, which indicates that taxi drivers are very sensitive to the reliability of travel time. Similarly, delays caused by the intersection are also drivers concerned about. Especially from period 2 to 4, PV has higher negative coefficient than other factors. This means, due to the increase of travel time in

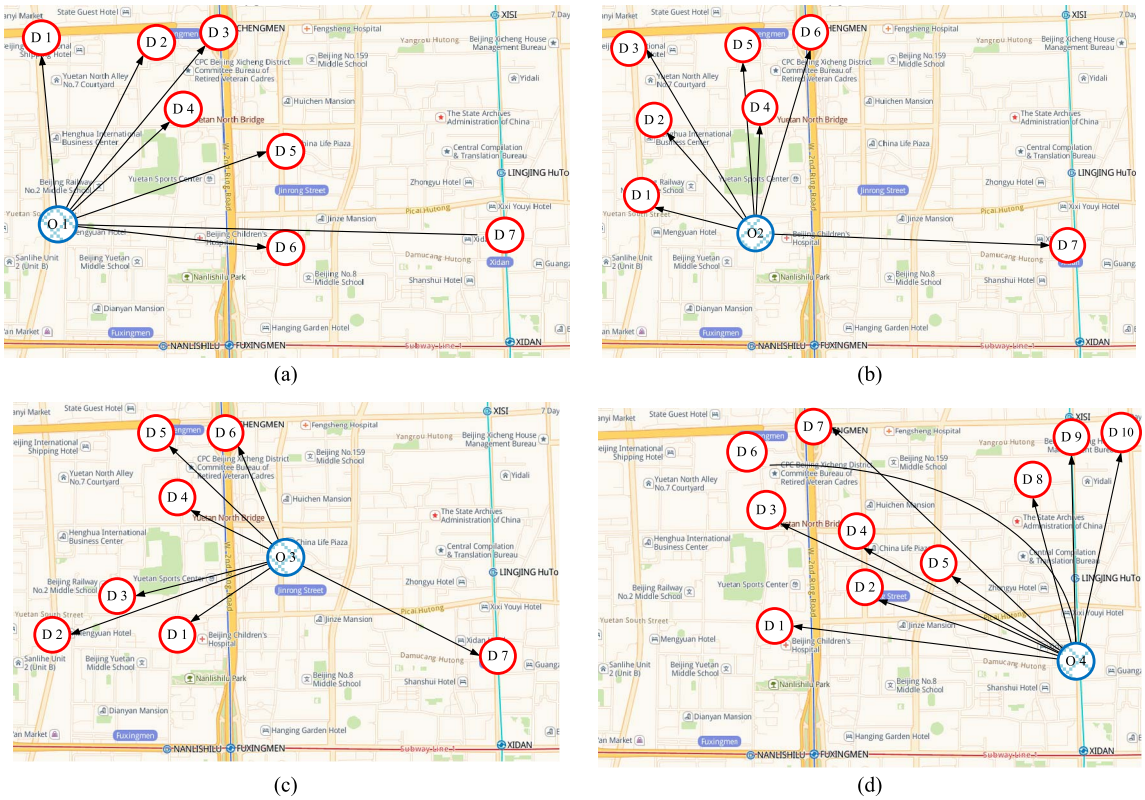

Fig. 3. Distribution of OD pairs in the case study. (a) Origin 1. (b) Origin 2. (c) Origin 3. (d) Origin 4.

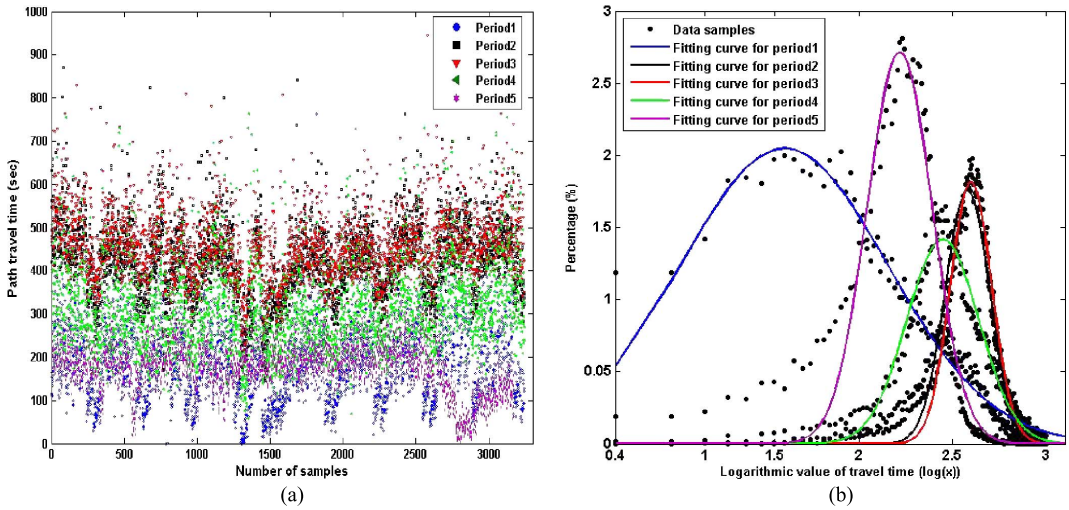

Fig. 4. Path travel time in five periods. (a) Travel time samples. (b) Fitting results.

these periods, drivers will be more sensitive to the delays added in intersections. Taxi drivers are different from private vehicle drivers as they have abundant driving experience and cognitive ability on the local traffic conditions. In order to maximize their profits, they try their best to shorten customer-search time, and they are more sensitive to the time values compared to regular drivers. Thus, the time travel and intersection delays have higher negative coefficient values than the other two factors in PSL and MNL.

The path distance has low negative coefficient value except in the period 1 and 5. In these two periods, as traffic system be in good condition and travel time express high reliability, the path distance become another important factor in routes choice of the vacant taxi drivers. Although the coefficient value of PS is less negative than that of other three factors, it reflects the cumulative experience during driver's long-term driving behavior. In all the candidate paths, the path with more links shared with other paths will be chosen with higher probability. In order to compare the calibration performance, we use final log-likelihood, adjusted R-squared and fitting errors (mean square error between actual routes choice probability and estimated probability using the calibrated coefficients) to measure the calibration accuracy. The log-likelihood and adjusted R-squared of PSL are higher than that of MNL, meanwhile, the fitting errors are lower, which

TABLE II
CALIBRATION RESULTS OF PSL AND MNL IN PERIOD 1

Factors	Period 1 (5:00-9:00) PSL Coefficient	PSL Std. Err	MNL Coefficient	MNL Std. Err
Travel time (min)	-0.589	0.020	-0.594	0.021
ln(PS)	-0.156	0.052	NA	NA
PV	-0.416	0.022	-0.374	0.024
Distance (km)	-0.478	0.021	-0.543	0.021
Statistical indicators				
No. of observations	4602		4602	
LL(0)	-6731.6		-6731.6	
LL(γ)	-5528.4		-5559.8	
Likelihood Ratio	2406.4		2343.6	
R-squared	0.947		0.936	
Adjusted R-squared	0.945		0.935	
Fitting errors (MSE)	0.096		0.099	

MSE: Mean Square Error

TABLE III
CALIBRATION RESULTS OF PSL AND MNL IN PERIOD 2

Factors	Period 2 (9:00-13:00) PSL Coefficient	PSL Std. Err	MNL Coefficient	MNL Std. Err
Travel time (min)	-0.509	0.025	-0.532	0.027
ln(PS)	-0.344	0.053	NA	NA
PV	-0.491	0.031	-0.355	0.032
Distance (km)	-0.283	0.028	-0.301	0.027
Statistical indicators				
No. of observations	6183		6183	
LL(0)	-8551.0		-8551.0	
LL(γ)	-7199.7		-7258.4	
Likelihood Ratio	2702.6		2585.2	
R-squared	0.923		0.913	
Adjusted R-squared	0.921		0.912	
Fitting errors (MSE)	0.099		0.102	

TABLE IV
CALIBRATION RESULTS OF PSL AND MNL IN PERIOD 3

Factors	Period 3 (13:00-17:00) PSL Coefficient	PSL Std. Err	MNL Coefficient	MNL Std. Err
Travel time (min)	-0.411	0.023	-0.442	0.023
ln(PS)	-0.377	0.055	NA	NA
PV	-0.614	0.027	-0.463	0.026
Distance (km)	-0.192	0.025	-0.289	0.025
Statistical indicators				
No. of observations	5926		5926	
LL(0)	-7114.5		-7114.5	
LL(γ)	-5869.9		-5901.8	
Likelihood Ratio	2489.2		2425.4	
R-squared	0.942		0.937	
Adjusted R-squared	0.940		0.936	
Fitting errors (MSE)	0.098		0.099	

proves the PSL model are superior to the MNL model in each time period.

C. Model Validation

To validate the prediction performance of Huff model and PSL model, we use data collected from last two months to test models. The average prediction results of five experiments for Huff model in first layer decision are shown in Table VII, which shows good prediction accuracy based on low residual. Table VIII provides average prediction results of five experiments for two route choice models, in which the PSL model outperforms the MNL model by obtaining higher LL(γ) and R-squared.

TABLE V
CALIBRATION RESULTS OF PSL AND MNL IN PERIOD 4

Factors	Period 4 (17:00-21:00) PSL Coefficient	PSL Std. Err	MNL Coefficient	MNL Std. Err
Travel time (min)	-0.515	0.016	-0.418	0.017
ln(PS)	-0.434	0.044	NA	NA
PV	-0.603	0.020	-0.537	0.021
Distance (km)	-0.369	0.022	-0.494	0.021
Statistical indicators				
No. of observations	4729		4729	
LL(0)	-5841.3		-5841.3	
LL(γ)	-4652.3		-4692.1	
Likelihood Ratio	2378.0		2298.4	
R-squared	0.951		0.945	
Adjusted R-squared	0.950		0.944	
Fitting errors (MSE)	0.095		0.097	

TABLE VI
CALIBRATION RESULTS OF PSL AND MNL IN PERIOD 5

Factors	Period 5 (21:00-24:00) PSL Coefficient	PSL Std. Err	MNL Coefficient	MNL Std. Err
Travel time (min)	-0.542	0.037	-0.520	0.035
ln(PS)	-0.331	0.071	NA	NA
PV	-0.521	0.044	-0.413	0.045
Distance (km)	-0.483	0.042	-0.348	0.040
Statistical indicators				
No. of observations	3417		3417	
LL(0)	-4521.8		-4521.8	
LL(γ)	-3739.9		-3762.8	
Likelihood Ratio	1563.8		1518.0	
R-squared	0.911		0.906	
Adjusted R-squared	0.909		0.905	
Fitting errors (MSE)	0.111		0.114	

TABLE VII
PREDICTION RESULTS OF THE HUFF MODEL

Models	α	β	squared sum of the residual
Huff	0.8467	-0.1316	1.0834

TABLE VIII
PREDICTION RESULTS OF PSL AND MNL MODELS

	PSL Time Period 1	2	3	4	5
No. of observations	2172	2983	2762	2540	1823
LL(0)	-2640	-3431	-3357	-3175	-2288
LL(γ)	-1573	-2243	-2069	-1915	-1367
R-squared	0.932	0.922	0.935	0.937	0.921
	MNL Time Period 1	2	3	4	5
No. of observations	2172	2983	2762	2540	1823
LL(0)	-2640	-3431	-3357	-3175	-2288
LL(γ)	-1617	-2276	-2094	-1967	-1388
R-squared	0.925	0.917	0.923	0.928	0.915

V. Conclusion

This study proposes a two-layer approach to model customer searches behavior of vacant taxi. The GPS trajectories of 36,000 taxis in urban areas of Beijing city were collected to support the model development. In the first layer, the historical drop-off and pick up locations are identified from the GPS data, and then DBSCAN model is used to group those locations into different clusters. Assuming all trips start from a drop-off cluster center and end at a pick-up cluster center, a Huff model is developed to describe the attractiveness of each pick-up cluster. Four types of cost and attraction function combinations are calibrated, and the classical Huff model obtains the highest accuracy. In the second layer, a PSL model is adopted to address the path overlapping issue in the route choice behaviors. The path size, delay in intersection, path travel time, and path distance are considered as important factors to predict the route choice decision. The model calibration results show that the two-layer decision effectively model the customer-searching behavior of taxi drivers. We also compare the performance of PSL model with the traditional MNL model in the five time periods. The results suggest that PSL model proposed in our study outperforms MNL model in terms of a larger log-likelihood and adjusted R-squared values, and a smaller fitting error.

The study also has several limitations. (1) We only consider the area with high density based on historical pick-up or drop-off locations. However, in actual scenarios, taxi drivers sometimes can pick up passengers on the way they head to hot areas. It necessary to focus on the impact of middle sources to route choice behavior in further work. (2) In the utility function, we define a penalty function to evaluate the intersection delay. However, the actual calculation of intersection delay is more complicated. So, the improvement of the delay calculation considering signal timing and management can make proposed model be practical in application. (3) In the utility function, we only consider four factors. Some other attributes or factors [21] such as road classification, traffic controlling and managing measures, weather condition, incidents, etc. should be considered in the modeling process.

References

[1] D. H. Lee, H. Wang, R. L. Cheu, and S. H. Teo, "Taxi dispatch system based on current demands and real-time traffic conditions," *Transp. Res. Rec.*, vol. 1882, pp. 193–200, 2004.

[2] W. C. Lee and B. W. Cheng, "Incorporating e-technology to advantage in a greener taxi industry and its impact on driving performance and safety," *Transp. Plan. Technol.*, vol. 31, no. 5, pp. 569–588, 2008.

[3] J. Tang, F. Liu, Y. Wang, and H. Wang, "Uncovering urban human mobility from large scale taxi GPS data," *Phys. A, Statist. Mech. Appl.*, vol. 438, no. 15, pp. 140–153, Nov. 2015.

[4] H. W. Chang, Y. C. Tai, and Y. J. Hsu, "Context-aware taxi demand hotspots prediction," *Int. J. Bus. Intell. Data Mining*, vol. 5, no. 1, pp. 3–18, Dec. 2010.

[5] R. Arnott, "Taxi travel should be subsidized," *J. Urban Econ.*, vol. 40, no. 3, pp. 316–333, Nov. 1996.

[6] R. D. Cairns and C. Liston-Heyes, "Competition and regulation in the taxi industry," *J. Public Econ.*, vol. 59, no. 1, pp. 1–15, Jan. 1996.

[7] R. C. P. Wong, W. Y. Szeto, and S. C. Wong, "A cell-based logit-opportunity taxi customer-search model," *Transp. Res. C, Emerging Technol.*, vol. 48, pp. 84–96, Nov. 2014.

[8] H. Yang, S. C. Wong, and K. I. Wong, "Demand–supply equilibrium of taxi services in a network under competition and regulation," *Transp. Res. B, Methodol.*, vol. 36, no. 9, pp. 799–819, Nov. 2002.

[9] B. D. Ziebart, A. L. Maas, A. K. Dey, and J. A. Bagnell, "Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior," in *Proc. 10th Int. Conf. Ubiquitous Comput.*, Seoul, South Korea, 2008, pp. 1–10.

[10] R. M. N. T. Sirisoma *et al.*, "Empirical evidence for taxi customer-search model in Hong Kong," *Proc. Inst. Civil Eng., Transp.*, vol. 163, no. 4, pp. 203–210, Dec. 2010.

[11] W. Y. Szeto, R. C. P. Wong, S. C. Wong, and H. Yang, "A time-dependent logit-based taxi customer-search model," *Int. J. Urban Sci.*, vol. 17, pp. 184–198, 2013.

[12] R. Tan, M. Adnan, and D. H. Lee, "A new path size formulation in path size logit for route choice modeling in public transport networks," presented at the Proceedings 94th Transportation Research Board Annual Meeting, Washington, DC, USA, 2015.

[13] M. Ben-Akiva and S. Ramming, "Lecture notes: Discrete choice models of traveler behavior in networks," presented at the Advanced Methods Planning Management Transportation Networks, Capri, Italy, 1998.

[14] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for sparse representations," in *Proc. 2nd Int. Conf. Knowl. Discov. Data Mining*, 1996, pp. 226–231.

[15] D. L. Huff, "A probabilistic analysis of shopping center trade areas," *Land Econ.*, vol. 39, no. 1, pp. 81–90, 1963.

[16] M. E. O'Kelly, "Trade-area models and choice-based samples: Methods," *Environ. Plan. A*, vol. 31, no. 4, pp. 613–627, 1999.

[17] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer-Verlag, 2006, pp. 258–264.

[18] E. Frejinger and M. Bierlaire, "Capturing correlation with subnetworks in route choice models," *Transp. Res. B, Methodol.*, vol. 41, no. 3, pp. 363–378, Mar. 2007.

[19] P. H. Bovy, S. Bekhor, and C. G. Prato, "The factor of revisited path size: Alternative derivation," *Transp. Res. Rec.*, vol. 24, no. 1, pp. 132–140, 2008.

[20] M. Rahmani and H. N. Koutsopoulos, "Path inference from sparse floating car data for urban networks," *Transp. Res. C, Emerging Technol.*, vol. 30, pp. 41–54, May 2013.

[21] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PloS one*, vol. 10, no. 3, 2015, Art. no. e0119044.