Modeling Location Choice of Taxi Drivers for Passenger Pick-Up Using GPS Data

Merkebe Getachew Demissie, Lina Kattan, Santi Phithakkitnukoon, Gonçalo Homem de Almeida Correia, Marco Veloso, Carlos Bento

***Abstract*— Recently the traditional taxi industry is struggling to keep its market share, especially with the emergence of new transport network companies (e.g., Uber). One of the problems with the traditional taxi services is the difficulty of matching the taxi demand to its supply when there is no phone booking or another reservation system. In that perspective, the taxi driver's experience is important in reaching the next passenger. A taxi driver with limited experience may not know the high-demand locations and times of taxi stands or street sections to visit after dropping off a passenger. This causes a large number of vacant taxi drivers to regularly cruise the roads to search a passenger, contributing to congestion, pollution, and resource waste. We formulate the problem of a taxi driver's next passenger pick-up location as a destination choice problem. Vacant taxi trips between drop-off and pick-up points are extracted from GPS records obtained from a taxi operator in Lisbon, Portugal to understand the travel behavior of vacant taxi drivers. We have estimated destination choice models with a multinomial logit and with a nested logit structure. It was found that passenger demand at the pick-up area, hotspot locations, service location preference, and major transport hubs positively influence a taxi driver's next choice of passenger pick-up location. Results of this study provide insight regarding the factors that explain a taxi driver's probability to choose a certain zone within a set of passenger pick-up zones, contributing to a better understanding of taxi drivers travel behaviour.**

***Index Terms*—taxi GPS trajectory data, destination choice modeling, taxi travel demand, vacant taxi trip, multinomial logit, nested logit**

I. Introduction

Taxi services are globally available and account for a small but significant portion of daily trips [1]. The taxi industry is struggling to keep its market share. One of the reasons for this is the emergence of new transport alternatives such as peer-to-peer ridesharing and transportation network companies (TNCs) like Uber and Lyft [2].

One important issue for a taxi service is matching the taxi demand to its supply. Wong et al. [3] and Yang et al. [4] created equilibrium models to express the relationship between taxi demand and a taxi driver's search for passengers. These studies show that the absence of such equilibrium could lead to an excess of vacant taxis, which can create competitiveness to get the next passenger or longer wait times and unreliable taxi service [5]. The main innovation of a TNC is the development of a platform that connects passengers to drivers [2]. Most taxi agencies also provide telephone-based dispatch services, but the new TNC service refines the system by using geo-positioning to reduce the time a passenger must wait for a driver [6].

Very often, a taxi driver's mobility intelligence is important in reaching passengers. Experienced taxi drivers know the locations and times of high-demand taxi stands or street sections and will go to them after a passenger drop-off based on the day of the week and the time of day. Conversely, a taxi driver with limited experience faces difficulty in reaching the next passenger. This causes drivers of vacant taxis to cruise the road in search of passengers, which contributes to traffic congestion, air pollution, and resource waste [7],[8].

A variety of studies have been carried out to understand the temporal and spatial variations of taxi demand. One method of achieving a more streamlined flow of taxi services has come in the form of detection for pick-up hotspots to aid vacant taxis in finding passengers [8]–[10]. Another has identified efficient taxi service strategies based on revenue [11], [12]. The aforementioned studies primarily focused on the use of historical GPS data to study the factors that affect a taxi driver's mobility intelligence and consequently their choice regarding the best route and pick-up location. To improve taxi services, it is necessary to understand taxi demand, how that demand varies through space and time, and which attributes influence that demand. To achieve this goal, Lacombe et al. [7] and Yang et al. [13] developed two trip generation models, one for trip production and the other for trip attraction, and applied various explanatory variables such as demographics, land use, accessibility to transit, and weather conditions to those models to determine whether any of those were likely to influence taxi demand.

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery, the Urban Alliance Chair in Transportation Systems Optimization, and the Alberta Innovate Strategic Grant on Integrated Urban Mobility at the University of Calgary. This research work was partially supported by Chiang Mai University. (Corresponding authors: Merkebe Getachew Demissie; and Santi Phithakkitnukoon.)

M. G. Demissie and L. Kattan are with the Department of Civil Engineering, University of Calgary, Canada (e-mail: merkebe.demissie@ucalgary.ca; lkattan@ucalgary.ca).

S. Phithakkitnukoon is with the Excellence Center in Infrastructure Technology and Transportation Engineering, Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand (e-mail: santi@eng.cmu.ac.th).

G.H. Correia is with the Department of Transport and Planning, Delft University of Technology, The Netherlands (e-mail: g.correia@tudelft.nl).

M. Veloso and C. Bento are with Center for the Informatics and Systems, University of Coimbra, Coimbra, Portugal (e-mail: mveloso@dei.uc.pt, bento@dei.uc.pt).

In this paper, we intend to address the issue of vacant taxi trips by identifying factors that influence vacant taxi trips. We argue that determining the factors that influence a taxi driver's next choice of passenger pick-up location compared to their choice of whether to hunt or wait locally versus traveling a distance could provide additional insights on the characterization of vacant taxi trips. We formulate the problem of a taxi driver's next passenger pick-up location as a destination choice problem. Traditionally, multinomial responses have been analyzed using the multinomial logit (MNL) model, which is the most common implementation of discrete choice model [14]. A number of studies have been conducted using MNL model to analyze individuals' destination choice for leisure, tourism, and recreation [15], [16]; work, shopping, and other destinations [17]; and non-work related trips destinations [18]. The MNL model structure has also been applied to estimate destination choice models to show the distribution of occupied taxi trips [19], [20].

In practice, often the researcher is unable to capture all the sources of correlation, especially in the case of destination choice modeling where spatial units are presented as alternatives. In this case, a more general model than the simple MNL model is needed [18], [21]. A common solution is to relax the independent and identically distributed error structure and there are several different types of model structures which may be used to model destination choices. Of these model structures, the following models have received particular attention in location choice analysis: Bhat and Guo [22] proposed the use of a mixed spatially correlated logit (MSCL) model for household residential location choice. The MSCL model has the advantage of the generalized extreme value (GEV)-based structure to accommodate correlation in the utility of the household residential units, and a mixing distribution over the GEV model structure to accommodate unobserved response heterogeneity. Wang et al. [23] explored the use of a paired combinatorial logit model to analyze location choice of metro commuters for after-work activities. Hammadou et al. [18] applied a mixed nested logit model to estimate destination choice model for non-work intra-urban trips.

As there is a gap in the literature about how taxi drivers choose zones for passenger pickup purposes, two major contributions are made to the literature. First, we bring together taxi GPS trajectory data, open -and crowdsourced geospatial data (Foursquare check-in count, and Points of Interest (POIs)), Google Distance Matrix API, and census records to enrich the set of variables available for modeling multiple aspects of taxi travel demand. Second, to study the location choice of taxi drivers for passenger pickup, we have developed a choice modeling framework based on a nested logit and a multinomial logit models. To the best of our knowledge, this is the first study where a nested logit structure is used to model the location choice of taxi drivers for passenger pickup based on data obtained from multiple sources. In our attempt to develop the nested logit models, a k-means clustering technique is used to group destination zones that are similar in terms of trip generation roles. Then, destination zones in the same cluster are assumed to be in the same nest. Attempts are made to characterize the time of day profile of destination zones using POIs, Foursquare check-in count, and population density data. Such insight can be hardly obtained using a static information that mostly comes from traditional survey-based data. Thus, this is a timely study showing the opportunities of open and proprietary datasets and how effectively such datasets can be utilized to augment the capability of the traditional discrete choice models in vacant taxi travel demand modeling.

The remainder of the paper is structured as follows. Section II discusses works related to improving taxi drivers' passenger pick-up strategies, taxi travel demand models, and factors affecting taxi travel demand. Section III presents the methodology and data requirements, including model formulation, variable definitions, identification of explanatory variables, and case study area. Section IV presents and discusses the results of the models. Section V concludes and summarizes this paper's main findings and points for future research directions.

II. Related Work

The use of opportunistic sensing datasets produced from various sources has attracted a lot of attention from transport planners in recent years [12], [24], [25]. Some examples of analyses of this type of dataset are GPS data [26], [27], call detailed records data of mobile phones [28]–[32], and open and crowdsourced data [33], [34]. Transport planners now have new ways of providing insights regarding the spatial distributions and temporal evolutions of human and vehicular movements within cities.

A significant portion of the literature is dedicated to detecting the spatial and temporal variations of taxi activity at major taxi trip generation and attraction points[8], [10], [35], [36]. The pick-up and drop-off events can be inferred by analyzing the transition of a taxi meter between the vacant and occupied statuses. This information can be used to understand the different taxi trip generation and attraction roles of the neighboring areas. For example, Wan et al. [36] applied a DBSCAN algorithm to cluster pick-up and drop-off points with the aim of predicting an area of interest for passenger pick-up based on the time of day. Lee et al. [10] applied a K-means clustering algorithm to generate popular clusters and to design a location recommendation service for vacant taxis to reduce their idling times. Chang et al. [9] developed a taxi demand hotspot prediction method based on drop-off location, weather, time, and request history information.

Some existing works are intended to provide information to taxi passengers in addition to taxi drivers. Phithakkitnukoon et al. [37] developed a method to extract the number of vacant taxis in different areas of a city to assist passengers in finding taxi services with greater certainty. Yuan et al. [8] and Yuan et al. [38] developed recommendation systems to assist taxi drivers and passengers in their search for a pick-up location and a vacant taxi, respectively. Jianxin et al. [39] developed real-time dispatch services where users can follow the location and ETA of the dispatched taxi. Moreira-Matias et al. [5] applied a time series forecasting technique to predict taxi demand for selected taxi stands in 30-minute intervals. Two classes of artificial neural networks, convolutional neural network [40],

and long short-term memory [41], have been proposed for ride hailing demand predictions based on historical trip request data.

The service strategies adopted by taxi drivers have direct influences on generated revenue. Veloso et al. [42] explored passenger searching and delivery strategies in Lisbon, Portugal and discovered that the preferred passenger pick-up strategy in the urban area was waiting at an adjacent taxi stand. Liu et al. [26] revealed that efficient and high revenue taxi drivers in the city of Shenzhen, China operate in different parts of the city based on the time of the day and avoid congested roads. Rong et al. [43] modeled a passenger searching strategy as a Markov decision process to optimize taxi driver revenue efficiency.

Li et al. [11] and Zhang et al. [12] took a different tact by analyzing taxi service based on three strategies: passenger searching, passenger delivery, and service area preference. The revenue generated by each is used as indicator to differentiate between efficient and inefficient taxi service strategies. Very few studies have investigated the influences of different factors on taxi travel demand. Knowing why, when, and how people travel helps transportation planners identify travel patterns and trends, which are important pieces of information to inform future planning [44]. Previous studies by Lacombe et al. [7] and Yang et al. [13] developed taxi trip generation models that can be used to estimate the total number of pick-up and drop-off events, where the focuses of these studies are taxi movements during passenger delivery. There is a clear need to improve the knowledge on taxi travel demand estimation especially in what regards to the characterization of taxi movements associated to passenger searching.

III. Methodology and Data Requirements

This study's methodology aims to model the pattern of trips generated by vacant taxis. We develop models to explain the way in which vacant taxi drivers choose among different passenger pickup zones (destination choice). The overall approach taken by this study has four main components: (i) Overall model design; (ii) Data processing; (iii) Model specification; and (iv) Data and case study area description.

A. Overall model design

Fig. 1 shows an overview of passenger searching strategies that may be employed by a taxi driver. After dropping off passengers, the driver must choose from N number of locations to search for a new passenger. In our study, these locations are assumed to be centroids of zones. The choice could be local (if drop-off and pick-up locations are within the same zone: e.g., Zone 1) or going farther ($N-1$ number of zones). The choice of destination (pick-up location) can be treated as a discrete choice problem and can be addressed with models at the individual level [44].

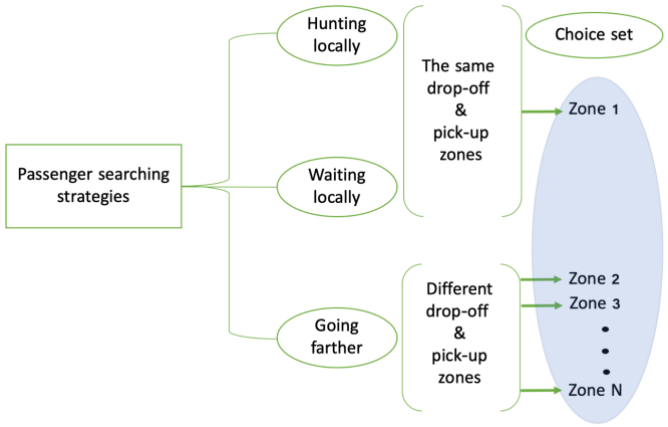


Fig. 1. Overview of passenger searching strategies.

B. Data processing

The passenger-searching strategies of taxi drivers are observed based on the location and time of consecutive drop-off and pick-up events extracted from taxi GPS trajectory data. In the context of taxi operation, we assume the GPS trajectory data represents all movements and activities. Fig. 2 shows a three-level data processing framework to generate vacant and occupied taxi trips. The framework encompasses data cleaning, activity detection, and vacant and occupied taxi trip extraction. The data cleaning process includes the removal of GPS pings outside the study region. We also remove occupied taxi trips with trip lengths over 30 km and trip durations of over 2 hours since the longest trip from one side of the city of Lisbon to the other side is around 22 km [42].

We must calculate event indicators such as time and distance gaps between GPS pings to obtain the components of taxi operations: trips and activities. Activities are drop-off, pick-up, and passenger waiting events. Drop-off and pick-up events are detected when a taxi meter transitions between the vacant and occupied statuses. Trips are connections between the drop-off and pick-up activity locations.

Fig. 3 shows the trajectory of a randomly selected taxi. The red line indicates passenger delivery (an occupied taxi trip). The green line indicates passenger searching (a vacant taxi trip). The change from red to green (#2) represents a passenger drop-off event. The change from green to red (#4) represents a passenger pick-up event. Zhang et al. [12] showed that a taxi driver's initial passenger searching strategy may not always be successful. For instance, Fig. 3 shows a sequence of decisions made by the taxi driver between the passenger drop-off (#2) and passenger pick-up (#4) events. After the drop-off event, the driver initially moved to location #3, where he/she waited for 23 minutes without finding a passenger (e.g., unsuccessful passenger pick-up attempt, driver resting).

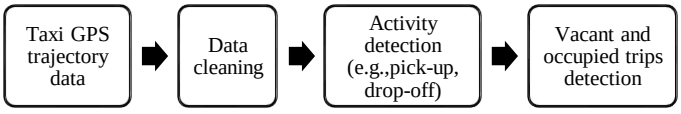


Fig. 2. Processing of taxi GPS trajectory data.

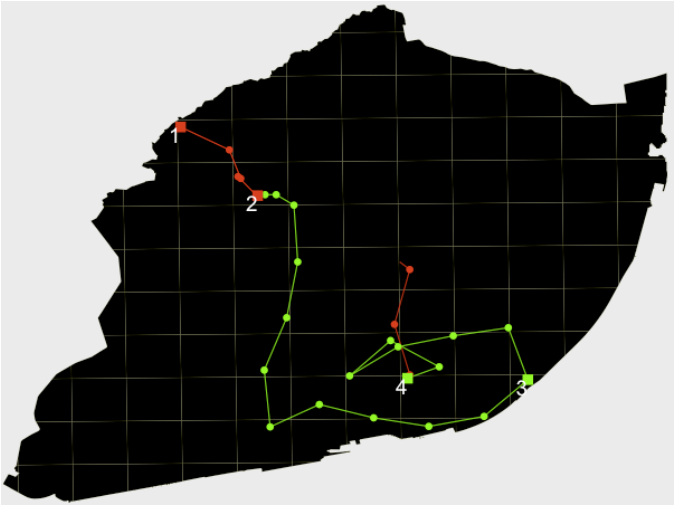


Fig. 3. Example of a single taxi's GPS trajectory and status: occupied (red) and available (green)

The driver then moved to location #4, where he/she succeeded in finding a passenger. The trip between location #2 and location #4 is used in our passenger pick-up location choice modeling. However, the chosen passenger pick-up location may not be the driver's initial intended destination for passenger pick-up.

C. Model specification

The taxi passenger pick-up location choice problem faced by a taxi driver will forthwith be referred to as a destination choice problem. With respect to the model specification, a discrete choice model has been the most widely used method to model the choice of a location among a set of mutually exclusive alternatives based on the principles of utility maximization [18], [22], [23]. We use discrete choice models in this study because the taxi passenger pick-up location choices are discrete and mutually exclusive. Traditionally, multinomial responses have been analyzed using the MNL model, which is the most common implementation of discrete choice model [21]. Zones are mutually exclusive and primarily created based on criteria that suggests homogeneous land use within a zone, but there is some level of correlation between zones located within mixed urban land use areas that share soft boundaries. These zones are likely to have similar unobserved attributes which introduces a dependency. To represent this dependency, a more general model than the simple MNL model is needed [18], [21].

1) Multinomial Logit Model

The probability (P_{im}) that a taxi driver from zone i chooses destination zone m is given by the utility of zone m and the utility of all other possible pick-up zones. The model's general form is shown in Equation (1). The attractiveness of alternatives is represented using the concept of utility, as described in Equation (2).

$$P_{im} = \frac{e^{V_{im}}}{\sum_{z=1}^{N} e^{V_{iz}}} \quad (1)$$

$$U_{nm} = V_{nm} + \varepsilon_{nm} \quad (2)$$

Where, V_{nm} is the measurable conditioning component of the utility individual n associates with alternative m; ε_{nm} is the error component of the utility individual n associates with alternative m; and N is the total number of pick-up zones in the study area, which is 108 in this case.

In our case (108 zones), the study region is suitable for estimating the destination choice models with the full set of alternatives. However, the computational requirements of estimating destination choice models typically rise for a study area with a large number of zones (alternatives). Thus, to make the modeling framework study more general so it can be transferable for a study area with a large number of zones, we conduct a choice set formation method as suggested by Ben-Akiva and Lerman [14]. In fact, because of the Independence from Irrelevant Alternatives (IIA) property of MNL, Ben-Akiva and Lerman [14] suggested using a restricted set of zonal alternatives rather than a full set when estimating a destination choice model. This study uses importance-based sampling with replacement procedure as in [14], [45], [46] to develop attractiveness indices for zones and thus calculates the probability of being included in the choice set.

The importance-based sampling approach involved the following steps: (i) calculate selection weights and selection probabilities; and (ii) sample possible alternative destinations for the observed choice and select a final choice set that contains both the chosen zone and sample zones drawn from the full set of zones. The selection weight of destination zone j relative to origin zone i (W_{ij}) is calculated using Equation (3).

$$W_{ij} = A_j \times e^{\left(-2\times\frac{D_{ij}}{D_{avg}}\right)} \quad (3)$$

Where, A_j is destination zone's size variable; D_{ij} is the travel impedance between the origin zone and the destination zone; and D_{avg} is average travel distance in the study region. This study uses the average number of Foursquare chick-in counts as size variable instead of the total number of trip ends, which was employed by [45].

The selection probabilities are estimated using the formula in Equation (4):

$$\mathrm{SP_{ij}} = \frac{\mathrm{W_{ij}}}{\sum_{z=1}^{N} \mathrm{W_{iz}}} \quad (4)$$

Where, $\mathrm{SP_{ij}}$ is the selection probability of destination zone j for a vacant taxi trip starting from zone i.

Once the selection probabilities were calculated, the next step is to select the destination zones that will be part of the choice set. Using the selection probabilities, the cumulative selection probability (cP_{ij}) are calculated by adding the selection probabilities of each origin zone i to all the possible N destination zones, which is 108 in this case. The cP_{ij} of each destination zone j from origin zone i has a range. The lower limit was the cumulative sum of the selection probabilities

($\sum_{z=1}^{j-1} SP_{iz}$) excluding the selection probability of zone j, and the value of the upper limit of the range equal to the selection probability of zone j plus the lower limit ($\sum_{z=1}^{j} SP_{iz}$). The upper limit of the last zone (j = 108) is equal to one.

A numerical experiment was carried out by Nerella and Bhat [47] to study the effect of the sample size of alternatives on model performance for an MNL model. The study suggested a minimum threshold of an eighth of the size of the full choice set to estimate an MNL model and a fourth of the full choice set as a desirable target. This study uses a half of the full choice set such as 54 zones. Fifty four random numbers between 0 and 1 were generated for each vacant taxi trip extracted from the taxi GPS data. The values of these random numbers were compared to the cumulative selection probabilities for the corresponding origin of the trip. If the random number fell in the range of cP_{ij}, the destination zone j was selected to be part of the choice set. This step was repeated for each of the random numbers, where the destination was chosen with replacement each time. Because of the elimination of these duplicates, most of the choice sets has less than 54 zonal alternatives.

A correction factor (CF) was added to reduce any bias that might occur in the model due to using a restricted set of zonal alternatives. The CF is only used in model estimation but not model application. The coefficient of this factor was constrained to 1. Correction factors take the following form $CF_{ij} = -\ln q_{ij} = -\ln (SP_{ij} \times n)$, where, CF_{ij} is correction factor of zone j for a trip starting from zone i; q_{ij} is overall probability of zone j being included in the sample set for model estimation; SP_{ij} = selection probability for a trip from origin zone i to destination zone j ; and n is the number of alternative zones selected.

2) Nested Logit Model

In practice, often the researcher is unable to capture all the sources of correlation, especially in the case of destination choice modeling where spatial units are presented as alternatives. For instance, a taxi driver could be faced with a choice set comprising several equally attractive zones for passenger pick-up. These equally attractive zones can be adjacent to each other or can be found at different parts of the city that are likely to have similar unobserved attributes. This introduces a dependency that conflicts with the IIA assumptions of the MNL functional form. In this case, a more general model than the simple MNL model is needed [21].

We have estimated destination choice models with a nested logit structure. In our attempt to develop a choice modeling framework based on a nested logit, a K-means clustering technique is used to group destination zones that are similar in terms of trip generation roles (land use densities). In our analysis, different clustering techniques are possible candidates to segment destination zones based on their time of day profile represented by POIs, Foursquare check-in count, and population density data. K-means is a simple unsupervised machine learning algorithm and is chosen because of its simplicity in implementation. K-means clustering algorithm identifies clusters of behavior and returns a typical member of that cluster represented by the mean behavior in that group. Previous studies have also shown that K-means clustering technique can be used to identify clusters of locations with similar zoned uses based on activity patterns generated from opportunistic datasets [48]–[50].

The K-means clustering method is applied on three zonal variables: Foursquare users check-in count, number of POIs, and population density. The values of POIs and population density do not change over time. In the case of Foursquare check-in count, we use the total hourly counts for each zone. A cluster may be comprised of zones that are adjacent to each other or can be found at different parts of the city. Then, choice alternatives (zones) in the same cluster are assumed to be in the same nest. This formulation assumes that a taxi driver first chooses an urban area of certain land use type and then, within that land use category, he/she will choose a specific passenger pick-up zone. In this study, we apply a nested logit formulation with two levels of decision for passenger pick-up location. To the best of our knowledge, this is the first study where a nested logit structure is used to model the location choice of taxi drivers for passenger pickup based on data obtained from multiple sources.

Using a similar notation to Train [51], the mathematical formulation of the nested logit with two levels of decision can be described as follows. Let the set of pick-up zones j be partitioned into K non-overlapping nests represented by B_1, B_2, …, B_K. The utility that is derived from the bundle of attributes that describe alternative j in nest B_k as perceived and valued by a taxi driver n is denoted as $U_{nj} = V_{nj} + \varepsilon_{nj}$, where V_{nj} is a measurable conditioning component which is observed by the researcher and ε_{nj} is a random variable (error term) whole value is not observed by the researcher. It can then be shown that the probability of choosing alternative m that belongs to nest B_k is given by Equation (5):

$$P_m = \frac{e^{V_m/\lambda_k}\left(\sum_{j\epsilon B_k} e^{V_j/\lambda_k}\right)^{\lambda_k - 1}}{\sum_{l=1}^{K}\left(\sum_{j\epsilon B_l} e^{V_j/\lambda_l}\right)^{\lambda_l}} \quad (5)$$

The parameter λ_k is a measure of the degree of independence in unobserved utility among the alternatives in nest k.

D. Data and case study area description

1) Case study area

Our methods are applied to a case study using GPS data from the municipality of Lisbon. Lisbon is the capital of Portugal and the center of the Lisbon Metropolitan Area (LMA). The LMA has a population of 2.3 million and is comprised of 18 municipalities (concelhos) that cover a total area of 2,958 km2. About 24.3% of the population of the LMA resides in the municipality of Lisbon [52].

Fig. 4a shows the LMA. Fig. 4b shows the municipality of Lisbon, representing an area of around 100.05 km2 and a population of 552,700. The central business district (CBD) includes the oldest and smallest parishes with high population densities. This area has also a large concentration of office

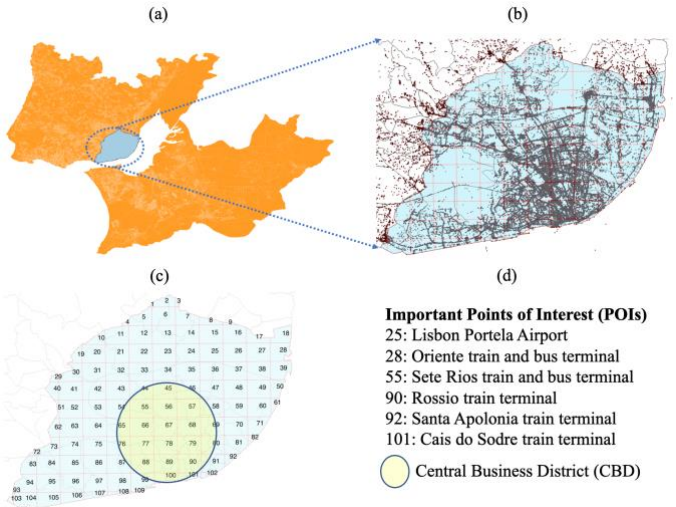


Fig. 4. Case study region showing (a) Lisbon Metropolitan Area, (b) Municipality of Lisbon with sample GPS records, (c) Cell/grid IDs, and (e) POIs

buildings, touristic and commercial activities, and transportation hubs for bus, metro, and ferry. A sample of GPS data is also displayed on Fig. 4b.

Defining passenger pick-up locations (destination zones) is one of the challenging tasks in the development of destination choice model. Modeling destination choices at a census block level is quite difficult because of the high number of alternatives (3712 census blocks in Lisbon). The choice of administrative districts in Lisbon such as freguesias (parish) results in large sized zones, and thus, the number of intra-zonal trips is substantial. This is especially important when a high share of vacant taxi trips is short and could result in a significant number of intra-zonal trips. To address this issue, we divided the municipality of Lisbon into a 1 km x 1 km grid/cell, shown in Fig. 4c. The aim is to generate reasonable number of passenger pick-up locations by boosting homogeneous land use within a zone. Some of the important POIs are listed in Fig. 4d.

2) Dataset

The datasets used to understand taxi travel demand are grouped into five categories:

***GPS records*:** A taxi GPS record dataset covering a period of two months (September 2009 and October 2009) was obtained from a company called GeoTaxi, which holds around a 15% market share in Portugal [53]. The dataset consists of the taxi's location (latitude, longitude), time, heading direction, and occupancy status (vacant, occupied). The GPS data were obtained from 253 taxis.

Data on where and when people checked in (Foursquare check-in data): Foursquare collects data on where and when its users check into a place (check-in). The Foursquare API can be used to obtain that data. Depending on the search area and criteria, the API returns a list of venue records with the following information: venue name, venue category, georeferenced location, number of unique visitors, and number of total check-ins. This study uses Foursquare check-in data collected by Yang et al. [54] between April 2012 and September 2013.

***Trip length and trip time matrix (Google Distance Matrix API)*:** To calculate the travel time and distance between each origin (TAZ centroid) and each destination, an HTTP request interface was used to access the Google Distance Matrix API. These values were obtained for a matrix of origins and destinations (108 x 108 = 11,664), which is based on the recommended routes between the start and end locations [55].

Point of interest (POI) data: POI data provides contextual information about a place and represents the location's characteristics or activity. POI data were acquired from Servidor de Apontadores Portugueses (SAPO). There are a total of 5,471 points located within the municipality of Lisbon.

Census data: The Instituto Nacional de Estatistica (INE) provided the census of demographic, economic, social, and housing information. The data was based on the 2011 Portuguese census [52].

IV. Results and Discussion

A. Results of exploratory data analysis

Fig. 5 shows the variability of a normalized average of the number of occupied (Fig. 5a) and vacant (Fig. 5b) taxi trips throughout a week during the study period. Weekdays show similar patterns, with a high intensity of taxi activity during the day and a low intensity of taxi activity late at night and in the early morning hours. Saturdays and Sundays have different patterns that exhibit a peak in taxi service activities around 12pm. The amount of weekend trips is higher than weekday trips between midnight and 5am but lower from 6am to 6pm.

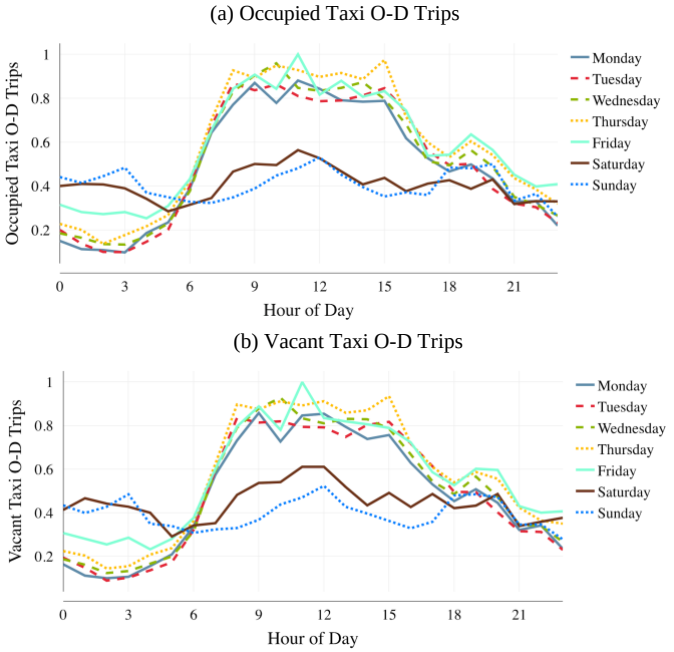


Fig. 5. Citywide occupied and vacant taxi O-D trip patterns throughout the week.

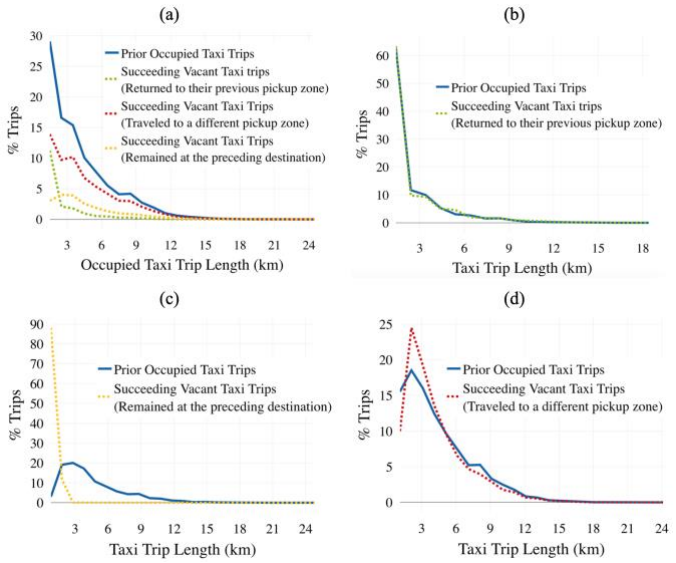


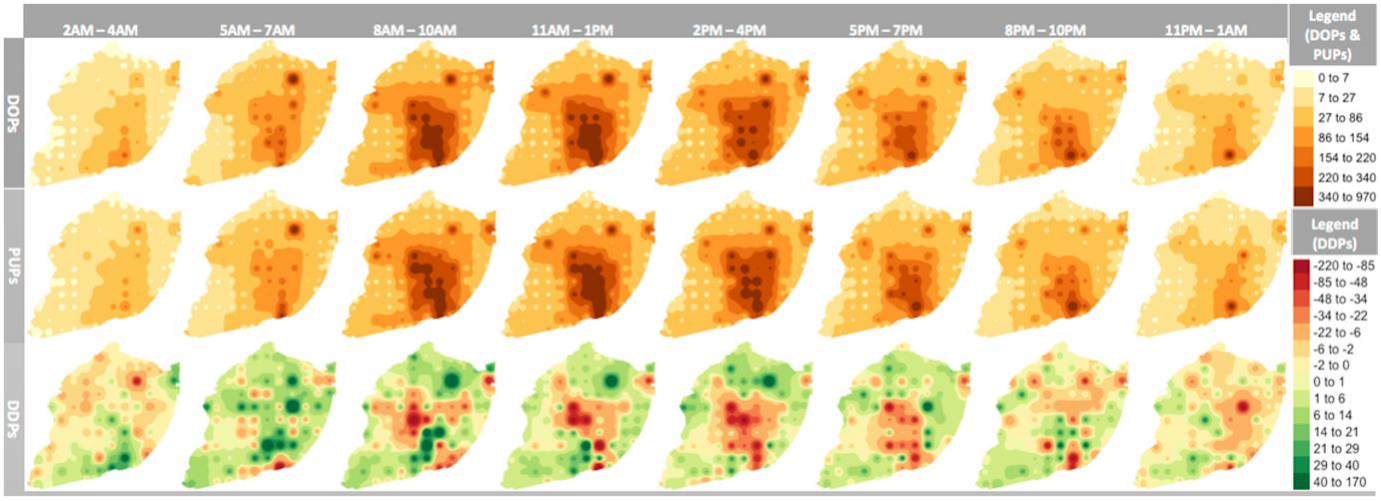
Fig. 6. Vacant and Occupied Taxi trip lengths for different passenger searching strategies.

The vacant and occupied taxi trips are further examined to understand the effect of prior occupied taxi trip length on the succeeding vacant taxi trip time/length. This analysis can shed some light on the searching strategies of vacant taxi drivers after a drop-off event. We started our analysis by examining successive passenger pick-up locations. For each occupied taxi trip, the previous and the next passenger pickup locations are recorded.

Fig. 6a shows the proportion of three customer searching strategies that was calculated for a range of prior occupied taxi trip lengths. After a passenger drop-off in a given location, 17.98% of the taxi drivers circulated within or waited at the area of the preceding destination (Fig. 6a orange color); 20.45% of the taxi drivers returned to their previous pickup location (Fig. 6a green color); and 61.57% of the taxi drivers traveled to a different location (Fig. 6a red color) to look for the next passenger.

Fig. 6b shows the trip length frequency distributions for the occupied and vacant taxi trips. In this category, vacant taxi drivers have returned to their previous pickup location to find their next customer. The average succeeding vacant taxi trip length is 2.524km, which is slightly longer than the prior average occupied taxi trip length (2.439km). A similar analysis is shown in Fig. 6c for the taxi drivers that remained at the preceding destination to look for their next customer. The highest average prior occupied taxi trip length is recorded in this group (4.570km) as well as the shortest average succeeding vacant taxi trip length (0.999km). Fig. 6d shows the trip length frequency distributions for the occupied and vacant taxi trips of the taxi drivers who travelled to areas other than the aforementioned two pickup locations to find their next customer. The average occupied taxi trip length in this category is 4.456km. The taxi drivers in this category faced the longest average vacant taxi trip length (3.410km) compared to the aforementioned customer searching strategies. In general, a large portion of taxi drivers tended to return to their previous pickup zone to find their next customer if the prior occupied trip length is short.

The spatial-temporal distributions of taxi passenger drop-off and pick-up events are further examined using zonal data, as shown in Fig. 7. The average hourly rate of taxi passenger drop-off and pick-up events are calculated for each TAZ over eight periods that represent morning/afternoon peak and evening/night off-peak times. In the daytime (8 am to 4 pm), a high number of taxi passenger drop-off and pick-up events are observed across the city, especially in the central part of the city. There are also a significant number of taxi activities outside the city center, especially at the Lisbon International airport, the Oriente train station and bus terminal, and the ferry dock located in cell #101 (see Fig. 4c, cell ID). The spatial-temporal distribution of taxi passenger drop-off and pick-up events is further examined with an additional metric obtained by subtracting the number of pick-up events from the number



DOPs: drop-off events; PUPs: pick-up events; DDPs = DOPs – PUPs; Red DPPs: more taxi passenger departures; Green DDPs = more taxi passenger arrivals

Fig. 7. Citywide taxi passenger drop-off and pick-up intensities during eight-time intervals (weekday).

TABLE I
EXPLANATORY VARIABLES

Variable Name	Description
Travel time (tt_{ij})	Travel time (minutes) from origin zone to destination zone
Travel distance (td_{ij})	Travel distance (km) from origin zone to destination zone
Combined travel time (ttc_{ij})	For each driver, the average travel time from the centroids of all the other zones to the centroid of the driver's preferred pick-up zone is calculated and multiplied by tt_{ij}.
Combined travel distance (tdc_{ij})	For each driver, the average travel distance from the centroids of all the other zones to the centroid of the driver's preferred pick-up zone is calculated and multiplied by td_{ij}.
Waiting time (wt_j)	Represents average waiting time in each zone a taxi driver faces before succeeding in finding the next passenger.
# Employees	Number of employees of the destination zone
# POIs	Number of Points of Interest of the destination zone
# Hourly trip ends	Hourly number of trip destination ends (# pickups) of the destination zone
Hotspot	Describes the passenger pick-up intensity of the destination zone (has three levels)
Major transport hub	A binary variable indicating destination zone is a major transportation hub
Service location preference	A binary variable indicating the driver's preferred pick-up zone

of drop-off events (DDP). The DDP metric reveals major taxi trip departure and arrival locations for various times of day. For example, the DDP metric shown in Fig. 7 reveals there are more taxi passenger arrivals at the airport during the day (5 am to 4 pm) and more pick-up events in the evening and at night (5 pm to 4 am). Major trip departure and arrival locations are also more noticeable around the central business district during the morning and afternoon peak hours.

Fig. 8 shows the paths for average vacant taxi trips between drop-off and pick-up locations on weekdays during eight different daily time windows. A similar pattern that can be seen throughout the day is the high intensity of taxi activity in the city center and at major transportation hubs. This is expected, as these locations have high human activity and people who are more likely to be using taxi services. We also developed a visualization that runs in a 2D map to display the vacant and occupied taxi Origin-destination flows. The visualization is generated based on week-long taxi GPS trajectory data from September 7, 2009 to September 13, 2009. The visualization is available on YouTube at:
https://www.youtube.com/watch?v=gLvo6RvaaWg

B. Destination choice model estimation results

1) Variables definition

Two types of data that are relevant for the destination choice model were obtained:

Observed choice data:

Observed choice data describes vacant taxi trips between the drop-off and the pick-up zones. The choice is among passenger destination zones in the city of Lisbon. Out of a total of 109 destinations, 108 are considered. One of the zones has no data and is not included as a choice. The destination choice models are created with a total of 29,053 observed choices on weekdays.

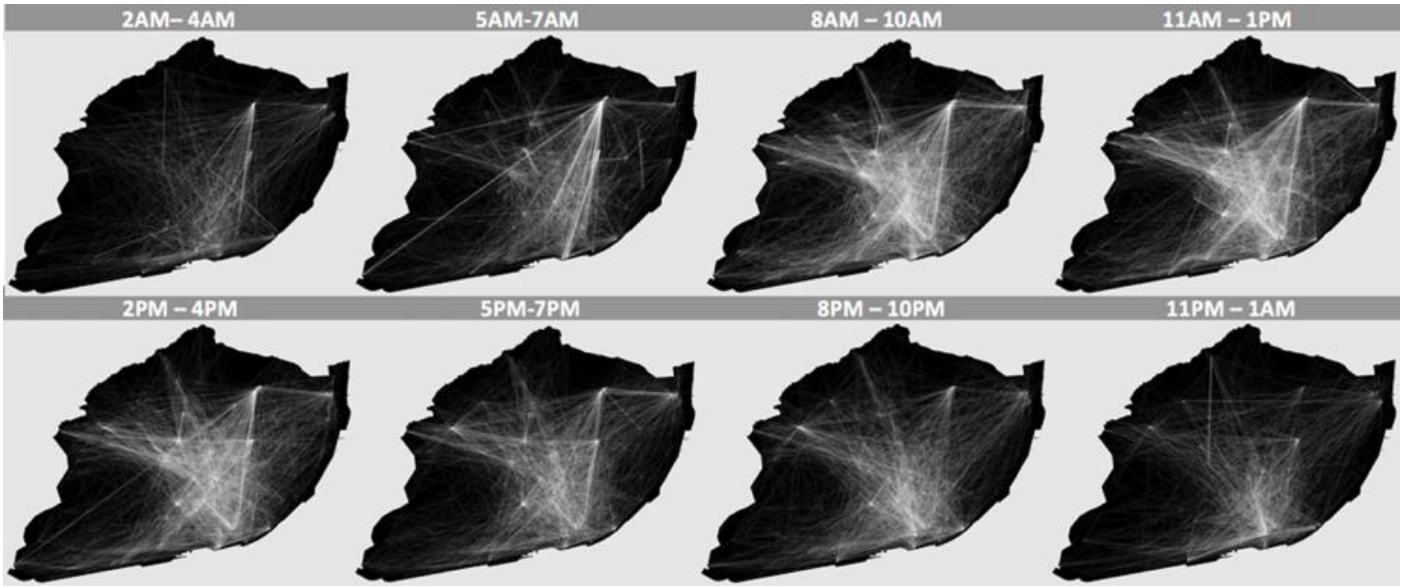


Fig. 8. Average weekday vacant taxi OD flow patterns (Origin: drop-off, and destination: pick-up).

TABLE II
ESTIMATION RESULTS OF GROUP 1 MODELS

Variable	Model 1.1 (5Am to 7AM)		Model 1.2 (8AM to 4PM)		Model 1.3 (5PM to 7PM)		Model 1.4 (8PM to 10PM)		Model 1.5 (11PM to 4AM)	
	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
# Employees	7.7886e-05	< 0.001	4.2508e-05	< 0.001	2.5092e-05	0.033	--	--	--	--
# POIs	1.0913e-03	< 0.001	7.8555e-04	< 0.001	--	--	--	--	--	--
# Hourly trip ends:5AM to 7AM	2.1185e-02	< 0.001	--	--	--	--	--	--	--	--
# Hourly trip ends:8AM to 4PM	--	--	3.5447e-01	< 0.001	--	--	--	--	--	--
# Hourly trip ends:5PM to 7PM	--	--	--	--	3.9698e-03	< 0.001	--	--	--	--
# Hourly trip ends:8PM to 10PM	--	--	--	--	--	--	4.2864e-03	< 0.001	--	--
# Hourly trip ends:11PM to 4AM	--	--	--	--	--	--	--	--	3.0073e-02	< 0.001
Waiting time (wt_j):8AM to 4PM	--	--	-2.3706e-02	< 0.001	--	--	--	--	--	--
Combined travel time (ttc_{ij})	-6.2551e+00	< 0.001	-7.7112e+00	< 0.001	-7.3912e+00	< 0.001	-6.1797e+00	< 0.001	-5.1653e+00	< 0.001
High hotspot	8.5299e-01	< 0.001	1.4295e+00	< 0.001	1.7274e+00	< 0.001	2.2478e+00	< 0.001	1.8654e+00	< 0.001
Medium hotspot	3.9195e-01	< 0.001	9.7819e-01	< 0.001	8.9255e-01	< 0.001	1.2174e+00	< 0.001	1.2871e+00	< 0.001
Major transport hub	4.5020e-01	< 0.001	2.6005e-01	< 0.001	2.1818e-01	< 0.001	--	--	--	--
Dissimilarity parameter (λ_k)	--	--	--	--	--	--	--	--	--	--
Sample Size	2692		17028		4044		2841		2448	
Percent correct index	51.26%		63.96%		62.39%		60.19%		48.33%	
Null log-likelihood (L(0))	-10635.0		66858.0		-15860.0		-11114.0		-9570.9	
Final log-likelihood (L(β))	-6329.6		-33849		-7743.2		-6000.7		-6008.7	
Rho-squared (ρ^2(0))	0.4048		0.4937		0.5118		0.4601		0.3722	

Explanatory data:

The explanatory variables considered in the developed destination choice models are grouped into three classes (TABLE I): (i) impedance variables like travel time, and travel distance describe the connectivity between drop-off and pick-up pairs; (ii) zonal variables like size variables regarding the number of employees, number of hourly trip ends, number of points of interests in the destination zone represent the number of opportunities available in the destination zone; and (iii) user variables like service location preference represent the user's characteristics (in this case, a taxi driver). A more detailed discussion of the explanatory variables is available in the Appendix Section.

TABLE III
ESTIMATION RESULTS OF GROUP 2 MODELS

Variable	Model 2.1 (5Am to 7AM)		Model 2.2 (8AM to 4PM)		Model 2.3 (5PM to 7PM)		Model 2.4 (8PM to 10PM)		Model 2.5 (11PM to 4AM)	
	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
# Employees	9.0293e-05	< 0.001	4.9125e-05	< 0.001	2.7989e-05	0.018	--	--	--	--
# POIs	1.1673e-03	< 0.001	6.3142e-04	< 0.001	--	--	--	--	--	--
# Hourly trip ends:5AM to 7AM	2.2423e-02	< 0.001	--	--	--	--	--	--	--	--
# Hourly trip ends:8AM to 4PM	--	--	3.7544e-01	< 0.001	--	--	--	--	--	--
# Hourly trip ends:5PM to 7PM	--	--	--	--	4.1286e-03	< 0.001	--	--	--	--
# Hourly trip ends:8PM to 10PM	--	--	--	--	--	--	4.3946e-03	< 0.001	--	--
# Hourly trip ends:11PM to 4AM	--	--	--	--	--	--	--	--	3.0674e-02	< 0.001
Waiting time (wt_j):8AM to 4PM	--	--	-2.5643e-02	< 0.001	--	--	--	--	--	--
Combined travel time (ttc_{ij})	-6.5820e+00	< 0.001	-7.1957e+00	< 0.001	-7.6459e+00	< 0.001	-6.4365e+00	< 0.001	-5.4776e+00	< 0.001
High hotspot	1.6928e+00	< 0.001	1.8192e+00	< 0.001	2.4783e+00	< 0.001	3.0955e+00	< 0.001	2.7764e+00	< 0.001
Medium hotspot	1.0668e+00	< 0.001	1.3161e+00	< 0.001	1.4957e+00	< 0.001	1.8860e+00	< 0.001	2.0318e+00	< 0.001
Major transport hub	4.5750e-01	< 0.001	1.9253e-01	< 0.001	2.1513e-01	< 0.001	--	--	--	--
Dissimilarity parameter (λ_k)	--	--	8.8202e-01	< 0.001	--	--	--	--	--	--
Sample Size	2692		17028		4044		2841		2448	
Percent correct index	50.22%		63.58%		62.12%		60.16%		49.10%	
Null log-likelihood (L(0))	-12604.3		-79727.3		-18934.5		-13301.9		-11462.0	
Final log-likelihood (L(β))	-6612.3		-34871.0		-7953.9		-6153.3		-6198.7	
Rho-squared (ρ^2(0))	0.4754		0.5626		0.5799		0.5374		0.4592	

The model specification of the utility function containing all the variables discussed in TABLE I is shown in Equation (6):

$$V_{ij} = \alpha_1 tt_{ij} + \alpha_2 td_{ij} + \alpha_3 ttc_{ij} + \alpha_4 tdc_{ij} + \alpha_5 wt_j + \beta S_j + \gamma_1 dummy_{high\_hotspot} + \gamma_2 dummy_{medium\_hotspot} + \gamma_3 dummy_{major\ transport\ hub} + \gamma_4 dummy_{service\ location\ preference} \quad (6)$$

Where α, β, γ are coefficients for their corresponding explanatory variables.

2) Estimation results

Taxi drivers use different passenger searching strategies in terms of zonal choices, in that they are reluctant to serve a specific area of the city during day time. They tended to circulate within the downtown area of Lisbon, which has very high taxi demand. They also tended to travel to high demand areas in the outskirts of the city during the evening and night periods (Fig. 7). We estimated destination choice models for five different periods of the day to explain the varying passenger searching behaviour of taxi drivers, which is strongly related to taxi passenger demand over time and space: Model 1 (5AM to 7AM); Model 2 (8AM to 4PM); Model 3 (5PM to 7PM); Model 4 (8PM to 10PM); and Model 5 (11PM to 4AM). Destination choice models were estimated using mlogit, a package for the R-programming environment.

Our first step is to analyze correlation within the explanatory variables. We found a high positive correlation between the following explanatory variables: travel time, travel distance, combined travel time, and combined travel distance. For the purposes of the models we developed, a correlation coefficient greater than 0.4 is considered strong.

A combination of different variables is examined to estimate the destination choice models for five different periods. A Bayesian Information Criterion (BIC) evaluation was performed to choose the appropriate utility function. The function with the lowest BIC value was applied for the estimation of the models for each period. First, we have estimated five models using MNL model structure (Group 1 Models). Group 1 Models are estimated using a restricted set of zonal alternatives rather than a full set. The estimation results of Group 1 Models is presented in Table II, where only statistically significant estimates are retained (P-value < 0.05). Second, we have estimated ten models using nested logit model structure (Group 2 Models, and Group 3 Models). Using nested logit structure, we tested how the different ways of defining driver's service location preference would influence model fit. The estimation results of the models using combined travel time (ttc_{ij}) variable is presented in Table III (Group 2 Models). We test how the introduction of service location preference dummy variable would lead to different models estimates and results are presented in Table IV (Group 3 Models).

The estimation results of Group 1 Models, which is based on MNL model structure with a restricted set of zonal alternatives, are presented in Table II. We used Rho-square ($\rho^2(0)$) to evaluate the overall quality of fit of the estimated models: $\rho^2(0) = 1-(L(\beta)/L(0))$. The $\rho^2(0)$ value compares the fit of the model with the vector of parameters β against the model with all parameters set to 0. The model fit measured in terms of $\rho^2(0)$ varied between 0.3722 (Model 1.5) and 0.5118 (Model 1.3), which is indicative of a very good fit for the models [17]. The models contain between four and eight parameters. All models contain at least an impedance variable and a size variable. The

TABLE IV
ESTIMATION RESULTS OF GROUP 3 MODELS

Variable	Model 3.1 (5Am to 7AM)		Model 3.2 (8AM to 4PM)		Model 3.3 (5PM to 7PM)		Model 3.4 (8PM to 10PM)		Model 3.5 (11PM to 4AM)	
	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
# Employees	9.1821e-05	< 0.001	5.6913e-05	< 0.001	2.4636e-05	0.022	--	--	--	--
# POIs	7.2627e-04	< 0.001	2.8672e-04	0.001	--	--	--	--	--	--
# Hourly trip ends:5AM to 7AM	1.7247e-02	< 0.001	--	--	--	--	--	--	--	--
# Hourly trip ends:8AM to 4PM	--	--	3.8454e-01	< 0.001	--	--	--	--	--	--
# Hourly trip ends:5PM to 7PM	--	--	--	--	3.6743e-03	< 0.001	--	--	--	--
# Hourly trip ends:7PM to 10PM	--	--	--	--	--	--	3.3398e-03	< 0.001	--	--
# Hourly trip ends:11PM to 4AM	--	--	--	--	--	--	--	--	2.3346e-02	< 0.001
Waiting time (wt_j):8AM to 4PM	--	--	-4.3646e-02	< 0.001	--	--	--	--	--	--
Travel time (tt_{ij})	-4.1789e-01	< 0.001	-4.8956e-01	< 0.001	-4.8641e-01	< 0.001	-3.9577e-01	< 0.001	-3.3065e-01	< 0.001
Service location preference	1.7500e+00	< 0.001	1.3908e+00	< 0.001	1.6895e+00	< 0.001	1.6596e+00	< 0.001	2.0058e+00	< 0.001
High hotspot	1.6477e+00	< 0.001	1.9310e+00	< 0.001	2.4119e+00	< 0.001	3.0664e+00	< 0.001	2.7543e+00	< 0.001
Medium hotspot	1.0534e+00	< 0.001	1.4322e+00	< 0.001	1.5113e+00	< 0.001	1.9079e+00	< 0.001	2.0549e+00	< 0.001
Major transport hub	4.3607e-01	< 0.001	1.4561e-01	0.009	--	--	--	--	--	--
Dissimilarity parameter (λ_k)	--	--	9.5001e-01	< 0.001	--	--	--	--	--	--
Sample Size	2692		17028		4044		2841		2448	
Percent correct index	53.01%		61.96%		60.26%		56.51%		51.51%	
Null log-likelihood (L(0))	-12604.3		-79727.3		-18934.5		-13301.9		-11462.0	
Final log-likelihood (L(β))	-6356.2		-34033		-7767.6		-6071.4		-5829.7	
Rho-squared ($\rho^2(0)$)	0.4957		0.5731		0.5898		0.5671		0.4914	

size variables (#Employees, #POIs, and #Hourly trip ends) are significant in all models. The parameter's sign for #Employees and #Hourly trip ends is positive, indicating that taxi drivers are more willing to choose a destination with high human activity where more people are likely to be using taxi services. The negative and significant coefficients for the combined travel time indicated that the attractiveness of a destination decreases with longer travel time.

Group 2 Models and Group 3 Models are estimated based on a full set of alternatives using nested logit model structure. Hypothesis tests on the correlations within the ten nested logit models are used to examine whether the correlations in unobserved factors over alternatives within each nest are zero. We perform hypothesis test that the dissimilarity parameter is 1, which is the value that it takes for a standard logit model. Except for Model 2.2 and Model 3.2, we are not able to reject the hypothesis that the true model is a standard logit at 95% confidence. Thus, for the remaining eight models, the nested logit structure collapses to the multinomial logit model.

The estimation results of Group 2 Models are presented in Table III. Except Model 2.2, the remaining four models are equivalent to their corresponding models in Table II (models estimated with a restricted set of zonal alternatives). However, compared to their corresponding models in Table II, Model 2.1, Model 2.3, Model 2.4 and Model 2.5 in Table III have high $\rho^2(0)$ values. The coefficient for the dissimilarity parameter (λ_k) in Model 2.2 is 0.88202, which is designed to be equal across nests and capture the general correlation between alternatives. The correlation is approximately 1- 0.88202 = 0.11798, which is a small correlation. The nested logit model is compatible with the random utility maximization behaviour for all possible values of the explanatory variables if $\lambda_k \forall\, k$ is between zero and one [56].

We estimated Group 3 Models by adding a dummy variable to measure the influence of a driver's preference of service area in their choice of passenger pick-up location. The estimation results of these models are presented in Table IV. Compared to their corresponding models in Table II and Table III, all the five estimated models in Table IV have fairly high $\rho^2(0)$ values, ranging from 0.4914 to 0.5898, which is an indication of better model fit to the data. The positive and significant service location preference dummy variable coefficient indicated a preference for destinations that the drivers are usually visiting for that purpose. It is also noted that the day time model (Model 3.2) has smaller service location preference dummy variable coefficient relative to the night time and early morning models. This indicates that taxi drivers are reluctant to service only a specific part of the city during day time, which is consistent with what is observed from traditional taxi operational modes. Shapiro [6] noted that hailing can quickly result in a match in cities with high population densities. The stand and dispatching modes are more common when the demand for taxis is low, which in this case is the night time and early morning (Fig. 5). The travel time variable reflects the travel cost between the drop-off and pick-up zones. The travel time parameter's sign is negative in all models, which means that taxi drivers are more willing to travel to the nearest zone than they are to travel to a farther zone to pick up their next passenger.

The size variable (#POIs) parameter's sign is negative and statistically significant for Group 3 (5pm to 7pm) and Group 4 (8pm to 10pm) models, and negative and statistically insignificant for Group 5 models (11PM to 4AM). The size variable measures the number of opportunities for passenger pick-up at each destination, which suggests that the POIs variable should positively influence vacant taxi trips. Thus, the POIs variable is not included in the aforementioned models. POIs data are mainly composed of service, recreation, office, education, health, and shopping facilities. These facilities are deserted during the evening and night times especially in the downtown area, which has a large concentration of office buildings. This could be one of the reasons for the negative parameter sign. In this study, POIs classes are not analyzed explicitly because of lack of POIs labels. Future studies should explore the inclusion of different POIs classes in the models (e.g., predominantly "office building" POIs for day time model, and predominantly service POIs for evening time models).

The positive signs for the $high\_hotspot$, $medium\_hotspot$, and $major\ transport\ hub$ parameters indicate that when the transportation demand and supply is high, the corresponding zone's utility will also be high. In other words, taxi drivers prefer a passenger pick-up destination in busy transportation cores.

We estimated three destination choice models by adding a waiting time variable to measure the influence of intra-zonal waiting time on a driver's choice of passenger pickup location. The intra-zonal waiting time is calculated by averaging all the waiting time the taxi drivers face before succeeding in finding the next passenger within the boundaries of each destination zone. There was lack of observations for some of the destination zones in four of the model estimation periods. To estimate the logit models, a complete waiting time variable for all the destination zones is required. Thus, we have only estimated three models (Model 1.2, Model 2.2, and Model 3.2) by adding the waiting time variable.

We also performed the likelihood ratio (LR) test to examine whether the observed difference in model fit is statistically significant between the final model and the null model (model with no parameters). The LR test for each model shows that the final model fits significantly better than the model with no parameters.

We also added the percent correct index, which is the percentage of observations where the model assigns the highest probability of choice to the alternative actually selected. While this has appeal because it is easily appreciated intuitively, it may be misleading. For example, compared to their corresponding models in Table II and Table III, Model 3.2 and Model 3.3 in Table IV have high Rho-squared values. Rho-squared is very sensitive, even the differences in its value as small as 0.01 can be indicative that one model has a better fit than another. However, Model 1.2 and Model 1.3 in Table II; and Model 2.2 and Model 2.3 in Table III have high percent correct indexes. Percent correct index is often included for information purpose only and it should not be used to make decisions about the appropriateness of utility function

specification. In cases where many choices are at stake requiring a high percent correct index makes no sense since there are many competitive alternatives with similar estimated probabilities of being chosen. Naturally using the highest probability for estimating a choice is going to be misleading. In a context of a Monte Carlo process then the use of the estimated probabilities will be able to reproduce the aggregated number of visits of taxis to different areas of the city.

One of the major benefits of the developed models is to explain the way in which taxi drivers choose among different passenger pickup zones. Modeling taxi driver's passenger pickup location choice behaviour is important to the evaluation of taxi driver's perceptions of pickup location characteristics. The developed models can also be used to forecast the taxi drivers behaviour under hypothetical scenarios. An example of such a scenario is the impact of new zonal attributes (e.g., a newly opened major transport hub) on vacant taxi traffic to the area. There has been a number of studies aimed at modeling occupied taxi trips. For instance, Liu et al. [57], Werabhat et al. [58], and Zhang et al. [59] estimated the O-D trips of occupied taxis using GPS data. Further improvements on the aforementioned studies were achieved through the development of trip distribution models [20][19]. However, the aforementioned models do not account for the traffic generated by vacant taxi movements. The developed destination choice models not only helping analyzing and understanding taxi drivers' behaviour, but also constitute an essential part of trip distribution modeling methods [20].

3) Model evaluation

The performances of the estimated models are evaluated using trip length (in minutes) frequency distribution and Coincidence Ratios (CR). Fig. 9 shows a comparison of the estimated and observed trip lengths for all the models. In general, all the models show good estimation results in terms of reproducing the observed vacant taxi trip lengths. Compared to the other models (models estimated without the waiting time variable), the day time models (8AM – 4PM) perform well in terms of reproducing the short vacant taxi trips (Fig. 9b).

The CR is used to quantitatively measure how well the estimated trip length frequency distribution overlaps with the observed trip length frequency distribution. The CR can be calculated from Equation (7).

$$CR = \frac{\sum_t \min\left(obs_t, est_t\right)}{\sum_t \max\left(obs_t, est_t\right)} \tag{7}$$

TABLE V
COINCIDENCE RATIOS

Model	Average Trip Length (in minutes)								
	Group 1 Models			Group 2 Models			Group 3 Models		
	$\sum_t min(obs_t, est_t)$	$\sum_t max(obs_t, est_t)$	CR	$\sum_t min(obs_t, est_t)$	$\sum_t max(obs_t, est_t)$	CR	$\sum_t min(obs_t, est_t)$	$\sum_t max(obs_t, est_t)$	CR
Model 1 (5AM to 7AM)	0.77	1.23	0.63	0.77	1.23	0.63	0.80	1.20	0.67
Model 2 (8AM to 4PM)	0.81	1.19	0.69	0.83	1.17	0.72	0.84	1.16	0.73
Model 3 (5PM to 7PM)	0.79	1.21	0.65	0.80	1.20	0.67	0.80	1.20	0.67
Model 4 (8PM to 10PM)	0.77	1.23	0.63	0.78	1.22	0.64	0.82	1.18	0.70
Model 5 (11PM to 4AM)	0.82	1.18	0.70	0.80	1.20	0.67	0.84	1.16	0.72

Where, CR is the coincidence ratio; obs_t is the proportion of observed distribution in interval t; est_t is the proportion of estimated distribution in interval t.

Table V shows that the estimated models perform well, with an average CR value of 0.66, 0.67, and 0.70 for Group 1, Group 2, and Group 3 Models, respectively. Although the Group 3 Models perform well overall, a few of them apparently will need some calibration work. For instance, Model 3.1 and Model 3.3 overestimate the short vacant taxi trips. Comparison of our CR values to the CR values of earlier studies reveals that the estimated destination choice models perform well [45]. In addition, the CR value for each model is well above the minimum threshold of 0.6 prescribed by the travel demand model report in [60].

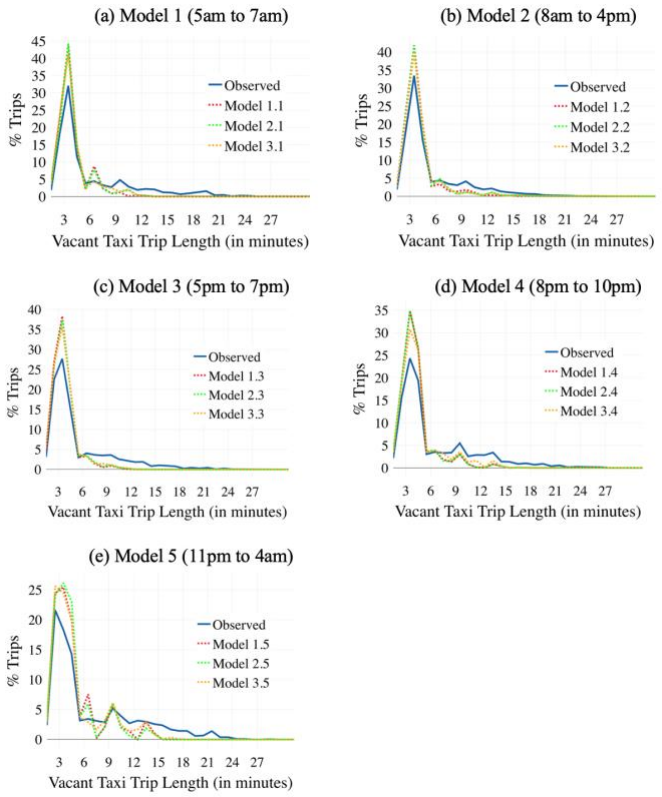


Fig. 9. Frequency distributions of observed and estimated vacant taxi trip lengths (in minutes).

V. Conclusion

This study attempts to understand taxi travel demand from the perspective of modeling vacant taxi trips that are made between passenger drop-off and passenger pick-up locations. Vacant taxi trips are the result of passenger searching attempts. Thus, a taxi driver's next passenger pick-up location choice can be framed as destination choice problem. We explored the possibility of using vacant taxi trips extracted from taxi GPS trajectory data to develop destination choice models with the discrete choice model structure such as nested logit and multinomial logit.

Modeling taxi driver's passenger pickup location choice behaviour is important to explain the way in which taxi drivers choose among different passenger pickup zones. For example,

several factors that are likely to influence a taxi driver's next choice of passenger pick-up location are identified. Variables that positively influence vacant taxi trips include size (#Employees and #Hourly trip ends), hotspot locations for taxi pick-up, service location preference dummy variable, and major transport hubs. The behavior model shows us where taxi drivers would like to go next and not where exactly they should go given existing competition. A potential future improvement include developing an intelligent taxi management system based on model prediction information.

The results of this study can also be used to support long-term strategic planning specially to model the pattern of trips generated by vacant taxis. Usually, the focus of a trip distribution model is to distribute occupied taxi trips (from a trip generation model) among destinations. The results of the destination choice models provide insights regarding the factors that explain the taxi driver's probability to choose a certain zone within a set of passenger pick-up zones, contributing to a better understanding of taxi drivers travel behaviour. Hence, results of this study can be used to develop a trip distribution model to distribute vacant taxi trips in the City of Lisbon.

Despite the relevance of our analysis, we should emphasize some limitations of the study. In our approach, we make an assumption that the observed passenger pick-up location is similar to where the driver intended to go right after he/she dropped off passengers. In reality, for example, a taxi driver could pick-up a passenger while traveling to a high demand area but then he/she finds someone on the way in a low demand area. In Fig. 3, we showed that it is difficult to explain some of the decisions that are made by a driver simply based on the GPS trajectory data.

One of the main challenges of studies that merge data from multiple sources is the reconciliation of the spatial and the temporal dimensions of the data. One of the limitations of this study is the discrepancy in time between the taxi dataset and the Foursquare datasets. The benefits of using multiple data sources depend on what they add to a particular piece of research. For this study, despite the discrepancy in time between the taxi and the Foursquare datasets, we believe that the insights gained from the Foursquare dataset are informative and useful for the proposed modeling framework. In addition, the destination choice models were calibrated with taxi GPS data collected in the year 2009. However, a lot has changed in the last decade in terms of urban mobility such as the emergence of new transportation network companies (e.g., Uber), new urban mobility concepts like mobility as a service, etc. Areas of future improvement include exploring the inclusion of variables related to recent urban mobility trends and realities in the model to improve model's explanatory power.

REFERENCES

[1] J. M. S. Grau and M. A. E. Romeu, "Agent based modelling for simulating taxi services," in *Procedia Computer Science*, 2015.

[2] R. Hughes and D. MacKenzie, "Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators," *J. Transp. Geogr.*, 2016.

[3] K. I. Wong, S. C. Wong, M. G. H. Bell, and H. Yang, "Modeling the bilateral micro-searching behavior for Urban taxi services using the absorbing Markov chain approach," in *Journal of Advanced Transportation*, 2005.

[4] H. Yang, C. W. Y. Leung, S. C. Wong, and M. G. H. Bell, "Equilibria of bilateral taxi-customer searching and meeting on networks," *Transp. Res. Part B Methodol.*, 2010.

[5] L. Moreira-Matias, J. Gama, M. Ferreira, and L. Damas, "A predictive model for the passenger demand on a taxi network," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2012.

[6] M. H. Shapiro, "Density of Demand and the Benefit of Uber," *Work. Pap.*, 2017.

[7] A. Lacombe and C. Morency, "Modeling taxi trip generation using GPS data: the Montreal case," *Transp. Res. Board 95th Annu. Meet.*, 2016.

[8] J. Yuan, Y. Zheng, L. Zhang, Xi. Xie, and G. Sun, "Where to find my next passenger," 2011.

[9] H. wen Chang, Y. chin Tai, and J. Y. jen Hsu, "Context-aware taxi demand hotspots prediction," *Int. J. Bus. Intell. Data Min.*, 2009.

[10] J. Lee, I. Shin, and G. L. Park, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *Proceedings - 4th International Conference on Networked Computing and Advanced Information Management, NCM 2008*, 2008.

[11] B. Li *et al.*, "Hunting or waiting? Discovering passenger-finding strategies from a large-scale real-world taxi dataset," in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops, PERCOM Workshops 2011*, 2011.

[12] D. Zhang *et al.*, "Understanding taxi service strategies from taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, 2015.

[13] C. Yang and E. J. Gonzales, "Modeling Taxi Trip Demand by Time of Day in New York City," *Transp. Res. Rec. J. Transp. Res. Board*, 2014.

[14] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Predict Travel Demand*. 1987.

[15] A. Simma, R. Schlich, and K. W. Axhausen, "Destination choice modelling for different leisure activities," *Brisk Bin. Robust Invariant Scalable Keypoints*, 2001.

[16] M. A. Pozsgay and C. R. Bhat, "Destination choice modeling for home-based recreational trips: Analysis and implications for land use, transportation, and air quality planning," in *Transportation Research Record*, 2001.

[17] S. Mishra, Y. Wang, X. Zhu, R. Moeckel, and S. Mahaparta, "Comparison between Gravity and Destination Choice Models for Trip Distribution in Maryland," in *Transportation Research Board*, 2013.

[18] H. Hammadou, I. Thomas, A. Verhetsel, and F. Witlox, "How to incorporate the spatial dimension in destination choice models: The case of Antwerp," *Transp. Plan. Technol.*, 2008.

[19] J. Tang, S. Zhang, X. Chen, F. Liu, and Y. Zou, "Taxi trips distribution modeling based on Entropy-Maximizing theory: A case study in Harbin city—China," *Phys. A Stat. Mech. its Appl.*, 2018.

[20] J. Zhu and X. Ye, "Development of destination choice model with pairwise district-level constants using taxi GPS data," *Transp. Res. Part C Emerg. Technol.*, 2018.

[21] D. McFadden, "Modelling the choice of residential location," *Spatial Interaction Theory and Planning Models*. 1978.

[22] C. R. Bhat and J. Guo, "A mixed spatially correlated logit model: Formulation and application to residential choice modeling," *Transp. Res. Part B Methodol.*, vol. 38, no. 2, pp. 147–168, 2004.

[23] Y. Wang, G. H. de A. Correia, E. de Romph, and H. J. P. (Harry. Timmermans, "Using metro smart card data to model location choice of after-work activities: An application to Shanghai," *J. Transp. Geogr.*, 2017.

[24] M. G. Demissie, "Combining datasets from multiple sources for urban and transportation planning: Emphasis on cellular network data," Coimbra University, 2014.

[25] M. G. Demissie, S. Phithakkitnukoon, T. Sukhvibul, F. Antunes, R. Gomes, and C. Bento, "Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal," *IEEE Trans. Intell. Transp. Syst.*, 2016.

[26] L. Liu, C. Andris, and C. Ratti, "Uncovering cabdrivers' behavior patterns from their digital traces," *Comput. Environ. Urban Syst.*, 2010.

[27] Z. Yang, M. L. Franz, S. Zhu, J. Mahmoudi, A. Nasri, and L. Zhang, "Analysis of Washington, DC taxi demand using GPS and land-use data," *J. Transp. Geogr.*, 2018.

[28] M. G. Demissie, G. H. de A. Correia, and C. Bento, "Exploring cellular network handover information for urban mobility analysis,"

J. Transp. Geogr., 2013.
[29] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin-destination trips by purpose and time of day inferred from mobile phone data," *Transp. Res. Part C Emerg. Technol.*, vol. 58, pp. 240–250, 2015.
[30] M. G. Demissie, S. Phithakkitnukoon, and L. Kattan, "Trip Distribution Modeling Using Mobile Phone Data: Emphasis on Intra-Zonal Trips," *IEEE Transactions on Intelligent Transportation Systems*, 2018.
[31] S. Phithakkitnukoon, T. Sukhvibul, M. Demissie, Z. Smoreda, J. Natwichai, and C. Bento, "Inferring social influence in transport mode choice using mobile phone data," *EPJ Data Sci.*, 2017.
[32] M. G. Demissie, G. H. de Almeida Correia, and C. Bento, "Intelligent road traffic status detection system through cellular networks handover information: An exploratory study," *Transp. Res. Part C Emerg. Technol.*, 2013.
[33] A. Vaccari *et al.*, "A holistic framework for the study of urban traces and the profiling of urban processes and dynamics," in *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, 2009.
[34] N. Kunama, M. Worapan, S. Phithakkitnukoon, and M. G. Demissie, "GTFS-Viz: tool for preprocessing and visualizing {GTFS} data," in *UbiComp/ISWC Adjunct*, 2017.
[35] J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.
[36] X. Wan, J. Kang, M. Gao, and J. Zhao, "Taxi origin-destination areas of interest discovering based on functional region division," in *2013 3rd International Conference on Innovative Computing Technology, INTECH 2013*, 2013.
[37] S. Phithakkitnukoon, M. Veloso, C. Bento, A. Biderman, and C. Ratti, "Taxi-aware map: Identifying and predicting vacant taxis in the city," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6439 LNCS, pp. 86–95.
[38] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," *IEEE Trans. Knowl. Data Eng.*, 2013.
[39] Yu Jianxin, Zhou Xiaomin, and Zhao Hongyu, "Design and implementation of taxi calling and dispatching system based on GPS mobile phone," 2009.
[40] C. Wang, Y. Hou, and M. Barth, "Data-Driven Multi-step Demand Prediction for Ride-Hailing Services Using Convolutional Neural Network," in *Advances in Intelligent Systems and Computing*, 2020.
[41] Wang Chao, Hao Peng, Wu Guoyuan, Qi Xuewei, Barth Matthew, "Predicting the Number of Uber Pickups by Deep Learning," in *Transportation Research Board 2018*, 2018.
[42] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," 2011.
[43] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, "The Rich and the Poor: A Markov Decision Process Approach to Optimizing Taxi Driver Revenue Efficiency," in *CIKM '16 Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 2016.
[44] J. de D. Ortúzar and L. G. Willumsen, *Modelling Transport*. 2011.
[45] B. Mei, "Destination Choice Model for Commercial Vehicle Movements in Metropolitan Area," *Transp. Res. Rec. J. Transp. Res. Board*, 2013.
[46] A. Kinjarapu, "Analysis of Truck Travel behaviour using passive GPS data – Case study of Calgary Region, Alberta," University of Calgary, 2018.
[47] S. Nerella and C. R. Bhat, "Numerical analysis of effect of sampling of alternatives in discrete choice models," in *Transportation Research Record*, 2004.
[48] R. A. Becker *et al.*, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Comput.*, 2011.
[49] B. C. Csáji *et al.*, "Exploring the mobility of mobile phone users," *Phys. A Stat. Mech. its Appl.*, 2013.
[50] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in Urban data collection," *IEEE Pervasive Comput.*, 2007.
[51] K. E. Train, *Discrete choice methods with simulation*. 2003.
[52] S. Portugal, "Statistics Portugal," *The Instituto Nacional de Estatística (INE)*, 2018. [Online]. Available: http://www.ine.pt/xportal/xmain?xpgid=ine_main&xpid=INE. [Accessed: 28-May-2018].
[53] Geotaxi, "Geotaxi," 2012. [Online]. Available: http://www.geotaxi.com/. [Accessed: 28-May-2012].
[54] D. Yang, D. Zhang, B. Qu, and P. Cudré-Mauroux, "PrivCheck: Privacy-Preserving Check-in Data Publishing for Personalized Location Based Services," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, 2016.
[55] Google, "Distance Matrix API," *Google*, 2018. .
[56] B. Lee, "Calling patterns and usage of residential toll service under self-selecting tariffs," *J. Regul. Econ.*, vol. 16, pp. 45–82, 1999.
[57] Y. Liu, C. Kang, S. Gao, Y. Xiao, and Y. Tian, "Understanding intra-urban trip patterns from taxi trajectory data," *J. Geogr. Syst.*, 2012.
[58] W. Mungthanya *et al.*, "Constructing Time-Dependent Origin-Destination Matrices With Adaptive Zoning Scheme and Measuring Their Similarities With Taxi Trajectory Data," *IEEE Access*, 2019.
[59] W. Zhang, S. Li, and G. Pan, "Mining the semantics of origin-destination flows using taxi traces," 2012.
[60] C. of Peterborough, "Travel demand model report. City of Peterborough comprehensive transportation plan update supporting document.," 2012.

APPENDIX

Explanatory Variables:

Travel time (tt_{ij}) and travel distance (td_{ij}) are used as impedance variables that are calculated based on the time and distance between the drop-off and pick-up locations of vacant taxi trips. There was lack of observations between some origin and destination pairs. To estimate the logit models, a complete impedance variable matrix is required. We address this problem by substituting the travel time and travel distance values with values obtained from the Google Distance API.

Service location preference*:* three variables are developed to incorporate the effect of Service location preference variable on the location choice of taxi drivers for passenger pick-up. First, each driver's preferred pick-up location is calculated based on their most frequently-visited locations for passenger pick-up service. The service location preference is an individual specific variable that does not vary across destinations. One of the common methods to represent the effect of the service location preference variable on destination choice is to interact it with the alternative specific variable. Then, a generic coefficient is estimated for each interacted variable.

To start the process, we interact the service location preference variable with impedance variables to generate the first two variables, such as combined travel time, and combined travel distance, as shown in the following. To obtain these variables, each driver's average travel time (att_{ij}) and average travel distance (atd_{ij}) from the centroids of all the other zones to the centroid of their preferred pick-up zone are calculated. These variables are used to represent a driver's preferred pick-up location or zone. The att_{ij} and atd_{ij} variables are specific to an individual. Thus, combined travel time ($\mathrm{ttc}_{ij} = \mathrm{tt}_{ij} \times \mathrm{att}_{ij}$) and combined travel distance ($\mathrm{tdc}_{ij} = \mathrm{td}_{ij} \times \mathrm{atd}_{ij}$) variables are used to incorporate a driver's preferred pick-up location for the model's estimation purposes.

The third variable is a dummy variable called $service\ location\ preference$. This dummy variable takes a value of 1 if the passenger pick-up zone is the driver's preferred

service location and 0 otherwise.

Size variable (S_j) First, the sampled vacant taxi trip ends (pick-ups) are expanded to represent the mobility behavior of the total vacant taxi population. Then, for each time interval, the number of hourly vacant taxi trip ends per TAZ are calculated. Number of employees and number of POIs are the other size variables considered.

Hotspot describes the passenger pick-up intensity of a zone. The busiest zones are labeled $high\_hotspot$ and the rest are labeled as $medium\_hotspot$ or $low\_hotspot$ based on the assigned cut-off points. In this study, a preliminary data analysis of the number of pick-up events shows that the 85th and 50th percentiles are reasonable threshold values to group the variables into three categories. The $low\_hotspot$ variable is used as reference variable.

Major transport hub is a dummy variable representing major transport hubs in the municipality of Lisbon. This dummy variable has two levels: 1 if a vacant taxi trip ends in a zone with a major transport hub and 0 otherwise.

Merkebe Getachew Demissie received the M.Sc. degree in transport systems from Royal Institute of Technology (KTH), Stockholm, Sweden, in 2009; and the Ph.D. degree in transportation systems from the MIT-Portugal program, in 2014. He is currently a Research Associate at the Department of Civil Engineering, University of Calgary, Canada. Before joining Calgary University, he was a Postdoctoral Fellow at the University of Coimbra, and at the Instituto Pedro Nunes, Portugal. His main research interests are transport demand modeling, intelligent transportation systems, data mining, and machine learning.

Lina Kattan is a Professor of Civil Engineering of University of Calgary, Canada. She is also holds an Urban Alliance Professorship in Transportation systems Optimization. Lina's research program focuses on advanced traffic management and information systems, including, Intelligent Transportation Systems (ITS), traffic control, application of artificial intelligence to ITS, Connected and Autonomous Vehicle, network microsimulation modeling and analysis, dynamic traffic assignment, dynamic demand modeling and traveler behavioral modeling in response to Traffic and Transit information.

Santi Phithakkitnukoon received B.S. and M.S. degrees in electrical engineering from Southern Methodist University, Dallas, USA in 2003 and 2005 respectively, and Ph.D. in computer science and engineering from the University of North Texas, USA. He is currently an Associate Professor at the Department of Computer Engineering, Chiang Mai University, Thailand. Before joining Chiang Mai University, he was a Lecturer in Computing at The Open University, UK, a Research Associate at Newcastle University, UK, and a Postdoctoral Fellow at the SENSEable City Laboratory of the Massachusetts Institute of Technology, USA. His research is in the area of urban informatics.

Gonçalo H. A. Correia received his Ph.D. degree in Transportation from the University of Lisbon, Portugal, in 2009. From 2008 to 2014, he has been an Assistant Professor with the Civil Engineer Department, University of Coimbra, Portugal. Since 2014 he started as an Assistant Professor at the Department of Transport & Planning at the Faculty of Civil Engineering and Geosciences at the Delft University of Technology, Netherlands. He is currently leading the hEAT lab (Research on Automated and Electric Transport). He is an invited Lecturer in Beijing Jiaotong University, China, where he teaches operations research on the TUDelft+BJTU joint bachelor in Transportation. His main research interest is in the planning and operations of transport systems in urban environments with the objective of sustainable development. He focuses particularly on the use of Transport Demand Management strategies and innovative services, such as ridesharing and carsharing, to tackle urban congestion, which he studies using mainly operations research with a particular focus on optimization and simulation techniques. At TU Delft he is looking at the impacts of automated driving and electric vehicles on mobility and urban development. He is currently serving as Associate Editor for the IEEE-ITS-M.

Marco Veloso is an Adjunct Professor at Polytechnic Institute of Coimbra, Portugal, and Researcher at Center for Informatics and Systems of University of Coimbra, where he is a member of the Ambient Intelligence laboratory. His research explores the use of data mining techniques on big data for Smart Cities and Intelligent Transportation Systems. He received Ph.D., M.S., and B.S. in Informatics Engineering from University of Coimbra.

Carlos Bento is currently Associate Professor with Habilitation at University of Coimbra, Coimbra, Portugal, where he is the Director of the Ambient Intelligence Laboratory, Centre for Informatics and Systems. He is also currently the Director of the Laboratory on Informatic Systems at Instituto Pedro Nunes (IPN), Coimbra. He has over 100 publications comprising papers in international journals and conferences and book chapters. In the past years, his research has addressed the role of urban data on improving decisions for better quality of life in urban areas.