A Data-Driven Customer-Search Modeling With the Consideration of Traffic Environment

frontiers in Public Health

ORIGINAL RESEARCH
published: 17 March 2022
doi: 10.3389/fpubh.2022.848748

A Data-Driven Customer-Search Modeling With the Consideration of Traffic Environment

Lan Yu[1], *Zhuo Sun*[1], *Lianjie Jin*[1] *and Chao Chen*[1,2*]

[1] College of Transportation Engineering, Dalian Maritime University, Dalian, China, [2] State Key Laboratory of Structural Analysis of Industrial Equipment, School of Automotive Engineering, Dalian University of Technology, Dalian, China

In order to explore the determinants of vacant taxi drivers' customer-search behavior, this paper intends to calibrate a time-dependent Multinomial Logit (MNL) model by mining over 1.6 billion GPS records from about 8,400 taxis in Shanghai, China. First, based on the ordering points to identify the clustering structure (OPTICS) algorithm, the downtown area of Shanghai city is divided into 47 hotspots to identify the hot areas of customer delivery and searching. Then, by investigating a typical search delivery process of a vacant taxi, five candidate factors that may affect the customer-search behavior are summarized and defined. Using the maximum likelihood method, the significant factors are finally found. The results reveal that the relative passenger demand, the regional likelihood of pick-ups as well as the expected rate of return are the most significant factors influencing customer-search behavior. Although the impact of traffic situation (i.e., the en-route delay and traffic condition of the target hotspot) is not particularly significant, service providers and policymakers should still take full advantage of it to schedule taxi service and mitigate the traffic congestion caused by the circulation of vacant taxis. Besides, this paper also shows that the customer-search behavior of a vacant taxi driver varies with the time of day. Findings in this paper are expected to provide comprehensive insights about factors that should be considered in the future operation pattern of a taxi service system where human driver taxis and self-driving taxis are mixed.

Keywords: trajectory extraction, clustering algorithm, logit-based model, customer-search behavior, time-varying, en-route delay

OPEN ACCESS

Edited by:
Yibin Ao,
Chengdu University of Technology, China

Reviewed by:
Yiming Bie,
Jilin University, China
Haobin Li,
National University of Singapore, Singapore

****Correspondence:***
Chao Chen
chenchaocc@mail.dlut.edu.cn

Specialty section:
This article was submitted to Original Research Article, a section of the journal Frontiers in Public Health

Received: *01 January 2022*
Accepted: *17 February 2022*
Published: *17 March 2022*

Citation:
Yu L, Sun Z, Jin L and Chen C (2022) A Data-Driven Customer-Search Modeling With the Consideration of Traffic Environment. Front. Public Health 10:848748. doi: 10.3389/fpubh.2022.848748

1. INTRODUCTION

By providing demand-responsive, privacy, and flexible transport, taxi service plays a vital role in satisfying travel demand within an urban area. Compared with other public travel modes, taxis could offer a more comfortable and fast service and thus have been expanding their mode share of urban trips in recent years (1, 2). However, there are several tricky problems associated with the taxi service. First, during peak hours, it is difficult for a passenger to take a taxi in densely populated areas. The spatial-temporal mismatch between taxi demand and supply makes the service level of taxis in satisfying travel demand low.

Second, as taxis are an essential component of the traffic-flow in urban roads, the route choice behavior of taxi drivers has a significant effect on traffic conditions of the urban transportation network. Although the online taxi-hailing service has been growing sharply in recent years, most vacant taxi drivers still cannot accurately find a new passenger to serve after finishing an order. Based on past operational experiences, they usually circulate in the areas with high travel demand in search of passengers. The inefficient movement of vacant taxis could further increase traffic congestion and air pollution (3, 4).

In addition, conventional human-driven taxis are regarded as unsustainable due to the emission of carbon, invalid circulation, and the declining service level. By contrast, self-driving taxis, one of the hot topics in recent years, have been a suitable alternative. In the near future, the taxi service will be provided where human-driven taxis and self-driving taxis are mixed. Many potential benefits, such as the increase of service level and the improvement of urban traffic conditions, can be achieved if human-driven taxis work well with self-driving taxis. Therefore, understanding the customer-search behavior of human driver taxis is also critical to facilitating the deployment of self-driving taxis in the future taxi industry. Furthermore, it could also benefit ride-hailing types of taxis where people search a customer depending on their own experiences.

To tackle the issues involved in the taxi service, researchers have put forward various studies from different perspectives and practical methods. For the review of problems or models involved in the taxi service, interested readers could refer to Yang et al. (5) and Salanova et al. (6). The related works can be divided into three parts, 1) studies in regulatory policies, 2) studies in taxi network modeling, and 3) studies in customer-search behaviors.

Studies that relate to regulatory policies (7–10) often investigate the impact of the implementation of such policies, such as price control or entry restriction. However, with the assumption of the idealized taxi market, these studies were all analyzed from a macroscopic view. The spatial structure of the taxi market is not considered (11–14). In addition to such studies in regulator-aspect, many studies also paid attention to the taxi network modeling intending to capture the spatial structure of the taxi market (5, 15–19). It is worthy of note that the studies on taxi network modeling are mainly based on the assumption that taxi drivers search for passengers to minimize their searching time. However, this assumption indeed cannot reflect the real situation because the search time is not the only determinant affecting customer-search behavior. To better model the behavior of vacant taxi drivers, an increasing number of researchers paid great attention to finding factors that influence the customer-search behavior (11–14, 20–23).

The previous studies on search behaviors of vacant taxi drivers are mainly based on logit-based or probability-based form. Sirisoma et al. (11) conducted a stated preference survey of 400 taxi drivers in Hong Kong and developed a Multinomial Logit (MNL) model to analyze the choice mechanism. The results revealed that the trip time, toll, and waiting time are all significant factors affecting customer-search behavior. Besides, the demographic of drivers, such as age, marital status, driving experience, and vehicle ownership, also affected the choice mechanism. In recent years, the cost of GPS data collection has been deeply discounted with the rapid development of technologies and information systems. After collating and analyzing taxi GPS data, the route choices of taxi drivers in a real-world situation can be extracted, providing the essential data to understand the customer-search behavior of a vacant taxi driver.

Based on the GPS data of 460 taxis in Hong Kong, Szeto et al. (12) developed a time-depend logit-based model to study the customer-searching strategies of vacant taxi drivers over a day. Their results indicated that the passenger demand and rate of return are two significant factors that affect the searching behavior. Also, using the GPS data from 460 taxis in Hong Kong, Wong et al. (13) intended to seek the underlying mechanism of taxi customer-search behavior and validated a logit model, in which several factors were explored, including the cross-zonal travel distance, intra-zonal circulating distance, relative passenger demand and rate of return. To capture the effect of the cumulative probability of successfully picking passengers up in the customer-search process, Wong et al. (14) proposed a cell-based model combining the logit-based model and intervening opportunity model. In their paper, the GPS data of the 460 taxis in Hong Kong was also used to calibrate the model.

Apart from the theoretical works based on the theory of modeling, some empirical studies were also conducted to study the customer-search behavior of a vacant taxi driver. By using the continuous digital traces of more than 3,000 taxis in over 48 million trips, Liu et al. (20) made an empirical analysis of the operation patterns of taxi drivers. In their paper, the transportation congestion condition is a critical determinant for taxi drivers to search for customers. Veloso et al. (24) performed an exploratory analysis to understand the customer-search behavior in suburban areas by using the taxi-GPS traces collected in Lisbon, Portugal. They found taxi drivers prefer to search customers in areas with a higher probability of picking passengers up even if they need to travel a longer distance to that location. Through the taxi GPS data collected in Shenzhen, China, Zong et al. (23) developed a zero-inflated negative binomial model to identify the impacts of external and internal information (e.g., previous pick-up experience) on the cruising behaviors of taxi drivers. Their results indicated that external factors such as land use, traffic conditions, and road grade have a more significant influence than the internal ones (i.e., previous pick-up experience).

In summary, existing literature specifically focused on the study of customer-search behaviors has explored several factors affecting the search mechanism of vacant taxi driver, including search time/distance (13, 24, 25), revenue (20, 24), rate of return (13), passenger demand (13), pick-up likelihood (14, 24), land use (23), and transportation congestion condition (20). Though studies based on the modeling theory could provide the quantitative influence of those factors, they were still constrained by the small sample size. Quite the opposite, some empirical studies considered abundant factors in analyzing customer-search behavior; however, the significance of each factor cannot be quantified (20, 25). With the rapid development of vehicle-loaded GPS tracking devices, the available taxi data has been greatly enriched, providing the basis for deepening the

understanding of customer-search behavior. In this paper, a data-driven taxi customer-search modeling is developed to fill the gap between the theoretical works and empirical studies on the customer-search behavior of a vacant taxi. Besides, to study the choice decisions of vacant taxi drivers, previous studies usually segmented the region into many small cells (14, 23, 24, 26) or different administrative districts (12, 20). To better reflect the demand distribution and search area preferences of vacant taxis, this paper utilizes the OPTICS algorithm to divide the study area into different customer hotspots. Furthermore, because of the temporal variations of passenger demand, the customer-search strategies of vacant taxi drivers indeed vary over the time of day (13, 14, 20, 23). To study the customer-search behavior of vacant taxi drivers in different periods, this paper develops and calibrates several time-dependent MNL models by mining over 1.6 billion GPS records from about 8400 taxis in Shanghai. Based on the Watson and Westin's pooling test (27), the time-varying search strategies are finally investigated and approved. An overview of the study methodology is also presented in **Figure 1**.

The contributions of this paper lie in the following.

1. Using the OPTICS algorithm, the study area of this paper is segmented into different customer hotspots. Unlike previous studies that usually divided a mega-region into many small cells or administrative districts, this method could better reflect taxi drivers' demand distribution and searching area preferences.
2. By investigating a typical search delivery process of a vacant taxi, five candidate factors that may have an influence on the customer-search behavior, including the relative passenger demand, regional likelihood of pick-ups, expected rate of return, en-route delay, as well as the traffic condition of target hotspot are defined.
3. Several time-dependent MNL models are developed to provide empirical evidence about the significant factors in the customer-search mechanism of a vacant taxi driver. Besides, the Waston and Westin pooling test is also conducted in this paper to reveal the time-varying characteristic of customer-search behavior.

The rest of this paper is structured as follows. Section 2 describes the raw data, initial data processing algorithms, as well as the study area and time. Section 3 presents the methodology on the determination of customer hotspots and logit-based taxi customer-search behavior model. Section 4 provides the detailed results. Finally, concluding remarks and outlooks for future work are drawn in Section 5.

2. DATA PREPARATION

2.1. Data Acquisition

Taxi GPS data used in this paper is collected from the Shanghai Qiangsheng Taxi Company. The number of taxis operated by Shanghai Qiangsheng Taxi Company occupies approximately 30% of the entire Shanghai taxi population, and it can adequately represent the behaviors of taxi drivers in Shanghai (3). The raw data collected every 10s for over a month (March 2016) contain about 8400 taxis and more than 1.6 billion GPS records in total. The database records the location in terms of longitude and latitude, vehicle identification, spot speed (km/h), timestamp, and operation status (empty or occupied). We should mention that the data includes both occupied and empty running trips and could describe the full operation behaviors of taxi drivers.

2.2. Data Clean, Processing, and Extraction

To obtain the pick-up and drop-off points, as well as those occupied and empty trips, some initial data processing procedures are firstly carried out. These procedures include data cleaning, coordinate transformation, and map matching. Then an Origin-Destination (OD) extraction and mapping algorithm based on time series are designed. This algorithm provides the

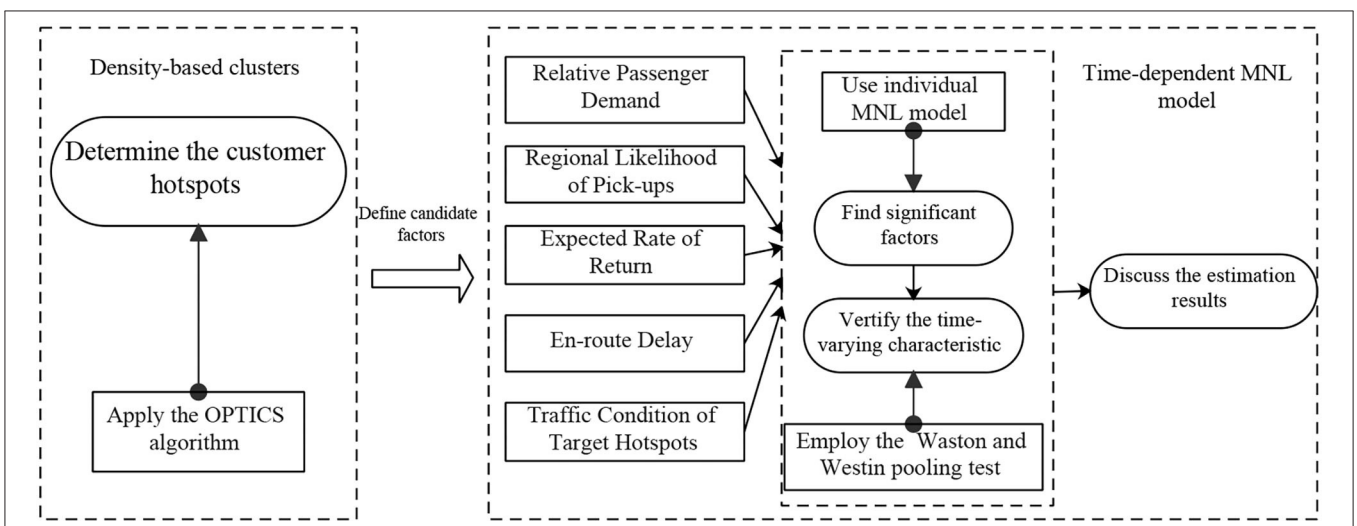

FIGURE 1 | Overview of study methodology.

data foundation for this paper. The trip distance, travel time, and other essential parameters thus can be ultimately calculated.

In this paper, the pick-up point can be considered as a point where the state of the taxi is translated from empty to occupied. Similarly, the drop-off point is a transition of the state of a taxi from occupied to empty. Thus, based on the status change, the procedure of the OD extraction and mapping algorithm can be designed and is shown in **Appendix A**.

2.3. Travel Fare and Operational Cost

To develop the MNL model, the taxi fare and operational cost should also be mentioned. The fare structure adopted in Shanghai is based on a non-linear fare structure. In the day-time, the initial charge for the first 3 km is 13 CNY and an additional 2.4 CNY charge for every subsequent 1 km or every 5 min waiting time. Trips after 11 pm and before 5 am will start at 16 CNY and will be charged an additional 3.1 CNY compared to the day-time 2.4 CNY. Besides, passengers also must pay 1 CNY for the fuel surcharge in both day-time and mid-night.

Before introducing the operational cost, it is necessary to describe the work-pattern of taxi drivers. In Shanghai, most taxis run 24h a day but in two shifts: The day shift is from 4 am to 4 pm, and the night shift is from 4 pm to 4 am. Thus, each vehicle has two drivers. Based on the report provided by the Shanghai Qiangsheng Taxi Company, the monthly taxi rental cost in 2016 for a single driver is 4,700 CNY. With the shift of 12h and the operating days of 30 days, the rental cost can be calculated as 0.22 CNY/min.

Apart from the rental cost, the fuel cost also accounts for a large proportion of the total operational cost of taxis. In 2016, all taxis of Shanghai Qiangsheng Taxi Company were running on petrol. Based on a survey of taxi drivers, the fuel consumption of a regular taxi is approximately 8 to 10 liters per 100 kilometers. Therefore, the fuel consumption of taxis per 100 kilometers in this paper is set as 9 liters. The unit price of petrol in 2016, Shanghai, is 6.87 CNY/liter. Thus, the unit fuel cost of taxis is 0.6 CNY/km. Although there are no data on precise fare collection, the approximate fare can be calculated by the vehicle trajectory (trip distance and travel time) deduced by the taxi location and speed information. The approximate operational cost can be calculated in the same way.

2.4. Study Region and Time

As shown in **Figure 2**, the distribution of pick-ups and drop-offs of taxis from QiangSheng Company during a typical workday covers almost the entire city of Shanghai. Furthermore, the Inner Ring Road, denoted by the red circle, encompasses the downtown area of Shanghai. The pick-ups and drop-offs within the Inner Ring Road contribute approximately up to 89.85 and 87.89% of daily pick-ups and drop-offs, respectively. This area can be regarded as the active service region of taxis. Thus, in this paper, the area of Shanghai within the Inner Ring Road is selected as the study area.

The customer-search behavior of vacant taxis indeed varies with time (12). It depends on the balance of the demand and taxis in operation (i.e., the level of competition). To accurately find factors affecting customer-search behavior, it is necessary to divide the time into several spans. In this paper, the ratio of empty/total service time (RETST) is adopted to show the level of competition. **Figure 3** shows the trends of the RETST and the average number of taxis in operation over a typical day in March 2016.

The low RETST during peak hours (7:00–10:00 and 16:00–19:00) means that the supply falls nearly short of demand. During this period, vacant taxi drivers could quickly get new orders. By contrast, the occurrence of high RETST after the evening rush hour and before dawn (23:00–7:00) denotes that the supply exceeds the demand. During this period, taxi drivers prefer not to provide service, and the number of taxis in operation reaches the minimum at 5 a.m. It should also be noted that the average number of taxis in operation falls slightly from 10:00 to 13:00 and then increases from 13:00 to 16:00. That may be attributed to that in Shanghai, taxi drivers usually have a flexible meal schedule and shift change during that time. However, it is difficult to distinguish betweenthe status of taking a meal, having a shift change, or waiting for new orders. To avoid the bothers brought by the small sample size, break-time, and work shift, we omit data collected during 23:00–7:00 and 10:00–16:00 and eventually select the continuous periods 7:00–10:00, 17:00–19:00, and 20:00–23:00 as our study time.

Thus, the customer-search behaviors of vacant taxi drivers during peak period (i.e., 7:00–10:00 and 17:00–19:00) and off-peak period (i.e., 20:00–23:00) are studied, respectively.

3. METHODOLOGY

3.1. Determination of Customer Hotspots

In order to study the customer-search behavior of a vacant taxi driver from a microcosmic perspective, it is necessary to separate the whole city area into many relatively small ones. Previous studies usually segmented the study region into small cells (14, 26, 28, 29) or different administrative districts (12, 14, 20, 30). However, these subdivision methods can not reflect the demand distribution and searching area preferences of vacant taxis. Unlike previous studies, in this paper, to alleviate the burden of data processing, the study area (i.e., the Shanghai city) is divided into different customer hotspots based on the OPTICS algorithm. The hotspot in this paper refers to a specific region where vacant taxis prefer to search for customers.

The OPTICS algorithm, first proposed by Ankerst et al. (31), is one of the most popular density-based clustering methods. The principle of the OPTICS is equivalent to that of an extended density-based spatial clustering of application with noise (DBSCAN) algorithm. Similarly, the OPTIC algorithm depends on the three parameters defined in the DBSCAN algorithm, including the generation distance (d), the radius of cluster (ε), and the minimum number of points required for a cluster (*MinPts*). However, unlike the DBSCAN explicitly generating the clusters for data set, this method produces a density-based clustering structure of the data in the order of the points. Besides, the OPTIC can also avoid the trouble of parameter selection in the DBSCAN algorithm (31, 32). It should be mentioned that a set of information, including the core distance (cd) and the reachability distance (rd), is stored in the

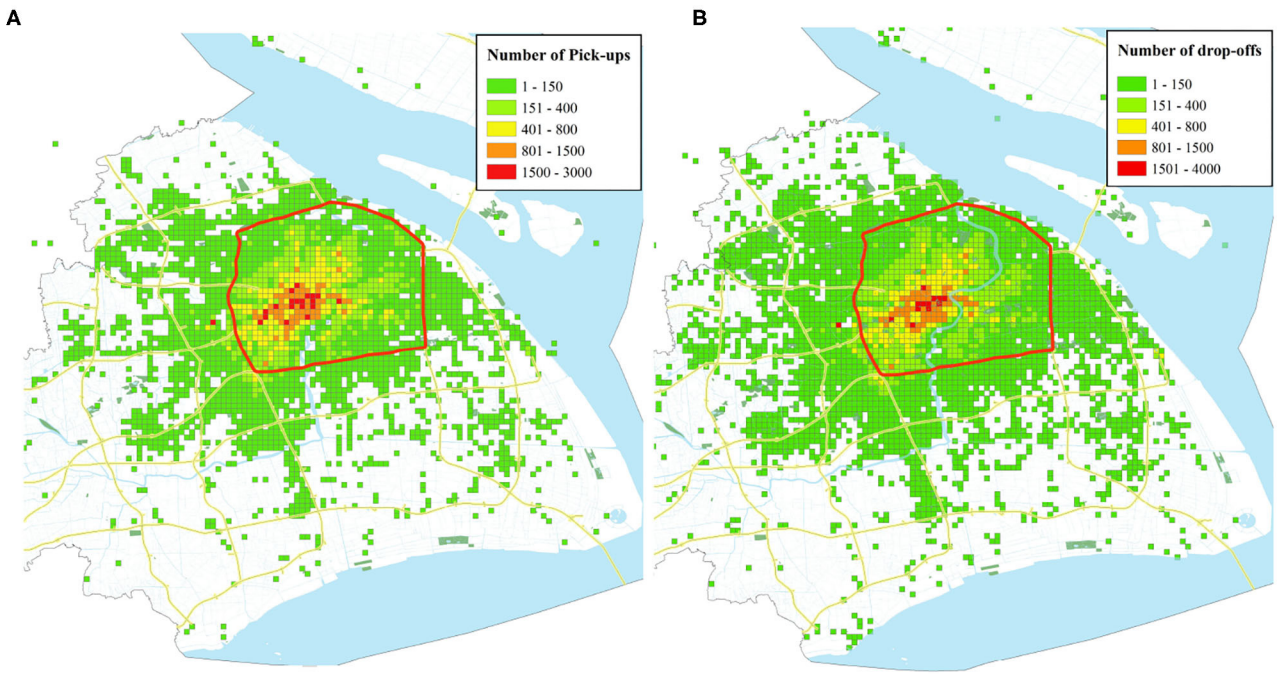

FIGURE 2 | The study area and demand distribution during a typical workday. **(A)** The distribution of pick-ups. **(B)** The distribution of drop-offs.

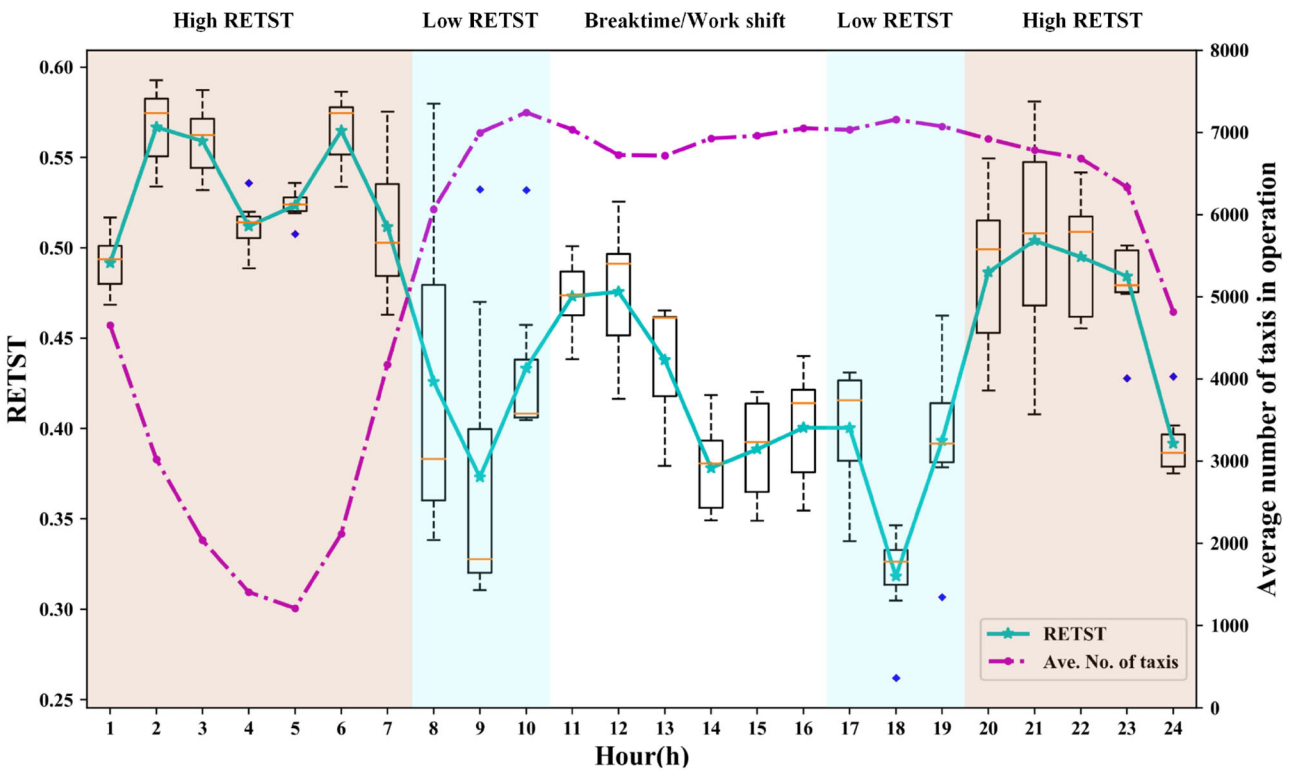

FIGURE 3 | Profiles of the ratio of empty/total service time (RETST) and the average number of taxis in operation.

density-based cluster structure. This kind of information can be finally utilized to extract the clusters and identify noises. For more details of the OPTICS algorithm, interested readers could refer to Ankerst et al. (31).

Based on the concepts of the OPTICS algorithm discussed above, the procedure of the clustering algorithm for our problem is designed. However, before giving the pseudo-code of the OPTICS algorithm, related symbols and functions should be first introduced. I is the data set, $N_\varepsilon(i)$ is the ε-neighborhood of data point i, c_i and r_i are the core distance and reachability distance of data point i, respectively. v_i is the sign variable indicating whether data point i is traversed or not. *Seedlist* is the temporary set of data points that have been processed and ordered by the reachability distance in ascending. P is the sequence of all data points in ascending order of reachability distance, and M is the clustering result. Then, the procedure of the OPTICS is shown in the following.

The function *insertlist*() is indeed a sub-procedure of the **Algorithm 1**. It is used to update the reachability distance of the data that have been processed. This function is repeated called by the OPTICS algorithm until all data points are traversed, and the pseudo-code is shown as follows.

Thus, the study of vacant taxi customer-search behavior in this paper turns into the analysis of the customer hotspot choice decision. Based on the historical taxi trajectory data and the algorithm, the spatial characteristics of pick-ups and drop-offs can also be explored.

Algorithm 1 : OPTICS algorithm.

```
//Initialization//
Given parameter ε, MinPts
Generate N_ε(i) and c_i, i = 1, 2..., N
Set k = 1, v_i = 0, r_i = UNDEFINED, i = 1, 2..., N
Set I = {1, 2, ..., N} and seedlist = ∅
//OPTICS//
While I ≠ ∅ do
    Get an element i ∈ I and set I = I \ {i}
    If v_i = 0 then
        Set v_i = 1, p_k = i, and k = k + 1
        If |N_ε(i)| ⩾ MinPts then
            insertlist
            (N_ε(i), {v_l}_{l=1}^N, {r_l}_{l=1}^N, c_i, seedlist)
            While seedlist ≠ ∅ do
                Get the first element j ∈ seedlist
                Set v_j = 1, p_k = j, and k = k + 1
                If |N_ε(j)| ⩾ MinPts then
                    insertlist
                    (N_ε(j), {v_l}_{l=1}^N, {r_l}_{l=1}^N, c_j, seedlist)
                EndIf
            EndWhile
        EndIf
    EndIf
EndWhile
Output P = {P_i}_{i=1}^N
//OPTICS Clustering Extracting//
Given parameter ε̃ (ε̃ ⩽ ε)
Set clusterID = −1, and n = 1
For i = 1, 2, ..., N do
    t = p_i
    If r_t > ε̃ or r_t = UNDEFINED then
        If c_t ≠ UNDEFINED and c_t ⩽ ε̃ then
            clusterID = n
            n = n + 1
            m_t = clusterID
        Else
            m_t = −1
        EndIf
    Else
        m_t = clusterID
    EndIf
EndFor
Output M = {m_i}_{i=1}^N
```

3.2. Candidate Factors Affecting the Vacant Taxi Customer-Search Behavior

Vacant taxi customer-search is a high-level human behavior and complex decision-making process that incorporates the drivers' past and unspoken experiences. Previous studies have explored several factors that may influence the vacant taxi customer-search behavior (11, 12). Concerning these factors, the en-route delay and the traffic condition of the target hotspot are always ignored. To find factors affecting customer-search behavior comprehensively, the overall customer-search and customer-delivery process should be studied. A typical process of customer search and delivery is shown in **Figure 4**.

As illustrated in **Figure 4**, a studied taxi finishes the current order at drop-off location α and is going to search for the next passenger. The demand in the target area influences the customer-search decision. Due to the high demand, searching for a customer successfully in a short period becomes promising. In other words, relative passenger demand is an essential factor that may affect customer-search behavior. However, in the search process ($\alpha \rightarrow \beta$), high passenger demand cannot ensure the probability of successfully picking passengers up. Vacant taxi drivers prefer to search through an area, in which passenger demand is great, and quantity supplied (i.e., the number of vacant taxis) is small. The probability of a vacant taxi driver successfully picking a passenger up decreases if there are many vacant taxis nearby. Thus, the regional likelihood of pick-ups is also a critical candidate factor. The traffic condition along the way of searching customers is also likely to affect customer-search behavior. The low speed caused by traffic jams will deter vacant taxi drivers from searching for passengers. Similarly, the traffic condition of the target area also needs to be considered. Finally, the fare, travel expense, and time (i.e., the expected rate of return) should also be considered because vacant taxi drivers prefer passengers with high travel fare and low searching time and cost.

In this paper, the vacant taxi customer-search behavior is considered as the choice to select a particular customer hotspot in the city region. The choice can either be intra-hotspot circulating or cross-zonal traveling. Besides, these choices are assumed to

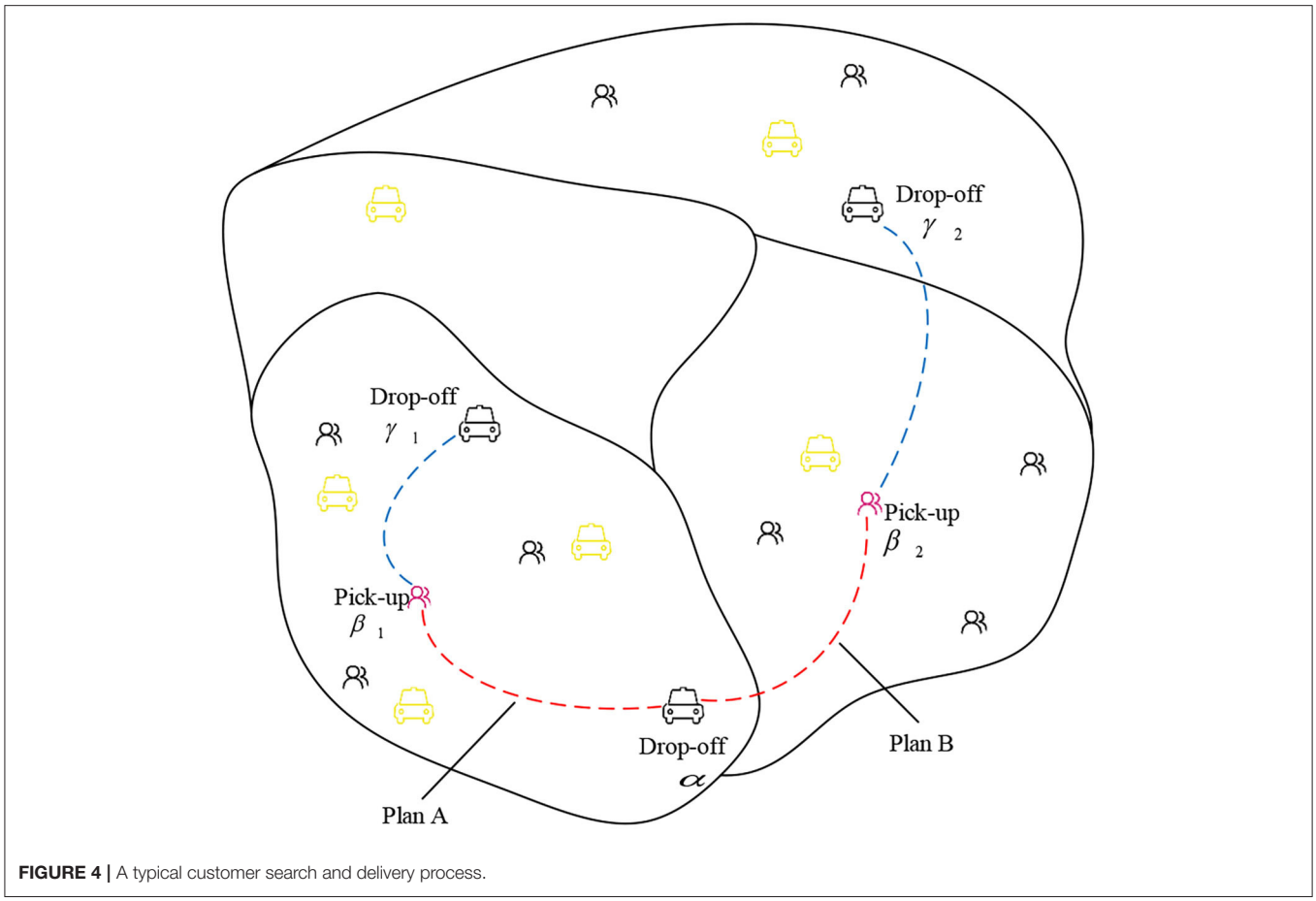

FIGURE 4 | A typical customer search and delivery process.

Algorithm 2 : The procedure of *insertlist* ().

For $h \in N_\varepsilon(H)$ **do**
  If $v_h = 0$ **then**
    rd=$max\left\{cd\_H, d\left(x^{(H)}, x^{(h)}\right)\right\}$
    If $r_h = UNDEFINED$ **then**
      Set $r_h = rd$
      Insert h into *seedlist* and sort *seedlist* in accending order
    Else
      If $rd < r_h$ **then**
        Set $r_h = rd$
        Insert h into *seedlist* and sort *seedlist* in accending order
      EndIf
    EndIf
  EndIf
EndFor

be independent. By analyzing the typical search and delivery process of a vacant taxi driver, five factors that may have an effect on the customer-search behavior are summarized and defined. The associated definitions and formulates for the five factors, including the relative passenger demand, regional likelihood of pick-ups, expected rate of return, en-route delay, as well as traffic condition of target hotspot, are presented in the following.

3.2.1. The Relative Passenger Demand

Passenger demand is the most important factor considered by vacant taxi drivers in finding customers. They always prefer to cruise in areas with higher passenger demand according to their experience. In this paper, the relative passenger demand of hotspot (i) is used to capture the attractiveness of high demand, and the definition is shown as follows.

Definition 1. Given the number of passengers picked up in hotspot i at time interval t, P_i^t, and the total number of occupied trips in the entire study area, the relative passenger demand E_i^t is defined as follows.

$$E_i^t = \frac{P_i^t}{\sum_{i \in N} P_i^t} \tag{1}$$

3.2.2. The Regional Likelihood of Pick-Ups

Higher passenger demand in one area cannot ensure a reasonable probability of successfully picking passengers up. It can be offset by the heavy competition. Therefore, the regional likelihood of pick-ups should also be a factor that needs to be considered.

Definition 2. Given the total number of vacant taxi V_i^t and occupied taxi O_i^t in hotspot i at time interval t, the regional

likelihood of pick-ups, LP_i^t, can be formulated as:

$$LP_i^t = \frac{V_i^t}{V_i^t + O_i^t} \tag{2}$$

3.2.3. Expected Rate of Return

When searching for the next customer, the vacant taxi driver will consider not only the time and cost consumed from the current location to the next pick-up location but also the expected profit brought by the next order, which will eventually affect their incomes. Thus, the expected rate of return is a candidate factor to be considered. Before calculating the expected rate of return, three related definitions, including the travel time, travel expense, and fare, are first given below.

Definition 3. The total travel time of trip k of an individual vacant taxi driver starting from hotspot i and aiming j to search for the next customer $T_{k,i,j}^t$.

$$T_{k,i,j}^t = ST_{k,i,j}^t + DT_{k,i,j}^t \tag{3}$$

where $ST_{k,i,j}^t$ is the searching time of trip k of an individual vacant taxi driver from hotspot i to j at time interval t; and $DT_{k,i,j}^t$ is the service time (i.e. delivery time) of trip k.

Definition 4. For a typical customer search and delivery process, the travel expense for a vacant taxi driver is comprised of two parts: one is the rental cost, and the other is the fuel cost. Thus, the travel expense of trip k of an individual vacant taxi driver from hotspot i to j at time interval t, $C_{k,i,j}^t$, can be calculated as follows.

$$C_{k,i,j}^t = C^r \cdot T_{k,i,j}^t + C^f \cdot \left(SD_{k,i,j}^t + DD_{k,i,j}^t\right) \tag{4}$$

where C^r is the rental cost per unit of time; C^f is the fuel cost per unit of distance; $SD_{k,i,j}^t$ is the search distance of trip k in customer-search process from hotspot i to j; $DD_{k,i,j}^t$ is the delivery distance of the occupied trip k from hotspot j.

Definition 5. The fare of trip k of an individual vacant taxi driver from hotspot i to j at time interval t can be calculated based on time interval, fixed fuel surcharge fare, and the delivery distance, which is formulated as follows.

$$F_{k,i,j}^t = Fc + F(s) \tag{5}$$

Where Fc is the fixed fuel surcharge fare, s denotes the delivery cost, and $F(\cdot)$ is the function of travel fare calculation discussed in Section 2.3.

We can, therefore, define, hereafter, the expected rate of return as follows.

Definition 6. The expected rate of return (EROR) is calculated based on travel fare, expense, travel time, and the time interval, which is denoted as follows.

$$R_{k,i,j}^t = \frac{F_{k,i,j}^t - C_{k,i,j}^t}{T_{k,i,j}^t} \tag{6}$$

The numerator of the Equation 6 shows the expected profit of trip k of an individual vacant taxi driver from hotspot i to j at time interval t. Furthermore, the denominator represents the expected time taken to obtain that profit. The two numbers are calculated based on an intact customer search and delivery process. The value of the ratio describes the effect of perceived profit in the decision-making process when vacant taxi drivers search for the next customer. A higher EROR may result in a higher probability of a vacant taxi driver going from one hotspot to the designated hotspot.

3.2.4. En-route Delay

The en-route delay reflects the traffic condition along the way of searching for customers. With the same searching distance, severe en-route delay may increase the searching time and cost, even destroying the light mood. Thus, it should be one another candidate factor affecting the customer-search behavior of a vacant taxi driver. En-route delay is indeed correlated with several factors, such as type of road, road capacity, and traffic accident. However, it is challenging to incorporate all factors into one parameter to represent the en-route delay. In practice, experienced drivers are more likely to search passengers through a route with a stable travel speed. Thus, in this paper, the en-route delay is represented by the standard deviation of customer-search time between the current hotspot and the designated hotspot, and the definition is given below. Note that the en-route delay in this paper does not intend to represent traffic conditions of a specific road. Instead, it aims to reflect the general road traffic conditions between two hotspots partly.

Definition 7. Given the average customer-search time, and the total number of search trips (i.e., empty trips) from hotspot i to j, the en-route delay can be mathematically expressed as follows.

$$ED_{i,j}^t = \sqrt{\frac{1}{N_{i,j} - 1} \sum_k (ST_{k,i,j}^t - \overline{ST}_{i,j}^t)^2} \tag{7}$$

Where N_{ij} indicates the total number of search trips from hotspot i to j. $ST_{k,i,j}^t$ is the search time of trip k of an individual vacant taxi driver from hotspot i to j at time interval t. $\overline{ST}_{i,j}^t$ is the average customer-search time from hotspot i to j at time interval t, which is calculated as $\sum_k \frac{ST_{k,i,j}^t}{N_{i,j}}$.

3.2.5. Traffic Condition of Target Region

Similarly, the traffic condition of target hotspot j will also affect the vacant taxi customer-search behavior. A congested traffic condition may cause taxi drivers to move slowly and consume more time to deliver the passengers. In this paper, the average speed is adopted to reflect the traffic condition of the target region j.

Definition 8. The average speed of target region j at time interval t, V_j^t, is a ratio of search and delivery distance to the search and delivery time.

$$\overline{V}_j^t = \frac{\sum_k (SD_{k,j}^t + DD_{k,j}^t)}{\sum_k (ST_{k,j}^t + DT_{k,j}^t)} \tag{8}$$

where $SD^t_{k,j}$ is the search distance before trip k. In contrast, $DD^t_{k,j}$ is the delivery distance of trip k starting from hotspot j at time interval t. And $ST^t_{k,j}$ and $DT^t_{k,j}$ denote the corresponding search time and delivery time, respectively.

For every single search-delivery trip, the five factors are calculated using the data introduced in Section 2. Then 22,507 samples are drawn for the off-peak period, 25,923 samples are drawn for the morning peak hours, and 16,864 samples for the evening peak hours.

3.3. Multinomial Logit Model for Customer-Search Behavior

In this paper, the MNL model is employed to analyze the customer-search behavior of a vacant taxi driver. Based on the theory of discrete choice, a vacant taxi driver chooses the designated hotspot, which brings him the highest utility. The utility (U_{ijk}) when the vacant taxi driver is currently in hotspot i and about to go to hotspot j for searching the next passenger in trip k, consists of two parts: measurable utility (V_{ijk}) and unmeasured utility (ε_{ijk}):

$$U_{ijk} = V_{ijk} + \varepsilon_{ijk} \tag{9}$$

Based on the candidate factors presented in Section 3.2, the measurable utility function can be mathematically expressed as follows.

$$V_{ijk} = \delta_j + \beta^E E^t_j + \beta^{LP} LP^t_j + \beta^R R^t_{k,i,j} + \beta^{ED} ED^t_{ij} + \beta^{\bar{V}} \bar{V}^t_j \tag{10}$$

Where, δ_j is the utility constant of hotspot j; β^E is the coefficient associated with the relative passenger demand; β^{LP} is the coefficient associated with the regional likelihood of pick-ups; β^R is the coefficient associated with the expected rate of return; β^{ED} is the coefficient associated with the en-route delay and $\beta^{\bar{V}}$ is the coefficient associated with the traffic condition of target hotspot.

When the factors of measurable utility and the unmeasured utility are independent, and the random utility (ε_{ijk}) is the unobserved error component obeying Gumbel distribution, the probability that the vacant taxi driver currently in hotspot i chooses hotspot j to start trip k can be obtained by the following form.

$$P_k(j \mid i) = \frac{e^{V^{ijk}}}{\sum_j e^{V_{ijk}}}, i = 1, 2, ..., N; j \in K_i \tag{11}$$

Where, K_i is the set of optional hotspots for the vacant taxi located in hotspot i.

3.4. Multicollinearity Detection

As the classical linear regression, the multicollinearity problem is also involved in the MNL model. It affects the stability of models and may cause the incorrect identification of influence factors (33). In this paper, two methods, Pearson product-moment correlation coefficient (PPMCC) and the variance inflation factor (VIF), are introduced to detect the presence of multicollinearity. The former method is to identify the linear correlation between any two variables in the five-candidate factors. Variables with the coefficients (absolute values) higher than 0.7 are considered to have the multicollinearity and will be removed in the MNL model. The latter method is used to reduce the likelihood of any multicollinearity further. VIFs equal to 1 indicates that there is no collinearity existing in the variables of the model. By contrast, VIFs larger than 10 present the severe collinearity (34). Thus, in this paper, variables with VIFs higher than 10 will also be eliminated in the model.

TABLE 1 | Combined MNL models with different time periods.

Model	Time periods		
	Morning-peak hours	Off-peak hours	Evening-peak hours
1	√	√	
2	√		√
3		√	√
4	√	√	√

3.5. Watson and Westin Pooling Test

The customer-search behavior of a vacant taxi driver appears not to be constant for different periods (12, 13, 20, 26). Taxi drivers will circulate the city area in different patterns in 1 day to maximize their profits. Thus, it is necessary to understand the underlying behavior patterns for different periods to mitigate the traffic congestion caused by the unnecessary circulation of vacant taxis and improve the income of taxi drivers. In this paper, four combined MNL models with different periods were first developed and listed in **Table 1**.

Then the Watson and Westin pooling test approach was applied to study the sensitivity of the periods (12, 13, 27, 35, 36). This method is based on the log-likelihood ratio, and the associated equation is shown as follows.

$$LR = -2(L_R - L_U) \tag{12}$$

Where L_R is the log-likelihood of the combined MNL model with two or three different periods, and L_U is the sum of the log-likelihoods of the individual MNL models developed for different periods. The null hypothesis of the test approach is that there is no difference between the individual MNL model and the combined MNL model. If the test statistic is larger than the threshold value (i.e., the value specified by the chi-squared distribution at a chosen level of significance), the null hypothesis is rejected. Besides, the degree of freedom can be calculated by subtracting the number of variables of the combined MNL models from the sum of the number of individual MNL models.

4. RESULTS AND DISCUSSION

4.1. Characteristics of Trips and Candidate Factors

Figure 5 shows the distribution of trip characteristics at different times of a typical day. In the morning (or evening) peak hours (7:00–10:00 or 17:00–19:00), vacant taxi drivers spent less time and traveled less distance to find the next customers than off-peak

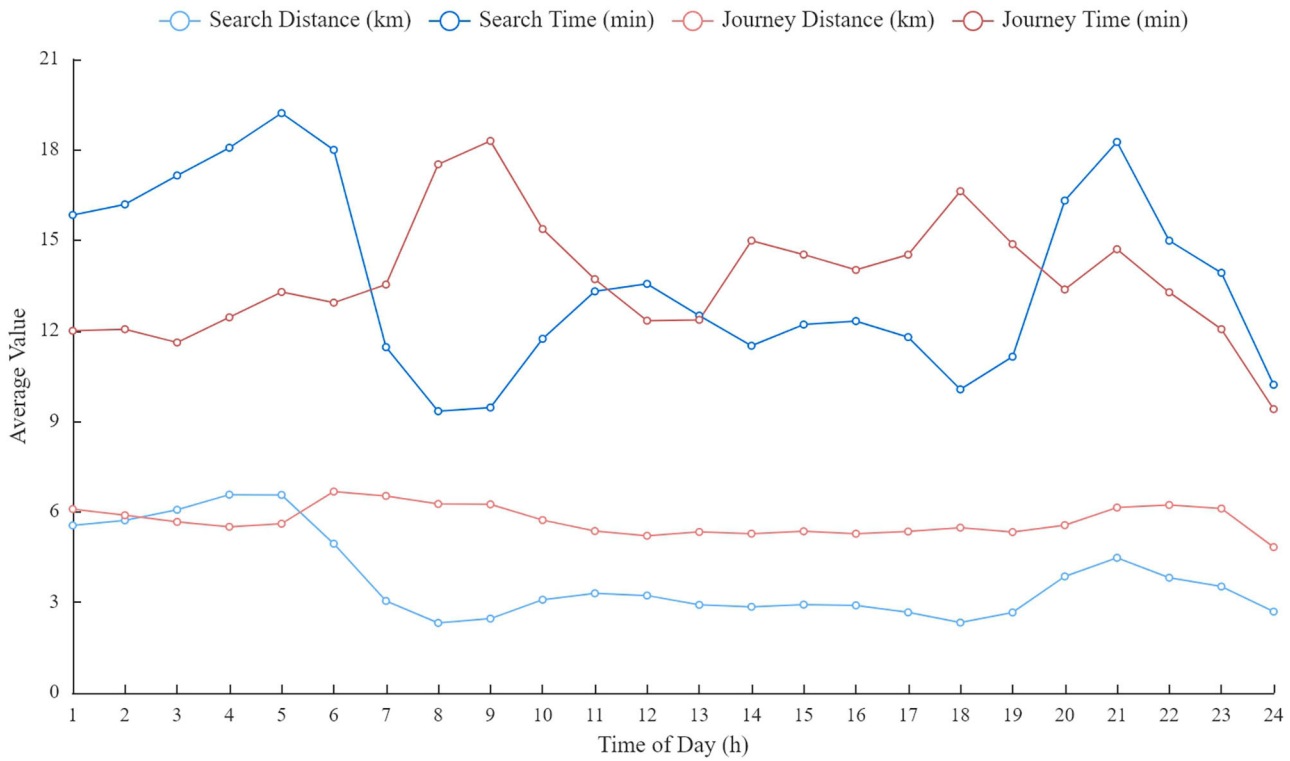

FIGURE 5 | The distribution of trip characteristics on the time of a typical day.

hours due to the high demand. By contrast, the search distance in the midnight period, on average, triples during peak hours. This demonstrates that vacant taxi drivers need to travel more distance to find their next customer in midnight period. Additionally, the journey distances present no appreciable difference in different periods, which is inconsistent with findings in previous research (13). Even more adverse, the journey distance in morning peak hours is higher than other hours. It is potentially related to the spatial division of residential and working areas in Shanghai City. However, in off-peak and mid-night periods, the journal time is notably short than peak hours. It can be attributed to the relatively good traffic condition, enabling taxi drivers to have a higher traffic speed. More details of the trip characteristics can be found in **Appendix B**.

Based on the analysis in **Figure 5**, the taxi market shows a significant fluctuation across a day. The vacant taxi drivers may use different customer-search strategies to increase their incomes. Therefore, the customer-search behaviors of vacant taxi drivers in different periods are studied separately. The behavior-changing across the day will also be explored.

4.2. Clustering Results Based on OPTICS

As discussed, the hotspot refers to a specific region where vacant taxis prefer to search for customers. In this paper, these hotspots are identified using the data of pick-ups in peak hours (including morning and evening peak hours). Note that the commonly used method in the OPTICS algorithm to identify clusterings is to utilize the reachability plot, a scatter diagram showing the reachability distance of each point in the ascending order. Using this plot, the hierarchical structure of the clusters can be easily obtained. In this figure, the x-axis represents the ordering of the points processed by OPTICS, while the y-axis represents the reachability distance.

To determine the best parameters, the range of the two parameters is set as $MinPts \in \{20, 30, 40, 50, 60, 70, 80, 90\}$ and $\varepsilon \in [0.01, 0.06]$. Then, a grid search is performed. Results indicate that the value of the parameter, *MinPts*, is 50. Given the value of *MinPts* is 50, the reachability plot for pick-up points in peak hours can be calculated (as shown in **Figure 6**).

Indeed, the clusters can be represented and extracted based on the dents of the reachability plot (31, 37). However, the number of clusters (i.e., the hotspots in this paper) and the number of points contained in each hotspot are critical for the computing accuracy and computation time. In this paper, the cumulative percentage distribution of pick-ups is used to identify the best neighborhood radius. The results are shown in **Figure 6**.

As we can see in **Figure 7**, the number of clusters with less than or equal to 100 pick-ups accounts for 90% of the whole clusters when the neighborhood radius $\varepsilon = 0.015$. This indicates that many clusters with very few pick-up points are generated.

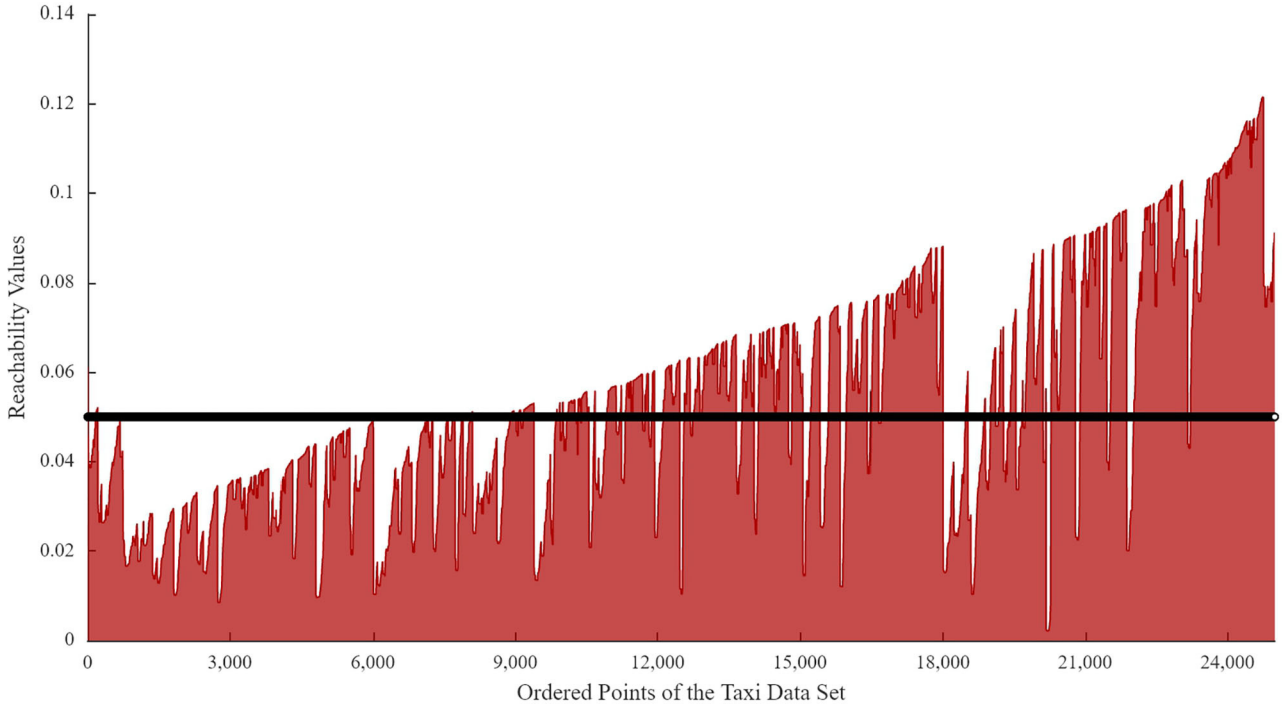

FIGURE 6 | The reachability plot for the data set of pick-ups in peak hours.

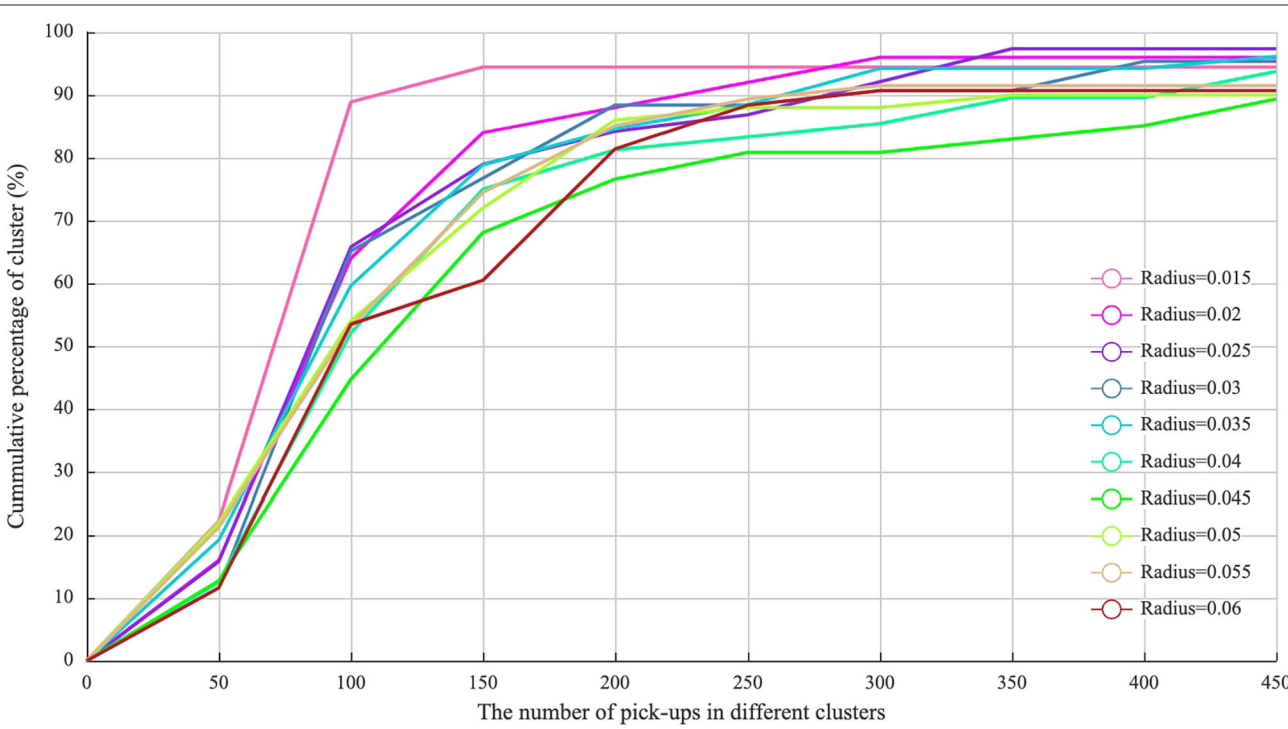

FIGURE 7 | Cumulative proportion of clusters of pick-ups.

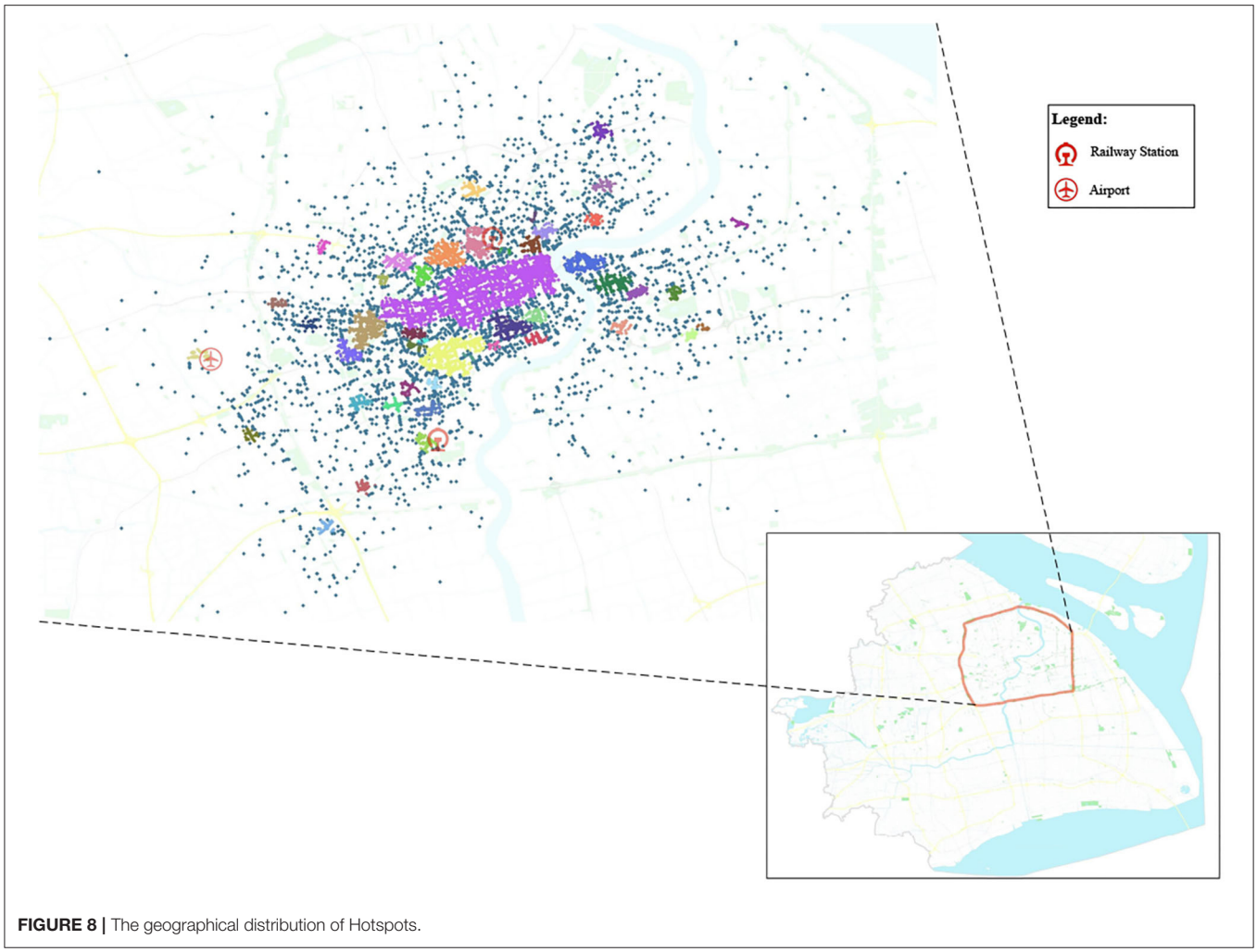

FIGURE 8 | The geographical distribution of Hotspots.

Some central business districts or highly populated areas cannot be efficiently reflected. However, when ε ranges from 0.04 to 0.06, the percentage of clusters with less than 450 pick-up points reaches 90%, illustrating the most attractive hotspots in the urban area can be identified. Besides, when $\varepsilon = 0.05$, more than 70% of pick-ups distributed within the Inner Ring of Shanghai City are clustered. Thus, in this paper, the neighborhood radius is set as $\varepsilon = 0.05$. The value of the neighborhood radius is also presented in the black horizontal line of **Figure 6**. Based on the reachability plot and neighborhood radius, 47 hotspots are finally identified by the OPTICS algorithm and visualized in **Figure 8**.

4.3. Multicollinearity Detection

Before the calibration of the MNL model, multicollinearity detection is firstly conducted. In this section, the PPMCC between any two factors considered in the utility function of the MNL model for off-peak hours is calculated and shown in **Figure 9**. As shown in **Figure 9**, the maximum value among the correlation coefficients is 0.388, smaller than the threshold value of 0.7. Thus, results reveal that there is no presence of collinearity for any pairwise correlations between any two candidate factors.

To further detect the presence of multicollinearity, the VIF is also computed using Stata 14.0, and the results are presented in **Table 2**. The VIF values for all candidate factors, ranging from 1.027 to 1.385, are approximately equal to 1 and much less than 10. The results further verify that no collinearity is presented, and the candidate factors are reasonably selected in the utility function of the MNL model. The multicollinearity detection method for the morning-peak and evening-peak models is the same as discussed above. The full results for all periods can be found in **Appendix C**.

4.4. Calibration Results

Using the maximum likelihood method, the coefficient of each variable in the MNL model is calibrated, and the results are shown in **Table 3**. First, for morning peak hours, the Cox & Snell Rho-squared and Nagelkerke Rho-squared of the MNL model are 0.525 and 0.706, respectively. The Cox & Snell Rho-squared and Nagelkerke Rho-squared of the MNL model for off-peak hours are 0.429 and 0.55. Moreover, the values for the evening peak hours' model are 0.24 and 0.32. The values of the two parameters demonstrate that the models have a good fit.

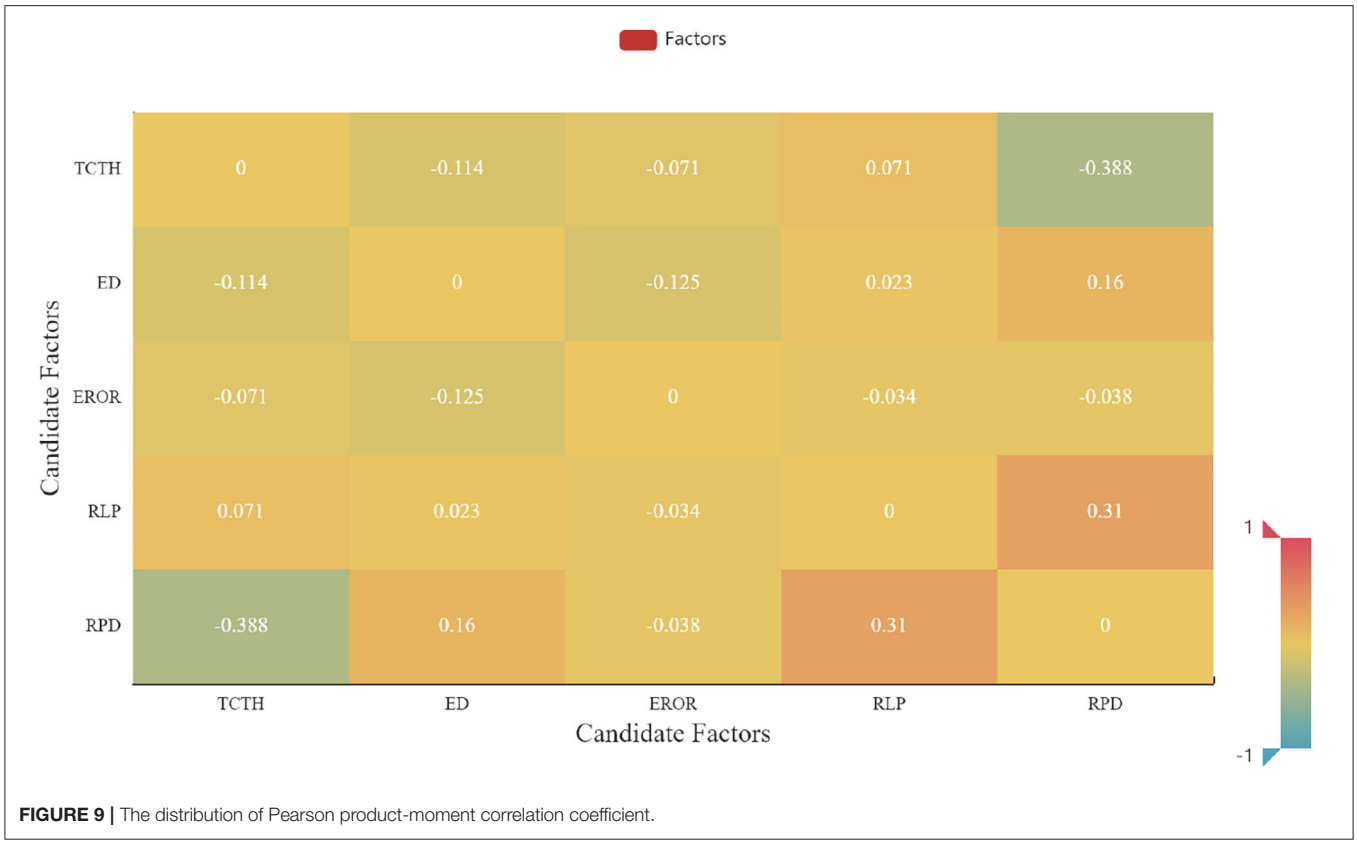

FIGURE 9 | The distribution of Pearson product-moment correlation coefficient.

TABLE 2 | The variance inflation factors.

Candidate factor	VIF	Sqrt(VIF)	Tolerance
RPD	1.385	1.18	0.722
RLP	1.162	1.08	0.861
EROR	1.027	1.01	0.974
ED	1.046	1.02	0.956
TCTH	1.251	1.12	0.799

RPD, Relative Passenger Demand; RLP, Regional Likelihood of Pick-ups; EROR, Expected Rate of Return; ED, Enroute Delay; TCTH, Traffic Condition of Target Hotspot.

TABLE 3 | Estimation Results of the time-dependent MNL model.

Factors	Coefficient[a] [Std.Err][b]		
	Morning-peak hours	Off-peak hours	Evening-peak hours
Constant	–5.536 [0.335]	4.619 [0.321]	2.422 [0.262]
EROR	2.83 [0.12]	2.435 [0.096]	2.338 [0.073]
RPD	3.549 [0.24]	2.219 [0.0301]	3.421 [0.273]
TCTH	–0.217 [0.013]	–0.368 [0.015]	–0.122 [0.012]
RLP	16.802 [0.405]	3.405 [0.558]	3.926 [0.251]
ED	–0.000255 [0.000014]	–0.001 [0.000086]	–0.000107 [0.000022]
Number of observations	25,923	22,507	16,864
–2 Log likelihood	5324.984	6028.273	9220.762
Cox&Snell R Square	0.525	0.429	0.240
Nagelkerke R Square	0.706	0.550	0.320

[a] *All parameters are at a 99% confidence level.*
[b] *The values in bracket represent the standard errors of the candidate variables.*

Second, the sign of the constant of the utility function indicates the preferences of vacant taxi drivers to make intra-hotspot circulating or cross-hotspot traveling. The negative constant in the morning peak hours' model shows that the vacant taxi drivers prefer the intra-hotspot circulating to find the next passenger. However, for the off-peak hours model and evening-peak hours model, the constants of the utility functions are both positive. This indeed concurs with the fact that due to the high travel demand in morning peak hours, vacant taxi drivers could find a nearby passenger in a short time because nearly every person needs to go to work in the morning (see **Figure 5**). By contrast, vacant taxi drivers may need to make cross-hotspot traveling to find the next order during the off-peak period in the face of shrinking demand. An interesting point worth pointing out is the constant of the utility function of the evening-peak period modeling. Typically, the travel demand in evening-peak hours is also relatively higher than in the off-peak period. Vacant taxi drivers should be able to find the next passenger after finishing an order quickly. However, the constant of the utility function in evening-peak hours is found to be positive in this study. This could be explained by the mismatch between the workplace and the dwelling. After delivering the passengers to

TABLE 4 | The results of the pooling test.

Model	Log likelihood					LR [a]	Threshold value	Conclusion of the hypothesis test [d]
	Morning-peak model	Off-peak model	Evening-peak model	Sum	Combined model			
1	–2662.492	–3014.1365	-	–5676.6285	–6234.2135	1115.17	15.086 [b]	Reject
2	–2662.492	-	–4610.381	–7272.873	–8662.6525	2779.559	15.086 [b]	Reject
3	-	–3014.1365	–4610.381	–7624.5175	–8010.607	772.179	15.086 [b]	Reject
4	–2662.492	–3014.1365	–4610.381	–10287.0095	–11984.344	3394.669	23.209 [c]	Reject

[a] *LR is the abbreviation of the log likelihood ratio.*
[b] *The threshold value represents the Chi-square critical value with 5 degrees of freedom at 99% confidence level.*
[c] *The threshold value represents the Chi-square critical value with 10 degrees of freedom at 99% confidence level.*
[d] *Null hypothesis test at 99% confidence level.*

their homes, vacant taxi drivers cannot find nearby passengers heading downtown. Thus, they prefer cross-hotspot traveling to find the next passenger during evening-peak hours.

Third, the model-fitting results show that all candidate factors affect customer-search behavior and are significant at the 99% level. It also proves that the proposed model is reasonable to describe the customer-search behavior of a vacant taxi driver. Besides, the coefficients of the expected rate of return, relative passenger demand, and the regional likelihood of pick-ups in all periods are positive, and the remaining two coefficients, traffic condition of target hotspot and en-route delay, are negative. All signs of the coefficients are logical because the area with a higher expected rate of return, relative passenger demand, and the regional likelihood of pick-ups attract more vacant taxi drivers to search for customers. However, the abominable traffic condition of the target hotspot and severe en-route delay suppress the customer-search willingness of taxi drivers.

Table 3 also shows that the magnitudes of the EROR for different periods are relatively close. It suggests that the perceptions of the EROR for taxi drivers are similar across the day. The results make sense because taxi drivers are always the pursuers of profit maximization. Similar to the perception of the EROR, the perception of the ED also has a minor fluctuation across periods. The coefficients of the ED are all negative, which suggests that the ED has a negative effect on the customer-search behavior in all modeling periods.

However, unlike the EROR and ED, the differences in the coefficients of RPD for different periods are more prominent. The magnitudes of the coefficients of RPD for the peak periods (i.e., the morning and evening peak hours) are more significant than that for the off-peak period. This could be explained by the relatively high travel demand during peak periods, and vacant taxi drivers could find the next customer in a relatively short time. However, for the off-peak period, taxi drivers need to travel a long distance to the hotspot with relatively high passenger demand to find the next order.

The coefficients of the TCTH for different periods are all negative, as we discussed above. However, the critical point worth pointing out is that the coefficient of the off-peak period is lower than those for the morning and evening peak hours. It indicates that vacant taxi drivers are unwilling to circulate within those hotspots with adverse traffic conditions during the off-peak period. Though the abominable traffic condition during peak periods also suppresses the customer-search willingness of taxi drivers, the effect is indeed much smaller. It may be attributed to the generally poor traffic condition in the entire urban area.

Regarding the RLP, its coefficient values range from 3.405 to 16.802, showing a fluctuation across different periods. Notably, the coefficient of the morning-peak model is much larger than those of evening-peak and off-peak models. It shows that vacant taxi drivers make a greater emphasis on the regional likelihood of pick-ups when they search for passengers during the morning peak period. Based on the Standard Error given in **Table 3**, it can conclude that the RLP, RPD, and EROR are the most significant factors affecting the customer-search behavior of vacant taxi drivers, followed by the importance of the TCTH and ED. Finally, to examine whether the customer-search behavior of vacant taxi driver changes across different periods, the Watson and Westin pooling test is applied, and the results are shown in **Table 4**.

Columns 5–6 present the log-likelihoods of the sum of individual period models and the four combined models. Based on Equation 12, the log-likelihood ratio (LR) can also be calculated and shown in column 7. For combination models 1–3, the degree of freedom is 5, and the Chi-square critical value is 15.086 at 99% confidence level. While for combination model 4, the Chi-square critical value with 10 degrees of freedom at 99% confidence level is 23.209. By comparing the values in columns 7 and 8, it can be found that the Chi-square critical values are significantly lower than the LRs. Thus, the null hypothesis that there is no difference between the individual period model and the combined model will be rejected. In other words, the customer-search behavior of vacant taxi drivers varies with the time of day, and we cannot pool the whole day data to mine the underlying patterns of the customer-search behaviors of taxi drivers. The results also implicate that the regulatory restraints for limiting taxi circulation or improving the taxi system performance should not be static but vary depending on the time of day.

Finally, some conclusions and interesting insights can be drawn from this paper, and they are consistent or inconsistent with some ideas in previous studies.

1. The regional likelihood of pick-ups is the most significant determinant affecting the customer-search behavior of a

vacant taxi driver. This finding is consistent with the assumption in previous related research that vacant taxi driver prefers to travel toward areas with the highly successful probability of picking a passenger up (14, 38).

2. Previous studies found those taxi drivers are susceptible to the reliability of travel time (4, 20, 23, 39). This paper proposed two factors to portray the reliability of travel time. One is the traffic condition of the target hotspot, and the other is the en-route delay. Compared with the significance of the traffic condition of the target hotspot, the en-route delay shows a minimal effect on customer search behavior.

5. CONCLUSION

This paper intends to study the customer-search behavior of a vacant taxi driver by mining over 1.6 million GPS records from about 8,400 taxis. To achieve the goal, we first divide the Inner Ring of Shanghai City into 47 hotspots according to the pick-up points in peak hours. Then five candidate factors that may affect the customer-search behavior are defined. To identify the quantitative relations between the factors and the customer-search decision, several time-dependent MNL models are developed and calibrated. By using the Watson and Westin pooling test, the time-varying characteristics of the customer-search behavior of a vacant taxi driver are also presented. Some interesting insights and conclusions are found in this paper and summarized as follows.

1) The hotspots with high-density pick-ups could adequately reflect the customer-search decisions of vacant taxi drivers. Specifically, the number of pick-ups in the identified hotspots accounts for around 70% of the total demand in the Inner Ring of Shanghai City. This finding indicates that vacant taxi drivers prefer to circulate airports, railway stations, hospitals, large-scale shopping centers, and other densely populated locations but rarely search for their passengers in remote areas.
2) The regional likelihood of pick-ups (RLP), relative passenger demand (RPD), expected rate of return (EROR), traffic condition of target hotspot (TCTH), and en-route delay (ED) are all significant factors affecting the customer-search behavior of vacant taxi driver. However, the standard error of the ED is quite small, indicating that ED has a minimal effect on the customer-search decision of a vacant taxi driver. Based on the knowledge of experiences, vacant taxi drivers usually search passengers through the routes they are most familiar with.
3) The customer-search behavior of a vacant taxi driver varies with the time of day. The influence factors have a similar but not identical impact on the customer-search behavior of a vacant taxi driver. In particular, the traffic condition has much more effect on the search strategies during the off-peak period. However, during peak periods, the relative passenger demand is the most significant factor. The expected rate of return and en-route delay is, however, the factors vacant taxi drivers greatly emphasize on during the whole day. Finally, taxi drivers tend to be more susceptible to the regional likelihood of pick-ups.

The findings in this paper provide comprehensive insight into significant factors affecting the customer-search behavior of a vacant taxi driver. Against the background of the rapid development of autonomous vehicles, the results will also be useful for facilitating the commercial deployment of self-driving taxis in the future.

In this paper, only GPS records were used to study the customer-search behavior. However, this work could be further extended in further studies by incorporating other data sources, such as field survey data and loop detected data, to cross-validate the conclusions drawn in this study. In addition, due to the data protection regulation and availability, the data used in this manuscript are a little bit old. Future studies may break the limitation if the latest data are available. With the acquisition of these data, it is also interesting to analyze the evolution of the traditional taxi industry and the differences of influencing factors in these years. Furthermore, within the groups of taxi drivers, there is a substantial variation in weights attached to these candidate factors. However, in this study, the heterogeneity within the population is ignored. In future studies, it will be more interesting to investigate the heterogeneous effects of these factors.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

LY: conceptualization, methodology, and writing–original draft. ZS: supervision and methodology. LJ: programming and writing–improvement. CC: data analysis and writing–review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

The work described in this article was supported by National Natural Science Foundation of China (U1811463). Finally, the authors gratefully acknowledge the support from China Scholarship Council (No. 201906060030).

ACKNOWLEDGMENTS

We would like to express our sincere thanks to Shanghai Qiangsheng Taxi Company for providing the raw data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2022.848748/full#supplementary-material

REFERENCES

1. Hensher DA. Some insights into the key influences on trip-chaining activity and public transport use of seniors and the elderly. *Int J Sustain Transp*. (2007) 1:53–68. doi: 10.1080/15568310601097004
2. Su F, Schmöcker JD, Bell MG. Mobility scooters on loan-A scheme complementing the existing special transport services in London. *Int J Sustain Transp*. (2010) 4:95–111. doi: 10.1080/15568310802449392
3. Luo X, Dong L, Dou Y, Zhang N, Ren J, Li Y, et al. Analysis on spatial-temporal features of taxis' emissions from big data informed travel patterns: a case of Shanghai, China. *J Clean Prod*. (2017) 142:926–35. doi: 10.1016/j.jclepro.2016.05.161
4. Tang J, Zhu Y, Huang Y, Peng ZR, Wang Z. Identification and interpretation of spatial-temporal mismatch between taxi demand and supply using global positioning system data. *J Intell Transport Syst*. (2019) 23:403–15. doi: 10.1080/15472450.2018.1518137
5. Yang H, Wong K, Wong S. Modeling urban taxi services in road networks: progress, problem and prospect. *J Adv Transp*. (2001) 35:237–58. doi: 10.1002/atr.5670350305
6. Salanova JM, Estrada M, Aifadopoulou G, Mitsakis E. A review of the modeling of taxi services. *Procedia-SocBehav Sci*. (2011) 20:150–61. doi: 10.1016/j.sbspro.2011.08.020
7. Yang H, Ye M, Tang WH, Wong SC. Regulating taxi services in the presence of congestion externality. *Transp Res Policy Pract*. (2005) 39:17–40. doi: 10.1016/j.tra.2004.05.004
8. Ch JDC, Briones J, et al. A diagrammatic analysis of the market for cruising taxis. *Transp Res Logist Transp Rev*. (2006) 42:498–526. doi: 10.1016/j.tre.2005.05.001
9. Loo BP, Leung BS, Wong S, Yang H. Taxi license premiums in hong kong: can their fluctuations be explained by taxi as a mode of public transport? *Int J Sustain Transp*. (2007) 1:249–66. doi: 10.1080/15568310600737600
10. Gholami A, Mohaymany AS. Analogy of fixed route shared taxi (taxi khattee) and bus services under various demand density and economical conditions. *J Adv Transp*. (2012) 46:177–87. doi: 10.1002/atr.157
11. Sirisoma R, Wong S, Lam WH, Wang D, Yang H, Zhang P. Empirical evidence for taxi customer-search model. *Proc Inst Civ Eng Transp*. (2010) 163:203–10. doi: 10.1680/tran.2010.163.4.203
12. Szeto WY, Wong RCP, Wong SC, Yang H. A time-dependent logit-based taxi customer-search model. *Int J Urban Sci*. (2013) 17:184–98. doi: 10.1080/12265934.2013.776292
13. Wong R, Szeto W, Wong S, Yang H. Modelling multi-period customer-searching behaviour of taxi drivers. *Transportmetr Transp Dyn*. (2014) 2:40–59. doi: 10.1080/21680566.2013.869187
14. Wong R, Szeto W, Wong S. A cell-based logit-opportunity taxi customer-search model. *Transp Res Emerg Technol*. (2014) 48:84–96. doi: 10.1016/j.trc.2014.08.010
15. Yang H, Wong SC, Wong KI. Demand-supply equilibrium of taxi services in a network under competition and regulation. *Transp Res Methodol*. (2002) 36:799–819. doi: 10.1016/S0191-2615(01)00031-5
16. Kim H, Oh JS, Jayakrishnan R. Effect of taxi information system on efficiency and quality of taxi services. *Transp Res Record*. (2005) 1903:96–104. doi: 10.1177/0361198105190300111
17. Wong K, Wong S, Yang H, Wu J. Modeling urban taxi services with multiple user classes and vehicle modes. *Transp Res Methodol*. (2008) 42:985–1007. doi: 10.1016/j.trb.2008.03.004
18. Hu X, Gao S, Chiu YC, Lin DY. Modeling routing behavior for vacant taxicabs in urban traffic networks. *Transp Res Record*. (2012) 2284:81–8. doi: 10.3141/2284-10
19. Li B, Szeto W. Taxi service area design: formulation and analysis. *Transp Res Logist Transp Rev*. (2019) 125:308–33. doi: 10.1016/j.tre.2019.03.004
20. Liu L, Andris C, Ratti C. Uncovering cabdrivers' behavior patterns from their digital traces. *Comput Environ Urban Syst*. (2010) 34:541–8. doi: 10.1016/j.compenvurbsys.2010.07.004
21. Wong R, Szeto W, Wong S. A two-stage approach to modeling vacant taxi movements. *Transp Res Procedia*. (2015) 7:254–75. doi: 10.1016/j.trpro.2015.06.014
22. Wong RC, Szeto W, Wong S. Behavior of taxi customers in hailing vacant taxis: a nested logit model for policy analysis. *J Adv Transp*. (2015) 49:867–83. doi: 10.1002/atr.1307
23. Zong F, Wu T, Jia H. Taxi drivers' cruising patterns-insights from taxi GPS traces. *IEEE Trans Intell Transp Syst*. (2018) 20:571–82. doi: 10.1109/TITS.2018.2816938
24. Veloso M, Phithakkitnukoon S, Bento C. Urban mobility study using taxi traces. In: *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*. Beijing: ACM (2011). p. 23–30.
25. Zhang D, Sun L, Li B, Chen C, Pan G, Li S, et al. Understanding taxi service strategies from taxi GPS traces. *IEEE Trans Intell Transp Syst*. (2014) 16:123–35. doi: 10.1109/TITS.2014.2328231
26. Qin G, Li T, Yu B, Wang Y, Huang Z, Sun J. Mining factors affecting taxi drivers' incomes using GPS trajectories. *Transp Res Emerg Technol*. (2017) 79:103–18. doi: 10.1016/j.trc.2017.03.013
27. Watson PL, Westin RB. Transferability of disaggregate mode choice models. *Reg Sci Urban Econ*. (1975) 5:227–49. doi: 10.1016/0166-0462(75)90005-8
28. Ao Y, Zhang Y, Wang Y, Chen Y, Yang L. Influences of rural built environment on travel mode choice of rural residents: the case of rural Sichuan. *J Trans Geogr*. (2020) 85:102708. doi: 10.1016/j.jtrangeo.2020.102708
29. Yang L, Liu J, Liang Y, Lu Y, Yang H. Spatially varying effects of street greenery on walking time of older adults. *ISPRS Int J Geo Inf*. (2021) 10:596. doi: 10.3390/ijgi10090596
30. Yang L, Ao Y, Ke J, Lu Y, Liang Y. To walk or not to walk? Examining non-linear effects of streetscape greenery on walking propensity of older adults. *J Trans Geogr*. (2021) 94:103099. doi: 10.1016/j.jtrangeo.2021.103099
31. Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Rec*. (1999) 28:49–60. doi: 10.1145/304181.304187
32. Benmahdi D, Rasolofondraibe L, Chiementin X, Murer S, Felkaoui A. RT-OPTICS: real-time classification based on OPTICS method to monitor bearings faults. *J Intell Manuf*. (2019) 30:2157–70. doi: 10.1007/s10845-017-1375-6
33. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*. (2013) 36:27–46. doi: 10.1111/j.1600-0587.2012.07348.x
34. Myers RH, Myers RH. *Classical and Modern Regression With Applications, Vol. 2*. Belmont, CA: Duxbury Press (1990). p. 488.
35. Loo BP, Wong S, Hau TD. Introducing alternative fuel vehicles in Hong Kong: views from the public light bus industry. *Transportation*. (2006) 33:605–19. doi: 10.1007/s11116-006-7947-5
36. Wong S, Wong C, Sze NN. Attitudes of public light bus drivers to penalties to combat red light violations in Hong Kong. *Transp Policy*. (2008) 15:43–54. doi: 10.1016/j.tranpol.2007.10.009
37. Paral P, Chatterjee A, Rakshit A. OPTICS-based template matching for vision sensor-based shoe detection in human-robot coexisting environments. *IEEE Trans Instrum Meas*. (2019) 68:4276–84. doi: 10.1109/TIM.2018.2890400
38. Qu M, Zhu H, Liu J, Liu G, Xiong H. A cost-effective recommender system for taxi drivers. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: ACM (2014). p. 45–54.
39. Yu B, Wu S, Yao B, Yang Z, Sun J. Dynamic vehicle dispatching at a transfer station in public transportation system. *J Transp Eng*. (2012) 138:191–201.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.