Mining efficient taxi operation strategies from large scale geo-location data

Huigui Rong[1], *(Member, IEEE),* Zepeng Wang[1], Hui Zheng[2], Chunhua Hu[2], *(Member, IEEE)*
Li Peng[1], Zhaoyang Ai[3], Arun Kumar Sangaiah[4],

***Abstract*—Taxi drivers always look for strategies to locate passengers quickly and therefore increase their profit margin. In reality, the passenger seeking strategies are mostly empirical and substantially vary among taxi drivers. From the history taxi data, the top performing taxi drivers can earn 25% more than the ones with mediocre seeking strategy in the same period of time. A better strategy not only helps taxi drivers earn more with less effort, but also reduce fuel consumption and carbon emissions. It is interesting to examine the influential factors in passenger seeking strategies and find algorithms to guide taxi drivers to passenger hotspots with the right timing. With the abundant availability of history taxicab traces, the existing methods of doing taxi business have been radically changed. This paper focuses on the problem of mining efficient operation strategies from a large scale history taxi traces collected over one year. Our approach presents generic insights into the dynamics of taxicab services with the objective of maximizing the profit margins for the concerned parties. We propose important metrics such as trip frequency, hot spots and taxi mileage, and provide valuable insights towards more efficient operation strategies. We analyze these metrics using techniques like Newton's polynomial interpolation and Gamma distribution to understand their dynamics. Our strategies use the real taxicab traces from the city of Changsha (P.R.China), may predict the taxi rides at different times by 90.68% per day, and increase the taxi drivers income levels up to 19.38% by controlling appropriate mileage per trip and following the route across more urban hot-spots.**

***Index Terms*—Taxi operation strategy, Trip frequency, Hot-spot metric, Trip mileage.**

I. Introduction

A. Background and Motivation

TAXICAB services have become the de-facto alternative to public transportation due to their easy availability, comfortable travel experience and flexibility in reaching any chosen destination. Similar to other service-oriented businesses, taxicab companies depend on the revenues generated by the individual cab drivers and are constantly exploring avenues for their profit margin increase. Therefore, taxicab companies have shown great interest in recommendation techniques for dealing with problems such as identifying a sequence of suitable pickup points and determining the fastest driving routes. However, they are in lack of insight into useful operation strategies which can benefit the taxicab companies in the long run.

Broadly speaking, taxicab services can be availed in two different modes: centralized –where a centralized booking office allocates taxicabs to customers, and ad-hoc –where taxicabs can be hailed directly from the road. The major advantage of taxicab services is their universal availability at any given time and location, unlike public transportation, which operates on fixed routes and shuts down during late hours. There are several situations where taxicab services are indispensable, such as medical emergencies, traveling with heavy luggage, traveling with old or physically challenged people, and traveling with infants. However, despite the tremendous commercial potential and business opportunities, taxicab services find it difficult to generate adequate profit margins.

B. Problem Overview

We focus on the problem of identifying efficient operation strategies for taxicab services, which not only increases the profit margins, but also enables the taxicab services to quantify their operational capabilities appropriately. Informally stated, an operation strategy is an insight towards improving the efficiency of the operations involved in running the taxicab business. As such, taxicab services operate in a highly dynamic and complex environment that creates many operational challenges, which creates considerable difficulties in identifying useful operational strategies. Towards this, we focus on two directly responsible operational challenges in taxicab services and quantify the problem in terms of these two challenges. The first important operational challenge is to identify the relation between the time, location and customer demand at any given point of time. Such a relation will be able to assist the taxicab services in focusing on specific locations at a given time and matching the customer demand. The second operational challenge is to be able to comprehend the interplay of taxicab trip frequency, the mileage covered and the corresponding income levels. This understanding will be of great help for the taxicab companies in designing efficient operational strategies targeting the increase of income levels of the taxicabs.

The recent years have witnessed the rapid development of wireless sensor technologies in mobile environments, such as GPS and RFID, which sense and record various metrics of a taxicab trip. Such fine-grained sensing allows us to

1. College of Computer Science and Electronic Engineering, Hunan University, HN, China, e-mail: {ronghg, zepeng, lipeng}@hnu.edu.cn.
2. Key Laboratory of Hunan Province for Mobile Business Intelligence, Hunan University of Commerce, HN, China, e-mail: {zhdilly, huchunhua777}@163.com.
3. Institute of Cognition and Language, College of Foreign Languages, Hunan University, HN, China, e-mail: aizhaoyang@hnu.edu.cn.
4. School of Computing Science and Engineering,VIT University,Tamil Nadu, India, e-mail: arunkumarsangaiah@gmail.com.
Corresponding author: Hui Zheng(zhdilly@163.com)
Manuscript received June 26, 2017; revised July 6, 2017.

record the number of taxicab trips on a daily basis and appropriately tag each trip by location, time, mileage and cost, which can provide valuable insights into the dynamics of the taxicab operations. Therefore, there is a critical technique for mining approaches that can analyze such trip data and identify efficient operation strategies for taxicab services. This paper discovers that our work fits in well with such an attempt in such a complex problem space.

C. Our contributions

Obviously, recent discusses mainly focus on finding the next passengers or reducing the waiting time for drivers. However, in the real world, the income of taxi drivers is correlated with some factors including trip frequency, hot spots and trip efficiency, and we need to find the relations among these aspects and present some efficient driving strategies. Our contributions are summarized as follows:

1) Firstly the trip frequency in a unit time is closely related to the income of taxi drivers and the urban hot spots, then we build an urban trip frequency model based on Newton interpolation polynomial, which may predict the taxi rides at different times by 90.68% per day and by a staggering 99% per week.

2) Secondly we divide equally the urban area of Changsha into 56 * 32 grids and then define a metric method for hot spots (HSM)-based trip frequency. Based on our proposed approach, we present some further analysis on hot distribution and migration, and increase taxi drivers income levels up to 19.38% by following the route across more urban hot spots.

3) Thirdly we identify efficient operational strategies by using the trip mileage data combined with the above two metrics. First, we notice that the cab driver income depends on the trip frequency of picking up passengers as well as on the driving route selection. Second, taxi drivers should reduce the mileage per trip and increase trip frequency per day by following the urban hot spots for improving the total income.

The rest of this paper is organized as follows. In Section II, we formally introduce the source of taxi data and propose some preprocessing strategies. Section III builds a taxi trip frequency model and forecasts the frequency curve in a unit time. Section IV analyzes the hot spot distribution and migration based on trip frequency and per taxi mileage. Section V presents the operational strategy for improving the taxi income based on taxi trip and mileage. Finally, we present conclusions in Section VI.

II. RELATED WORK

Taxicab service, as one of the most common choice among all kinds of transportation tools, plays an important role in public transportation and people lifes for its convenience, flexibility and shortcut. Recent years, we have witnessed the rapid development of wireless sensor technologies in mobile environments, such as GPS, Wi-Fi and RFID, which provide us with unprecedented opportunities to automatically discover useful knowledge [1]. Then, these useful findings in turn deliver intelligence for making real-time and efficient driving decision, and mining-efficient driving strategies from the history taxi data may be the most promising approach to improving the taxi driver revenue as well as saving time and energy consumption.

Recent efforts have been made on providing efficient driving strategy for taxi drivers through using the large scale geo-location data. There is some existing work for determining the sequence of picking up for cab driver, and finding the fastest driving route from the current location to the destination [2]–[4]. The previous work can help cab drivers to reduce waiting time and miles when picking up the next passenger to a greater extent. J.Powell and Y.Huang, et al. believes the tax drivers should reduce the number of cruising miles while increasing the number of live miles, thus it may improve profitability without systematic routing [5]. Then they presents a simple yet practical method for reducing cruising miles by suggesting profitable locations to taxicab drivers, which is proved to help increase profitability obviously by using a large Shanghai taxi GPS data set.

Indeed, the contribution of before-mentioned research is essentially to find the next passenger in a short time from the current location to the destination by optimizing the driving route. Then, some researchers propose an adaptive route method in the cruising taxis [6], where the driver will be assigned the passenger with the most probability of matching the pathways customers are expected to take. And the simulation experiment shows that the proposed method is able to gain more customers than the existing means of cruising taxis. For the taxi drivers, hunting or waiting in vacant cab may be a puzzle problem. B.Li and D.Zhang et al. [7] discover passenger finding strategies from a large-scale real-world taxi dataset and present the selected route patterns which can well interpret the empirical study results derived from raw data analysis and even reveal interesting hidden facts. Obviously, it is more critical for taxi drivers to know the actual driving routes to minimize the driving time before finding a customer. Some other work on route optimization and picking up passengers is also proved to have a better contribution for reducing the waiting time and improving the drivers income [8]–[10].

However, the above work is mostly based on user ratings or interactions, but for taxi drivers, these strategies are in lack of real time effect and difficult to follow in driving a taxi. Therefore, the mobile recommender systems are more popular for smart mobile devices, such as the PDA, cell phones. Y. Ge, et al. [11] design and implement a mobile recommender system based on Potential Travel Distance (PT-D) function for evaluating each candidate sequence. They design two algorithms, ie., LCP and SkyRoute, from this proposed PTD function in addition to finding the recommended routes. This design has been proved effective in providing effective mobile sequential recommendation. The knowledge extracted from location traces can be used for coaching drivers, leading to the efficient use of energy. They also propose a taxi business intelligence system to improve taxi business performance and passenger experience [12].

Zheng, Y et al. [13] try to mine the relation between users and their locations and report on a personalized friend and location recommender for the geographical information sys-

tems (GIS) on the Web. In this framework, they incorporated a content-based method into a user-based collaborative filtering algorithm to estimate the rating of a user on an item and get the better recommendation results. A recommender system was presented for both taxi drivers and people expecting to take a taxi, using the knowledge of passengers mobility patterns and taxi drivers picking-up/dropping-off behaviors [14]. They build our system using historical trajectories generated by over 12,000 taxis during 110 days and validate the system with extensive evaluations including in-the-field user studies. M. Qu, et al. propose the recommender system which is capable of providing an entire driving route, and the drivers are able to find passengers with the most potential profit by following the recommendations [1]. Especially, they design for the first time a net profit objective function for evaluating the potential profits of the driving routes, and it is more practical and profitable than other existing recommender systems through the real-world dataset collected from the San Francisco Bay area.

Moreover, with the large scale taxi traces available, the abundant taxi traces have provided new ways of doing business from other valuable perspectives. Taxi ridesharing can be of significant social and environmental benefit by saving energy consumption and meeting the needs of residents. Ma Shuo et al. [15] proposed a large-scale taxi ride sharing service, which serves real-time requests sent by taxi users and generates ride sharing schedules that reduce the total travel distance significantly. Y. Zheng et al. [16] regard urban computing for city planning as one of the most significant applications in Ubiquitous computing, and they detect flawed urban planning using the GPS trajectories of taxicabs traveling in urban areas. They conduct the proposed method using the trajectories generated by 30,000 taxis in Beijing, and evaluate the results with the real urban planning of Beijing. Rong et al. [17] model the passenger seeking process as a Markov Decision Process (MDP), and learn a different set of parameters for the MDP from data for finding the best move for a vacant taxi to maximize the total revenue in that time slot. For better prediction of cab drivers future behaviors, a comprehensive study is carried out to reveal how the social propagation affects, and comprehensive experiments on a real-world data set collected from the New York City clearly validate the effectiveness of their proposed framework on predicting future taxi driving behaviors [18]. To support marketers in improving marketing effectiveness, a new application of recommender systems that select vehicles as recommenders is proposed by Ting Li et al [19]. They proposed algorithms to enhance the marketers' marketing effectiveness based on different evaluation standards. Some other works [15], [20]–[24] focus on specific types of recommendation scenarios such as fast routing, ride-sharing, or fair-recommendation, while they are not so related with our work.

III. DATA DESCRIPTION AND PREPROCESSING

The taxi dataset we are using comes from the capital city of a central province in China and covers almost thirteen months' taxi operation records from 2012 to 2013. There are totally about 1400 taxi cabs in the data. The taxi geo-location data is mainly stored in primary table T_BUSINESS HISTORY, described as the TABLE I.

TABLE I: Primary structure of taxi data

Number	Name	Type	Description
1	BUSINESSHIS_ID	Number	unique ID of the taxi in the dataset
2	STAMP	DATE	The time of inserting data
3	NUMBER_PLATE	Varchar	Number plate of taxi
4	ONTIME	DATE	Get-on time of the taxi
5	ONLON	Number	Get-on longitude of the taxi
6	ONLAT	Number	Get-on latitude of the taxi
7	OFFTIME	DATE	Get-off time of the taxi
8	OFFLON	Number	Get-off longitude of the taxi
9	OFFLAT	Number	Get-off latitude of the taxi
10	PERPRICE	Number	Price of every mileage
11	RUNLEN	Number	Run length of each time
12	RUNTIME	Number	Run time of each time
13	RUNMONEY	Number	Run money of each time

The table structure covers mainly Get-on/off time, Get-on/off latitude and longitude, Price per mileage, running time per time, running money per time, and etc. There are approximately 35 million taxi trip records (85,000 per day), where each record has the latitude-longitude coordinates and timestamps of the pick-up and drop-off events, along with total traveled distance and the fare of the corresponding trip.

However, there are some errors of GPS data used in this paper because of GPS satellite positioning, operational mistakes, atmosphere influence, GPS multipath problems and so on. Therefore, we need to pretreat this raw GPS data for facilitating the late data analysis. Our pre-processing methods used for identifying and repairing GPS raw data are usually based on GPS system principle. In this paper, we should take a full consideration of GPS data characteristics and taxi operational status to achieve a better preprocessing purposes. The pre-processing of taxi GPS data should comply with the following strategies:

1) The cross-border cleaning of latitude and longitude

Our research covers the main scope in the capital city of Changsha. The scope of longitude coordinates is from 112.854452 to 113.083556 and its scope of Latitude is from 28.149096 to 28.239767. Therefore, the GPS location coordinates outside this range should be eliminated.

2) The error revision and map matching

Since this paper uses Baidu map to analyze taxi operations and positioning data, and this map on the selected coordinate system may be somewhat different from the taxi GPS coordinate system. Therefore, This deviations need to be revised through latitude and longitude location. Map matching is the process of associating a sorted list of user or vehicle positions to the road network on a digital map, which is considered as an effective and widely used method for error correction. We use the map matching method to correct deviations and determine the location of road vehicles by transverse match and longitudinal match.

We use the average distance method to determine the road vehicle by considering a simple road network conditions, and

then map the taxi GPS data into Baidu map of road network.

According to the above rules, we may determine the deviation distance of all GPS anchor points on x and y axis. Firstly, we should eliminate the deviation before the taxi GPS data is associated with Baidu map of road network, and then let one taxi anchor point shown on Baidu map. Therefore, we can get the taxi GPS data that highly match Baidu map through this method.

IV. TAXI TRIP FREQUENCY MODELING AND ESTIMATION

We may get the period of taxis demand by analyzing the travel frequency of urban residents in a time slot, which helps to reduce the empty ride ratio and facilitates the taxi driver to prepare picking up passengers well in advance.

A. Building a taxi trip frequency

The frequency forecast will increase the taxi rides and enhance the income of taxi drivers. This paper chooses taxi dataset of 24 hours per day in Changsha, as shown in Fig 1:

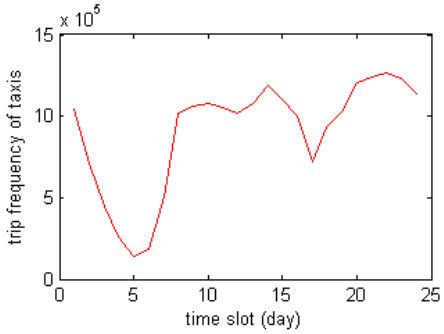


Fig. 1: Taxi trip frequency per day (24 hours)

We firstly use the discrete data to construct a simple function *f(x)* as the approximate function of original function *g(x)*. Then we may make predictions through manipulating *f(x)* to get the approximate result of *g(x)*. Interpolation method uses a number of points in a certain range of function value generated by function *f(x)* to construct the appropriate specific function, and then selects the known values from these points as the approximation value of the function *f(x)*. If this specific function is a polynomial, then it is called interpolation polynomial. With the increase or decrease of interpolation nodes, the interpolation primary function needs to change with them. We choose Newton interpolation polynomial to fit the pick-up times of Changsha taxis.

Firstly, we introduce the definition of *difference quotient* calculated by the following Eq (1):

$$f[x_i, x_j] = \frac{f(x_i) - f(x_j)}{x_i - x_j} = \frac{y_i - y_j}{x_i - x_j} \tag{1}$$

As the first-order difference quotient of function $f(x)$ about x_i and x_j, which is calculated by the following Eq (2):

$$f[x_i, x_j, x_k] = \frac{f[x_i, x_j] - f[x_j, x_k]}{x_i - x_k} \tag{2}$$

As the second-order difference quotient of function $f(x)$ about x_i , x_j and x_k, which is generally calculated by following Eq(3):

$$f[x_0, x_1, ..., x_k] = \frac{f[x_0, x_1 ..., x_{k-1}] - f[x_1, x_2, ..., x_k]}{x_0 - x_k} \tag{3}$$

Finally, the kth-order difference quotient of function $f(x)$ about x_0, x_1, ...,x_k can be calculated by the following the Table II:

Now, Newton's interpolation polynomial is calculated by following Eq (4):

$$\begin{aligned} f(x) &= f[x_0] + f[x_0, x_1](x - x_0) \\ &+ f[x_0, x_1, x_2](x - x_0)(x - x_1) + ... \\ &+ f[x_0, x_1, ...x_n](x - x_0)(x - x_1)...(x - x_{n-1}) \end{aligned} \tag{4}$$

Now, we show how to construct the trip frequency model based on Newton interpolation polynomial. First, we list difference quotient table as following Table III:

The corresponding interpolation polynomial is calculated by following Eq (5):

$$\begin{aligned} f(x) &= 1042183 - 224391.0(x - 1) \\ &+ 73792.8(x - 1)(x - 5) \\ &+ 73792.8(x - 1)(x - 5)(x - 8) \\ &- 9072.7(x - 1)(x - 5)(x - 8)(x - 13) \\ &+ 789.7(x - 1)(x - 5)(x - 8)(x - 13)(x - 16) \\ &- 52.2(x - 1)(x - 5)(x - 8)(x - 13)(x - 16) \\ &+ 2.2(x - 1)(x - 5)(x - 8)(x - 13)(x - 16)(x - 22) \end{aligned} \tag{5}$$

We can represent the fitting polynomial as follows.

$$\begin{aligned} f(x) &= a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 + a_2x^2 \\ &+ a_1x^1 + a_0 \end{aligned} \tag{6}$$

Simplifying the above results by computing the values for a_0 to a_6, we get the following polynomial shown in Eq(7):

$$\begin{aligned} f(x) &= -0.0035x^6 + 0.2212x^5 - 4.524x^4 + \\ &25.14x^3 + 203.3x^2 - 2073x + 4916 + g(x) \end{aligned} \tag{7}$$

The fitting curve based on real taxi data is showed in Fig 2,which can be seen to be close to the discrete data points.

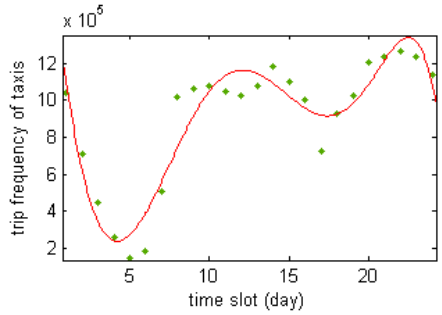


Fig. 2: The fitting curve of real taxi data per day

As we know from Fig 2, taxi rides sharply rise from 7 to 8 PM in Changsha because the urban residents trip for work

TABLE II: Difference Quotient Table

x_k	y_k	first-order	second-order	third-order	fourth-order	...
x_0	y_0					
x_1	y_1	$f[x_0, x_1]$				
x_2	y_2	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$			
x_3	y_3	$f[x_2, x_3]$	$f[x_1, x_2, x_3]$	$f[x_0, x_1, x_2, x_3]$		
x_4	y_4	$f[x_3, x_4]$	$f[x_2, x_3, x_4]$	$f[x_1, x_2, x_3, x_4]$	$f[x_0, x_1, x_2, x_3, x_4]$	...
...	...	...	...	...	...	...

TABLE III: Difference Quotient Table

x_i	$f[x_i]$	$f[x_{i-1}, x_1]$	$f[x_{i-2}, x_{i-1}, x_i]$	$f[x_{i-3}, ..., x_i]$	$f[x_{i-4}, ..., x_i]$	$f[x_{i-5}, ..., x_i]$	$f[x_{i-6}, ..., x_i]$
1	1042183 5	1042183	-224391.0				
8	144619	292158. 7	73792.8				
13	1021095	11516.8	-35080.2	-9072.7			
16	1078679	-25073.3	-4573.7	2773.3	789.7		
20	1003459	51179.5	10893.2	1288.9	-98.9	-52.2	
22	1267934	29878.5	-3550.1	-1604.8	-206.6	-5.6	2.2

TABLE IV: Comparison of the prediction results based on real data

Time	2	5	8	11	14	18	21	24
Model Prediction Results	1719	585	2015	3362	3544	2630	3600	3130
Real Data	1896	357	2697	3162	3431	2615	3588	3152

in this rush hours lasts from 7 to 8 AM and from 15 to 16 PM. The pick-up number in this period is at a high level. Due to the shifting of duty among taxi drivers, the taxi ride times begins to fall sharply from 16 to 17 PM. From 19 to 20 PM in the night, although people already get off work, the taxi rides begins to raise up because the public transportation is out of service and the rich nightlife of Changsha begins. This case usually lasts until 24 PM or even 1 AM next morning. We randomly select some taxi ride data in 2013 for verifying the prediction accuracy of our proposed model and the results are shown in Table.IV.

Obviously, our proposed model may predict precisely the taxi rides at different times per day and the average accuracy reaches by 90.68%, which provides a credible basis for the daily schedule of taxi drivers. So, we recommend that the taxi drivers may rest a while for relieving stress during the day time of 14 to 15 PM, which is good for taxi drivers to raise revenue and ensure sufficient rest time at the same time.

B. Taxi trip frequency per week

Then, we present the distribution of taxi rides in a week from Monday to Sunday to run statistics on the taxi trip frequency per week and exploit the urban travel habits weekly, as shown in Fig 3 .

Similarly, we want to predict pick-up times per week from Monday to Sunday for urban residents by building the mathematical model of Newton interpolation. We predict the taxi trip frequency of different working days and holidays in a week by using the full year data. We use the numbers 1-7 to represent Monday to Sunday to verify the results with the perfect combination of mathematical model.

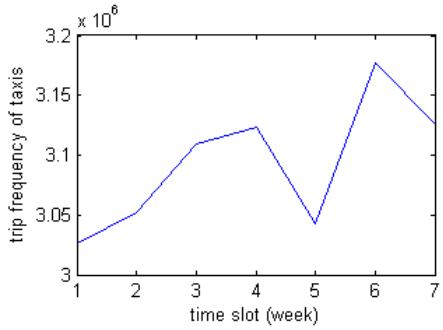


Fig. 3: Taxi trip frequency per week from Monday to Sunday

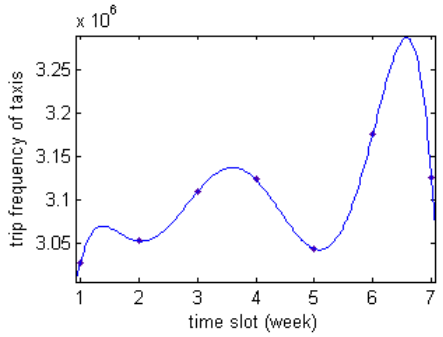


Fig. 4: The fitting curve of taxi trip frequency from Monday to Sunday

The interpolation polynomial is calculated as in the follow-

ing Eq (8):

$$\begin{aligned} N_6(x) &= 3026566 + 25385.0(x-1) \\ &+ 16031.5(x-1)(x-2) \\ &+ 16031.5(x-1)(x-2)(x-3) \\ &- 12596.1(x-1)(x-2)(x-3)(x-4) \\ &+ 1036.7(x-1)(x-2)(x-3)(x-4)(x-5) \\ &+ 2782.2(x-1)(x-2)(x-3)(x-4)(x-5) \\ &- 1943.2(x-1)(x-2)(x-3)(x-4)(x-6)(x-6) \end{aligned} \tag{8}$$

Then, we design the fitting binomial formula as follows:

$$\begin{aligned} f(x) = a_6x^6 + a_5x^5 + a_4x^4 + a_3x^3 \\ + a_2x^2 + a_1x^1 + a_0 \end{aligned} \tag{9}$$

The experimental data to obtain the value is described as below:

$$\begin{aligned} a_6 &= -1943 \\ a_5 &= 4.359 * 10^4 \\ a_4 &= -3.808 * 10^5 \\ a_3 &= 1.642 * 10^6 \\ a_2 &= -3.654 * 10^6 \\ a_1 &= 3.977 * 10^6 \\ a_0 &= 1.401 * 10^6 \end{aligned} \tag{10}$$

The final expression for obtaining the value is described as follows:

$$\begin{aligned} f(x) &= -1943x^6 + 4.359 * 10^4x^5 + -3.808 * 10^5x^4 \\ &+ 1.642 * 10^6x^3 + -3.654 * 10^6x^2 + 3.977 * 10^6x^1 \\ &+ 1.401 * 10^6 \end{aligned} \tag{11}$$

By fitting the experimental data and the proposed model, and the accuracy reaches by a staggering 99%, just in Fig.4, which provides a support for the prediction in the rest of this paper.

We notice that the taxi ride times have a sharp rise on Saturday and Sunday by comparing the taxi rides times on the weekend and weekdays. Obviously, on Saturday, the first day of weekend, the residents usually go out for shopping, social networking, entertainment and other activities. So, we recommend that taxi drivers spend more time and effort in driving a taxi on Saturday and Sunday, while they take rest on Monday to ensure the balance of rest and work, and increase revenue simultaneously.

V. HOT SPOT MODELING AND ESTIMATION

Urban hot spots refer commonly to such certain areas with the most concentrated pick-up area of taxis, and they are usually the urban core functional areas with more choices for picking up passengers. In addition, the ride mileage is always concentrated in a reasonable range from the whole city taxi mileage. These studies usually are very valuable in facilitating residents travel and improving the taxi efficiency.

A. Hot spot definition

We divide Changsha city map area equally into 56 * 32 grids for finding the distribution of taxi hot spots. We define function $F(k_i)$ as the aggregated trip frequency in the ith grid in a unit time, i.e., $\sum f_i$ of all f_is, across all taxis, in a given T, which can be an hour, day or week. We regard a trip as an effective trip when the getting on spot or the getting off one is located within the same grid. Now, to get the distribution of hot spots, in Eq.12 we define our hot spot metric (HSM), $H(R,T)$ where R denotes the bound for choosing the top-R hot spots.

$$H(R,T) = TOP\left(\left\{\frac{F(k_i)}{T}\right\}_{i=1}^{n}, R\right) \tag{12}$$

where n represents the total number of grids, k_i denotes the trip frequency of ith grid, T denotes the unit of time and R represents the number of desired hot spot grids from the whole metropolitan area. For instance, if we wish to focus on the few important city hot spots, we may set R to be around 20. According to our metric method for hot spots, we get the hot spot distribution as shown in Fig. 5. The $y-axis$ is the trip frequency and the $x-axis$ denotes the aggregate number of grids that registered this frequency in unit of time T. Now, we introduce a three-dimensional structure for describing 56 * 32 grids as following Eq. (13):

$$Grid(i) = [x_i, y_i, f_i]_{i=1}^{n}) \tag{13}$$

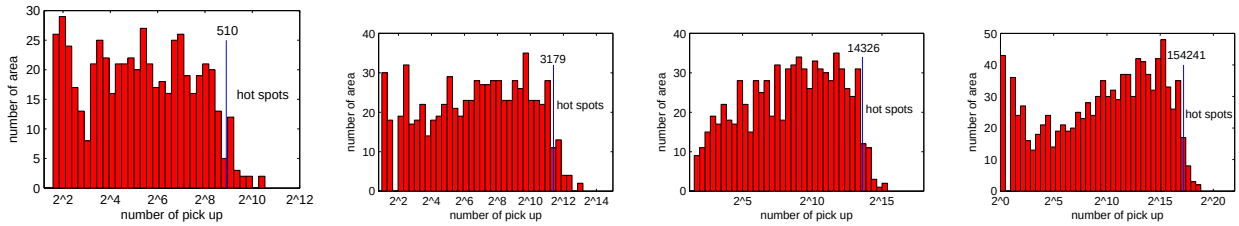
where, x_i represents the X coordinate of ith grid, y_i denotes the Y coordinate of ith grid and f denotes the trip frequency in ith grid.

From Fig.5, we arrive at the statistics that the top 20 hot spots per day and per week are almost the same, only the grid with co-ordinates [20,29,6484] in Fig.5(b) replacing the grid [18,16,943]. Obviously the urban hot area is relatively stable in the short term. The hot grid with [15,12] ranks the highest in hot spot distribution per day, per week and per month, while it ranks second in the annual hot map. On other hand, some hot grids just appear in hot map for a short time, such as [15,12,2892], [20,12,2224], [18,19,1088], [28,20,1028], [18,21,957] in Fig.5(a), [15,12,17563], [6,9,8380], [20,29,6484], [18,21,6425], [18,19,6204] in Fig.5(b) and [20,11,63962], [18,13,38375], [28,20,31455] in Fig.5(c). These grids usually disappear in the hot map of the whole year.

B. Hot spot distribution and migration

We use matlab to display the taxi hot spots of getting-on on the map of Changsha. Fig. 6 can clearly show the regional distribution of hot spots of getting-on.

From the distribution of the hot spots in one year, as shown in Fig 6, the most popular getting-on spots in Changsha cross from the Huang Xing Road(West Commercial Street) to Wuyi Square(the core commercial area of Changsha). This area is the dining, entertainment, shopping center of Changsha, and therefore, this region has become the hot spots area for various social activities. The train station is one of the most



(a) Day (b) Week (c) Month (d) Year

Fig. 5: Hot spot distribution

intensive areas of population mobility, which drives economic development by attracting the commercial plaza, restaurants, hotels and other tertiary industries.

The hot spot analysis only for the whole year is not enough to illustrate the problem. Then we further focus on hot spots of the taxi rides and divide the time period into three time periods including the morning (4:00 -12:00), the afternoon (12:00 -20:00) and the evening (20:00-4:00 the next day). This analysis provides more accurate description with time-related passengers hot spots for the taxi drivers, as shown in Fig. 7(a), Fig. 7(b) and Fig. 7(c):

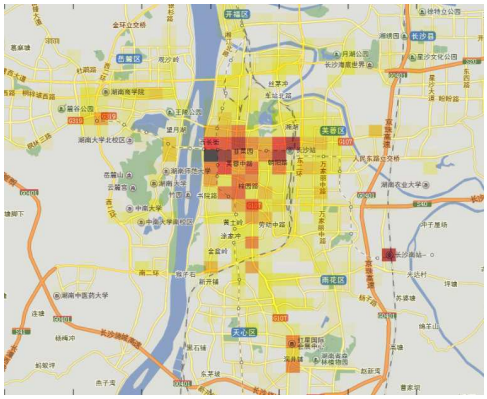


Fig. 6: Taxi hot spots of getting-on/off in full year

From the distribution of taxi hot spots in the morning, the taxi hubs of picking up such as the train station, South Railway Station, West Bus Station and Central Bus Station, et al. are the hottest areas. Compared with the business district at this time, the hot spots of getting-on and off in the business district are less than that of transportation hubs because the public usually go to work during this time.

From the analysis of hot spots map for the afternoon, the railway station becomes a hot spot in this period because more rail passengers need to take taxi for leaving, which shows a clear difference compared with other open areas.

In the hot spot map for the evening, there are more people getting off/on from Wuyi commercial area than elsewhere due to a variety of bars, KTV, shopping malls, plazas, etc. located along the commercial streets, which greatly contributes to the consumption in this region, and also leads to more taxi operation income in this area.

From the above Fig. 7, we observe that several neighboring hot spots are close to each other, such as [15,10], [15,11], [15,12], [15,13] and [16,10], [16,11], [16,12], which shows that the urban development generally expands around the center. There are some occasional hot grids, such as [6,9], [20,29], and [28,20], which means that these grids correspond to a new business district, and gradually form a large hot area.

C. Statistical analysis of per taxi mileage

There is a carrying mileage for each taxi ride, and we can find the taxi mileage distribution by analyzing each taxi mileage. In this paper, we classify taxi mileage into multiple mileage periods by setting the mileage interval as mi, and show mileage distribution by the mileage interval valued 2 Km, as shown in Fig 8.

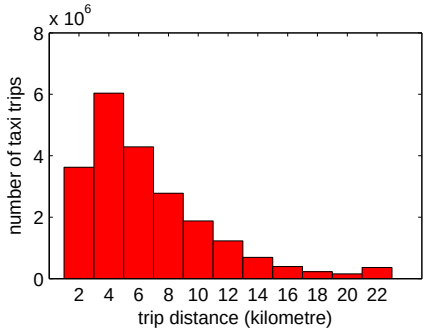
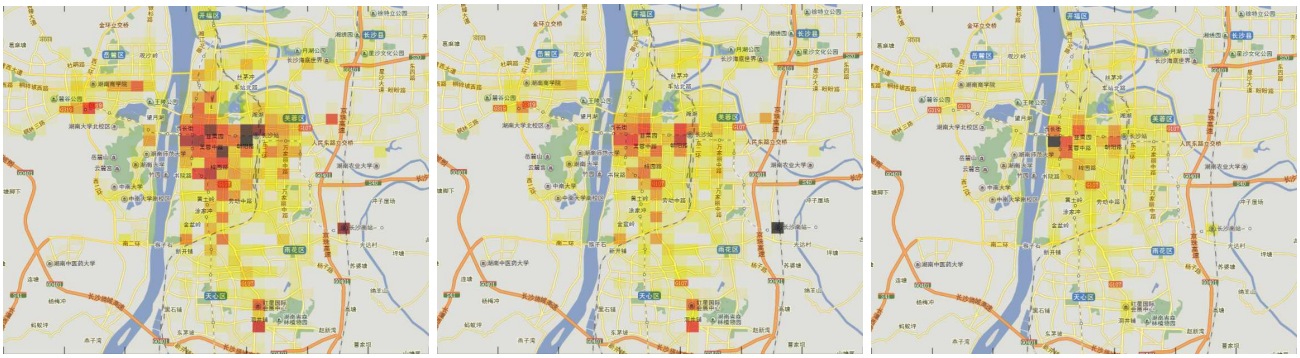


Fig. 8: The taxi driver trip distance interval distribution

We may learn that the taxi mileage period from 2 to 4 km reaches six millions and it is the most common taxi mileage period from Fig.8. At the same time, the mileage distribution decreases at both sides. Less people take taxi with a mileage less than 2km or more than 12km. Obviously, the taxi mileage distribution is consistent with the basic characteristics of the gamma distribution. The authors build a mathematical model of the gamma distribution to simulate the taxi mileage distribution. The distribution curve inosculates with this model curve by adjusting the parameters of κ and θ. Ultimately, this paper draws a mathematical model of a taxi mileage distribution through data fitting, as shown in Fig 9, and κ=β, θ=1/α.

The moment generating the function of Gamma distribution is described as below:

$$M_x(t) = E(e^{xt}) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{xt} x^{\alpha-1} e^{-\lambda * x} dx = (\frac{\lambda}{\lambda - 1})^\alpha \quad (14)$$



(a) Morning (b) Afternoon (c) Evening

Fig. 7: Taxi hot spots of getting-on/off in (a) Morning (b) Afternoon and (c) Evening

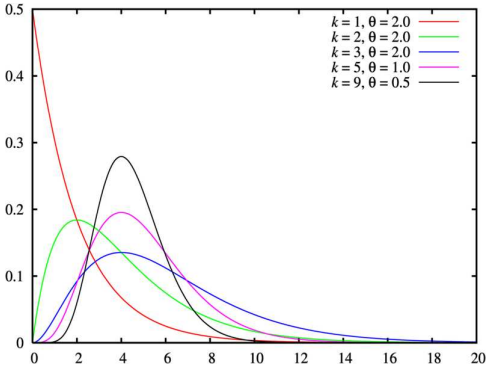


Fig. 9: Gamma distribution

The probability generating function is described as below:

$$K_x(t) = lnM_x(t) = \alpha[ln\lambda - ln(\lambda - t)] \tag{15}$$

The expected value is calculated as follows:

$$\frac{(dK_x(t))}{dt} = \frac{\alpha}{\lambda - 1}, when(t = 0), E(x) = \frac{\alpha}{\lambda} \tag{16}$$

The variance value is calculated as follows:

$$\frac{d^2K_x(t)}{dt^2} = \frac{\alpha}{(\lambda - t)^2}, \quad when(t = 0), \sigma^2(X) = \frac{\alpha}{\lambda^2} \tag{17}$$

The experimental results reveal that when α values 2.6 and β values 1.3, the curve obtained can coincide well with the actual mileage distribution.

$$f(x) = \frac{\alpha^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-x\alpha} \tag{18}$$

And

$$\Gamma(\beta) = \int_0^\infty e^{-t} * t^{x-1} dt \tag{19}$$

Therefore, we suggest taxi companies enhancing the starting price and starting distance appropriately in order to facilitate residents travel and improve the taxi efficiency, such as the starting price being raised to 8 yuan and starting distance being increased to 3 km. In this way, this strategy will not only prevent passengers from abusing taxi, but also ensure more passengers can find a taxi in emergency quickly and easily.

VI. MINING EFFICIENT TAXI OPERATION STRATEGIES

After the above analysis of relevant sections, we prove that there is some implicit association between the proposed models and the taxi operation strategy. We analyze and compare the different operation strategies among the different income spectra and present an optimized operation strategy to improve the income of taxi companies and drivers.

A. Taxi revenue analysis of taxi drivers

As there is commonly different taxi revenue for different taxi drivers, we may find the taxi revenue distribution through analyzing the total taxi revenue.

In this paper, we classify taxi revenue into multiple revenue ranges by setting the interval as *tr*, and show the total revenue distribution by the revenue interval valued 2, In this paper, a total taxi revenue interval distribution about taxi drivers in 2012 is showed as below in Fig.10:

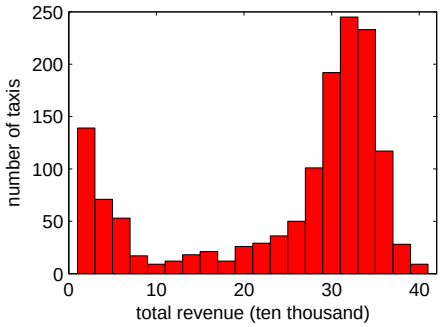


Fig. 10: The taxi driver total revenue interval distribution

For a more scientific analysis of the taxi annual income, we adopt Fourier transform to establish a mathematical model for forecasting the income of taxi drivers.

Trigonometric Fourier series expansion may be described below:

$$f(t) = a_0 + \sum_{n=1}^{\infty}(a_n * \cos nw_0 t + b_n * \sin nw_0 t) \quad (n = 1, 2, 3...) \tag{20}$$

Fourier coefficients are listed below:

$$a_0 = \frac{1}{T}\int_{-T/2}^{T/2} x(t)dt \tag{21}$$

$$a_n = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos nw_0 tdt \tag{22}$$

$$b_n = \frac{2}{T}\int_{-T/2}^{T/2} x(t)\sin w_0 tdt \tag{23}$$

According to this Fourier model, we fit the experimental data by taking n as 3, and then get the final expression for the Fourier formula:

$$\begin{aligned} f(t) &= a_0 + a_1 * \cos(wt) + b_1 * \sin(wt) + \\ &a_2 * \cos(2wt) + b_2 * \sin(2wt) + \\ &a_3 * \cos(3wt) + b_3 * \sin(3wt) \end{aligned} \tag{24}$$

The coefficient values are described below:

$$a_0 = 62.9, a_1 = -24.44, a_2 = 29.41, a_3 = 56.83,$$

$$b_1 = -36.67, b_2 = 61.74, b_3 = -19.34, w = 0.1268$$

We get the following equation by calculating the actual coefficient values, as shown in following formula Eq 25:

$$\begin{aligned} f(t) &= 62.9 - 24.44\cos(0.1268t) - 36.67\sin(0.1268t) \\ &+ 29.41\cos(0.2536t) + 61.74\sin(0.2536t) \\ &+ 56.83\cos(0.3804t) - 19.34\sin(0.3804t) \end{aligned} \tag{25}$$

The fitting curve is showed in Fig. 11:

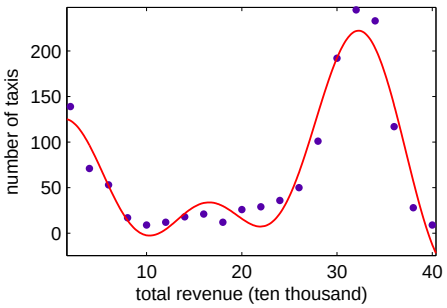


Fig. 11: The model fitting diagram based on real taxi income data

By fitting the experimental data and the prediction accuracy of the proposed model reaches by 95%, which will help to accurately predict the taxi income in the coming years. According to our predicted results, taxi companies should increase the serving taxi number by reducing the barriers to entry threshold.

B. Efficient driving strategy from trip frequency

To clearly analyze the impact of each trip mileage on total revenue, we rank all the drivers (around 1000) by revenue into a set $S = \{s \mid \text{sorting by revenue from high to low}\}$, and then select two sub-sets of the drivers from this set denoted by S_1 and S_2. The total income of the first part lists top 50 of all drivers ($S_1 = \{s_i \mid s \in S, i = 1, 2, 3, \cdots, 50\}$), and another part ranks from 500 to 550 ($S_2 = \{s_j \mid s \in S, j = 500, 501, 502, \cdots, 550\}$). Next, we divide the trip mileage into n equidistant intervals from 0 to 20: $\{[\delta_1, \delta_2], [\delta_2, \delta_3][\delta_{n-1}, \delta_n]$ where n is set as 12 for the present analysis. Without loss of generality, in the following definitions, we consider each pickup of a passenger to be equivalent to a trip. We use M_p to represent the sum total of the number of pick-ups in the pth interval of different income drivers. We use M_{total} to denote the total number of pick-ups of all drivers across all the n intervals. The ratio $\frac{M_p}{M_{total}}$ gives the proportion of pickups in a given interval p. These definitions allow us to analyze the situation of each pick-up mileage by differently ranked drivers.

Fig.12 shows the total pick-up times across taxi drivers with different income levels in 2012. The blue columns refer to the drivers with the income ranked by top-50, while the red columns indicate drivers with income ranked from 500 to 550.

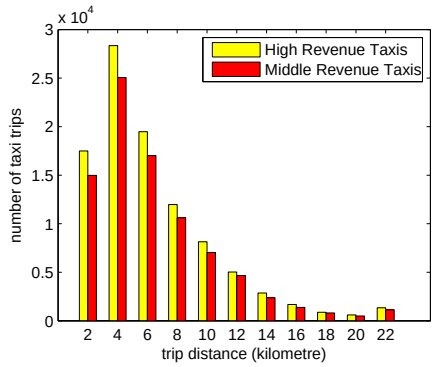


Fig. 12: The total pick-up times by different income drivers in 2012

From Fig.12, for any mileage driving interval, there are always greater pick-up times of taxi drivers in S_1 than that of taxi drivers in S_2. So, we can believe that the total income of taxi drivers has a positive correlation with the pick-up times, as shown in Eq (24):

$$f(M) = \alpha * M + \beta, (\alpha > 0, \beta > 0) \tag{26}$$

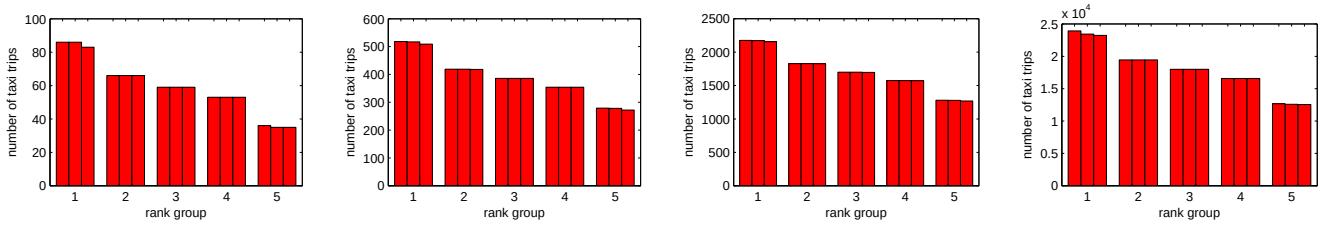
From Eq (24), we know the more pick-up time M, the more income $f(M)$.

Therefore, we believe that picking up only long-distance passenger or short-distance ones is not able to increase the total income of taxi drivers. To increase the total revenue, the taxi drivers should increase pick-up times, which may not only raise the total revenue of taxi drivers, but also facilitates passengers in finding taxis.

C. Improving income strategy

To identify a suitable income strategy, we perform a comparative analysis of trip frequencies of drivers at varied income levels. We compare the total income SUM of taxi drivers ranked by different incomes and get the following inequality:

$$SUMp(p \in S_1) > SUMq(q \in S_2) \tag{27}$$



(a) Day (b) Week (c) Month (d) Year

Fig. 13: Rank group

From Fig.13, there is a large gap of the taxi total income between the top 50 and the ranked income from 500 to 550. So it is significant to explore the impact of the total taxi mileage on the total income. On the average, the total mileage D s ss1 of top 50 drivers is 115048.7(km) and the total income SUMp(pS1) of top 50 drivers is 368474.4(RMB). Obviously, the total income of drivers is in a descending order, while the mileage does not strictly decrease, and it fluctuates irregularly. The total mileage of drivers Ds ss2 ranking from 500 to 550 is 115048.7(km), while the total income SUMq(qS2) of top 50 drivers is 308667.5 (RMB), as shown in Table V: The total

TABLE V: the average income and mileage by different income of drivers

Taxi driver	Average income(RMB)	Average mileage
Top 50	368474.4	136121.7
Rank 500 to 550	308667.5	115048.0

income of top 50 taxi drivers is higher than that of the income ranked from 500 to 550 taxi drivers by 19.38%, as shown in the following equation:

$$\frac{[SUMp1(p1 \in S_1) - SUMp2(p2 \in S_2)]}{SUMp2(p2 \in S_2)} = 19.38\% \quad (28)$$

While the total mileage of top 50 taxi drivers is higher than that ranked from 500 to 550 taxi drivers by 18.32%, as shown in the following equation:

$$\frac{(M_n/M^{s1}_{total} - M_n/M^{s2}_{total})}{M_n/M^{s2}_{total}} = 18.32\% \quad (29)$$

From the analysis of the above results, the income of different drivers depends on the frequency of picking up passengers as well as the driving route selection, and especially, the former can increase the taxi revenue significantly. The taxi drivers with higher income may consider a more optimized route, factoring in traffic jams, situation lines, traffic lights and so on, to realize a higher income level. To demonstrate this, we select five groups of taxi drivers (three drivers per group) from the lowest income to the highest one ranked by trip frequency, and show these results in Fig.13.

So, we conclude that the taxi drivers with higher income are running higher trip frequencies than that of lower income drivers. The drivers with the highest income have a 60% higher trip frequency than that of ones with the lowest income.

VII. CONCLUSIONS

In this work, we present the first systematic approach towards mining efficient operation strategies from large scale taxicab traces. From the analysis of the real taxi data, we mine some efficient strategies for the city of Changsha, which allows us to look at the taxi businesses from a fresh angle. Our experimental results based on real taxi data from the capital city of Changsha show that the trip frequency and urban hot spot distribution play a major role in the income levels of the drivers and taxicab companies. Our proposed models and metrics give insights on taxi operation strategies that (1) the frequency model may predict precisely the taxi rides at different times per day and per week, and the average accuracy increases 90.68% and 99% respectively; (2) the taxi companies appropriately enhance the starting price and starting distance in order to facilitate the passenger traveling and improve the taxi efficiency; (3) the income of different drivers depends on the frequency of picking up passengers as well as the driving route selection, and following our strategies will increase the taxi drivers' income levels by up to 19.38%.

ACKNOWLEDGMENT

This work is partially supported by National Natural Science Foundation of China under Grant Numbers 61672221, 61304184 and 61273232, by the Program for New Century Excellent Talents in University under Grant Number NCET-13-0785.

REFERENCES

[1] M. Qu, H. Zhu, J. Liu, G. Liu, and H. Xiong, "A cost-effective recommender system for taxi drivers," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014, pp. 45–54.

[2] G. Nagy and S. Salhi, "Heuristic algorithms for single and multiple depot vehicle routing problems with pickups and deliveries," European journal of operational research, vol. 162, no. 1, pp. 126–141, 2005.

[3] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems. ACM, 2010, pp. 99–108.

[4] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011, pp. 109–118.

[5] J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps." in SSTD. Springer, 2011, pp. 242–260.

[6] K. Yamamoto, K. Uesugi, and T. Watanabe, "Adaptive routing of cruising taxis by mutual exchange of pathways," in International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. Springer, 2008, pp. 559–566.

[7] B. Li, D. Zhang, L. Sun, C. Chen, S. Li, G. Qi, and Q. Yang, "Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on. IEEE, 2011, pp. 63–68.
[8] J.-F. Cordeau, "A branch-and-cut algorithm for the dial-a-ride problem," Operations Research, vol. 54, no. 3, pp. 573–586, 2006.
[9] J.-L. Lu, M.-Y. Yeh, Y.-C. Hsu, S.-N. Yang, C.-H. Gan, and M.-S. Chen, "Operating electric taxi fleets: A new dispatching strategy with charging plans," in Electric Vehicle Conference (IEVC), 2012 IEEE International. IEEE, 2012, pp. 1–8.
[10] H. Jeung, M. L. Yiu, X. Zhou, and C. S. Jensen, "Path prediction and predictive range querying in road network databases," The VLDB JournalłThe International Journal on Very Large Data Bases, vol. 19, no. 4, pp. 585–602, 2010.
[11] Y. Ge, H. Xiong, A. Tuzhilin, K. Xiao, M. Gruteser, and M. Pazzani, "An energy-efficient mobile recommender system," in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010, pp. 899–908.
[12] Y. Ge, C. Liu, H. Xiong, and J. Chen, "A taxi business intelligence system," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011, pp. 735–738.
[13] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W.-Y. Ma, "Recommending friends and locations based on individual location history," ACM Transactions on the Web (TWEB), vol. 5, no. 1, p. 5, 2011.
[14] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-finder: A recommender system for finding passengers and vacant taxis," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 10, pp. 2390–2403, 2013.
[15] S. Ma, Y. Zheng, and O. Wolfson, "T-share: A large-scale dynamic taxi ridesharing service," in Data Engineering (ICDE), 2013 IEEE 29th International Conference on. IEEE, 2013, pp. 410–421.
[16] Y. Zheng, Y. Liu, J. Yuan, and X. Xie, "Urban computing with taxicabs," in Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011, pp. 89–98.
[17] H. Rong, X. Zhou, C. Yang, Z. Shafiq, and A. Liu, "The rich and the poor: A markov decision process approach to optimizing taxi driver revenue efficiency," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016, pp. 2329–2334.
[18] T. Xu, H. Zhu, X. Zhao, Q. Liu, H. Zhong, E. Chen, and H. Xiong, "Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 1285–1294.
[19] A. L. C. H. T. Li, M. Zhao, "On selecting vehicles as recommenders for vehicular social networks," IEEE Access, no. 5, pp. 5539–5555, 2017.
[20] S. Liu, S. Wang, C. Liu, and R. Krishnan, "Understanding taxi drivers routing choices from spatial and social traces," Frontiers of Computer Science, vol. 9, no. 2, pp. 200–209, 2015.
[21] G. L. Q. Z. H. Y. H. C. Rong, Huigui, "A personalized recommendation approach based on content similarity calculation in large-scale data," in Proceedings of 15th International Conference on Algorithms and Architectures for Parallel Processing. Springer, 2015, pp. 509–516.
[22] C. H. T. Li, A. Liu, "A similarity scenario-based recommendation model with small disturbances for unknown items in social networks," IEEE Access, no. 4, pp. 9251–9272, 2016.
[23] S. Qian, J. Cao, F. L. Mouël, I. Sahel, and M. Li, "Scram: a sharing considered route assignment mechanism for fair taxi route recommendations," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 955–964.
[24] W. Zhang, S. Li, and G. Pan, "Mining the semantics of origin-destination flows using taxi traces." in UbiComp, 2012, pp. 943–949.

Huigui Rong received the Ph.D degree in information management from Wuhan University, China, in 2010. He was a visiting scholar in computer science department of Michigan State University (MSU) from 2014 to 2015. He is currently a full associate professor of computer science department at Hunan University, and he is also the system analyst and senior project manager certified by The Ministry of Industry and Information. His major research areas include data mining, cloud computing and network optimization. He has published more than 20 research papers in international conferences and journals, such as ACM-CIKM, ACM-TIST, IEEE-HPCC. He is a member of IEEE and CCF.

Zepeng Wang Zepeng Wang is a master in College of Computer Science and Electronic Engineering of Hunan University, under the supervision of vice Prof. Huigui Rong, who is a master supervisor. He is a student member of CCF and his main research interests include software engineering, Machine learning and data mining. During the past years, he has published an invention patent and has won three first-class scholarships as well as one national inspirational scholarship.

Hui Zheng received the master degree from Central South University, China, in 2009. She was a visiting scholar in computer science department of Michigan State University (MSU) from 2014 to 2015. She is currently a full lecturer of tourism management department at Hunan University of Commerce. Her major research areas include tourism data mining, cloud computing and electronic business. She has published more than 10 research papers in international conferences and journals, such as FEBM2016, Tourism Overview, Journal of CSU.

Chunhua Hu received the Ph.D. degree in computer science from Central South University, Changsha,China, in 2007.He is currently a Professor from the School of Computer and Information Engineering, Hunan University of Commerce, Changsha. Up to now, he has chaired two National Natural Science Foundation of China projects and published more than 20 research papers in international journals and international conferences. In 2012, he has been selected into the Program of New Century Excellent Talents in University. His research interests include cloud computing, service computing, and dependability computing.