Title**Mining factors affecting taxi drivers' incomes using GPS trajectories****Author Names and Affiliations**

Guoyang Qin ^a, Tienan Li ^a, Bin Yu ^b, Yunpeng Wang ^b, Zhenhua Huang ^c, Jian Sun ^{a, *}

Abstract

Taxis provide essential transport services in urban areas. In the taxi industry, the income level remains a cause of concern for taxi drivers as well as regulators. Mining underlying factors affecting the income level will not only benefit the newcomers and low-income drivers but also assist in developing effective optimization algorithms for taxi operations. This paper intends to disclose the factors affecting incomes along with their quantitative influence by mining over 167 million GPS records from nearly 8000 taxis in Shanghai. We first identify a marked difference in drivers' incomes and categorize drivers into three income levels accordingly. We next investigate the overall search-delivery process, thereby defining several factors that may affect the income level. We then develop a generalized multi-level ordered logit (GMOL) model to find the significant factors that influence incomes. Finally, we compute the elasticity for those significant factors and present their contributions, as well as challenge some preconceived ideas regarding how to earn high incomes.

Keywords

Taxi drivers; GPS trajectories; Level of income; Influencing factor; Ordered logit model; Elasticity analysis

1. Introduction

The taxi plays a significant role in transportation systems in the urban areas across the world, providing flexible and demand-responsive transport service with widely spatiotemporal coverage. The taxi industry is often the largest employer in an urban community; therefore it will touch the lives of a significant proportion of a city population (Cooper et al., 2010). However, inefficient regulations, intense external competitions as well as soaring costs have caused constant concern among taxi drivers in many cities around the world, thereby inducing strikes, protests or demonstrations over decades (BBC News, 2013; Branigan, 2008; The Irish Times, 2009; Untapped Cities, 2015). These actions can be mainly attributed to concerns about low incomes.

To tackle the tough issue, many researchers have examined the economic consequences of regulatory restraints, thereby providing information for government decision-making regarding specific regulations (Yang et al., 2002). In addition to such regulator-side studies, an increasing number of driver-side studies emerged in the wake of available taxi GPS data (Cao et al., 2005; Huang et al., 2015; Li et al., 2009; Veloso et al., 2011; Yuan et al., 2013; Zheng et al., 2011; Zheng et al., 2009). Some of them specifically focused attentions on the low-income problem and intended to seek underlying crowd intelligence from GPS patterns, in which several factors affecting incomes were explored respectively, including duration and distance (Liu et al., 2010; Sirisoma et al., 2010) of passenger search/delivery, fare (Liu et al., 2010), likelihood of pickups (Qu et al., 2014; Rong et al., 2016), operational region (Dong et al., 2016; Gao et al., 2012; Zhang et al., 2015), ridership (Gao et al., 2012), and delivery speed (Zhang et al., 2015). Despite abundant factors, there is little knowledge of how much these factors will affect the income level, hence unknown significance of each factor. However, understanding the significance of each factor and its quantitative influence will not only offer detailed instructions to taxi drivers but also help researchers select most suitable factors in their optimization algorithms for taxi operations.

In this paper, we intend to move forward to explore the quantitative influence of factors affecting incomes, thereby providing a comprehensive as well as model-based insight into this critical issue. Unlike the exploratory data analyses in previous related studies, we aim to develop a regression model to interpret how much the factors will affect the incomes, and furthermore how responsive the income level is to a change in each factor. To achieve this objective, we need to tackle four challenges as follows. First, we need to separate over 167 million GPS records into individual trajectories and calculate drivers' incomes so as to categorize them into varying levels by incomes. Second, we need to discuss the overall operation process and explore main factors that may affect the income level. Third, we need to select a proper model to correlate the income level and the factors, and find their significance. Fourth, we need to understand contributions of each factor to each income level.

Three major contributions are made in this paper. First, we identify a considerable difference in taxi drivers' incomes on condition of sufficient competitiveness and equal work duration, as well as categorize drivers into three levels by their average incomes and standard deviations of rankings. Second, we investigate the overall search-delivery operation process and propose five main factors that may affect the income level, including *supply-demand ratio*, *search distance*, *search intensity* as well as *fare* and *delivery speed*, and also provide mathematical definitions of these factors. Third, we develop a generalized multi-level ordered logit (GMOL) model to find the significant factors and conduct an elasticity analysis to reveal the quantitative associations between the income level and the factors, as well as challenge some preconceived ideas of how to earn high incomes.

The rest of this paper is structured as follows. The next section discusses some studies related to taxi drivers' incomes. Section 3 describes the data and the research scope. Section 4 presents the categorization of the income level, definitions and mathematical expressions of factors affecting incomes, as well as the development of a logit model of the income level to find significant factors. In section 5, results are interpreted using the elasticity analysis and contributions of factors for each income level are also presented. This paper is concluded in section 6.

2. Overview of taxi drivers' income studies

Over the past two decades, researchers from different fields have undertaken various studies regarding taxi drivers' income problem using different methods. Since the emergence of taxi GPS tracking has facilitated and accelerated the research process, related work can be divided into two parts, *i.e.* studies in the pre-GPS-tracking and studies in the post-GPS-tracking period.

Back to the pre-GPS-tracking period, researchers mainly from economics and transportation fields examined this issue. In 1997, [Camerer et al. \(1997\)](#) embarked on an exploration of the relationship between taxi drivers' working hours and the average daily incomes using trip sheets data of New York City taxis. This empirical study refuted the economic hypothesis of positive intertemporal substitution. And they consequently proposed a daily targeting theory to explain the negative wage elasticities. The new theory has sparked a lot of discussions ever since ([Chou, 2002](#); [Crawford and Meng, 2011](#); [Farber, 2005](#); [Kőszegi and Rabin, 2006](#)). In the transportation field, Yang et al. have made persistent efforts to develop and modify their passenger search/delivery models since 1998. In an empirical study published in 2010 ([Sirisoma et al., 2010](#)), they validated their model using data from a stated preference survey of 400 Hong Kong taxi drivers' searching strategies. Analysis indicated that the trip time, toll, and waiting time are significant factors in the choice decision at the 1% level, whereas distance is significant at the 5% level. Studies in this period carried out some valuable exploration but were still constrained by the small sample size and the subjectivity of drivers' responses.

In the post-GPS-tracking period, mass taxi GPS data has reformed the study methods and attracted the focus of researchers from more fields. Related work involves mining individual travel patterns ([Zheng et al., 2011](#); [Zheng et al., 2009](#)), travel time estimation ([Zhan et al., 2013](#)), travel speed variation ([Wang et al., 2016](#)), mobile sequential recommendation ([Cao et al., 2005](#); [Huang et al., 2015](#); [Li et al., 2009](#); [Yuan et al., 2013](#)), taxi service modelling ([Wong et al., 2015](#); [Yang and Gonzales, 2017](#)), understanding taxi service strategies ([Veloso et al., 2011](#)), siting of electric taxi charging stations ([Tu et al., 2016](#); [Yang et al., 2017](#)) and so on. A few income-related studies were undertaken at the same time. [Gao et al. \(2012\)](#) developed a visualization system to analyze spatial-temporal trajectory data and identify factors that differentiate the top drivers from the ordinary drivers by incomes. They first assumed high-income drivers are more intelligent, serve more passengers, search actively, and change strategies less often when vacant. They then visualized trajectory data of taxi drivers in Shanghai using two novel visual encoding schemes, and they finally validated their assumptions. Unlike the previous researchers, [Liu et al. \(2010\)](#) compared the annual average incomes directly. They first defined two classes of drivers according to average incomes and standard deviations of daily rankings calculated by GPS data from Shenzhen taxis. From taxi driver's activity history, they then revealed that the trip distance and time of top drivers were significantly higher than ordinary drivers', while idle trip distance and time of the top drivers were significantly lower. [Zhang et al. \(2015\)](#) provided a new way to study taxi drivers' incomes. After investigating the locally-hunting and going-distant search strategies, they first proposed an optimal search strategy to increase incomes. By inspecting the delivery process, they then found that income during peak hours is more likely to increase while driving on routes with good traffic conditions (permitting a higher speed); they finally predicted the taxi drivers' incomes using their strategies. The results demonstrated that the extracted strategies well characterized the taxi drivers' driving behavior and operation performance. These studies have presented diverse perspectives and methods to investigate taxi drivers' incomes. The considered factors encompass the duration ([Liu et al., 2010](#); [Rong et al., 2016](#); [Sirisoma et al., 2010](#)) and distance ([Liu et al., 2010](#); [Sirisoma et al., 2010](#)) of passenger search/delivery, fare ([Liu et al., 2010](#)), likelihood of pickups ([Qu et al., 2014](#); [Rong et al., 2016](#)), operational region ([Dong et al., 2016](#); [Gao et al., 2012](#); [Zhang et al., 2015](#)), ridership ([Gao et al., 2012](#)), and delivery speed ([Zhang et al., 2015](#)). Despite abundant factors affecting incomes being considered, further work is still required to be conducted with regard to summarizing these factors in a comprehensive perspective, finding significant factors, and quantifying how much the incomes are in association with these significant factors.

3. Data description, study area and time span

3.1. Data acquisition

Pre-GPS-tracking-period studies were typically conducted by distributing questionnaires to taxi drivers and collecting their answers, the personal information was well retrieved and could promote significant findings. In the post-GPS-tracking period, personal information that matches the mass trajectory data is hardly retrievable due to privacy concerns. Therefore, the factors affecting incomes during the drivers' search-delivery process are paid more attention in our paper.

Taxi GPS data in this study are retrieved from Shanghai Qiangsheng Taxi Company, spanning entire 20 weekdays across May 2011. The data contain sequential trajectories and operation status for approximately 9000 taxis, including taxi ID, date, time, longitude, latitude, speed (km/h), bearing and operation status identifier (1 for hired/0 for vacant). The GPS data are collected every 10s and contain nearly 1 billion GPS records in total. Valid data of about 8000 taxis are kept after data cleaning over erroneous positioning information and operation status. Despite missing personal information, reliable conclusions could still be drawn due to the large sample size and good quality of GPS trajectories. Further work can be done to retrieve the anonymous personal information so as to report more findings.

3.2. Study area and time span

Our study area covers the entire city of Shanghai (approx. area: 6340km²), thereby containing any region of pickups and drop-offs. As shown in Fig. 1, the downtown area (approx. area: 660km²) within the Outer Ring Road contributes up to 82.25% of daily pickups and dominates the active operation regions of taxis.

PLEASE PLACE Fig. 1 ABOUT HERE

To successfully reveal factors affecting the drivers' income level on condition of sufficient competitiveness and equal work duration, a proper study time span is expected.

As for ensuring sufficient competitiveness, we pick a frequently-used taxi performance indicator, the ratio of empty/total mileage (RETM), to identify sufficient competitiveness. If RETM runs low, the supply falls nearly short of the demand, which often occurs during peak hours (*i.e.* 8:00-10:00, 17:00-19:00 as shown in Fig. 2). During this competitiveness-lacking period, it is prone to track biased ability of drivers due to easy pickups. If RETM runs high, the supply exceeds the demand, which often occurs after midnight and before dawn (*i.e.* 23:00-7:00 as shown in Fig. 2). During this period, most drivers stop running taxis; it is also prone to track biased ability of drivers due to the inadequate sample size. Only when RETM runs at a medium level does exist sufficient competitiveness. We hence omit data collected during 23:00-7:00, 8:00-10:00, and 17:00-19:00.

PLEASE PLACE Fig. 2 ABOUT HERE

As for ensuring equal work duration, our inquiry found that taxi drivers in Shanghai have a flexible meal schedule with varying length during the daytime (scattering across 11:00-14:00 and 16:00-19:30), it is thus very complicated to distinguish between the status *waiting for passengers* and the status *taking a meal*, hence hard to calculate how long the drivers are working in this period. Therefore, selecting data during the period does not ensure the equal work duration, we also omit data collected during 11:00-14:00 and 16:00-19:30.

We eventually select the residual and continuous period 20:00-23:00 as our study time. The selected data collected during 20:00-23:00 across the working days in May 2011 contains a considerable number of over 167 million GPS records.

Spatial areas are commonly segmented into small cells in specific dimensions according to related studies (Dong et al., 2016; Liu et al., 2010; Powell et al., 2011). In this paper, we segment the study area into a significant number of 100m×100m cells, and separate the time span into six 30-min intervals, as a means to obtain summary statistics (*e.g.* average pickups, historical search distance), which smooths out the data noise

and restricts the computation overhead. It is worth mentioning that the taxi routes remain a set of GPS coordinates and each coordinate obtains its spatial attribute (*e.g.* average counts of pickups) mapped from the summary statistics.

4. Methodology

In this section, we first tackle GPS trajectories from over 8000 taxis in Shanghai before quantifying the difference in drivers' incomes; we then define some candidate factors that may explain the income difference; we finally develop a logit model to find significant influencing factors and their elasticity of incomes.

4.1. Categorizing taxi drivers by the income level

It is necessary to categorize taxi drivers by the income level so as to obtain each taxi driver's performance. To draw driver samples with consistent performance, we expect to eliminate the randomness of income difference over a short period. For this purpose, we categorize taxi drivers by predefined thresholds of average incomes and standard deviations of income ranking.

The income calculation is conducted as follows. Longitudes, latitudes, and speeds of taxis are first extracted from the GPS records. Fares are then calculated based on delivery mileages and lower speed durations according to taxi rates in Shanghai. Summation of the fare for each delivery trip is thus the total income. For our study time span 20:00~23:00, Fig. 3 indicates that incomes within this three hours have a mean of 168 yuan and a standard deviation of 58 yuan. Incomes of lowest 25% of drivers remain below 129 yuan, while above 210 yuan for highest 25%. The latter group outperforms the former one by at least 71 yuan, which indicates a significant income difference.

PLEASE PLACE Fig. 3 ABOUT HERE

As presented in Fig. 4, the drivers are then categorized by average incomes and standard deviations of rankings. In this paper, taxi drivers whose standard deviations of rankings within 50th percentile are considered being consistent in performance. These consistently performed drivers are subsequently categorized into low-, medium- or high-income level according to whether their average incomes are between 0~20th, 40~60th, or 80~100th percentiles ($Y=1,2,3$ for high-, medium-, and low-income level). Gaps between each category are set to resolve the ambiguity of categorization boundary. Almost 3000 drivers are consequently drawn and categorized into low-, medium-, and high-income level (accounting for 33%, 27%, and 40% respectively). The income gap is near 45 yuan between any adjacent levels, namely an increase of 45 yuan corresponding to one level rise.

PLEASE PLACE Fig. 4 ABOUT HERE

We have so far successfully labeled each selected driver as either low-, medium-, or high-income level. We will define factors affecting the income difference according to the search-delivery operation process in the next part of this section.

4.2. Defining and calculating the factors affecting the income level

Previous studies have reported abundant factors that influence the income level, including the duration and distance of passenger search/delivery, fare, the likelihood of pickups, operational region, ridership, and delivery speed. Concerning these reported factors, we intend to discuss the overall search-delivery process in order to define factors affecting the income level in a comprehensive manner.

Fig. 5 exemplifies a typical search-delivery process ($\alpha \rightarrow \beta \rightarrow \gamma$). A studied taxi is about to search for next passengers from the drop-off location α . Supply and demand of the taxi service around the location α may impact on subsequent search time and incomes. If quantity supplied falls nearly short of quantity demanded (indicating low supply-demand ratio), searching for next passengers in a short period is more promising because of a high demand. Namely, the supply-demand ratio is an important candidate factor that may affect

incomes. In the search process ($\alpha \rightarrow \beta$), the search route choice is another factor that would affect incomes. If the driver is experienced, he/she is more likely to search through a route with greater pick-up likelihood and pick up passengers in shorter time and distance ($\alpha \rightarrow \beta_1$), while the inexperienced driver might otherwise spend more time in searching passengers ($\alpha \rightarrow \beta_2$). The search distance and pick-up likelihood are thus critical candidate factors. After picking up passengers at the end of search (β), the studied taxi starts delivery ($\beta \rightarrow \gamma$). The fare and delivery speed are likely to have an impact on the income level, since which would be limited by lower speed due to congested traffic conditions and lower fare caused by near destination ($\beta_1 \rightarrow \gamma_1$) or conversely be raised by good road conditions or longer rides ($\beta_2 \rightarrow \gamma_2$).

PLEASE PLACE Fig. 5 ABOUT HERE

In the light of the above scenario, we formulate the factors affecting incomes according to the typical operation process, including the supply-demand ratio (around the drop-off location), search intensity and distance (during the search period), as well as fare and delivery speed (during the delivery period). Mathematical definitions and formulae for these factors are presented as follows.

4.2.1 Supply-demand ratio around the drop-off location

In the search-delivery scenario as illustrated in Fig. 6, supply-demand ratio around the drop-off location is the first factor the drivers will encounter. Quantity demanded and the number of taxis in the vicinity would both affect the subsequent search time. The ratio of quantity supplied of vacant taxis to quantity demanded is, therefore, a potential factor to be considered.

PLEASE PLACE Fig. 6 ABOUT HERE

To limit the area of a region where supply and demand will impact on the studied taxi, we set a geographical boundary in which supply and demand has an effect on the taxi, while out of which does not. This boundary is called a buffer zone. The size of the buffer zone is determined according to the historical search distances of all taxis with the same drop-off locations. The buffer zone shaped to a simple square rather than round or any other irregular shape is for the convenience of computation; each piece of cell information has its specific buffer zone.

Definition 1: Given a search distance record $L_{s_{x,y}}^t(i)$ of all taxis with the same drop-off location (x, y) at time interval t , sorted ascending ($i = 1, 2, \dots, n$), the side length of the square buffer zone centered at (x, y) is:

$$b_{x,y}^t = 2 \left\lceil \frac{L_{s_{x,y}}^t(k)}{g} \right\rceil, \text{ where } k = \left\lceil \frac{N \times n}{100} \right\rceil \quad (1)$$

where g denotes the side length of cells (km), and N denotes the percentile (0~100) to omit extreme values. N is set to 85th in this paper, which indicates $b_{x,y}^t$ is set by the 85th percentile of historical search distances.

We can define the supply-demand ratio in the buffer zone as follows.

Definition 2: Given total time (min) of vacant taxis $V_{x,y}^t$ and hired taxis $H_{x,y}^t$ in the buffer zone centered at (x, y) at time interval t , the supply-demand ratio is stated mathematically as:

$$SD_{x,y}^t = \frac{\sum_{(x,y) \in \mathbb{B}_{x,y}^t} (V_{x,y}^t + H_{x,y}^t)}{\sum_{(x,y) \in \mathbb{B}_{x,y}^t} H_{x,y}^t} \quad (2)$$

where total time of hired taxis $H_{x,y}^t$ indicates consumption of time induced by passenger delivery, while total time of vacant taxis $V_{x,y}^t$ indicates the spare time.

Therefore, the numerator of the Eq. 2 describes the total time of supply or quantity supplied, and the denominator represents the used time or quantity demanded. The ratio of these two numbers approximates to the supply-demand ratio. The value is greater than or equal to 1. Namely, the supply meets or exceeds the

demand. Lack of range less than 1 or supply short of demand case does not affect this study since the study time span is not among peak hours, so that the quantity supplied typically exceeds the quantity demanded.

Take Fig. 6 as an example, following equation presents the calculation of the supply-demand ratio for the cell (4,6).

$$SD_{4,6}^t = \frac{(V_{2,4}^t + H_{2,4}^t) + (V_{3,4}^t + H_{3,4}^t) + \dots + (V_{6,8}^t + H_{6,8}^t)}{H_{2,4}^t + H_{3,4}^t + \dots + H_{6,8}^t} \quad (3)$$

4.2.2 Search distance and intensity of pickups

Fig. 7 illustrates a typical search route. If the supply-demand ratio around the drop-off location can be thought as an uncontrollable factor, drivers at the high-income level might subjectively find routes with a high pick-up likelihood according to their experience and shorten search time and distance consequently.

PLEASE PLACE Fig. 7 ABOUT HERE

Definition 3: the actual distance from drop-off location to next pick-up location is defined as the search distance (L_s , km).

$$L_s = \sum_i d([lat_i, lon_i], [lat_{i+1}, lon_{i+1}]) \quad (4)$$

where $[lat_i, lon_i]$ and $[lat_{i+1}, lon_{i+1}]$ denote the longitudes and latitudes of every two consecutive records during the search period, and $d(\cdot)$ is a function of calculating the orthodromic distance.

Definition 4: Given average number of pickups $I_{x,y}^t$ in the cell (x, y) at the time interval t among several specific days, for arbitrary search route, search intensity of pickups (I_s) is a weighted average of numbers of pickups in all cells,

$$I_s = \frac{\sum_{(x,y) \in \text{SearchingPath}} I_{x,y}^t \times T_{s,x,y}^t}{\sum_{(x,y) \in \text{SearchingPath}} T_{s,x,y}^t} \quad (5)$$

where $T_{s,x,y}^t$ indicates the time consumption of searching passengers in the cell (x, y) at the time interval t .

Take Fig. 7 as an example, following equation presents the calculation of the search intensity of pickups.

$$I_s = \frac{I_{4,8}^t \times T_{s,4,8}^t + I_{4,7}^t \times T_{s,4,7}^t + \dots + I_{5,3}^t \times T_{s,5,3}^t}{T_{s,4,8}^t + T_{s,4,7}^t + \dots + T_{s,5,3}^t} \quad (6)$$

4.2.3 Fare and delivery speed

The delivery route is presented in Fig. 8, fare and delivery speed can also be the factors drivers will encounter after they pick up passengers, which will eventually affect their incomes.

Definition 5: Fare is calculated according to the time interval, delivery distance, speed as well as local taxi rates, which is denoted as F (yuan).

$$F = F_1(s) + F_2(t_{\text{waiting}}) \quad (7)$$

where s denotes the delivery distance (km), and t_{waiting} denotes waiting time (min) at low speeds, $F_1(\cdot)$ and $F_2(\cdot)$ are the functions of calculating corresponding charges.

PLEASE PLACE Fig. 8 ABOUT HERE

Also, the average delivery speed reflects the drivers' experience to some extent and indicate the delivery time indirectly, which is thought to be one another candidate factor.

Definition 6: For arbitrary delivery route, delivery speed (km/h) is a weighted average of speeds through the delivery route,

$$V_d = \frac{\sum v(j) \times T_d(j)}{\sum T_d(j)} \quad (8)$$

where $T_d(j)$ indicates the duration of speed $v(j)$.

Take Fig. 8 as an example, following equation presents the calculation of delivery speed.

$$V_d = \frac{v(1) \times T_d(1) + v(2) \times T_d(2) + \dots + v(n) \times T_d(n)}{T_d(1) + T_d(2) + \dots + T_d(n)} \quad (9)$$

In summary, we have so far summarized and defined candidate factors that would affect incomes by investigating the overall search-delivery process. For every single search-delivery route of each driver, those candidate factors are calculated. And they additionally are averaged because the dependent variable Y is an aggregate value. 2938 samples are drawn as a result.

PLEASE PLACE Fig. 9 ABOUT HERE

Fig. 9 presents the distributions of the five factors among three income levels. In Fig. 9(a) and Fig. 9(c), the supply-demand ratio (SD) and search intensity (I_s) fail to show any significant trend as the income level changes, which indicates the effects of these two factors are evenly distributed among the different income levels. While in Fig. 9(b), search distance (L_s) reduces as income level rises, and the variance of search distance of the low-income level exceeds the higher income level, which indicates the inconsistent performance of search distance exists at the low-income level. In Fig. 9(d) and Fig. 9(e), fare (F) and delivery speed (V_d) increase as the income level rises, high-income drivers perform most consistently in these two factors. These revealed trends also comply with the conclusions drawn by Liu et al. (2010).

4.3. Developing a generalized multi-level ordered logit model

Logistic regression (also known as logit model) is widely used to model the outcomes of a categorical dependent variable (Czepiel, 2002). In contrast to conventional regression models, the left-hand side of logit model is a log-transformed odds ratio of the dependent variable (the right-hand side remains the same). Due to this particular form, the left-hand side brings a probabilistic perspective to the analysis of factor impact on incomes, namely, a question such as “whether the probability of earning high incomes will increase or not if some factor changes” turns to be answerable, which ensures a straightforward and intuitive conclusion.

Empirically, the impact of some factor may vary for drivers at different income levels. That is, a factor that has a great impact on X -level drivers may have less impact on their Y -level counterparts. For this factor, varying coefficients are therefore required for each income level respectively. Thus, the generalized multi-level ordered logit is specified. This model will exclude the constrictive parallel-lines assumption, which is commonly applied to simple logit (this assumption leads to equal coefficients among ordered levels), and make it possible to detail the discrepancy in different levels as well as offer level-oriented advice accordingly.

Provided that the income level is an ordered-response discrete variable y_i (the income level for the driver i) with factors $\mathbf{X}_i = (x_{i1}, x_{i2}, \dots, x_{i5})$ as independent variables, an ordered logit model can be written in terms of cumulative probability of a specific level for a given set of factors $\mathbf{X}_i$

$$P(y_i \leq j) = \frac{e^{-(\alpha_j + \mathbf{X}_i \boldsymbol{\beta}^T)}}{1 + e^{-(\alpha_j + \mathbf{X}_i \boldsymbol{\beta}^T)}} \quad j = 1, 2, \dots, M - 1 \quad (10)$$

where j is a number for the income level (j ends at $M - 1$ since $P(y_i \leq M) \equiv 1$), and the level falls as the number increases; α_j are cut points for the thresholds of the ordered logit model, and $\boldsymbol{\beta}$ is a vector of coefficients to be estimated; M is the total number of income levels; and T is a transpose sign.

As noted above, the conventional logit models are constricted by the parallel-lines assumption that same β are estimated for the arbitrary income level j . In fact, factor impact among different income levels are inclined to violate the assumption, thus requiring different β estimation. The ordered logit is hence rewritten as

$$P(y_i \leq j) = \frac{e^{-(\alpha_j + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_j^{(2)T})}}{1 + e^{-(\alpha_j + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_j^{(2)T})}} = \frac{1}{1 + e^{\alpha_j + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_j^{(2)T}}} \quad (11)$$

where $\beta^{(1)}$ is a coefficient vector for factors meeting the assumption, which take same values among different levels, while $\beta_j^{(2)}$ is a coefficient vector for factors violating the assumption, which take varying values among different levels.

The probability for each level is computed as follows

$$P(y_i = 1) = \frac{1}{1 + e^{\alpha_1 + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_1^{(2)T}}} \quad (12)$$

$$P(y_i = j) = \frac{1}{1 + e^{\alpha_j + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_j^{(2)T}}} - \frac{1}{1 + e^{\alpha_{j-1} + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_{j-1}^{(2)T}}}, j = 2, \dots, M-1 \quad (13)$$

$$P(y_i = M) = 1 - \frac{1}{1 + e^{\alpha_{M-1} + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_{M-1}^{(2)T}}} \quad (14)$$

Applying logit transformation (inverse of the sigmoidal logistic function) to the cumulative probability we have

$$\text{logit}P(y_i \leq j) = \ln \frac{P(y_i \leq j)}{P(y_i > j)} = \alpha_j + X_i^{(1)} \beta^{(1)T} + X_i^{(2)} \beta_j^{(2)T} \quad (15)$$

where $\frac{P(y_i \leq j)}{P(y_i > j)}$ is called the odds, which represents the ratio of the probability of being less than or equal to j to the probability of being greater than j . Similarly, $\ln \frac{P(y_i \leq j)}{P(y_i > j)}$ is a log-odds. Having transformed the probability into a log-odds, the logit function linearizes the association between the probability and independent variables. In this context, the model is a generalized linear model.

Though as a generalized linear model, simple linear regression estimation methods like least-squares estimation are not applicable. The maximum likelihood estimation (MLE) is applied instead.

MLE is implemented by calculating the joint probability density.

$$f(\mathbf{y} | \beta^{(1)}; \beta^{(2)}) = \prod_{i=1}^N \left[\frac{n_i!}{\prod_{j=1}^J m_{y_{ij}}!} \prod_{j=1}^M p_{ij}^{m_{y_{ij}}} \right] \quad (16)$$

where n_i samples are drawn at each time of the N observations ($\mathbf{y}$); $m_{y_{ij}}$ of the n_i samples are in the category j , and p_{ij} [or $P(y_i \geq j)$] is a function of X_i , $\beta^{(1)}$ and $\beta^{(2)}$ to estimate the probability of falling into the category j . The product of joint probability densities for all samples are the probability of these N observations. The idea of MLE is to find the parameter values betas that maximize the probability.

Keeping $\mathbf{y}$ constant with $\beta^{(1)}$ and $\beta^{(2)}$ as variables, we deduce the likelihood function

$$\mathcal{L}(\beta^{(1)}; \beta^{(2)} | \mathbf{y}) = \prod_{i=1}^N \left[\frac{n_i!}{\prod_{j=1}^J m_{y_{ij}}!} \prod_{j=1}^M p_{ij}^{m_{y_{ij}}} \right] \quad (17)$$

The likelihood function (Eq. 17) is algebraically equivalent to the joint probability density function (Eq. 16), except the only difference being that the likelihood function expresses the unknown values of $\beta^{(1)}$ and $\beta^{(2)}$ in terms of known fixed constant values for y (Czepiel, 2002).

The log-transformation is widely used to deal with the likelihood function in practice. Therefore we have a log-likelihood function

$$\ln \mathcal{L}(\beta^{(1)}; \beta^{(2)} | y) = \sum_{i=1}^N \ln \frac{n_i!}{\prod_{j=1}^J m_{y_{ij}}!} \prod_{j=1}^J p_{ij}^{m_{y_{ij}}} \quad (18)$$

The MLE method is thus deduced as

$$\{\hat{\beta}_{\text{mle}}\} \subseteq \arg \max_{\beta^{(1)}, \beta^{(2)}} \ln \mathcal{L}(\beta^{(1)}; \beta^{(2)} | y) \quad (19)$$

Interpreting parameters beta is the following work. An elasticity analysis is often performed to explain the results by describing how responsive the income level is to a change in each factor. The equation for the elasticity is (Washington et al., 2010)

$$E_{X_{kj}}^{P_j} = \frac{\partial P_j}{\partial X_{kj}} \times \frac{X_{kj}}{P_j} \quad (20)$$

where P_j represents the probability of falling in the level j ; X_{kj} denotes the k th factor which responds to the level j .

Considering the logit transformation, the partial derivative of the equation is

$$E_{X_{kj}}^{P_j} = [1 - P_j] \beta_j X_{kj} \quad (21)$$

Elasticity refers to the degree of responsiveness of the income level in relation to 1% changes of independent variables X_{kj} . Generally, the level for a given factor is elastic when the elasticity is greater than one (in absolute value), namely a 1% change of the factor X_{kj} has a relative effect on a greater than 1% change of the income level; the income level for a given factor is inelastic when the elasticity is less than one (in absolute value), namely a 1% change of the factor X_{kj} has a relative effect on a less than 1% change of the income level.

5. Results and Discussion

The response variable in our model is the ordered income level (1 = high, 2 = medium and 3 = low), whereas the independent variables are measurement variables including supply-demand ratio, search distance and intensity, as well as fare and delivery speed. The generalized multi-level ordered logit model is finally developed using the gologit2 algorithm (Williams, 2006) on Stata 14.0.

5.1. Collinearity detection of the factors

Collinearity refers to the dependence of factors, usually in a regression analysis. It can be problematic for parameter estimations since it inflates the variance of regression parameters and hence potentially leads to the incorrect identification of relevant factors in a statistical model (Dormann et al., 2013). In this part, variance inflation factors (VIF) is computed for detecting collinearity among factors in Stata 14.0 before fitting the GMOL model. In statistics, variance inflation factor is a measure of collinearity among variables. It has a value greater than or equal to 1, where VIFs equal to 1 indicate no collinearity, and VIFs exceeding 4 warrant further investigation (Fox, 1997), while VIFs exceeding 10 are signs of severe collinearity requiring correction (Myers, 2000). Table 1 presents the results of collinearity detection of the factors affecting incomes, VIFs for all five factors range from 1.00 to 1.07, much less than 4; this indicates that there is no significant collinearity among the predictors either. Therefore, all five factors can be input to fit the GMOL model.

PLEASE PLACE Table 1 ABOUT HERE

5.2. Necessity of using GMOL model

As discussed in the previous section, the impact of some factor may vary among drivers at different income levels. The parallel-lines assumption test in this part is used to evaluate this claim. The test suggests that supply-demand ratio (SD , p -value = 0.6577), search intensity (I_s , p -value = 0.1447), fare (F , p -value = 0.8285) meet the parallel-lines assumption, coefficients for these three factors remain the same among different income levels. While search distance (L_s) and delivery speed (V_d) both violate the assumption (both p -value = 0.0000), coefficients for these two factors are varying among different income levels, namely, search distance (L_s) and delivery speed (V_d) have varying impacts on drivers in different income levels. The one probable explanation is that the factors supply-demand ratio, search intensity, and fare exert equivalent background impacts on taxi drivers' incomes, while the factors search distance and delivery speed supplement distinct impacts that in fact diversify the income levels. The parallel-lines assumption test validates the claim about different factor impacts among drivers at different income levels, which necessitates the use of a generalized multi-level ordered logit (GMOL) model in this study.

5.3. Significant factors affecting the income level

The results of the GMOL model are presented in Table 2. As noted in the previous section, $P(y_i \leq 3) \equiv 1$, the output thus presents only results of $P(y_i \leq 1)$ and $P(y_i \leq 2)$. The results show that the coefficients of supply-demand ratio and fare are .3945 and -.1531 respectively for any income level, while the coefficients of search distance and delivery speed differ across different levels, but signs of the coefficients remain the same.

PLEASE PLACE Table 2 ABOUT HERE

The test of statistical significance indicates that the supply-demand ratio (SD) is significant at the 10% level, the search intensity (I_s) is insignificant, and the other three factors are significant at the 1% level. The insignificance of search intensity implies no obvious difference of pick-up likelihoods of search routes for drivers at different income levels. A possible explanation for this outcome is that the high likelihood of pickups may attract a large number of taxis, it is thus difficult to increase incomes under this competitive environment. The positive sign of the coefficient of supply-demand ratio also backs this explanation, that is, the possibility of lower incomes rises by increasing supply-demand ratio. In the light of this explanation, searching passengers through routes in a high likelihood of pickups is not always a rule of thumb to increase incomes.

Mathematical expressions of probability for each level are given according to the estimated coefficients (the independent variable I_s is excluded due to its insignificance).

$$P(y_i = \text{high}) = \frac{1}{1 + e^{12.8570 + .3945(SD) + 2.2389(L_s) - .1531(F) - .5587(V_d)}} \quad (22)$$

$$P(y_i = \text{medium}) = \frac{1}{1 + e^{6.8156 + .3945(SD) + 1.7961(L_s) - .1531(F) - .3443(V_d)}} - \frac{1}{1 + e^{12.8570 + .3945(SD) + 2.2389(L_s) - .1531(F) - .5587(V_d)}} \quad (23)$$

$$P(y_i = \text{low}) = 1 - \frac{1}{1 + e^{6.8156 + .3945(SD) + 1.7961(L_s) - .1531(F) - .3443(V_d)}} \quad (24)$$

5.4. Elasticity analysis

Responsiveness of the income level to any single independent variable is illustrated by keeping the other three variables constant. As shown in Fig. 10, any three independent variables are held constant at their means, and the rest one changes across its range.

PLEASE PLACE Fig. 10 ABOUT HERE

Concretely, the curves for the high-income level fall away as the supply-demand ratio (Fig. 10(a)) or search distance (Fig. 10(b)) increases; this indicates that reducing these two factors will benefit drivers with a

higher likelihood of increasing incomes. The probability curves climb otherwise as fare (Fig. 10(c)) or speed (Fig. 10(d)) rises, this implies that raising these two factors is more likely to gain higher incomes.

The following analysis shows that growing supply-demand ratio is inclined to cause increasing difficulty in searching passengers, and consequently lower the probability of climbing to a higher income level. A possible measure is to keep away from those taxi-intensive regions and operate in an alternative area where supply and demand is in relative balance. Moreover, increasing search distance suggests longer time in searching passengers and a drop in hourly incomes as well as the likelihood of high-income level. Thus income-conscious drivers need to shorten their search distance as a response. Higher fare and speed are both beneficial to increasing incomes; this implies that high-income level tends to have a higher fare per trip and delivery speed, and they prefer roads in better traffic conditions. In summary, those attempt to increase incomes require to operate in regions with supply and demand in relative balance and try best to shorten search distance, increase fare and delivery speed, as well as run on routes in good traffic condition.

With an elasticity analysis, we can further illustrate the income level responsiveness to the change of factors and determine their contributions accordingly. Similar to Fig. 10, given an income level, any three factors are held constant at their average values, and the rest one changes across its range, probability responsiveness of high-income level to any single factor is presented in Fig. 11. The vertical solid lines in Fig. 11 denote the average values of the unfixed factors, and the two adjacent vertical dashed lines signify a value with $\pm 10\%$ of range change from the means. Consistent changes from the average values ensure a fair comparison. Namely, this paper assumes that every 10% positive change from its average value is of similar difficulty for any factor. It should be pointed out that the “positive change” means a change that causes a probability increase of high-income level. Each factor across the four subplots are in same values at the vertical solid lines, since they have the same probability (see white dots that intersect the solid line with the curve). And the white dots move to new intersections as the investigated factor gain a positive change; a black arrow illustrates this movement; the vertical uplift of the arrow presents how much probability of the high-income level changes on the 10% change of the investigated factor.

PLEASE PLACE Fig. 11 ABOUT HERE

Different lengths and slopes of the arrows in Fig. 11 indicate the dominant position of the search distance (Fig. 11(b)) to increase the probability of becoming high-income for a driver at the low-income level. Delivery speed (Fig. 11(d)), fare (Fig. 11(c)), and the supply-demand ratio (Fig. 11(a)) are the second to last factors affecting incomes, respectively. Similarly, Fig. 12 demonstrates the impact of factors for drivers at the medium-income level. Unlike the effects on drivers at the low-income level, both search distance (Fig. 12(b)) and delivery speed (Fig. 12(d)) have a nearly similar impact for drivers at the medium-income level to become high-income, while fare and supply-demand ratio have less effect.

PLEASE PLACE Fig. 12 ABOUT HERE

And Fig. 13 demonstrates the impact of factors for drivers at the high-income level, delivery speed (Fig. 13(d)) has a slightly stronger effect than that of search distance. Meanwhile, fare and supply-demand ratio have less effect.

PLEASE PLACE Fig. 13 ABOUT HERE

The ratio of probability change on the $\pm 10\%$ change of factors is the elasticity of the income level. Contributions can be thus computed by normalizing the absolute value of the elasticities (in percentage), the outcomes are summarized in Table 3.

PLEASE PLACE Table 3 ABOUT HERE

5.5. Discussion

Table 3 has quantified the high-income-level responsiveness to a change of any factor. Impacts of significant influencing factors on each income level are summarized as follows:

- 1) As for drivers at the low-income level, shortening the search distance is the optimal option, whose contribution accounts for up to 46.93%. For an average driver at the low-income level (supply-demand ratio = 2.11, search distance = 2.36 km, fare = 22 yuan, and delivery speed = 25 km/h), if he/she shortens the search distance by 10% (236 m), the probability to become a driver at the high-income level surges upward from 19% to almost 50%. Relevant authorities are suggested to take some measures to shorten search distance, such as disseminating information on nearby passenger demand or designating taxi stands where vacant taxis can reliably find passengers.
- 2) As for drivers at the medium-income level, shortening the search distance and increasing the delivery speed are both optimal options, whose contributions account for 43.75% and 36.76% respectively. For an average driver at the medium-income level (supply-demand ratio = 2.08, search distance = 2.14 km, fare = 23 yuan, and delivery speed = 25 km/h), if he/she shortens the search distance by 10% (214 m), the probability to become a driver at the high-income level climbs up from 31% to nearly 48%; if he/she increases his/her delivery speed by 10% (1.1 km/h), the probability to become a driver at the high-income level rises from 31% to almost 45%. In addition to measures noted above, relevant authorities can also take some measures to increase the delivery speed, for instance, broadcasting real-time traffic conditions and offering delivery route recommendation.
- 3) As for drivers at the high-income level, increasing the delivery speed is the optimal option, whose contribution accounts for 39.30%. For an average driver at the high-income level (supply-demand ratio = 2.09, search distance = 1.99 km, fare = 23 yuan, and delivery speed = 26 km/h), if he/she increases the delivery speed by 10% (2.6 km/h), the probability to maintain the high-income level rises from 58% to nearly 74%.

Additional interesting insights are also found, they are consistent or inconsistent with some preconceived ideas and related studies.

- 1) High-income drivers often have shorter search distances and higher delivery speeds, these results are consistent with findings in similar research (Liu et al., 2010; Zhang et al., 2015);
- 2) A related study (Qu et al., 2014) recommends search route according to the likelihood of pickups through the route, but this factor in our study does not show significance in affecting incomes. Therefore, suggesting search route based on not only likelihood of pickups but also search/delivery speed may improve the model performance;
- 3) A previous study (Zhang and Wang, 2016) found that high-income drivers rejected passengers whose destinations were beyond the drivers' expectations, which indicated that selecting passengers could guarantee higher incomes. In our study, the little contributions of supply-demand ratio and fare deduce that supply-demand ratio does not have an apparent impact on the income level. Therefore, we infer that entering low demand areas does not indicate low incomes. Similarly, the modest contribution of fare suggests that taking only long rides with high fares does not ensure higher incomes. Namely, higher incomes could be achieved even without selecting passengers.

6. Conclusions

In this paper, we first carefully tackle over 167 million GPS records from nearly 8000 taxis in Shanghai and categorize taxi drivers into three levels according to their incomes. We then investigate the overall search-delivery process and define five factors that may explain the income difference, as well as develop a GMOL model to find a quantitative association between incomes and service strategies. We finally conduct an elasticity analysis of the significant factors using the GMOL model and estimate contributions of each significant factors. The major conclusions derived from this study are summarized as follows:

- 1) **There exists a marked difference in taxi drivers' incomes.** The incomes within the study hours (20:00~23:00) have a mean of 168 yuan and a standard deviation of 58 yuan. Incomes of lowest 25% of drivers remain below 129 yuan, while above 210 yuan for highest 25%. The gap is larger than 80 yuan.
- 2) **Supply-demand ratio, search distance, fare and delivery speed are the four significant factors affecting incomes,** while search intensity is insignificant. The insignificance of search intensity indicates that searching passengers through routes in a high likelihood of pickups is not a rule of thumb to increase incomes.
- 3) **Search distance and delivery speed dominate contributions to the income level.** Specifically, for low-income drivers, the search distance is the key factor to promote (contribution: 46.93%); for medium-income drivers, the search distance and the delivery speed are both the key factors to promote (contribution: 43.75% and 36.76%); for high-income drivers, the delivery speed is the key factor to promote (contribution: 39.30%).
- 4) **Entering low demand areas does not indicate low incomes, and taking only long rides does not guarantee high incomes.** As for taxi drivers who enter an area in low demand, they still have a chance to earn high incomes if they perform well in the key factors affecting the income level. Similarly, as to taxi drivers who only take long rides, they may result in mediocre incomes if they do not have a good performance in the key factors affecting the income level.

These findings provide comprehensive and quantitative insight into factors affecting taxi drivers' incomes, as well as challenge some preconceived ideas of how to earn high incomes. Some measures are suggested from the findings to aid the newcomers and low-income drivers. Despite merely analyzing factors affecting the income level for taxi drivers in this paper, the overall methodology and framework can also be applied to other similar problems involving quantification of factors determining human capability.

Acknowledgments

We would like to thank Shanghai Qiangsheng Taxi Company for providing the data source. And this study was supported by National Science Foundation of China (51422812), and the Fundamental Research Funds for the Central Universities.

References

- BBC News, 2013. Public hire taxi drivers protest over fare changes.
- Branigan, T., 2008. China taxi drivers strike as economic unrest spreads. *The Guardian*.
- Camerer, C., Babcock, L., Loewenstein, G., Thaler, R., 1997. Labor supply of New York City cabdrivers: One day at a time. *The Quarterly Journal of Economics* 112(2), 407-441.
- Cao, H., Mamoulis, N., Cheung, D.W., 2005. Mining frequent spatio-temporal sequential patterns, *Fifth IEEE International Conference on Data Mining*. IEEE.
- Chou, Y.K., 2002. Testing alternative models of labour supply: Evidence from taxi drivers in Singapore. *The Singapore Economic Review* 47(1), 17-47.
- Cooper, J., Mundy, R., Nelson, J., 2010. *Taxi!: Urban economies and the social and transport impacts of the taxicab*. Ashgate Publishing, Ltd.
- Crawford, V.P., Meng, J., 2011. New York City cab drivers' labor supply revisited: Reference-dependent preferences with rational expectations targets for hours and income. *The American Economic Review* 101(5), 1912-1932.
- Czepiel, S.A., 2002. Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep.net/stat/mlelr.pdf.
- Dong, Y., Zhang, Z., Fu, R., Xie, N., 2016. Revealing New York taxi drivers' operation patterns focusing on the revenue aspect, *2016 12th World Congress on Intelligent Control and Automation (WCICA)*. IEEE, pp. 1052-1057.

- Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1), 27-46.
- Farber, H.S., 2005. Is tomorrow another day? The labor supply of New York City cabdrivers. *Journal of Political Economy* 113(1), 46-82.
- Fox, J., 1997. *Applied regression analysis, linear models, and related methods*. Sage Publications, Inc.
- Gao, Y., Xu, P., Lu, L., Liu, H., Liu, S., Qu, H., 2012. Visualization of taxi drivers' income and mobility intelligence, *International Symposium on Visual Computing*. Springer Berlin Heidelberg, pp. 275-284.
- Huang, J., Huangfu, X., Sun, H., Li, H., Zhao, P., Cheng, H., Song, Q., 2015. Backward path growth for efficient mobile sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering* 27(1), 46-60.
- Köszegi, B., Rabin, M., 2006. A model of reference-dependent preferences. *The Quarterly Journal of Economics* 121(1), 1133-1165.
- Li, Q., Zeng, Z., Yang, B., Zhang, T., 2009. Hierarchical route planning based on taxi GPS-trajectories, *2009 17th International Conference on Geoinformatics*. IEEE, pp. 1-5.
- Liu, L., Andris, C., Ratti, C., 2010. Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems* 34(6), 541-548.
- Myers, R.H., 2000. Classical and modern regression with applications (Duxbury Classic). *Duxbury Press, Pacific Grove*.
- Powell, J.W., Huang, Y., Bastani, F., Ji, M., 2011. Towards reducing taxicab cruising time using spatio-temporal profitability maps, *International Symposium on Spatial and Temporal Databases*. Springer Berlin Heidelberg, pp. 242-260.
- Qu, M., Zhu, H., Liu, J., Liu, G., Xiong, H., 2014. A cost-effective recommender system for taxi drivers, *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 45-54.
- Rong, H., Zhou, X., Yang, C., Shafiq, Z., Liu, A., 2016. The rich and the poor: A Markov decision process approach to optimizing taxi driver revenue efficiency, *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*. ACM, pp. 2329-2334.
- Sirisoma, R.M.N.T., Wong, S.C., Lam, W.H.K., Wang, D., Yang, H., Zhang, P., 2010. Empirical evidence for taxi customer-search model, *Proceedings of the Institution of Civil Engineers: Transport*. Thomas Telford (ICE Publishing).
- The Irish Times, 2009. Taxi drivers stage Dublin protest over regulation.
- Tu, W., Li, Q., Fang, Z., Shaw, S.-I., Zhou, B., Chang, X., 2016. Optimizing the locations of electric taxi charging stations: A spatial-temporal demand coverage approach. *Transportation Research Part C: Emerging Technologies* 65, 172-189.
- Untapped Cities, 2015. Today in NYC history: The taxi riots of 1934 start February 5, 1934.
- Veloso, M., Phithakkitnukoon, S., Bento, C., 2011. Urban mobility study using taxi traces, *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*. ACM, pp. 23-30.
- Wang, X., Fan, T., Li, W., Yu, R., Bullock, D., Wu, B., Tremont, P., 2016. Speed variation during peak and off-peak hours on urban arterials in Shanghai. *Transportation Research Part C: Emerging Technologies* 67, 84-94.
- Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. *Statistical and econometric methods for transportation data analysis*. CRC Press.
- Williams, R., 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal* 6(1), 58.
- Wong, R.C.P., Szeto, W.Y., Wong, S.C., 2015. A two-stage approach to modeling vacant taxi movements. *Transportation Research Part C: Emerging Technologies* 59, 147-163.
- Yang, C., Gonzales, E.J., 2017. Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data, in: Thakuriah, P., Tilahun, N., Zellner, M. (Eds.), *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*. Springer International Publishing, Cham, pp. 405-425.
- Yang, H., Wong, S.C., Wong, K.I., 2002. Demand-supply equilibrium of taxi services in a network under competition and regulation. *Transportation Research Part B: Methodological* 36(9), 799-819.
- Yang, J., Dong, J., Hu, L., 2017. A data-driven optimization-based approach for siting and sizing of electric taxi charging stations. *Transportation Research Part C: Emerging Technologies* 77, 462-477.

- 1 Yuan, J., Zheng, Y., Xie, X., Sun, G., 2013. T-Drive: Enhancing driving directions with taxi drivers'
intelligence. *IEEE Transactions on Knowledge and Data Engineering* 25(1), 220-232.
- 2 Zhan, X., Hasan, S., Ukkusuri, S.V., Kamga, C., 2013. Urban link travel time estimation using large-scale
3 taxi data with partial information. *Transportation Research Part C: Emerging Technologies* 33, 37-49.
- 4 Zhang, D., Sun, L., Li, B., Chen, C., Pan, G., Li, S., Wu, Z., 2015. Understanding taxi service strategies
5 from taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems* 16(1), 123-135.
- 6 Zhang, S., Wang, Z., 2016. Inferring passenger denial behavior of taxi drivers from large-scale taxi traces.
7 *PLOS ONE* 11(11), e0165597.
- 8 Zheng, Y., Zhang, L., Ma, Z., Xie, X., Ma, W.Y., 2011. Recommending friends and locations based on
9 individual location history. *ACM Transactions on the Web (TWEB)* 5(1), 5.
- 10 Zheng, Y., Zhang, L., Xie, X., Ma, W.Y., 2009. Mining interesting locations and travel sequences from
11 GPS trajectories, *Proceedings of the 18th International Conference on World Wide Web*. ACM, pp. 791-800.
- 12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

List of Table Titles

- **Table 1.** Collinearity detection of the factors affecting incomes
- **Table 2.** Outputs of the generalized multi-level ordered logit model for levels of incomes
- **Table 3.** Elasticities and Contributions of Determinants (X_{kj} in 10% change)

List of Figure Captions

- **Fig. 1.** Study area and distribution of pickups. The red closed curve denotes the Outer Ring Road, which encompasses the downtown area of Shanghai. Approx. 82.25% of daily pickups happen within the downtown area.
- **Fig. 2.** Trends of the ratio of empty/total mileage (RETM) and the number of taxis in service. The green box is a box and whisker diagram based on the five-number summary: minimum ($Q1 - 1.5IQR$), first quartile($Q1$), median($Q2$), third quartile($Q3$), and maximum ($Q3 + 1.5IQR$), where IQR is the interquartile range, equal to $Q3 - Q1$.
- **Fig. 3.** Distribution of taxi drivers' incomes.
- **Fig. 4.** Sample selection and categorization of taxi drivers. The candidate drivers are selected by two thresholds – standard deviation of rankings (horizontal dashed line, threshold: ≤ 1763.5) and average incomes (vertical dashed lines, thresholds include: ≤ 165.8 yuan for low-income; 184.7~199.5 yuan for medium-income; and ≥ 215.5 yuan for high-income).
- **Fig. 5.** Defining factors that might affect the income level based on a typical search-delivery scenario. Two alternative route plans that start from the same drop-off location are illustrated in the scenario. Each route has different search and delivery choices.
- **Fig. 6.** Analysis of supply-demand ratio around the drop-off location. In the context of the supply and demand, the studied taxi A is mainly affected by any other vacant taxis in the buffer area (the square in orange).
- **Fig. 7.** Extraction of features in a search route. The studied taxi is cruising through the search route (the consecutive light green cells), and the historical pick-up count in each cell may affect the time for searching next passengers and affect the incomes eventually.
- **Fig. 8.** Extraction of features in delivery routes. The studied taxi is delivering passengers through the delivery route (the consecutive red cells), and the speed in each GPS coordinates may affect the time for carrying passengers and affect the incomes eventually, the fare is also an intuitive factor that may have an impact on incomes.
- **Fig. 9.** The distributions of the five factors among three income levels. The five factors include (a) supply-demand ratio; (b) search distance; (c) search intensity; (d) fare; (e) delivery speed.
- **Fig. 10.** Probability curves for different income levels and their probability change due to a change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.
- **Fig. 11.** Probability curve of being high-income drivers (for average low-income drivers) and their probability increase due to change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.
- **Fig. 12.** Probability curve of being high-income drivers (for average medium-income drivers) and their probability increase due to change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.
- **Fig. 13.** Probability curve of remaining high-income drivers (for average high-income drivers) and their probability increase due to change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.

Tables

Table 1. Collinearity detection of the factors affecting incomes

	VIF	SQRT VIF	Tolerance	R ²
Supply-demand ratio (<i>SD</i>)	1.07	1.03	0.9353	0.0647
Search distance (<i>L_s</i>)	1.00	1.00	0.9972	0.0028
Search intensity (<i>I_s</i>)	1.07	1.03	0.9336	0.0664
Fare (<i>F</i>)	1.07	1.04	0.9310	0.0690
Speed (<i>V_d</i>)	1.14	1.07	0.8765	0.1235

Note: average VIF = 1.07

Table 2. Outputs of the generalized multi-level ordered logit model for levels of incomes

y_i	category	Coefficient	Std. Err.	2-tailed $P> z $	95% Conf. Interval	
≤ 1						
Drop-off	SD ratio (SD)*	.3945	.2211	.074	-.0389	.8280
Search	Distance (L_s ***	2.2389	.1092	.000	2.0248	2.4530
	Intensity (I_s)	2.0353	1.8229	.264	-1.5374	5.6081
Delivery	Fare (F ***	-.1531	.0157	.000	-.1829	-.1224
	Speed (V_d ***	-.5587	.0262	.000	-.6100	-.5074
	constant	12.8570	.7644	.000	11.3588	14.3553
≤ 2						
Drop-off	SD ratio (SD)*	.3945	.2211	.074	-.0389	.8280
Search	Distance (L_s ***	1.7961	.0984	.000	1.6032	1.9890
	Intensity (I_s)	2.0353	1.8229	.264	-1.5374	5.6081
Delivery	Fare (F ***	-.1531	.0157	.000	-.1839	-.1224
	Speed (V_d ***	-.3443	.0241	.000	-.3917	-.2970
	constant	6.8156	.7233	.000	5.3979	8.2333
Log likelihood = -2684.8196		LR χ^2 =1020.96	Prob.> χ^2 =0.0000	Pseudo R^2 =0.1598		

Note: significance codes: <0.01 ***, 0.05 **, and 0.1 *. Number of obs.=2938

Table 3. Elasticities and Contributions of Determinants (*X_{kj}* in 10% change)

Category	POH	<i>SD</i>		<i>L_s</i>		<i>F</i>		<i>V_d</i>	
		$E_{X_{kj}}^{P_j}$	impact	$E_{X_{kj}}^{P_j}$	impact	$E_{X_{kj}}^{P_j}$	impact	$E_{X_{kj}}^{P_j}$	impact
Low	18.61%	-0.869	2.40%	-16.978	46.93%	6.991	19.32%	11.339	31.34%
Medium	30.82%	-0.289	2.27%	-5.571	43.75%	2.193	17.22%	4.682	36.76%
High	57.62%	-0.173	2.37%	-2.733	37.36%	1.534	20.97%	2.875	39.30%

Note: POHs are the original probability of being at the high-income level.

Figures

Fig. 1. Study area and distribution of pickups. The red closed curve denotes the Outer Ring Road, which encompasses the downtown area of Shanghai. Approx. 82.25% of daily pickups happen within the downtown area.

Fig. 2. Trends of the ratio of empty/total mileage (RETM) and the number of taxis in service. The green box is a box and whisker diagram based on the five-number summary: minimum ($Q1 - 1.5IQR$), first quartile($Q1$), median($Q2$), third quartile($Q3$), and maximum ($Q3 + 1.5IQR$), where IQR is the interquartile range, equal to $Q3 - Q1$.

Fig. 3. Distribution of taxi drivers' incomes.

Fig. 4. Sample selection and categorization of taxi drivers. The candidate drivers are selected by two thresholds – standard deviation of rankings (horizontal dashed line, threshold: ≤ 1763.5) and average incomes (vertical dashed lines, thresholds include: ≤ 165.8 yuan for low-income; $184.7 \sim 199.5$ yuan for medium-income; and ≥ 215.5 yuan for high-income).

Fig. 5. Defining factors that might affect the income level based on a typical search-delivery scenario. Two alternative route plans that start from the same drop-off location are illustrated in the scenario. Each route has different search and delivery choices.

Fig. 6. Analysis of supply-demand ratio around the drop-off location. In the context of the supply and demand, the studied taxi A is mainly affected by any other vacant taxis in the buffer area (the square in orange).

Fig. 7. Extraction of features in a search route. The studied taxi is cruising through the search route (the consecutive light green cells), and the historical pick-up count in each cell may affect the time for searching next passengers and affect the incomes eventually.

Fig. 8. Extraction of features in delivery routes. The studied taxi is delivering passengers through the delivery route (the consecutive red cells), and the speed in each GPS coordinates may affect the time for carrying passengers and affect the incomes eventually, the fare is also an intuitive factor that may have an impact on incomes.

Fig. 9. The distributions of the five factors among three income levels. The five factors include (a) supply-demand ratio; (b) search distance; (c) search intensity; (d) fare; (e) delivery speed.

Fig. 10. Probability curves for different income levels and their probability change due to a change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.

Fig. 11. Probability curve of being high-income drivers (for average **low-income drivers**) and their probability increase due to change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.

Fig. 12. Probability curve of being high-income drivers (for average **medium-income drivers**) and their probability increase due to change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.

Fig. 13. Probability curve of remaining high-income drivers (for average **high-income drivers**) and their probability increase due to change of (a) supply-demand ratio; (b) search distance; (c) fare; and (d) delivery speed.