Understanding Taxi Service Strategies from Taxi GPS Traces
Daqing Zhang, Lin Sun, Bin Li, Chao Chen, Gang Pan, Shijian Li, and Zhaohui Wu

Abstract—Taxi service strategies, as the crowd intelligence of massive taxi drivers, are hidden in their historical timestamped GPS traces. Mining GPS traces to understand the service strategies of skilled taxi drivers can benefit the drivers themselves, passengers and city planners in a number of ways. This paper intends to uncover the efficient and inefficient taxi service strategies, based on a large-scale GPS historical database of approximately 7600 taxis over one year in a city in China. First, we separate the GPS traces of individual taxi drivers and link them with the revenue generated. Second, we investigate the taxi service strategies from three perspectives: passenger-searching strategies, passenger-delivery strategies, and service-region preference. Finally, we represent the taxi service strategies with a feature matrix and evaluate the correlation between service strategies and revenue, informing which strategies are efficient or inefficient. We predict the revenue of taxi drivers based on their strategies and achieve a prediction residual as less as 2.35RMB$^1$ per hour, which demonstrates that the extracted taxi service strategies with our proposed approach well characterize the driving behavior and performance of taxi drivers.

Index Terms—Taxi trajectory mining; taxi GPS traces; service strategies; revenue prediction

I. INTRODUCTION

With the ubiquity of mobile sensing devices, such as smart phones and GPS navigators, a large number of digital footprints characterizing people’s mobility behaviors have become available. These digital footprints provide us with a unique opportunity to understand human behaviors in various situations, and exploit the underlying intelligence [7], [12], [15], [20], [32].

In many cities, taxis are equipped with GPS devices, which periodically report to a central server the real-time information about the vehicle, including taxi locations and whether the taxi is occupied by a passenger. The collected GPS traces implicitly convey the service behaviors of the taxi drivers, such as where they pick up the passengers, how they find and deliver the next passengers, etc. These service behaviors vary from one driver to another, depending on one’s service strategy in a given situation. For example, after dropping off passengers, some drivers may prefer to wait for new passengers at some familiar places, while others may prefer to search for new passengers in a busy area.

The strategies employed by taxi drivers have a direct influence on the amount of time and distance the taxi is occupied/vacant, resulting in differences in revenue, fuel consumption and carbon emission. Good service strategies not only lead to high operating revenues but also improve efficiency of the entire taxi service system and bring better services to passengers. In addition, good service strategies can help reduce the carbon emission by either decreasing the vacant/occupied driving distance ratio or selecting energy-efficient routes. Thus, understanding the service strategies of taxi drivers can benefit the drivers themselves, passengers and city planners.

It is known that a number of factors affect the taxi drivers’ service strategies, including passenger demands along a hunting route, potential travel distance of passengers, waiting time, traffic conditions, competition from other taxis, and cost of fuel, etc. Most of the existing work intends to improve the taxi service performance by building models to explicitly take into account some of these factors. For instance, based on historical taxi GPS records, some recent work focuses on extracting passenger pick-up hotspots (i.e. location where the number of pick-up events is greater than a certain threshold) and urban human mobility patterns in order to aid vacant taxis in finding the next passengers [4], [13], [17], [21], [31]. Considering the competition introduced by other vacant taxis, Yuan et al. [31] propose to estimate the passenger searching possibilities of roads, which are used to infer the optimal passenger searching locations and routes for drivers. Even though increasing number of practical factors have been incorporated into the existing models, there is no analytical model proposed to systematically study which service strategies are efficient/inefficient for both passenger-searching and passenger-delivery.

In this paper, we intend to address the problem from a different perspective. As opposed to considering influencing factors separately, we argue that the taxi service strategies – the crowd intelligence of taxi drivers – are hidden in their GPS traces and are a combination of all the influencing factors. Given that the ultimate goal of taxi drivers is to achieve high revenue in their operations, we investigate the correlation between each service strategy and the resulting revenue in a specific context, and uncover which strategy is beneficial or harmful to the revenue, leveraging the GPS traces of a large number of taxis. Based on the existing work, we classify taxi service strategies as follows:

- **Passenger-searching strategies.** These strategies con-
sider the practical factors affecting a taxi driver’s decision making when searching for new passengers. The taxi driver may prefer to wait at a popular location nearby (e.g. grand hotels and railway stations) or go to a familiar area far away to search for new passengers.

- **Passenger-delivery strategies.** These strategies characterize the taxi drivers’ intelligence in delivering passengers. A driver may choose one route among several possible ones to deliver passengers, considering the traffic conditions, trip fares, etc.

- **Service area preference.** In a particular time period of the day, taxi drivers may prefer to serve in certain regions in a city by considering the traffic, passenger demands, and their familiarity with the regions.

In order to link the service strategies to the resulted revenue of taxi drivers, we need to address the following four challenges.

The first issue is to separate the GPS traces of individual taxi drivers so that each taxi driver’s service behaviors can be investigated and linked to the revenue. In Hangzhou, almost all the taxis are shared by two drivers, and these drivers take shift handover twice everyday (once in the morning and once in the afternoon), but the exact shift handover time changes from day to day and is not indicated in the GPS records.

The second issue is to measure a taxi driver’s performance in terms of the revenue per operation time slot, and investigate whether a taxi driver’s operation performance is consistent across all the time slots. This issue affects the selection of time granularity. Specifically, if a driver’s operation performance is consistent on a daily basis, we simply use the average daily revenue as the performance indicator; otherwise, we need to investigate the correlations between the service strategies and the average revenue at each time slot rather than on a daily basis.

The third issue is to extract and model the service strategies of taxi drivers. As the service strategies reflect the key decisions of taxi drivers at different time and locations, it is not always easy to describe and represent them, especially when they consist of a sequence of decisions. For example, for passenger searching strategies, a taxi driver may make a sequence of decisions after dropping off the current passengers and before picking-up the next passengers; we only investigate the first intention of the taxi driver and use it to represent his/her passenger-searching service strategy.

The last issue is to discover the efficient and inefficient service strategies. In our preliminary study presented in [14], we first built a feature matrix, in which each column represents a service strategy (a feature) and each row a taxi. L1-SVM highlights the salient strategies that can effectively separate good and ordinary taxis. This is because that L1-SVM can select strategies and assign the selected strategies with positive and negative weights to reflect their contributions in differentiating the two taxi groups. That is to say, the weights can be treated as the indicators of whether the corresponding strategies are efficient (ranked positive) or inefficient (ranked negative). However, as pointed out by Guyon and Elisseeff [10], a feature which is completely useless by itself can provide a significant performance improvement when taken with others. So those highly weighted strategies may be useless when working alone. Thus, it might be insufficient to merely use the L1-SVM weights to measure the efficiency of the strategies.

The main contributions of this paper can be summarized as follows:

- We propose a mechanism to identify when and where taxi drivers take shift handover in order to separate individual taxi drivers’ GPS trajectories. We further propose an algorithm to extract the initial intended location of taxi drivers after each drop-off event.

- We investigate the taxi service strategies from three perspectives: passenger-searching strategies, passenger-delivery strategies and service-region preference. In particular, we propose three concrete strategies for passenger searching, i.e. *hunting locally, waiting locally, and going distant*.

- We uncover the fact that most of taxi drivers do not perform consistently across all the time slots. Thus, the correlation between the service strategies and drivers’ performance in terms of revenue should be studied at the granularity of each time slot.

- We represent the taxi service strategies with a feature matrix, and use the correlation between different service strategies and their corresponding revenue to reveal which strategies are **efficient and inefficient**. By predicting the revenue of taxi drivers and achieving a prediction residual of less than 2.35 RMB per hour, we show that the extracted taxi service strategies with our proposed approach can well characterize the driving behavior and performance of taxi drivers.

The remainder of the paper is organized as follows. In Section II, we review the related work; in Section III, we introduce how to handle two data pre-processing tasks: separate the GPS traces of each taxi driver for the shared taxis and estimate the taxi revenue and performance consistency according to the recorded GPS traces. Section IV discusses how to extract and represent the taxi service strategies. In particular, we present a novel mechanism to extract the initial intention of a taxi driver after dropping off passengers. In Section V, after representing the service strategies using a feature matrix, we evaluate the correlation between each strategy and the generated revenue to uncover the efficient and inefficient service strategies. We further predict the revenue of taxi drivers based on their chosen strategies and demonstrate that the extracted service strategies with our proposed approach properly characterize the behaviors and performance of taxi drivers. Section VI concludes the paper.

**II. RELATED WORK**

Mining taxi GPS traces has received increasing attention from the data mining, intelligent transportation, database, and ubiquitous computing communities [2], [34]. A variety of research issues have been addressed leveraging large scale GPS traces, such as urban human mobility understanding [1], [11], [4], [8], [19], [26], [17], urban planning [30], [16], [29], [35], traffic prediction [24], [9], [23], anomalous trajectory
In this section we briefly present the related work on improving taxi drivers' performance, which can be roughly classified into the following three categories.

The first category of research focuses on identifying and recommending popular pick-up areas [13], [4], [21], [26], [17]. Since the change of taxi status from vacant to occupied is recorded in the traces and indicates a passenger pick-up event, the collection of such events from a large taxi fleet reflect the taxi demand in the city. By clustering the GPS locations of these pick-up events using $K$-means at different time of the day and in different days of the week. Lee et al. [13] analyzed the pick-up patterns of taxis in Jeju, Korea and recommended the popular clusters to vacant taxis to reduce the idling time of taxis. Chang et al. [4] gathered the demand requests according to the time, location and weather context, clustered these requests into hotspots by $K$-means, Agglomerative Hierarchical Clustering and DBSCAN, and recommended the top ranked places to taxi drivers. Li et al. [17] predicted the number of pick-up events in different hotspots based on the historical information recorded in the taxi GPS traces and provided guidance to vacant taxis drivers.

Besides taking into account the pick-up hotspots, the second category of research work also considers other practical influencing factors, such as human mobility patterns, passenger searching possibilities and potential trip lengths, and provides optimized passenger searching areas and routes.Unlike the methods focusing on hotspots globally, Powell et al. [21] examined only the surrounding areas. They measured the profitability of each area in terms of fare gains of all occupied trips originated from that area, the number of trips and the cost from the current location to that area. Exploiting the knowledge of passenger’s mobility patterns and taxi drivers’ pick-up/drop-off behaviors inferred from taxi GPS traces. Yuan et al. [31] used the historical probability of searching a passenger along a route to provide drivers with location/route suggestions.

Instead of providing explicit guidance about areas or routes for searching passengers, the last category of research tries to extract effective taxi service guidelines in a city. Veloso et al. [26] investigated the passenger delivery patterns and passenger searching processes, and revealed that in Lisbon, Portugal, a good passenger-searching strategy in urban areas was that taxis normally went to adjacent locations while in suburban areas taxis went to distant locations. By analyzing the taxi GPS traces, Liu et al. [18] uncovered that, in the city of Shenzhen, China, high revenue taxi drivers had good skills of choosing the right areas to serve in the city at different time of the day and selecting the delivery routes with less traffic congestion. These studies are closer to our work than the first two categories because they provide strategic service guidelines rather than concrete areas/routes to taxi drivers. However, [26] did not tackle the passenger-delivery and area preference strategies while [18] failed to differentiate taxi GPS traces for the paired drivers. Neither [26] nor [18] built the model characterizing the relationship between taxi service strategies and the resulted performance. Besides leveraging taxi GPS traces, Takayama et al. [25] used survey information from taxi drivers to recommend promising waiting/cruising locations, and Yamamoto et al. [27] suggested routing strategies for multiple taxis by mutual exchange of their pathways. However, both methods require human intervention.

Different from all previous work, we first identify a series of taxi service strategies for both passenger-searching and passenger-delivery. Then, we conduct a correlation analysis study to extract the efficient and inefficient taxi service strategies at different time and locations. Finally, we build a model based on L1-SVM using top-ranked service strategies, and validate the effectiveness of the extracted service strategies.

III. DATA PREPROCESSING AND EMPIRICAL STUDY

We begin this section by introducing the taxi GPS traces collected for our study. The large-scale taxi GPS data set was acquired from approximately 7600 taxis served in Hangzhou, a megacity in China, for one year (April 2009 ~ March 2010). Each taxi is equipped with a GPS device which acquires the real-time taxi information, including its longitude/latitude, the time-stamp, the passenger status (“occupied” or “vacant”), the driving speed and orientation. It uploads the data to a central server via telecommunication network at a sampling rate of once per minute (i.e. the GPS sampling frequency is about one sample per minute). To avoid mistakes caused by device error or network failure, we first filter out suspicious taxi records, such as taxi being vacant for the entire day, or lacking records for extended periods of time, and obtain 6863 taxis to conduct the research.

On a two dimensional plane, a taxi’s moving trajectory over a time interval can be depicted by connecting the GPS points. For instance, Fig. 1 shows one taxi’s trajectory for a complete passenger-searching and passenger-delivery cycle, where red lines and blue lines correspond to the passenger delivery stage; while the taxi is occupied, it is said to be in the passenger delivery stage; while the taxi is vacant, it is usually in the passenger searching stage. The change from red to blue corresponds to a passenger pick-up event and from blue to red indicates a drop-off event.

Before analyzing taxi driver’s service strategies, two data preprocessing tasks need to be done:

- **Individual Taxi Driver GPS Trace Extraction:** Almost all the taxis in Hangzhou are served by two drivers, one works in daytime and the other in nighttime. As the shift status of a taxi is not recorded in the taxi GPS traces, it is necessary to detect when the shift handover takes place in order to separate the taxi GPS traces of each driver. Only with each taxi driver’s GPS traces extracted, individual drivers’ service strategies and revenues in different timeslots can be known.

- **Taxi Driver’s Performance Quantification:** In order to discover what service strategies are efficient and inefficient, we should be able to identify “good” and “ordinary” drivers by linking the service strategies with the revenues generated. However, the taxi trip fares are not recorded in each taxi’s GPS traces, thus it’s necessary to quantify each taxi driver’s performance in each time slot according to the revenues calculated from the generated GPS traces.
In the following sections, we will present how to deal with the above two data preprocessing problems. We will also do an empirical study based on the processed data to visualize some statistical results.

A. Extracting Individual Driver's Digital Traces

For shared taxis, normally the two drivers set up an agreement on when and where to hand over the taxi. Through interviews with taxi drivers in Hangzhou, we are told that the afternoon shift handover usually happens between 16:00~18:00pm and the morning shift handover takes place around 06:00am. The morning shift handover is generally more flexible than the one in the afternoon, because many night shift drivers park the taxis at the agreed location around 03:00am so that the daytime drivers could take the taxis early if they wish. However, due to issues such as traffic, delay of last passenger deliveries, or personal reasons, it is hard to strictly obey the shift handover agreement every day, especially for the afternoon shift handover. Thus, the exact shift handover time detection becomes a non-trivial task in order to extract individual taxi driver’s GPS traces.

1) Shift Handover Event Detection: The detection of shift handover events is based on two observations: (1) shared taxis go to the same agreed location during the rough shift handover time slot of the day with vacant status; (2) shared taxis park at the agreed shift handover location for some time (with vacant status) to ensure taxi handover.

The detailed shift handover event detection process is elaborated as follows. First we identify the parking locations in a passenger searching trajectory using the same method reported in [31]. We divide the city map into grid cells of 50m × 50m and split a day into half-overlapping 20-minute time frames. In each time frame, we record the days that a taxi visits each grid cell. If the shift handover location falls near the edge of a cell, the shift handover events may scatter in the neighboring cells due to GPS error. To address this problem, besides counting the days a taxi visits each cell, we also count the days it visits the neighboring cells. Since we count the visit to neighboring cells, the same shift handover event may correspond to several candidate cells. We can combine these candidate cells into one as they might be produced from the same traces. In order to accommodate early or delayed shift handover events, we expand the time frame for shift handover event detection to 2 hours. In real life, there is anomalous shift handover that the two drivers change the shift handover time and location for reasons like personal issues or temporarily blocked roads. So we may be unable to find the shift handover event in all days. In practice, if the number of days that the taxi visits a grid cell during the rough shift handover time slot is greater than 25 in a month, we mark the grid cell as the candidate shift handover location. Furthermore, if it is the only candidate within the 15:00~19:00 time slot, it is said to be the afternoon shift handover location; if it is the only candidate within the 03:00~8:00am time slot, it is the morning shift handover location. If there are more than one candidate cells, we consider it as an uncertain (failed) case.

With the shift handover event detected, the GPS traces of the two shared taxi drivers are naturally separated. For the anomalous shift handover aforementioned, we simply choose the vacant parking location in the digital trace that is closest to the middle of the detected shift handover time slot and separate the digital trace of that day at that point.

One possibility that may cause false shift handover event detection is when some drivers wait for passengers everyday at certain popular locations (e.g., railway/bus stations, grand hotels, etc.) during a certain period of time. This appears to have a similar pattern as the shift handover event. However, we observe that the majority of these passenger waiting events are followed by an immediate “occupied” status (indicating that a passenger is picked-up successfully); while the shift handover event seldom takes place at a popular place and the taxi usually has vacant status after the taxi leaves the shift handover location. In such a way we can filter out most of the false shift handover events.

With the proposed algorithm, we successfully identified the shift handover events (including location and time) for 4773 taxis out of the 6863 taxis. The remaining 1957 taxis do not follow the fixed shift handover pattern, which may be due to the following two reasons: 1) the location and time to handover the taxi vary from day to day based on two drivers’ agreement; 2) there are some taxis served by only one driver. To exclude the error introduced by the shift handover event detection algorithm, we only keep the taxi traces from the taxis whose shift handover events were identified successfully (i.e., the 4773 taxis) for experiments. It is believed that the 9546 individual drivers’ traces from 4773 taxis are enough to illustrate our proposed idea and show meaningful statistical results for discovering efficient and inefficient taxi service strategies.

2) Statistical Study of Taxi Shift Handover Location and Time: The geographical distributions of the morning and afternoon taxi shift handover locations are shown in Fig. 2 (a) and (b), respectively. It can be observed that most of the taxi shift handover events are not taking place in popular areas such as the downtown and railway stations. Another interesting observation is that many drivers choose to handover the taxi near bus stops during afternoon shift handover time slot. This may be because the taxi drivers need to take buses for the taxi handover.
The time distribution of morning and afternoon taxi shift handover events are shown in Fig. 3 (a) and (b), respectively. The morning taxi handover time is almost evenly distributed between 4:00~7:20am; while the afternoon shift handover events mainly take place between 16:40~17:20pm, which explains why people in Hangzhou find it difficult to find taxis around this time period.

B. Analyzing Individual Driver’s Performance

Based on the separated digital traces for each individual driver, we can further extract the passenger delivery traces of each driver according to the passenger status (“occupied” or “vacant”). We can then estimate the taxi revenue in each time slot by considering the accumulated passenger delivery distance. By averaging the revenue in each time slot during a long time period (e.g., a month), we are able to measure the performance of a driver. In this subsection, we will investigate the performance of all the taxi drivers in different time slots and give an empirical study on the factors influencing the performance.

1) Taxi Performance Quantification: First we split a day into five time slots: late night (00:00~05:59), morning (06:00~09:59), noon (10:00~13:59), afternoon (14:00~17:59), and evening (18:00~23:59). Since workdays have quite different taxi service patterns from weekends and holidays, in the paper we only use the taxi GPS traces in workdays to illustrate our ideas. The performance of a taxi is measured by hourly revenue within each time slot. Since taxi drivers may only work part of the time within a time slot, for example, one night shift driver might work until 4:00am; another driver might work until 06:20am and hand over the taxi to the partner (exceeding 20 minutes in the morning time slot), we only count the taxis which serve more than half of the time slot and discard those taxi GPS traces which serve less than half of the time slot.

The distribution of taxi drivers’ performance in different time slots is shown in Fig. 4. We can observe that, generally speaking, taxi drivers perform best in the noon time slot (around 30~45 RMB/ Hour) and worst in late night time slot (around 5~20 RMB/ Hour); they also have better performance in the evening and afternoon time slots than the morning time slot. The distribution of taxi driver performance in each time slot roughly follows a normal distribution, which is compliant with the observation in [18]. Since we aim to investigate taxi service strategies in terms of drivers’ performance, we first investigate the consistency of individual driver’s performance over different time slots. People might think that good taxi drivers perform well in all the time slots and vice versa. Through our study, it is found that only 110 taxis performing best in the late night time slot are among the top 500 taxis performing well during the nighttime (containing both the evening and late night time slots), and only 67 taxis performing best in the daytime are among the top 500 taxis if we count the performance in the three time slots containing morning, noon, and evening. Our study shows that less than 22% of the taxi drivers consistently perform well in the nighttime and only 13.4% in the daytime. Thus, unlike the method proposed in [14], which chooses good/ordinary taxis according to the daily revenue, we investigate taxi drivers’ performance in each time slot and expect to obtain more reasonable and accurate results.

For each time slot, we select the 500 top performing taxis as good samples and 500 taxis in the middle range (around the mean of the normal distributions in Fig. 4) as ordinary taxis. The reason why we do not select the taxis in the bottom range is that many factors might cause taxis’ extra low performance and these factors are irrelevant to taxi service strategies. By comparing good taxis with ordinary ones, we can discover some simple taxi service patterns that are closely related to
revenue. In the following subsection, we will report a brief empirical study on two influencing factors.

2) The Number of Passenger Delivery Trips: Intuitively, a larger number of passenger delivery trips should give a higher overall revenue. We use the correlation coefficient between the number of passenger delivery trips ($P$) and revenue ($R$)

\[ corr(P, R) = \frac{\sum_{i=1}^{n}(p_i - \bar{p})(r_i - \bar{r})}{\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2} \sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}} \]  

(1)

to validate this intuition, where $p_i$ and $r_i$ are the total number of passenger delivery trips and the revenue of the taxi driver $i$, respectively, and $n$ denotes the number of taxi drivers. The results for all the time slots are shown in Table I. The correlation coefficient in most time slots is close to 1, suggesting that the revenue is highly correlated to the number of passenger delivery trips.

We also compare the average number of passenger delivery trips for good and ordinary taxis in Table II in all time slots. The number of passenger delivery trips per hour is displayed as “Mean±Std”. We can see that good taxi drivers complete more trips than ordinary ones, where the difference is at least 21% in the afternoon time slot and at most 87% in the night time slot. These results imply that good drivers are always more efficient in finding next passengers than ordinary ones.

3) Popular Passenger Pick-up and Drop-off Areas: Another straightforward but important influencing factor on taxi drivers’ performance is the taxi operation area. This factor has been considered in a number of studies [13], [4], [21] to guide taxi drivers to find passengers. In Fig. 5, we plot the top 99 pick-up and drop-off hotspots at different time slots. We can see that the railway station and the Zhejiang University are among the popular pick-up and drop-off areas across all the time slots. Residential areas become popular at night as many people are returning home. The pick-up number decreases greatly from the downtown to suburb areas. It is noted that the numbers in grid cells of Fig. 5 refer to the ranking of the area in terms of the number of pick-up or drop-off events. Apparently, the same grid cell may have different numbers in different time slots (night, morning, noon, afternoon, evening), due to the variation of the taxi and passenger demands.
IV. TAXI SERVICE STRATEGY FORMULATION

We investigate taxi drivers’ service strategies from three perspectives: 1) passenger-searching strategies, 2) passenger-delivery strategies, and 3) service area preference. For passenger-searching strategies, we group drivers’ preference into three possible behaviors, namely: hunting locally, waiting locally, and going distant (i.e. travelling to a distant location). For passenger-delivery strategies, we study their average passenger-delivery speed, which potentially reflects the drivers’ ability to choose clearer routes when delivering passengers. For service area preference, we study their preference of service areas in a city. The overview of the taxi service strategies is illustrated in Fig. 6.

In the following, we will first propose a method to identify a taxi driver’s initial intention right after dropping off the passengers. Based on the initially intended path, we then introduce a method to extract the three types of taxi service strategies from the path. Finally, we will construct a driver-strategy matrix for mining useful taxi service strategies.

A. Initial Intention Identification

Passenger searching strategies refer to taxi drivers’ behaviors after a drop-off event. However, a passenger searching process may consist of a series of decisions. When studying taxi drivers’ behaviors after dropping off the passengers, we need to first identify a driver’s initial intended location, before which the trace will be used for passenger-searching strategy extraction.

It is difficult to tell the initial intention of a driver merely based on his/her digital traces. Consider the trace in Fig. 1: After dropping off passengers at location A, the driver started to search for new passengers. He drove to location C, where he waited 10 minutes but failed to search passengers. Then he decided to go to location B, where he succeeded in finding passengers. In this example, the final pick-up location is not where the driver initially intended to go right after the last drop-off event; otherwise, he would have followed the most efficient way to location B. Thus location C should be the driver’s initial intention. If we simply extract the passenger-searching strategies based on the trace between the last drop-off location and the pick-up location as that in our previous work [14], we would regard that the driver chooses a “local hunting” strategy (since locations A and B are close) instead of his/her initial intention which is actually a “distant hunting” strategy (since location C is far from location A). As we want to study a driver’s decision at each drop-off location, we should investigate what the driver intended to do right after he dropped off passengers instead of simply observing the final pick-up location.

We consider the longest normal and non-waiting trajectory starting from the last drop-off event as the “initial intended path”. Here, a trajectory $t_f$ is a sequence of points $(p_1, p_2, \ldots, p_n)$, where $p_i$ is a GPS reading (latitude, longitude) sampled during a passenger searching or delivery process. For a passenger-searching trajectory $t_f$ whose source location is $S$ (where the corresponding drop-off event takes place), we first collect the passenger-delivery trajectories that either originate from or pass through $S$ to form a trajectory database $\mathcal{T}$. Intuitively, as most passenger-delivery trajectories follow efficient routes, we can detect the longest normal trajectory starting from $S$ by comparing it with the trajectories in $\mathcal{T}$.

Detecting the longest normal trajectory is equivalent to the problem of detecting the first anomaly starting from $S$ without a given destination. Anomalous trajectory detection has been well investigated in our previous work [33], [6], [5] when its destination is given. The proposed method in [5] is based on a trajectory data set that shares the same source and destination areas as the test trajectory. However, for this problem the destination is not provided. One intuitive way is to assume that, starting from $S$, each sampling point is a destination and then we see whether anomaly occurs at that point. However, it is quite time and resource consuming if we need to perform anomaly detection at all the sampled points on the trajectories of all the taxis.

Before proceeding to extract taxi service strategies, we first introduce a method to detect initial intended locations. We denote the $i$th trajectory in $\mathcal{T}$ as $t_i$. For each $t_i$, we map it into a sequence of adjacent cells, through which it traverses, using the same method as in [33]. We denote the mapped trajectory as $\bar{t}_i = \{g_1, g_2, \ldots, g_m\}$, where $g_i$ denotes a cell index. For a mapped test trajectory $t_f = \{g_1, g_2, \ldots, g_m\}$, if we can find $g_1, g_2, \ldots, g_m$ sequentially in $\bar{t}_i$, we say $t_i$ complies with $t_f$. We define $\text{hasPath}(\mathcal{T}, t_f)$ to indicate whether there exist a trajectory in $\mathcal{T}$ that complies with $t_f$.

\textbf{Definition 1:} Given a trajectory $t_f$, whose mapped trajectory is $\bar{t}_f = \{g_1, g_2, \ldots, g_m\}$, and a trajectory data set $\mathcal{T}$,

\begin{equation}
\text{hasPath}(\mathcal{T}, t_f) = \left\{ t' \in \mathcal{T} \left| \begin{array}{l}
(i) \forall g_i \in \bar{t}_f, \\
1 \leq i \leq m, \\
(ii) \exists \text{pos}(t', g_1) < \text{pos}(t', g_2) < \ldots < \text{pos}(t', g_m)
\end{array} \right. \right\}
\end{equation}

where $\text{pos}(\bar{t}', g_i)$ denotes the index of the cell in $\bar{t}'$ that matches the location of $g_i$.

We also define $\text{passTraj}(\mathcal{T}, g)$ as the trajectories in $\mathcal{T}$ that pass through cell $g$.

\textbf{Definition 2:} Given a trajectory data set $\mathcal{T}$ and a cell $g$,

\begin{equation}
\text{passTraj}(\mathcal{T}, g) = \left\{ t' \in \mathcal{T} \left| g \in \bar{t}' \right. \right\}
\end{equation}

Below we define what is an anomalous trajectory $t$ with respect to a data set $\mathcal{T}$.
Definition 3: Given a threshold $0 \leq \beta \leq 1$, a trajectory $tf$, whose mapped trajectory is $\overrightarrow{tf}=(g_1, g_2, \ldots, g_m)$, is $\beta$-anomalous with respect to a set of trajectories $T$ if

$$Support(T, tf) = \frac{|\text{hasPath}(T, tf)|}{|\text{passTraj}(T, g_m)|} < \beta \quad (4)$$

The difference between $\theta$-anomalous defined in [6], [5] and the proposed $\beta$-anomalous in the paper is that, $\theta$-anomalous determines whether the pattern of $tf$ is rare in the trajectories of $T$, while $\beta$-anomalous determines whether the pattern of $tf$ is rare in the trajectories that start from the source cell to the last cell of $tf$ in $T$.

We can then detect the initial intended path as follows: Starting from $g_2$ of $tf$, if the trajectory $\overrightarrow{t'}=(g_1, g_2)$ is not $\beta$-anomalous, we include the next cell into the trajectory until we find that $\overrightarrow{t'}=(g_1, g_2, \ldots, g_k)$ is $\beta$-anomalous or $\overrightarrow{t'}=tf$. We say $\overrightarrow{t'}$ is the initial intended path of $tf$.

B. Taxi Service Strategy Extraction

In the following, we introduce how to extract the taxi service strategies from the three perspectives outlined in the beginning of this section.

1) Passenger-Searching Strategies: After detecting the initial intended paths, we proceed to extract the passenger-searching strategies of individual drivers after dropping off passengers. We use the same partition of the city as that in Fig. 5. For each time slot, we select the top 99 busiest cells and treat the rest as an entire non-popular area; and finally obtain 100 location labels. For each location, we are interested in understanding drivers’ preferences on three types of passenger-searching strategies after dropping off passengers, i.e., hunting locally, waiting locally, and going to distant locations. The local and distant properties are determined by comparing the distance $d_{drop}$ between the first and the last cells of the initial intended path to a threshold $\tau_d$. The hunting and waiting properties depend on whether the trajectory is ended with a waiting event whose time duration $t_{wait}$ is longer than $\omega_d$. In this study, we empirically set $\tau_d=1.5$ kilometers and $\omega_d=5$ minutes. The criteria for determining these properties are as follows.

$$d_{drop} \begin{cases} > \tau_d & \text{going distant} \\ \leq \tau_d & \begin{cases} t_{wait} > \omega_d & \text{waiting locally} \\ t_{wait} \leq \omega_d & \text{hunting locally} \end{cases} \end{cases} \quad (5)$$

For a specific location, we count the number of going distant event $s_{dd}$, hunting locally event $s_{dh}$, and waiting locally event $s_{dw}$. In [14], we construct the taxi-pattern matrix directly with these statistics. However, as the total number of drop-off events of the good drivers are significantly larger than that of the ordinary drivers, in most locations, the number of committing specific strategies for the good drivers is generally larger than that of the ordinary drivers. This difference cannot reveal an individual driver’s preference on different strategies. Thus, we define a notion called strategy preference $SP$, which represents the proportions of different passenger-searching strategies that a driver adopts at location $l$ in time slot $t$.

$$SP(s^{l,t}) = \frac{s_{dd}^{l,t} + s_{dh}^{l,t} + s_{dw}^{l,t}}{s_{dd}^{l,t} + s_{dh}^{l,t} + s_{dw}^{l,t}} \quad (6)$$

where $SP(s^{l,t})$ indicates a driver’s inclination to certain strategies and this normalization avoids the bias of the model towards the strategies with large number of drop-off events.

As a driver may have three strategies at each location during different time slots after dropping off passengers, we build a 100 locations $\times$ 3 strategies $\times$ 5 time slots $=1500$-dimensional feature vector for each driver. Each dimension of the feature vector corresponds to a specific $(\text{location}, \text{time}, \text{strategy})$ combination.

2) Passenger-Delivery Strategies: We consider passenger-delivery speed, which is an average speed over all the passenger delivery trajectories of a driver, as a passenger-delivery strategy. This is based on the observation that a higher value implies that the driver has good skill to choose efficient passenger-delivery routes. We calculate the average passenger-delivery speed in each hour as follows

$$\text{Speed}_i = \frac{\sum d_{delivery}^i}{\sum t_{delivery}^i} \quad (7)$$

in which $d_{delivery}^i$ and $t_{delivery}^i$ are the delivery distance and time of a trip in hour $t$, respectively. Then for each driver, we build a feature vector with each dimension corresponding to one hour.

3) Service Area Preference: As shown in [18], high revenue drivers are capable of choosing to serve at certain city areas to make high profit and meanwhile avoid heavy traffic in Shenzhen, China. In this paper, we also investigate the preferences of passenger service areas of different taxi drivers. Unlike passenger-searching strategies, we divide the city into $10 \times 5$ areas, each of which is about $5km \times 5km$. For each driver, we count the number of passenger-searching trips $pft_i$ in area $i$ in each time slot. The preference of area $i$ is defined as:

$$P_i = \frac{pft_i}{\sum_i pft_i} \quad (8)$$

which measures a driver’s preference of service in each area. For each time slot, we build a 50-dimensional feature vector, with each dimension corresponding to the preference to a particular area.

C. Driver-Strategy Matrix Construction

For each driver, we combine the three types of feature vectors into one row vector. Then we construct a driver-strategy matrix $X$ by stacking individual drivers’ feature vectors in rows (as shown in Fig. 7). In the obtained matrix $X$, each row corresponds to one taxi and each column corresponds to the service strategy in a specific time slot and/or for a specific grid
V. UNDERSTANDING TAXI SERVICE STRATEGIES

In this section, we will leverage the driver-strategy matrix to investigate the taxi drivers’ service strategies. Specifically, two studies are conducted: 1) We attempt to uncover effective taxi service strategies and validate their effectiveness in classifying good and ordinary taxi drivers. The discovered service strategies can be used as guidance for taxi drivers. 2) We estimate individual taxi drivers’ revenues based on their taxi service strategies. This result can be used to predict a driver’s future performance based on the historical driving behaviors.

A. Discovering Good Taxi Service Strategies

For service strategy vector shown in each column of the driver-strategy matrix, we evaluate its impact on taxi drivers’ performance by computing its correlation with the revenue of the corresponding taxi drivers. A positive correlation value indicates that this strategy generally brings benefits to drivers’ performance, while a negative value indicates that the strategy will not help. In the following, we study the impact of the three types of strategies.

1) Passenger-searching strategies: To understand how different passenger-searching strategies would affect the drivers’ revenues generated, we fix the grid cell ID as shown in Fig. 8 (same as the ID numbers in the bottom figure of Fig. 5), then compute the correlation between each passenger-searching strategy vector at a specific grid cell and time slot after creating the driver-strategy matrix with 3 groups of grid cells.
For grid cells of Group 1, it can be seen from Fig. 9 that the revenue more significantly than the local waiting strategy. The correlation value of all the three strategies are zero. During these two periods, thus the correlation values for all the three strategies are zero.

For the local waiting strategy, the absolute correlation value is often negative except for the late night time slot. We can also see that it is positive at the late night time slot for all the selected grid cells, and is as big as that of the local hunting strategy, indicating that taxi drivers could either choose to travel to the nearby grid cells or go to a distant popular grid cell to search new passengers, especially for grid cells in Group 2. However, the local hunting strategy is more preferable since it consumes less fuel than going to distant.

2) Passenger-delivery strategies: For passenger-delivery strategies we aim to investigate the influence of the ability of choosing clear routes to the driver performance. It is easy to understand that both good and ordinary taxi drivers can choose clear routes when the traffic in all the road network is smooth, for example, at night. However, good drivers usually choose better routes when the road network is congested, which make their generated revenue higher than ordinary drivers. In this study, we use the average passenger delivery speed of the taxis to represent the capability of taxi drivers in choosing clear routes. We plot the correlation values of the average passenger delivery speeds in each time slot with drivers’ hourly revenue in Fig. 10, with respect to the average passenger-delivery speeds. We can easily see that, during late night, the traffic is clear and the correlation values are around zero (i.e. no much influence). In day time when the traffic is more congested, the correlation value is more than 0.1. More specifically, when the difference between the passenger delivery speed of good and ordinary taxi drivers is larger, the correlation value becomes bigger. This result confirms that good drivers do have better skills of choosing routes with light traffic.

3) Service area preference: The correlation value between the preference of service area and the hourly revenue rate generally reflects whether an area is worth serving for a taxi. The results are shown in Fig. 11. We can see that, in late night, taxi drivers had better serve in the entertainment area (the red zone in Fig. 11(a)), because there are more passengers who need to be delivered to their homes after night life. Meanwhile, the blue areas are those areas where the taxi demands are few (most of them are suburban areas), so it is better for taxis to go back to the top hotspots directly instead of staying in those areas.

Finally, we validate the effectiveness (and ineffectiveness)
of the discovered good (and bad) strategies by using them to classify good and ordinary taxi drivers with support vector machine (libsvm [3]). We use the features with the top 1/3 correlation values and those with the bottom 1/3 correlation values to classify the good and ordinary taxi drivers. We use 5-fold cross validation to evaluate the classification accuracy. Specifically, we randomly split the good and ordinary taxi drivers in 5 random groups and choose 4 groups as the training data set and the remaining one as the test data set. The evaluation results are shown in Fig. 12. We can see that the strategies with the top 1/3 correlation values have much better classification accuracy than those with the bottom 1/3 correlation values.

B. Performance Prediction Based on Historical Strategies

Taxi service strategies are key factors that influence the revenue performance of a taxi driver. Here we show that it is feasible to predict the performance of a taxi driver based on his/her historical strategies. We randomly split all the drivers in 5 groups. Again we use four groups as the training set and the remaining group as the test set. We use support vector regression (libsvm [3]) to train a regression model on the driver-strategy matrix of the training drivers. Then we use the obtained model to predict the revenue of a test driver in the test data set based on their strategies. For a driver \( d_i \), if the predicted hourly revenue rate is \( y_{t_i} \), and the true value is \( y'_t \), then the residual is \( r_i = |y'_t - y_{t_i}| \). We use two metrics to evaluate the prediction error: the mean residual \( r = \text{mean}(r_i) \), which reflects the average accuracy of the prediction; and the relative residual of taxi drivers, which is computed as \( \epsilon_r = \text{mean} \left( \frac{r_i}{y_{t_i}} \right) \), indicating the percentage of the mis-prediction against the true value. A smaller value corresponds to a more accurate prediction result.

The results in all the time slots are shown in TABLE III. Over all the time slots, we can obtain the residuals that are less than 2.35RMB/h, which suggests a quite accurate prediction result. The relative residuals in all the time slots except night time are quite small, which fully validate the feasibility of using the extracted taxi service strategies to predict taxi performance. The reason of the relatively low prediction accuracy in night time is probably because that, taxi drivers may choose to sleep for some time in late night and this behavior influences the revenue performance of the drivers.

<table>
<thead>
<tr>
<th>time slots</th>
<th>night</th>
<th>morning</th>
<th>noon</th>
<th>afternoon</th>
<th>evening</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \epsilon_r )</td>
<td>0.155</td>
<td>0.078</td>
<td>0.056</td>
<td>0.057</td>
<td>0.063</td>
</tr>
<tr>
<td>( r ) (RMB/h)</td>
<td>2.35</td>
<td>2.34</td>
<td>2.21</td>
<td>1.96</td>
<td>2.25</td>
</tr>
</tbody>
</table>

VI. Conclusion

Taxi GPS traces are valuable resources to investigate taxi drivers’ service behaviors. In this paper, we study the taxi drivers’ service strategies based on the activities of thousands of taxi drivers revealed in a real life large scale taxi GPS dataset collected from 7600 taxis in Hangzhou, aiming to provide useful guidance to taxi drivers. To understand the behaviors of a taxi driver, we first propose to separate the taxi GPS traces of each pair of shared taxi drivers based on two observations: 1) The shift handover location and time are more or less fixed on most days; 2) The taxi generally stops for a while in the shift handover location and shift handover time slot for handover, with vacant status before and after the shift handover event. We then convert each taxi delivery distance into corresponding taxi revenue in each time slot and find that, most of the taxi drivers don’t perform consistently well across different time slots. Thus we proposed to study taxi drivers’ behaviors in each time slot rather than on daily basis.

In this paper, we proposed to model the taxi driver’s service strategies from three perspectives: passenger-searching strategies, passenger-delivery strategies, and service area preference. We further represented the service strategies using a feature matrix. In order to accurately identify the passenger searching strategies, we derived a method to learn the initial intention of taxi drivers right after each drop-off event. Through measuring the correlation between each service strategy and the corresponding revenue, we reveal the efficient and inefficient strategies in each time slot and location. Generally speaking, it is found that hunting is usually more efficient than waiting in
order to find passengers locally, with a few exceptions, such as in airport where people go to fixed places to take taxis. Going distant becomes a preferable service strategy when a taxi drops-off passengers in the suburb area, where taxis usually spend less time on average finding the next passengers by moving from non-hot areas to hot ones. The correlation of the taxi average passenger delivery speed and the revenue shows that, when the traffic becomes congested at busy time slots, choosing the light-traffic route increases the driver’s revenue. We also found some service areas are more profitable than others for taxi drivers in different time slots of the day.

Taxi service strategies are key factors influencing the taxi drivers’ revenue. Based on the extracted historical service strategies, we predict the performance of a taxi driver with a support vector regression method. We obtain a residual of less than 2.35 RMB/hour, suggesting that the extracted taxi service strategies with our proposed approach well characterize the driving behavior and performance of taxi drivers.

In the future, we plan to broaden and deepen this work in two directions. First, we plan to conduct further research in characterizing subtle features of the taxi service behaviors and strategies, and to understand the human decision making process; Second, we attempt to explore the appropriate ways which provide more concrete instructions to taxi drivers according to the taxi service strategies.

ACKNOWLEDGEMENTS

The authors want to thank the anonymous reviewers and the editors for their helpful comments and suggestions. The authors also want to thank Dr. Pablo Castro for the proof-reading. This research was partly supported by the Institut Mines-TElECOM through the “Futur et ruptures” Program, the National Key Basic Research Program of China (2013CB329504), and New Century Excellent Talents in University Program (NCET-13-0521). C. Chen is the corresponding author of this paper.

REFERENCES

Daqing Zhang is a professor at Institut Mines-TELECOM/TELECOM SudParis, France. He obtained his Ph.D from University of Rome La Sapienza and the University of L’Aquila, Italy in 1996. His research interests include large-scale data mining, urban computing, context-aware computing, and ambient assistive living. He has published more than 180 referred journal and conference papers, all his research has been motivated by practical applications in digital cities, mobile social networks and elderly care.

Dr. Zhang is the Associate Editor for four journals including ACM Transactions on Intelligent Systems and Technology. He has been a frequent Invited Speaker in various international events on ubiquitous computing. He is the winner of the Ten Years CoMoRea Impact Paper Award at IEEE PerCom 2013, the Best Paper Award at IEEE UIC 2012 and the Best Paper Runner Up Award at MobiQuitous 2011.

Lin Sun received the Ph.D. degree from Pierre and Marie Curie University, Paris, France, and Institut Mines-TELECOM/TELECOM SudParis, Evry, France, in 2012. His research interests include context-aware computing, social and community intelligence, pervasive elderly care, and trajectory data mining.

Bin Li received his PhD degree in Computer Science from Fudan University, Shanghai, China, in 2009. He is currently a Researcher in Machine Learning Group, National ICT Australia (NICTA). Prior to this, he was a Lecturer in the Centre for Quantum Computation & Intelligent Systems (QCIS), University of Technology, Sydney (UTS), Australia (2011-2013) and a Postdoctoral Research Fellow at the Institut Mines-TELECOM/TELECOM SudParis, Evry, France (2009-2010). Dr Bin Li’s research interests include machine learning, and its applications to recommender systems and real-world big data analytics.

Chao Chen received the Ph.D degree from Pierre and Marie Curie University, Paris, France, and Institut Mines-TELECOM/TELECOM SudParis, Evry, France, in 2014. He received the B.Sc. and M.Sc. degrees in control science and control engineering from Northwestern Polytechnical University, Xian, China, in 2007 and 2010, respectively. He is currently an Associate Professor at Chongqing University, Chongqing, China. Mr. Chen was a co-recipient of the Best Paper Runner-Up Award at MobiQuitous 2011.

In 2009, he worked as a Research Assistant with Hong Kong Polytechnic University, Kowloon, Hong Kong. His research interests include pervasive computing, social network analysis, data mining from large-scale taxi data, and big data analytics for smart cities.

Gang Pan received the B.Sc. and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 1998 and 2004, respectively. He is currently a Professor with the College of Computer Science and Technology, Zhejiang University. He has published more than 90 refereed papers. He visited the University of California, Los Angeles, Los Angeles, during 2007-2008. His research interests include pervasive computing, computer vision, and pattern recognition. Dr. Pan has served as a Program Committee Member for more than ten prestigious international conferences, such as IEEE International Conference on Computer Vision and IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Shijian Li received his Ph.D. degree from College of Computer Science, Zhejiang University (2006). He is currently an associate professor in College of Computer Science, Zhejiang University. His research interests include Ubiquitous Computing, and Social Computing. He has published over 60 papers on above domains.

Zhaohui Wu received the B.Sc. and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 1988 and 1993, respectively. He is currently a Professor with the Department of Computer Science, Zhejiang University. His research interests include distributed artificial intelligence, semantic grid, and pervasive computing. Dr. Wu is a Standing Council Member of the China Computer Federation.