A CELL-BASED LOGIT-OPPORTUNITY TAXI CUSTOMER-SEARCH MODEL

R.C.P. WONG, W.Y. SZETO[*] and S.C. WONG

Department of Civil Engineering, The University of Hong Kong, Hong Kong
[*]E-mail: ceszeto@hku.hk
Phone: (852) 28578552
Fax: (852) 25595337

ABSTRACT

This paper proposes a cell-based model to predict local customer-search movements of vacant taxi drivers, which incorporates the modeling principles of the logit-based search model and the intervening opportunity model. The local customer-search movements were extracted from the global positioning system data of 460 Hong Kong urban taxis and inputted into a cell-based taxi operating network to calibrate the model and validate the modeling concepts. The model results reveal that the taxi drivers' local search decisions are significantly affected by the (cumulative) probability of successfully picking up a customer along the search route, and that the drivers do not search their customers under the random walk principle. The proposed model helps predict the effects of the implementation of the policies in adjusting the taxi fleet size and the changes in passenger demand on the customer-search distance and time of taxi drivers.

Keywords: Logit-opportunity Model; Probability of Success; Taxi Customer-search; Cell-based Network; Global Positioning System Data.

1.   INTRODUCTION

Taxis are an important transportation mode but they tend to circulate around a city to search for customers. The circulation activity occupies spare road spaces and worsens traffic congestion and air quality. Hence, advanced technologies and taxi regulation policies are established to tackle these problems. Studies which relate to advanced technologies (e.g., Horn, 2002; Lee and Cheng, 2008; Conway et al., 2012; da Costa and de Neufville, 2012; Miwa et al., 2013; Jung et al., 2014) aim to increase the customer satisfaction or improve overall system efficiency from the perspective of taxi operators. Studies which provide insights on taxi regulatory policies (Douglas, 1972; De Vany, 1975; Manski and Wright, 1976; Beesley and Glaister, 1983; Schroeter, 1983; Arnott, 1996; Cairns and Liston-Heyes, 1996; Xu et al., 1999; Yang et al., 2000, 2002, 2005a, 2010a; Flores-Guri, 2003; Fernández et al., 2006; Moore and Balaker, 2006; Loo et al., 2007; Wong et al., 2014c) often focus on examining the consequences on the implementation of such policies. However, these studies have been developed based on an idealized market of conventional economic analysis in which the spatial structure of the market was ignored.

In an attempt to capture the spatial structure of the market, Yang and Wong (1998) developed a model to determine the taxi movements on a given road network. In this model, the customer-search behavior of taxi services is based on the assumption that each vacant taxi driver attempts to minimize his or her expected search time to find a customer. Later, the model was further improved by Wong and Yang (1998), Wong et al. (2001, 2002, 2005,

2008), Yang et al. (2001, 2005b, 2008, 2012a), Yang and Yang (2011), Kim et al. (2005), and Hu et al. (2012) to capture congestion effects, multiple user classes, multiple taxi modes, customer hierarchical modal choice, taxi search behavior of taxi customers, day-to-day learning processes, stochastic travel time, and search frictions between vacant taxis and taxi customers. Yet, the assumption has ignored the case that vacant taxi drivers travel to remote areas to pick up customers because of the high profit return.

To take into account the profitability consideration, Wong et al. (2003) developed a taxi network model. They explicitly examined the effects of perceived profitability on the customer-search behavior of vacant taxis and the expected profit that taxi drivers could earn by picking up customers in particular zones. Yang et al. (2010b) further extended this concept to "profit per unit time" by incorporating the operational cost and time consumption involved in taxi trips to maximize profits from customer-search. However, the search behavior models in the preceding taxi network models are in logit form and have not been calibrated and validated.

To provide an empirical evidence to validate the search behavior models, Sirisoma et al. (2010) determined the significant factors affecting vacant taxi drivers' customer-search decisions based on the data collected from a stated preference survey. Wong et al. (2014a) calibrated and validated the logit models based on the global positioning system (GPS) data obtained from 460 urban taxis to predict the drivers' strategic zonal choice for searching for customers during both peak and off-peak periods. The zoning system follows the administrative district system with 18 zones, in which the zone sizes vary from 7 km^2 to 175 km^2 with an average value of 60 km^2. Szeto et al. (2013) further extended the consideration to every hour in a day. The preceding behavior models developed and calibrated are in logit form, which may not be the best form for depicting the search behavior. Moreover, these behavior models are zone-based. They ignore the fact that vacant taxi drivers can meet a customer on the way to their destination zones.

Wong et al. (2014b) considered sequential customer-search decisions of vacant taxi drivers on finding customers at intermediate zones while heading to their designated zones. Yet, the local search behavior of vacant taxi drivers within a zone has not been captured. It is not possible to study how the local search may worsen the congestion and air pollution problems within a zone by using their model. Numerous studies have been conducted to attempt modeling the route choices of drivers based on taxi GPS data (e.g., Ziebart et al., 2008; Jiang et al., 2009; Li et al., 2009; Morikawa and Miwa, 2010). They can be used to model the local movements of drivers within a zone. However, most of these efforts have been put on modeling the route choices of occupied taxi drivers with a known destination and these studies did not model vacant taxi drivers' route choice behavior for finding the next customers. Furthermore, the behavior model of Wong et al. (2014b) ignores the fact that as vacant taxi drivers search for customers, the probability of successfully meeting a taxi customer along the way increases. In addition, the probability of a vacant taxi driver meeting a taxi customer decreases if there are many vacant taxis nearby. This important fact has not yet been considered.

To address the preceding issues, this study proposes a local customer-search model combining both the modeling principles of the logit-based search model and intervening opportunity model. The logit-based modeling concept is used to handle multi-directional, discrete search choices for vacant taxi drivers and the intervening opportunity model is

reformulated and incorporated into the logit-based search model to capture the cumulative probability of successfully meeting a customer in each direction of customer-search. The proposed model assumes that vacant taxi drivers do not have clear destinations in customer-search after they drop off their preceding customers, but they travel towards an area with a high probability of meeting a customer. Similar to other studies (e.g., Ben-Akiva et al., 2012; Feng and Timmermans, 2013; Sun and Ban, 2013; Zhan et al., 2013), this study uses GPS data to develop and/or calibrate the proposed model. This study also adopts the search movement information extracted from the GPS data of 460 Hong Kong urban taxis in a week in order to establish a cell-based urban taxi operating network, which consists of thousands of cells with bi-directional links connected between each pair of adjacent cells. The cells are square-shaped and their sizes are identical. The length of each side of a square cell is set to be 200 m. Each taxi driver may decide to either circulate within his/her current cell or travel towards an adjacent connected cell as his/her local customer-search movement. A calibration process is also proposed in this paper by formulating the calibration problem as a bi-level program. A likelihood ratio test is conducted in this study to show that the proposed model is more reasonable, informative, and significantly different from a random walk model. A sensitivity analysis is also accomplished to illustrate the changes in customer-search distance and time with respect to the variation in the (cumulative) probability of meeting a customer. The sensitivity analysis results illustrate that the proposed model can predict the effects of the implementation of the policies in adjusting the taxi fleet size and the changes in passenger demand on the customer-search distance and time of vacant taxi drivers.

The contributions of this paper include the following:
- Introducing the concept of the cumulative probability of success of being hired and demonstrating that it is a significant factor that influences the local choices of vacant taxi drivers to find customers;
- Proposing a logit-opportunity modeling approach to modeling taxi customer-search behavior in a cell-based network; the logit-opportunity approach combines the modeling concepts of the logit-based search model and the intervening opportunity model to explain the local customer-search behavior of vacant taxi drivers;
- Proposing a calibration procedure for the model;
- Providing model results to illustrate the model properties and verify the proposed model is more reasonable, informative, and significantly different from a random walk model, and
- Helping to identify the local customer-search decision and the trajectories of vacant taxis, investigate the vacant taxi circulating behavior, and explain the effects of the taxi fleet size and passenger demand on the customer-search distance and time.

This paper comprises this Introduction and five further sections: Section 2 gives an introduction to the traditional intervening opportunity model and discusses the limitations of using this model directly in this study; Section 3 describes the formulation of our proposed logit-opportunity taxi customer-search model and proposed model evaluation procedure; Section 4 provides the empirical evaluation to validate the modeling concepts; Section 5 discusses the model results, and Section 6 summarizes and concludes the paper.

2. REVIEWS ON THE TRADITIONAL INTERVENING OPPORTUNITY MODEL

This paper proposes an enhanced modeling approach combining the modeling principles of the logit-based search model and the intervening opportunity model to depict the local customer-search behavior of vacant taxi drivers.

The formulation of opportunity-type model was originally introduced by Stouffer (1940). The main idea of this model is that trip making is not explicitly related to distance (or time) but to the relative accessibility of opportunities (e.g., car parking spaces) for satisfying the objective of the trip (e.g., finding an empty car parking space for a driver). The theory was further enhanced by Schneider (1959) by adding two hypotheses about human behavior: (1) The total travel time from an origin to a destination is minimized; and (2) the probability of choosing any potential destination is a constant and independent of the order of visit. Therefore, the model assumes that a traveler makes his/her trip as short as possible, and the trip is only lengthened if the traveler fails to find an acceptable destination at a lesser distance (Heanue and Pyers, 1966). This assumption implies that travelers consider their nearest destinations to their origins, and if this trip decision is unacceptable, they consider their next nearest destinations and so on. The proposed theory of intervening opportunity by Schneider (1959) implies that trip makers select destinations sequentially starting from the one closest to their origins. They initially create a hierarchy of the choice alternatives by ranking them in terms of increasing spatial separation from their origins, proceed step-by-step by following the hierarchical ranks, and make their decisions based on the number of opportunities given at each destination and the probability of the objective being satisfied.

We find that the behavioral assumptions of the intervening opportunity model have analogous characteristics to taxi customer-search behavior. The model assumes that the trip makers (vacant taxi drivers) have not decided a fixed destination when they are at the origin, but move forwards to potential destinations based on the number of opportunities (number of taxi customers) available in these destinations in fulfilling their trip objectives (finding a customer). The taxi customer-search behavior also matches with the first hypotheses of the intervening opportunity model proposed by Schneider (1959) who predicted the trip distributions with an assumption of time minimization. This is consistent with the fundamental concept of taxi search model proposed by Yang and Wong (1998) with the consideration of minimizing the customer-search time of vacant taxi drivers.

Yet, the second hypothesis of the intervening opportunity model is that the probability of selecting any potential destination is equal to a constant. This hypothesis has over-simplified the real situation of taxi markets as it has ignored the variation in the probability of successfully meeting a customer in different locations (i.e., main streets and inner roads). The probability of meeting a customer depends on the associated passenger demand and the availability of vacant taxis which vary in different areas. Moreover, the traditional intervening opportunity model has only formulated the problem as a one-dimensional search which cannot reflect the reality of the multi-directional customer-search of vacant taxi drivers in the road network. Therefore, this model cannot be directly applied to depict the local customer-search behavior of vacant taxi drivers in this study. To address these problems, the intervening opportunity model is reformulated and incorporated into a logit-based search model, which is depicted in the following section.

3.   CELL-BASED LOGIT-OPPORTUNITY SEARCH MODEL

3.1   Cell-based Taxi Operating Road Network

This paper proposes modeling the local customer-search behavior of vacant taxi drivers using a cell-based approach. To facilitate the purpose, we divided the study area into many identical squares to form a cell-based network for that area. Figure 1 shows an example of the cell-based network over a typical urban area, and the length of each square equals 200 m. The size of each cell can be adjusted subject to the required level of modeling accuracy. There are imaginary bi-directional links connected between each pair of adjacent cells, representing taxi movements between these cells. Each taxi driver may decide to either circulate within his/her current cell or travel towards an adjacent connected cell. Since the size of each cell is identical, the distance traveled and the time consumed in customer-search are assumed to be independent to the search direction but dependent on the number of cells traveled in the vacant taxi journey.
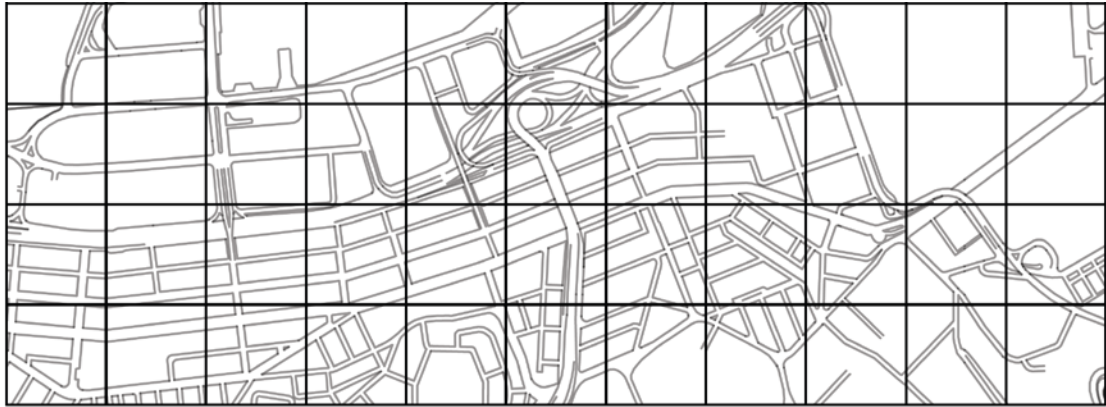

Figure 1. Example of the Cell-based Network over a Typical Urban Area

3.2   Local Customer-search Options in the Cell-based Network

When a vacant taxi driver is searching for his/her next customer within the cell-based network, he/she has two categories of options in each cell:

Search option (1):   Searching for a customer in the current cell i, and

Search option (2):   Traveling in one of at most four cardinal directions towards one of the adjacent cells j next to the current cell.

Based on these two categories of search options, a vacant taxi driver in each cell has five search options at most, as illustrated in Figure 2 below. All these options are assumed to be independent.
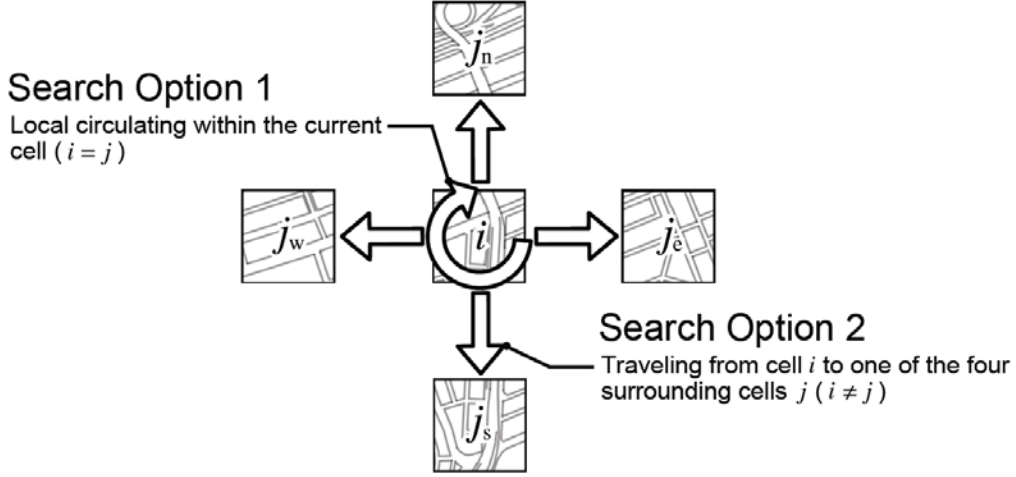
Figure 2. Local Customer-search Options for a Vacant taxi Driver in the Cell-based Network

A vacant taxi-driver may not meet a customer in the cell selected once he/she reaches that cell. If so, he/she makes another search decision in that cell based on the available search options at that cell. The search decision made in that cell is assumed to be independent of the decision made previously. This process is repeated until he/she meets a customer. This process implies that the driver circulates in the cell-based network cell by cell to search for customers and the driver is assumed to sequentially make a series of independent decisions to select the subsequent cells during his/her cruising within the cell-based network, with one decision made in each cell.

Experienced taxi drivers indeed know an approximately average distance and time they require to travel to meet the next customers. For this reason, they do not simply make their local customer-search decisions depending on the attractiveness of the adjacent cells, but they do also consider the subsequent possible choice of cells that can be reached later. Based on this reason, we introduce the following definitions, which will be used to develop the search choice model.

3.3   Variable Definitions

Probability of Success: We believe that vacant taxi drivers search the areas more with higher possibilities to meet the next customers. Therefore, we adopt the probability of success as an explanatory variable to depict their local search behavior. The probability of success (s_j) is the probability of a vacant taxi driver being presented to successfully pick up a customer in one cell. It is defined as the passenger demand generated from a cell over the availability of vacant taxis in that cell, subject to the condition that the probability is between zero and one inclusively. This can be expressed mathematically as

$$s_j = \min\left[\frac{O_j}{A_j}, 1\right],\tag{1}$$

where O_j denotes the number of recorded occupied trips starting in cell j and A_j denotes the number of recorded vacant taxis found in cell j. This probability differs from cell to cell and hence it can capture the variation of the probability of successfully meeting a customer in different cells.

6

Number of Search Decisions Anticipated: The number of search decisions anticipated (L) by vacant taxi drivers is defined as the average number of cells traveled before picking up a customer, because it is assumed that each driver makes exactly one search decision in each cell. The number L is interpreted as the long-term average for an individual driver in his/her mindset based on his/her past experience. If the number is small, then a vacant taxi driver expects to meet his/her next customer in a short distance (due to high passenger demand during peak hours), and the expected search area covered by this driver is small. This means that the area is defined by L. The boundary of this area is called the customer-search boundary. The customer-search boundaries for different number of search decisions anticipated are shown in Figure 3.
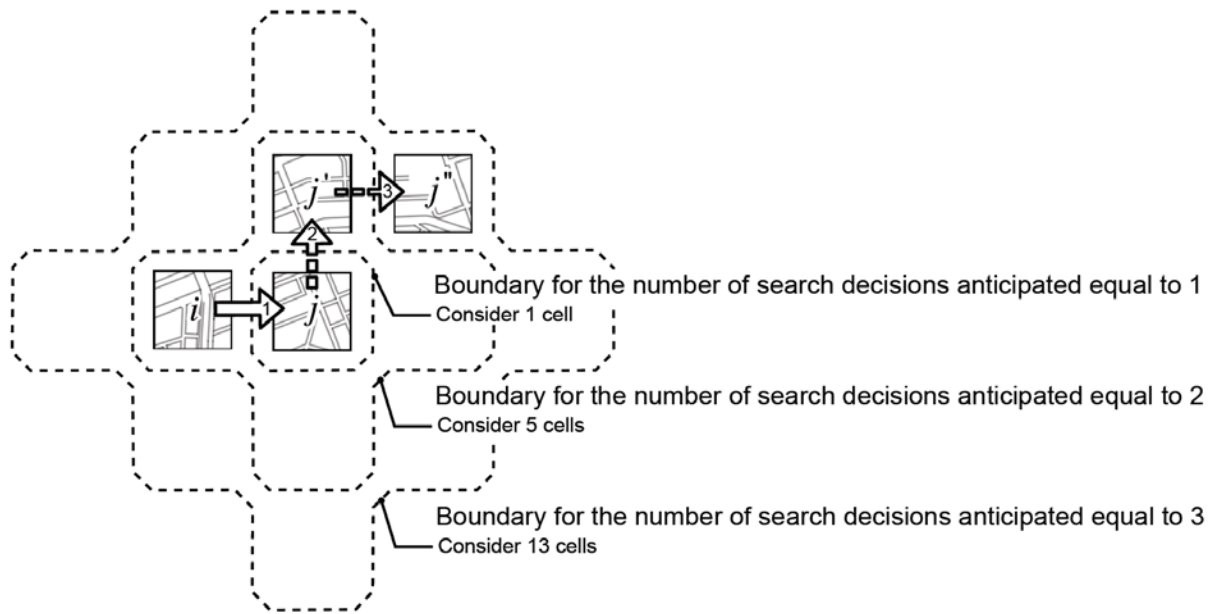

Figure 3. Customer-search Boundaries for Different Numbers of Search Decisions Anticipated

The numbers shown inside the arrows form the sample (planned) decision sequence and the broken line arrows denote the sample (planned) decisions after reaching the next cells. For example, if a vacant taxi driver starts in cell i and decides to travel eastbound to cell j for customer-search. When $L = 1$, it means that the driver is expected to search for and successfully meet a customer in that cell only with a probability s_j. When $L = 2$, it means that the driver is ready to make another local search decisions again right after he/she arrived at cell j. Hence, the expected search area is formed by five cells, including the cells that can be reached by a vacant taxi driver in the next move. The probability of successfully meeting a customer in this search area is not only simply due to the probability of success of cell j only, but also due to those of the adjacent cells of cell j. Therefore, we propose a concept of the cumulative probability of success for a search decision, which is a function of both the probability of success and the number of search decisions anticipated including the initial decision made, and aggregates the probabilities of success for vacant taxi drivers reaching more than one cell.

Cumulative Probability of Success: The cumulative probability of success (S_j^l) represents the accumulated probability of a vacant taxi driver successfully picking up a customer, if the driver initially selects cell j, and prepares to make a local search decision l times in total to meet a customer. The cumulative probability equals the probability of successfully picking up a customer in cell j plus the additional probability gained from the subsequent cells. This can be formulated based on both the logit and intervening opportunity concepts, and expressed as

$$S_j^l = \begin{cases} s_j & \text{, if } l = 1, \\ s_j + \left(1 - s_j\right) \sum_{m \in K_j} \left[p^{l-1}(m \mid j) S_m^{l-1} \right] & \text{, if } 1 < l \le L, \end{cases} \tag{2}$$

where K_j denotes the set of cells that vacant taxi drivers in cell j can select. The parameter L is the number of search decisions anticipated by the vacant taxi drivers and its value should be calibrated. $p^{l-1}(m \mid j)$ is the probability that an individual taxi driver in cell j selects subsequent cell m in the $(L - (l-1) + 1)$-th *planned decision* to search for the next customer. The probability can be expressed in the following multinomial logit form (McFadden, 1974):

$$p^{l-1}(m \mid j) = \frac{\exp\left(\alpha^L S_m^{l-1}\right)}{\sum_{n \in K_j} \exp\left(\alpha^L S_n^{l-1}\right)}, \tag{3}$$

where α^L denotes the non-negative coefficient associated with the cumulative probability of success, which depends on the number of search decisions anticipated L by the vacant taxi drivers.

In fact, the second and subsequent decisions are only the initial (or conceptual) plan of the taxi driver made at his/her current position. He/she may not necessarily follow the plan and confine his/her remaining search decisions to the subsequent cells in the plan. The driver re-evaluates his/her plan after reaching the next cell and his/her decision made in each cell is independent of the decision made in other cells. *In this sense, the local search decisions made in all cells are independent and are made sequentially.*

For the case of $L = 1$, not more than five cells (including the current cell and the cells next to it in four cardinal directions) are considered by the driver in cell i in his/her local customer-search plan and the associated probability of success for cell j is simply equal to the probability of being hired at that single cell given in Equation (1).

For the case of $L > 1$, when the number of search decisions anticipated is more than one, vacant taxi drivers do not make their local search decisions only based on the probabilities of success for the five cells, the drivers also consider the additional probabilities of success gained from the subsequent cells after leaving the five cells. At the lower right side of recursive Equation (2), the first term s_j represents the probability of a vacant taxi driver successfully picking up a customer in cell j. The probability $\left(1 - s_j\right)$ equals the probability of the vacant taxi driver failing to meet a customer in cell j and leave the cell. The function $\sum_{m \in K_j} \left[p^{l-1}(m \mid j) S_m^{l-1} \right]$ represents the cumulative probability of success of the driver after

8

leaving cell j. Therefore, the second term $\left(1 - s_j\right) \sum\limits_{m \in K_j} \left[p^{l-1}\left(m \mid j\right) S_m^{l-1} \right]$ represents the additional probability of success gained from the subsequent cells.

3.4 Logit-based Search Model

Vacant taxi drivers in a cell select one of their adjacent cells for customer-search according to the cumulative probabilities of success associated with each adjacent cell. These probabilities depend on the numbers of the search decisions anticipated (L) and are inputted into a multinomial logit discrete choice model expressed as:

$$P_q^L\left(j \mid i\right) = \frac{\exp\left(\alpha^L S_j^L\right)}{\sum\limits_{m \in K_i} \exp\left(\alpha^L S_m^L\right)}, \tag{4}$$

where $P_q^L\left(j \mid i\right)$ denotes the probability of an individual vacant taxi driver who is currently in cell i, makes a local movement q from cell i to cell j to search for the next customer, and prepares to make the local search decisions L times in total during his/her customer-search. Every taxi driver in cell i is assumed to have the same perception concerning the cumulative probability of success associated with the corresponding adjacent cell. Therefore, for simplicity, the subscript q is omitted in the attribute. The modeling form is identical to the selection probability of vacant taxi drivers in subsequent cells as indicated in Equation (3) and has the same calibrated coefficient (α^L) for each cell to maintain a consistent search decision everywhere in the network. According to Equation (4), a higher cumulative probability of success for cell j implies a higher probability of selecting that cell. When $\alpha^L = 0$, the proposed model becomes an equally-likely model. Equation (4) captures multi-dimensional, discrete search choices for vacant taxi drivers by using the logit model form and captures the concept of intervening opportunities by using the cumulative probability of success as inputs to the model.

3.5 Model Calibration

The coefficient α^L can be calibrated by solving the following optimization problem for each value of L:

$$\text{Maximize LL}^L = \sum_i \sum_j \ln P_q^L\left(j \mid i\right), \tag{5}$$

where LL^L is the log-likelihood of the estimated model for a given number of search decisions anticipated L. The optimization problem is formed using the maximum log-likelihood principle.

The optimal number of search decisions anticipated ($\hat{L}$) is determined based on the Bayesian information criterion (BIC) evaluation. The function with the lowest value of BIC is the most preferable as it has the best fit to the taxi trip data or involves the fewest explanatory variables, or both. Therefore, the objective (or the optimization problem) can be formulated into

$$\text{Minimize BIC} \quad = -\frac{2LL^L}{Q} + \frac{b \ln Q}{Q}, \tag{6}$$

where b is the number of parameters in the model and Q is the total number of local search movements between cells made by the vacant taxi drivers considered. By finding the minimum BIC value in Problem (6), we can obtain the optimal number of search decisions anticipated ($\hat{L}$).

3.6 Likelihood Ratio Test

In the previous study, Jiang et al. (2009) assumed that the taxi movements within the local network are following the random walk principle (i.e., without a definite pattern). To verify whether our proposed local customer-search model is more reasonable, informative, and significantly different from a random walk model, a likelihood ratio test (Ortúzar and Willumsen, 1999) is therefore conducted to compare the calibrated search model and an equally-likely model (i.e., a random walk model), which is based on the log likelihood ratio (LR):

$$LR = -2 \left[LL(0) - LL(\alpha^{\hat{L}}) \right], \tag{7}$$

where $LL(0)$ and $LL(\alpha^{\hat{L}})$ are the log-likelihoods for the equally-likely model and for our model with the calibrated coefficient at the optimal number of search decisions anticipated, respectively. The null hypothesis that there is no intervention between the two models is rejected, as the test statistic exceeds the threshold value that is specified for the chi-squared distribution at the chosen level of significance. The degree of freedom equals the number of variables in the calibrated model.

3.7 Model Performance Assessment

We propose a procedure for evaluating our developed logit-based local customer-search model to confirm the model accuracy in representing the local search decisions of vacant taxi drivers. The assessment is undertaken by comparing the observed average customer-search distance and time to the corresponding predicted values. The predicted distance and time (D^c and T^c) can be estimated by multiplying the mean values ($\bar{D}^c$ and $\bar{T}^c$) deduced from the GPS data in one cell by the expected (average) number of cells traveled ($\bar{C}$) to meet the next customer:

$$D^c = \bar{D}^c \bar{C} \text{ and } T^c = \bar{T}^c \bar{C} \tag{8}$$

The latter term $\bar{C}$ in Equation (8) is estimated by the sum of the product of the proportion of trips selecting cell j as the next cell, $\dfrac{Q_j}{Q}$, and the corresponding expected number of cells traveled to reach a customer after reaching that cell, $\left[s_j + \sum\limits_{l=1}^{\infty} (l+1)\left(\hat{S}_j^{l+1} - \hat{S}_j^l \right) \right]$:

10

$$\bar{C} = \sum_{j} \frac{Q_j}{Q} \left[s_j + \sum_{l=1}^{\infty} (l+1) \left(\hat{S}_j^{l+1} - \hat{S}_j^l \right) \right], \qquad (9)$$

where Q_j denotes the predicted number of local movements choosing cell j. It is calculated by summing up the products of the number of local movements started in cell m (H_m) times the probability of selecting cell j shown in Equation (4):

$$Q_j = \sum_{m \in K_j'} H_m P_q^{\hat{L}} (j \mid m), \qquad (10)$$

where K_j' denotes the set of all cells in which the drivers in those cells can select cell j as the next cell to visit.

The average number of cells traveled in Equation (9) should be based on the *predicted decisions* made by taxi drivers rather than the plans in their minds, because the drivers may change their mind once reaching the next cell due to having more accurate information for the next cell or those cells nearby. That is why $p^{l-1}(m \mid j)$ is no longer applicable for this case. Therefore, $p^{l-1}(m \mid j)$ in Equation (2) should be replaced by $P_q^{\hat{L}}(m \mid j)$ to compute the cumulative probability of success ($\hat{S}_j^l$) based on the predicted search decisions.

4. EMPIRICAL EVALUATION USING GPS DATA

4.1 Data Collection

In 2009, there were 18,138 taxis in Hong Kong, of which 15,250 were urban taxis, 2,838 were New Territories taxis and 50 were Lantau taxis. In this study, the survey data were collected from GPS devices previously installed in 460 urban taxis that tracked the drivers' daily activities. By using satellite communication, the database recorded the taxis' location in terms of longitude and latitude, travel speeds, and occupational statuses at 30-second intervals. The recorded activities included waiting at taxi stands for customers, circulating on roads for searching for customers, responding to call requests, waiting at call request locations, driving customers to their destinations, stopping for breaks, and traveling to get fuel up. However, the available information in the dataset is not enough to distinguish these activities. Although there was no available information on the collected fares, the location and speed information were used to estimate the collected fares by deducing the vehicle trajectory and waiting time. Moreover, because HK$5 is charged for telephone bookings and it is not difficult for a passenger to catch a vacant taxi along the urban roadside in Hong Kong, telephone bookings are uncommon. We consider such bookings as having a negligible to calibrate our developed customer-search model. The sample of 460 urban taxis represents approximately 3% of the entire Hong Kong urban taxi population which offer an adequate sample size to deduce the travel behavior of the Hong Kong urban taxi drivers. Hence, we believe that the dataset has provided sufficient information to track the urban taxis' daily operations in general to develop our model.

The information for vacant taxi operations was extracted from the GPS data. It is necessary to simplify the huge amount of detailed information in order to capture the main trends. In this study, the Hong Kong territory was divided into thousands of cells as mentioned in Section

3.1 with each pair of cells connected by a bi-directional link representing taxi movements between these cells. The vacant taxi trajectories extracted from the GPS data disclosed the cells and links involved during a customer search process. In the case if these vacant trips were operating at high travel speeds and thus bypassed the adjacent cells, these cells were estimated by the interpolation method to form a complete search trajectory in the cell-based network. Cells and links with no recorded taxi activity were considered as invalid and were excluded.

The GPS survey data was collected in the week staring from 16-23 August 2009 inclusively. The whole week trip data was extracted to develop a cell-based urban taxi operating road network, comprising 9,420 valid cells (with the dimension 200 m x 200 m) and 36,185 valid links, in which 2-hour (07:30 to 09:30) data was used for model calibration. This data was used to review the local taxi customer-search behavior during the morning peak period.

It is noted that some of the extracted occupied and vacant trips were actually non-existence, such as the trips with exceptionally long or short travel distances and times (i.e., occupied trips traveled less than 1 min, or vacant trips traveled for more than 30 km). The recording of these extracted trips could be due to the GPS device malfunctions, poor connectivity to satellites in the urban areas surrounded by high-raised buildings, or human error by taxi drivers who operated the devices. Therefore, a data screening process was carried out to eliminate such trips. However, some non-search-related unoccupied trips, such as the trips involving stopping at roadsides for a short break or trips to gas stations, could not be eliminated. Nevertheless, based on our knowledge obtained from taxi drivers, the number of these trips is relatively small. Hence, we considered these trips would have a negligible impact on verifying the accuracy of our customer-search model. After the data screening process, about 22,000 taxi trips and 157,000 local search movements were captured in the morning peak period.

Survey data indicates that the sampled urban taxis were mostly concentrated in the urban areas of Hong Kong Island and Kowloon Peninsula in customer-search. This could happen because the urban taxis charge a higher fare and are less competitive to operate in rural areas than New Territories and Lantau taxis (Wong et al., 2014a). Therefore, the developed cell-based urban taxi operating road network mainly covers the urban areas.
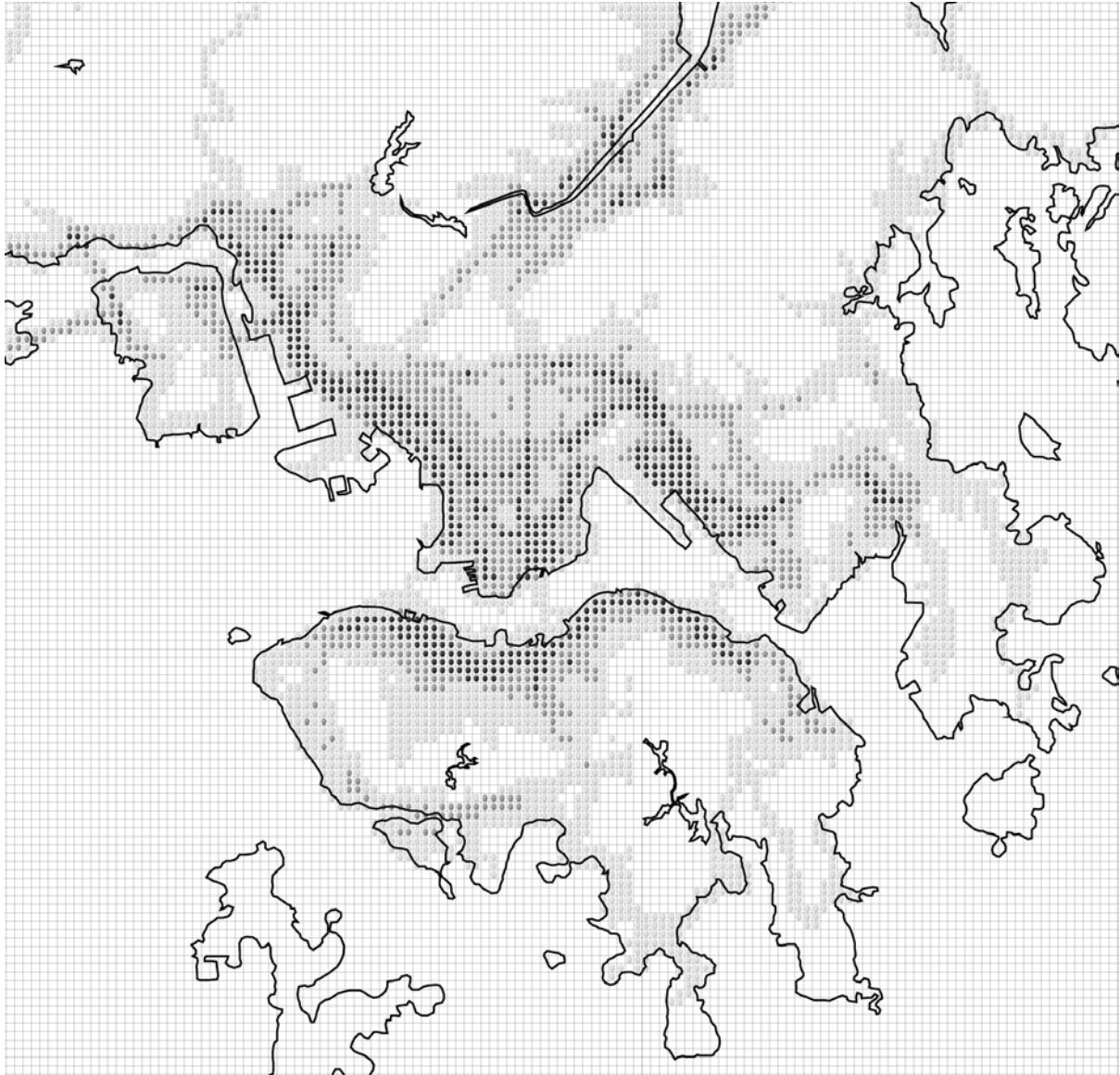
Figure 4. Part of the Cell-based Taxi Operating Road Network in the Urban Areas

Figure 4 shows a part of the cell-based urban taxi operating road network. The shaded cells represent the valid cells in the developed network. The cells without taxi service activities are left blank and were excluded from further analysis. Darker dots in the cells represent more trips involved in those cells. Based on the darkness of cell distribution indicated in this figure, it is noticed that most of the urban taxi drivers operated within city centers along the northern and southern shores of Hong Kong Island and Kowloon Peninsula, respectively. Those valid cells outside the shorelines represent the taxi trips traveling across the harbors or channels by tunnels or bridges.

4.2　Distribution of the Probability of Success

Figure 5 plots the cumulative percentage of valid cells against the probability of success associated with a cell in the network. The figure shows that 80% of the valid cells have a zero probability of success to meet customers, meaning that those cells had no recorded passenger demand during the study period. Majority of those cells (95%) have the probability of

success less than 0.18, while the mean value and the standard deviation of the probability of success for the remaining 20% of valid cells are 0.14 and 0.15, respectively. It is also noticed that only 26 cells (less than 0.5%) have a probability of success of one in customer-search. These particular cells could contain the popular taxi stands or pick-up/drop-off points with adequate passenger demands where the drivers probably stayed at the cells until they met a passenger.
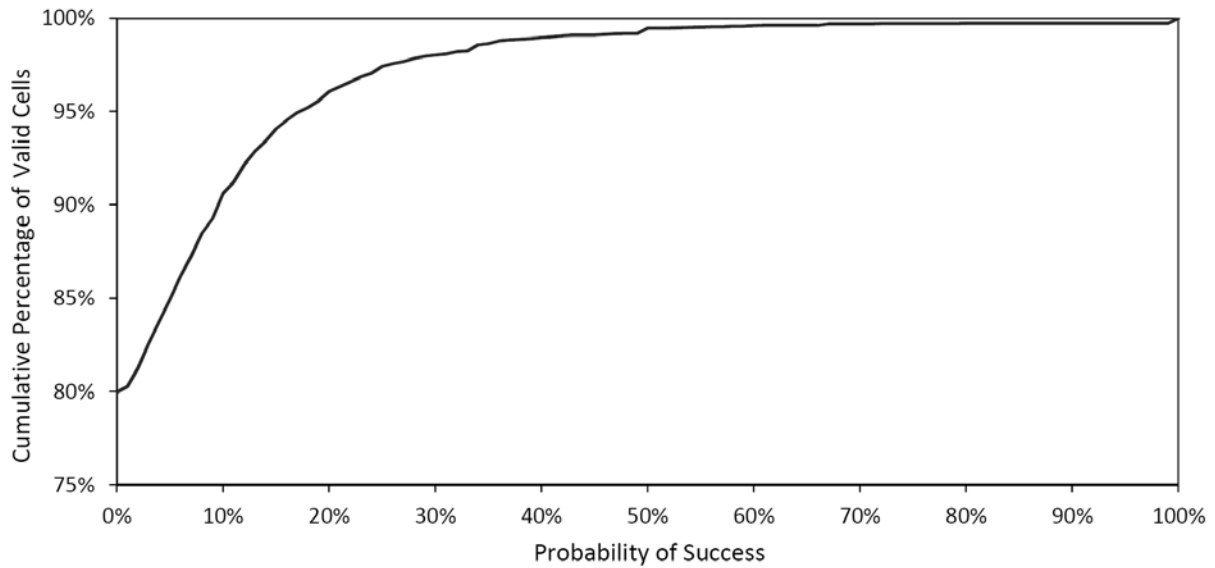

Figure 5. Cumulative Percentage of Valid Cells against the Probability of Success

4.3   Characteristics of Vacant Trips

The vacant trips extracted from the GPS data also provide information on the travel distance and time for vacant taxi drivers in customer-search during the morning peak period as shown in Table 1. Due to the high passenger demand, the vacant taxi drivers only required to spend approximately 10 min and travel about 2 km on average to find their next customers. In terms of the number of cells traveled, the mean value and the standard deviation are 15.07 and 20.46, respectively. The results show that the vacant taxi drivers could find a nearby customer within their current district easily and needed not travel across the district in a long distance for customer-search. Therefore, the data extracted from the morning peak period in one district are considered appropriate to examine the local customer-search behavior.

Table 1. Means and Standard Deviations of Vacant Trips

Vacant trip characteristics	Mean [Standard deviation][a]
Customer-search distance (km)	2.37 [3.44]
Customer-search time (min)	10.32 [11.23]

Note: [a] The values in brackets represent the standard deviations of the data.

Figure 6 illustrates how travel distance in terms of the number of cells traveled by the vacant taxi drivers in customer-search relates to the cumulative percentage of vacant taxi trips revealed from the GPS data. It is observed that the total percentage increases with the number of cells traveled but the slope of the curve is decreasing, meaning that the percentage of vacant taxi trips successfully picked up a customer gradually decreases with the number of

cells traveled in customer-search. The figure also reveals that 80% of the vacant taxi drivers could reach a customer within 22 cells, and the trend becomes flatter beyond this level. It is also noticed that only less than 5% of the vacant taxi drivers searched for more than 50 cells (i.e., about 10 km) for meeting their customers. Those trips were probably across districts for customer-search and considered to have a minimal impact to our model development for local customer-search.
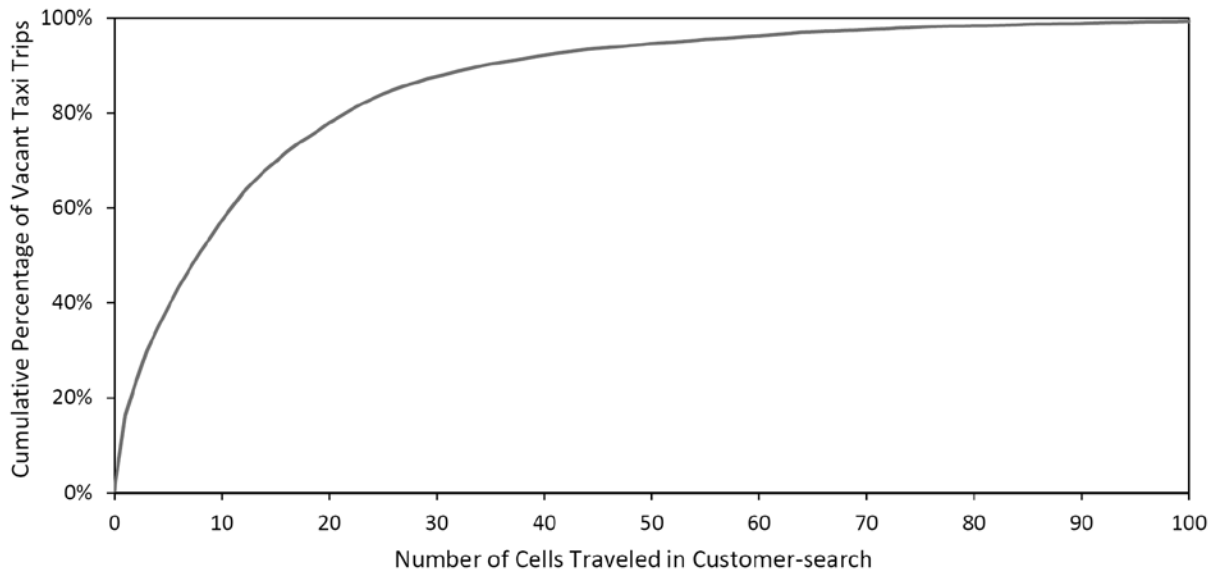

Figure 6. Number of Cells Traveled in Customer-search to the Cumulative Percentage of Vacant Taxi Trips

5. RESULTS AND DISCUSSION

The local taxi customer-search model (4) was calibrated by solving a bi-level non-linear optimization problem with objective (5) in the upper level problem and objective (6) in the lower level problem. Figure 7 illustrates the calibrated model coefficient and the corresponding BIC over different numbers of customer-search decisions anticipated (L). Table 2 illustrates the detailed model results at the optimal number of decisions anticipated $\hat{L}$ of 22.
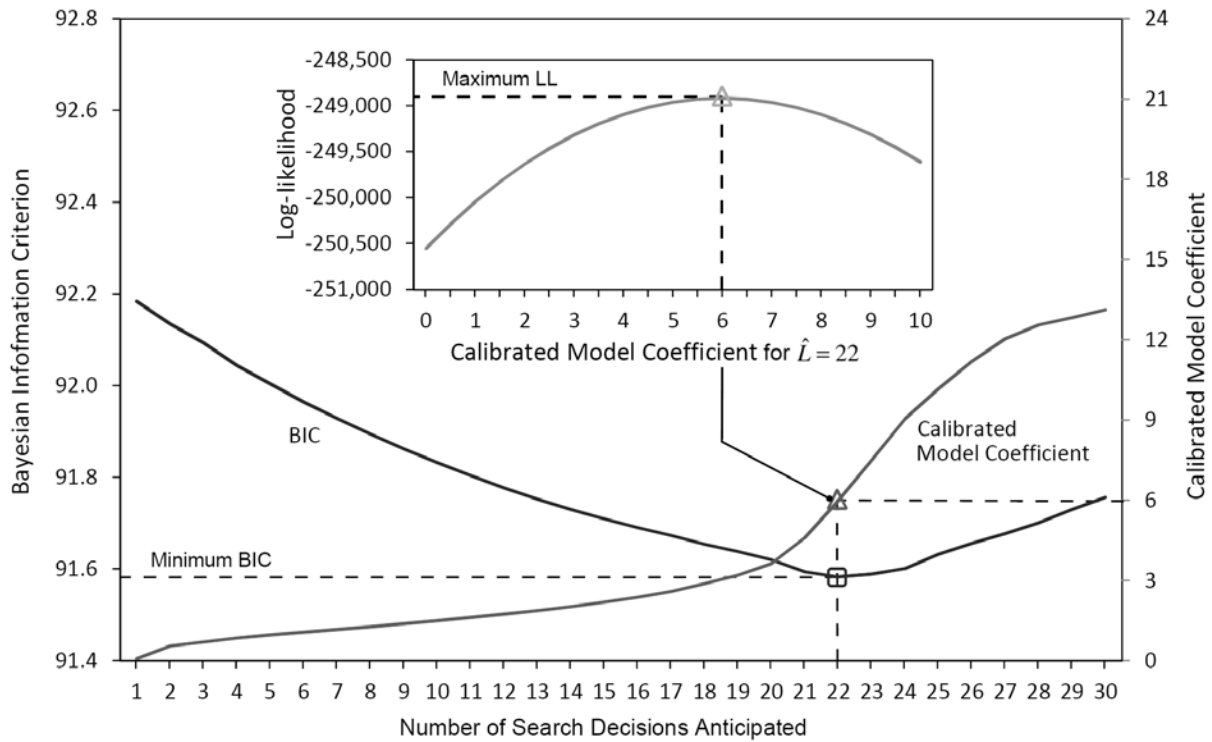
Figure 7. Calibrated Model Coefficient and Bayesian Information Criterion at Different Numbers of Search Decisions Anticipated

Table 2. Coefficient of the Utility Function and the t-statistics

Explanatory variable	Coefficient[a] [t-statistics][b]
Cumulative probability of success at the optimal number of decisions anticipated	5.99 [56.2]

Note: [a] The parameter is significant at the 1% level.

[b] The value inside a pair of brackets represents the t-statistics of the explanatory variable.

The above results are logic. The calibrated coefficients with all numbers of search decisions anticipated are positive at the 5% significance level. This implies that a cell with a higher cumulative probability of success has a greater attraction to the vacant taxi drivers. The calibrated coefficients, which can also be interpreted as the level of information quality received by the vacant taxi drivers, are gradually increased from 0.08 to 13.12. The trend makes considerable sense because a larger number of search decisions anticipated implies a larger spatial coverage and more information considered by the drivers, and hence implies a better level of information quality received by them.

The BIC value demonstrates the incremental benefits of considering more customer-search decisions is decreasing until it reaches its minimum of 91.58 at the optimal number of decisions anticipated of 22. The result demonstrates the spatial coverage considered by the vacant taxi drivers in customer-search during the morning peak period is about 4.4 km (22 x 200 m).

A key point which is worth to discuss is the calibrated model coefficient (0.08) when the number of search decisions anticipated equals 1. This value is close to zero and suggests that

16

the customer-search pattern tends to be random when an insufficient number of search decisions is considered. This could explain why some of the previous researches did not find a significant factor for a local customer-search and they assumed that taxis were operating in a random walk manner.

Table 3. The Results of the Likelihood Ratio Test

Measures/Conclusions		Results
Log likelihood	The equally-likely model	-250,554
	Our proposed model with the optimal parameter value	-248,918
Likelihood ratio (LR)		3,272
Chi-square critical value[a]		6.6
Conclusion of the hypothesis test[b]		Reject

Note: [a] The chi-square critical value when the degree of freedom is 1 and the significance level is 0.01.

[b] The null hypothesis test at the 99% confidence interval.

Table 3 shows the log-likelihood ratio test results for our calibrated search model at the optimal number of search decisions anticipated and the equally-likely model. Given that the degree of freedom is 1, the corresponding chi-square critical value at the 1% significance level is 6.6, which is significantly lower than the calculated log-likelihood ratio. Therefore, the null hypothesis that there is no intervention between the two models is rejected accordingly. We can conclude that the local customer-search behavior did not follow the random walk principle. Our proposed model is more reasonable, informative, and significantly different from a random walk model. The implication is that the driver's search strategies did consider the cumulative probabilities of success.

To validate our model, we compared the estimated and actual expected numbers of cells traveled in customer-search. Using Equation (9), we found that the expected number of cells traveled in customer-search equals 15.02. The prediction is close to the actual average number of cells traveled in customer-search of 15.07 discussed before, and is within the range of two standard deviations to the mean values at the 95% confidence interval. On the other hand, given that the mean observed customer-search distance and time in one cell deduced from the GPS data is, respectively, 0.17 km and 0.78 min, 15.02 cells implies that the predicted circulating distance and time are, respectively, 2.55 km and 11.72 min for each vacant taxi trip. By comparing these numbers with the corresponding numbers in Table 1, we notice that the predictions are close to the corresponding observed values and are within the range of two standard deviations to the corresponding mean values at the 95% confidence interval. This implies that our model reflects the actual situation.

Following the above evaluation for confirming the accuracy of model predictions, we conducted a sensitivity test to analyze the potential consequences of the implementation of the policies on regulating the taxi fleet size and the changes in passenger demand due to the population growth. As we have described in Equation (1), the probability of success for the vacant taxi drivers meeting their next customers on the road network is subject to the availability of vacant taxis (which is a function of the taxi fleet size) and the passenger demand (which is a function of the passenger trip generation rate, population size, and the

capacity provided by alternative transport modes). A larger taxi fleet size results in more vacant taxi available on streets. Higher trip generation rates, more populations, and lower capacity provided by alternative modes can result in higher passenger demand. To simplify analysis, we use a multiplier to capture their effects into the probability of success. Figure 8 shows the predicted changes in the expected number of cells traveled in customer-search under various multipliers to the probability of success (with multipliers between 0.5 and 1.5).

Figure 8. Changes in Expected Number of Cells Traveled in Customer-search under Various Multipliers to the Probability of Success

The sensitivity analysis results indicate that the expected number of cells traveled to meet a customer (which can be converted to the search distance or time by multiplying that number by the mean travel distance or time in one cell) keeps decreasing while factoring up the probability of success. It is logical because a larger multiplier implies either few vacant taxis available in the road network (less competition between taxis) or higher passenger demand, leading to a higher cumulative probability of successfully meeting a customer during the search and a shorter search distance and time.

It is necessary to clarify that a small number of cells could provide us an exceptionally high probability of success after they were factored up by 50%. Their probabilities of success were controlled by the upper limit of 100% in order to prevent the violation of the implicit model condition of $0 \leq S_j^l \leq 1$. Since the probability of success is bounded, increasing the multiplier makes no effect to those cells with the probability of success already equal to 100%. It explains the declining trend of the decreasing rate on the expected number of cells traveled in customer-search while factoring up the probability of success. In particular, the decreasing rate is 0.44 cell per 1 percent growth of the probability of success when the multiplier equals 0.50, and the rate drops to 0.06 cell per 1 percent growth of the probability of success when the multiplier equals 1.50.

6.   CONCLUSION

This study has developed, calibrated, and validated a logit-opportunity taxi customer-search model using the GPS data obtained from 460 urban taxis in Hong Kong. The proposed model combines the concepts of both the logit-based search model and the intervening opportunity model to explain the local customer-search behavior of vacant taxi drivers. The cumulative probability of success of being hired was used as the explanatory variable and found to be a significant factor that affects the search choice of vacant taxi drivers. The model parameters were calibrated by solving the proposed bi-level optimization problem. The model properties have also been illustrated. The result illustrates that the local customer-search behavior does not follow the random walk principle. The developed model can help identify the local customer-search decision and the trajectories of vacant taxis, investigate the vacant taxi circulating behavior, and predict the effects on the customer-search distance and time subject to the variations of the taxi fleet size and passenger demand.

REFERENCES

Arnott, R. (1996) Taxi travel should be subsidized. Journal of Urban Economics, 40: 316-333.

Beesley, M.E. and Glaister, S. (1983) Information for regulation: The case of taxis. The Economic Journal, 93: 594-615.

Ben-Akiva, M.E., Gao, S., Wei, Z. and Wen, Y. (2012) A dynamic traffic assignment model for highly congested urban networks. Transportation Research Part C, 24: 62-82.

Cairns, R.D. and Liston-Heyes, C. (1996) Competition and regulation in the taxi industry. Journal of Public Economics, 59: 1-15.

Conway, A., Kamga, C., Yazici, A. and Singhal, A. (2012) Challenges in managing centralized taxi dispatching at high-volume airports: Case study of John F. Kennedy International Airport. Transportation Research Record, 2300: 83-90.

da Costa, D.C.T. and de Neufville, R. (2012) Designing efficient taxi pickup operations at airports. Transportation Research Record, 2300: 91-99.

De Vany, A.S. (1975) Capacity utilization under alternative regulatory constraints: An analysis of taxi markets. Journal of Political Economy, 83: 83-94.

Douglas, G.W. (1972) Price regulation and optimal service standards: the taxicab industry. Journal of Transport Economics and Policy, 6: 116-127.

Feng, T. and Timmermans, H.J.P. (2013) Transportation mode recognition using GPS and accelerometer data. Transportation Research Part C, 37: 118-130.

Fernández, J.E., De Cea, J. and Briones, J. (2006) A diagrammatic analysis of the market for cruising taxis. Transportation Research Part E, 42: 498-526.

Flores-Guri, D. (2003) An economic analysis of regulated taxicab markets. Review of Industrial Organization, 23: 255-266.

Heanue, K.E. and Pyers, C.E. (1966) A comparative evaluation of trip distribution procedures. Highway Research Record, 114: 20-50.

Horn, M.E.T. (2002) Fleet scheduling and dispatching for demand-responsive passenger services. Transportation Research Part C, 10: 35-63.

Hu, X., Gao, S., Chiu, Y.C. and Lin, D.Y. (2012) Modeling routing behavior for vacant taxi cabs in urban traffic networks. Proceedings of the 91st Annual meeting of the Transportation Research Board, National Research Council, Washington, D.C., United States.

Jiang, B., Yin, J. and Zhao, S. (2009) Characterizing human mobility patterns in a large street network. Physical Review E, 80: 1-11.

Jung, J., Chow, J.Y.J, Jayakrishnan, R. and Park, J.Y. (2014) Stochastic dynamic itinerary interception refueling location problem with queue delay for electric taxi charging stations. Transportation Research Part C, 40: 123-142.

Lee, W.C. and Cheng, B.W. (2008) Incorporating e-technology to advantage in a greener taxi industry and its impact on driving performance and safety. Transportation Planning and Technology, 31: 569-588.

Li, Q, Zeng, Z., Yang, B and Zhang, T. (2009) Hierarchical route planning based on taxi GPS-trajectories. Proceedings of the 17th International Conference on Geoinformatics, Virginia, United States.

Loo, B.P.Y., Leung, B.S.Y., Wong, S.C. and Yang, H. (2007) Taxi license premiums in Hong Kong: Can their fluctuations be explained by taxi as a mode of public transport? International Journal of Sustainable Transportation, 1: 249-266.

Manski, C.F. and Wright, J.D. (1976) Nature of equilibrium in the market for taxi services. Transportation Research Record, 619: 296-306.

McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (ed.), Frontiers in econometrics. New York: Academic Press, 105-142.

Miwa, T., Ishiguro, Y., Yamamoto, T. and Morikawa, T. (2013) Allocation planning for probe taxi devices based on information reliability. Transportation Research Part C, 34: 55-69.

Moore, A.T. and Balaker, T. (2006) Do economists reach a conclusion on taxi deregulation? Economic Journal Watch, 3: 109-132.

Morikawa, T. and Miwa, T. (2010) Preliminary analysis on dynamic route choice behavior: Using probe-vehicle data. Journal of Advanced Transportation, 40: 140-162.

Ortúzar, J. de D. and Willumsen, L.G. (1999) Modeling transport. Second edition. Chichester: John Wiley & Sons.

Schneider, M. (1959) Gravity models and trip distribution theory. Papers and Proceedings of the Regional Science Association, V: 51-56.

Schroeter, J.R. (1983) A model of taxi service under fare structure and fleet size regulation. Journal of Economics, 14: 81-96.

Sirisoma, R.M.N.T., Wong, S.C., Lam, W.H.K., Wang, D., Yang, H. and Zhang, P. (2010) Empirical evidence for taxi customer-search model in Hong Kong. Proceedings of the Institution of Civil Engineers, Transport, 163: 203-210.

Stouffer, A. (1940) Intervening opportunity: A theory relating mobility and distance. American Sociological Review, 6: 845-867.

Sun, Z. and Ban, X. (2013) Vehicle classification using GPS data. Transportation Research Part C, 37: 102-117.

Szeto, W.Y., Wong, R.C.P., Wong, S.C. and Yang, H. (2013) A time-dependent logit-based taxi customer-search model, International Journal of Urban Sciences, 17: 184-198.

Wong, K.I., Wong, S.C. and Yang, H. (2001) Modeling urban taxi services in congested road networks with elastic demand. Transportation Research Part B, 35: 819-842.

Wong, K.I., Wong, S.C., Yang, H. and Tong, C.O. (2002) A sensitivity-based solution algorithm for the network model of urban taxi services. In: Taylor, M.A.P. (ed.), Proceedings of the 15th International Symposium on Transportation and Traffic Theory. Amsterdam: Elsevier Science, 23-42.

Wong, K.I., Wong, S.C., Yang, H. and Tong, C.O. (2003) The effect of perceived profitability on the level of taxi service in remote areas. Journal of the Eastern Asia Society for Transportation Studies, 5: 79-94.

Wong, K.I., Wong, S.C., Bell, M.G.H. and Yang, H. (2005) Modeling the bilateral micro-searching behavior for urban taxi services using the absorbing Markov chain approach. Journal of Advanced Transportation, 39: 81-104.

Wong, K.I., Wong, S.C., Yang, H. and Wu, J.H. (2008) Modeling urban taxi services with multiple user classes and vehicle modes. Transportation Research Part B, 42: 985-1007.

Wong, R.C.P., Szeto, W.Y., Wong, S.C. and Yang, H. (2014a) Modeling multi-period taxi customer-searching behavior. Transportmetrica B, 2: 40-59.

Wong, R.C.P., Szeto, W.Y. and Wong, S.C. (2014b) Sequential logit approach to modeling the customer-search decisions of taxi drivers. Asian Transport Studies, in press.

Wong, R.C.P., Szeto, W.Y. and Wong, S.C. (2014c). Bi-level decisions of vacant taxi drivers traveling towards taxi stands in customer-search: Modeling methodology and policy implications. Transport Policy, 33: 73-81.

Wong, S.C. and Yang, H. (1998) Network model of urban taxi services: Improved algorithm. Transportation Research Record, 1623: 27-30.

Xu, J.M., Wong, S.C., Yang, H. and Tong, C.O. (1999) Modeling the level of urban taxi services using a neural network. ASCE Journal of Transportation Engineering, 125: 213-216.

Yang, H., Fung, C.S., Wong, K.I. and Wong, S.C. (2010a) Nonlinear pricing of taxi services. Transportation Research Part A, 44: 337-348.

Yang, H., Lau, Y.W., Wong, S.C. and Lo, H.K. (2000) A macroscopic taxi model for passenger demand, taxi utilization and level of services. Transportation, 27: 317-340.

Yang, H., Leung, C.W.Y., Wong, S.C. and Bell, M.G.H. (2010b) Equilibria of bilateral taxi-customer searching and meeting on networks. Transportation Research Part B, 44: 1067-1083.

Yang, H. and Wong, S.C. (1998) A network model of urban taxi services. Transportation Research Part B, 32: 235-246.

Yang, H., Wong, K.I. and Wong, S.C. (2001) Modeling urban taxi services in road networks: Progress, problem and prospect. Journal of Advanced Transportation, 35: 237-258.

Yang, H., Wong, S.C. and Wong, K.I. (2002) Demand-supply equilibrium of taxi services in a network under competition and regulation. Transportation Research Part B, 36: 799-819.

Yang, H. and Yang, T. (2011) Equilibrium properties of taxi markets with search frictions. Transportation Research Part B, 45: 696-713.

Yang, H., Ye, M., Tang, W.H.C. and Wong, S.C. (2005a) Regulating taxi services in the presence of congestion externality. Transportation Research Part A, 39: 17-40.

Yang, H., Ye, M., Tang, W.H.C. and Wong, S.C. (2005b) A multi-period dynamic model of taxi services with endogenous service intensity. Operations Research, 53: 501-515.

Yang, T., Yang, H. and Wong, S.C. (2012a). Taxi services with search frictions and congestion externalities. Journal of Advanced Transportation. doi:10.1002/atr.1210.

Yang, T., Yang, H., Wong, S.C. and Sze, N.N. (2012b) Returns to scale in the production of taxi services: An empirical analysis. Transportmetrica A, doi:10.1080/23249935.2013.794174.

Zhan, X., Hasan, S., Ukkusuri, S.V. and Kamga, C. (2013) Urban link travel time estimation using large-scale taxi data with partial information. Transportation Research Part C, 33: 37-49.

Ziebart, B.D., Maas, A.L., Dey, A.K. and Bagnell, J.A. (2008) Navigate like a cabbie: Probabilistic reasoning from observed context-aware behavior. Proceedings of the 10th International Conference on Ubiquitous Computing, Seoul, South Korea.