Taxi Operation Optimization Based on Big Traffic Data

Qiuyuan Yang[1],Zhiqiang Gao[1],Xiangjie Kong[1,2],Azizur Rahim[1],Jinzhong Wang[1],Feng Xia[1]
[1]School of Software, Dalian University of Technology, Dalian 116620, China
[2]Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China
xjkong@ieee.org

***Abstract*—Motivated by the remarkable improvement of information and communication technologies, along with the rapid progress of urbanization, smart city has become a novel brand. By mining big traffic data generated by widely deployed GPS devices and sensors in modern cities, we can unlock the knowledge of human mobility patterns and social functional regions, and then apply it to tackle critical problems in city construction. One of the tough issues is the paradoxical situation in urban traffic control and management, which is the empty carrying phenomenon for taxi drivers and the difficulty of taking a taxi for passengers. In the paper, we propose a data-driven taxi operation strategy to maximize drivers' profit, reduce energy consumption, and decrease environment pollution. Specifically, we capture social properties of functional areas through integrating, processing and analyzing the big traffic data. Later, we introduce the Time-Location-Sociality model which can identify three dimensional properties of city dynamics to predict the number of passengers in different social functional regions. Furthermore, we recommend Top-N areas for drivers according to the prediction outcomes, which introduce more profitable opportunities to pick up passengers. We conduct extensive experiments using the real GPS data generated by 12,000 taxis during 10 weekdays and 8 weekends in Beijing, and achieve prediction accuracies of 90.14% on weekdays and 86.37% at weekends respectively, which implies the effectiveness of our optimizing taxi operation strategy by considering the three dimensional properties.**

***Index Terms*—big traffic data, functional region, human mobility, taxi, recommendation.**

I. Introduction

Motivated by the rapid development of urbanization and the tremendous advancement in the field of information and communication technologies, the concept of smart city has emerged as a novel paradigm to deal with existing problems and circumvent potential issues in modern cities, such as traffic congestion, energy consumption, and environment pollution [1]. In recent years, numerous cities have taken substantial steps towards the process of constructing smart cities, e.g., Vienna, Toronto and Paris [2].

Nowadays, large-scale computing devices and sensing technologies have produced a variety of big data (traffic data, social network data, geographical data, etc.) in the scope of city. Sensing, integration, and analysis of these big data play pivotal roles in life quality improvement, economic development and environmental sustainability [3]. Especially, taxicabs, buses, and logistics trucks have been equipped with GPS sensors in many metropolis like Beijing and New York [4], and each vehicle is capable of transmitting its real-time locations to certain systems at regular intervals. By mining these *big traffic data*, we can discover the underlying social dynamics of an individual, community and city, and then apply these powerful knowledge into city planning, intelligent transportation system, human mobility analysis, and so forth.

With the increasing scale and population of modern cities (such as New York, London and Beijing), there appears a paradoxical situation in urban traffic control and management, which is the empty carrying phenomenon for taxi drivers and the difficulty of taking a taxi for passengers. According to Beijing traffic development and construction report during the 12th five-year plan [5], each taxi runs around 400 kilometers a day in Beijing, while average empty carrying ratio of taxis is about 40%. It indicates that taxis occupy large amounts of transportation resources, and aggravate the environmental pollution at the same time. On the other hand, 66,600 taxis provide services for 2,000,000 passengers in Beijing, which far exceeds the national standard of per capita share of taxis. However, there are still many citizens annoying with the difficulty of taking a taxicab.

A number of methods have been proposed to maximize the profits of taxi drivers by using historical GPS trajectories of taxis, and enhance the opportunity of finding vacant taxis for passengers simultaneously [6]–[8]. These methods provide experimental proofs on how the spatio-temporal property of trajectory influence the efficiency of passenger-finding algorithms. Nonetheless, exploring social relationships and properties of trajectory to increase drivers' profit requires further research. On the other hand, modern cities are composed of diverse functional areas, such as entertainment area, residential area, and industrial zone. These functional areas are historically associated with city planning, but mostly are shaped by people's actual needs of social activities in a long period [9] [10]. For instance, entertainment areas are generally visited by people to relax themselves on weekday evenings and weekends. Furthermore, people commute between these functional areas to engage a sequence of related social activities. Intuitively, people usually leave their offices on weekdays or residential areas at weekends for entertainment areas. The social attributes of different regions extracted from trajectories are relatively stable and exclusive compared to individual mobility [11] [12], which can be used to optimize the taxi drivers' operation strategy.

In the paper, we aim to propose a data-driven taxi operation strategy using a Time-Location-Sociality model to tackle the tough issue above. By integrating, processing and analyzing the traffic big data, we identify different social attributes of city functional regions efficiently. People in each social functional region have their own mobility patterns. After that, we characterize the mobility pattern in the Time-Location-Sociality model, which considers three dimensional properties of city dynamics. Then we can predict the number of passengers in different social functional regions using the introduced model. Furthermore, we recommend Top-N areas for drivers according to the prediction outcomes, which enables drivers to pick up more passengers for energy saving, and lightens the transportation pressure for a city.

Our major contributions can be summarized as follows:

- We explore different social features of city regions efficiently by mining big traffic data collected by ubiquitous sensors. Each social functional region contain its specific human mobility pattern.
- We present a Time-Location-Sociality model, which can identify three-dimensional properties of city dynamics to predict the distribution of passengers for different social functional regions.
- We recommend Top-N areas to drivers based on the prediction outcomes of our model, so that they can decide where to pick up passengers to maximize their profits.
- We conduct extensive experiments using the real data collected by 12,000 taxis lasting 18 days in Beijing. The results achieve prediction accuracies of 90.14% on weekdays and 86.37% at weekends respectively, which imply the effectiveness of our proposed operation strategy.

The rest of this paper is organized as follows. In Section II, we have a brief review on the related work. Section III presents the proposed operation strategy based on Time-Location-Sociality model. Then we describe the dataset and the experiments we conduct to evaluate our proposed strategy in Section IV. At last, we conclude the work in Section V.

II. Related Work

In 2008, Samuel Palmisano, the IBM CEO, proposed the concept of smart city in his speech [13]. In recent years, researchers start to keep a watchful eye on big data generated by widely deployed sensors in smart city. Walravens *et al.* [14] provide insight of the current state of the mobile services, and regard the mobile sensors as the primary interface to modern cities. With the development of information and communication technology, the project of smart city has been constructed in New York and Singapore [4].

Modern cities develop with the generation of various functional areas and understanding functional areas can significantly shed light on studying urban dynamic issues. Mobile telephone position data can be used to classify land uses [15] [16], while the GPS trajectories of taxi traveling in cities provide detailed location information which is widely used in assessing the functions of urban spaces [17]. Qi *et al.* [18] discover that the get-on and get-off frequencies of passengers in a region can depict the social functions of the regions. Similarly, Liu *et al.* [19] observe temporal patterns of pick-ups and drop-offs differ greatly from place to place and are dependent on the functions of places. DRoF [20] is a topic-modeling-based approach to detect region functions with the help of points of interests (POIs) and GPS trajectories, and in [9] the authors further generalize the problem using location and mobility semantics mined from trajectories to improve the framework. Besides, Zhong *et al.* [10] utilize transportation data obtained from surveys and smart card systems to deduce social activities and infer the urban functions at the building level. In spite of the fact that the mentioned methods have explored the functional areas from trajectories and presented that discovering functional areas can enable a variety of valuable applications, they still failed in how to apply these knowledge into practice.

Taxis play a quite significant role in public transportation in modern cities, and several passenger-finding strategies have been introduced to connect passengers to vacant taxis. Most of these strategies focus on recommending some popular positions to drivers. Li *et al.* [21] utilize L1-Norm Support Vector Machine and the real-time information of location and time to advise drivers to wait in a local area or hunt in nearer hotspots. Li *et al.* [22] propose an improved ARIMA based prediction method to forecast the spatial-temporal variation of passenger pick-ups from urban hotspots for drivers. In [23], the authors mine the periodicity and seasonality of passenger demands to answer the choice question: which is the best taxi stand the drivers should head after a drop-off. As for the recommendation of a sequence of potential pick-up positions, Ge *et al.* [24] formulate the targeted problem as a mobile sequential recommendation problem to maximize the drivers profit. T-Finder [7] facilitate taxi drivers' services with a sequence of locations and the routes to these locations based on the behaviors of high profit drivers. In addition, how to cruise in the searching process is also a research topic. In HUNTS [25], the authors aim to find a connected trajectory of high profit rather than several POIs in real-time, which is defined as global-optimal trajectory retrieving problem. They introduce a dynamic scoring system to evaluate road segments, in which picking-up rate and average income are taken into consideration. Zhang *et al.* [6] present a distributed online scheduling strategy named pCruise to obtain efficient cruising routes with the help of a cruising graph, of which the weighting process takes the number of nearby taxis and passengers together into account.

Our strategy goes one step further than the above-mentioned methods by exploring social properties of functional regions upon big traffic data and applies the knowledge to improve drivers' income. To this end, a Time-Location-Sociality model is proposed in order to identify three-dimensional properties of city dynamics, which can predict the distribution of passengers for different social functional regions. Based on the outcomes of proposed model, we recommend profitable areas adjacent to the current driver's location, which would be more effective since drivers are not willing to follow a particular route or cruise a longer distance just for a best pick-up in practice.

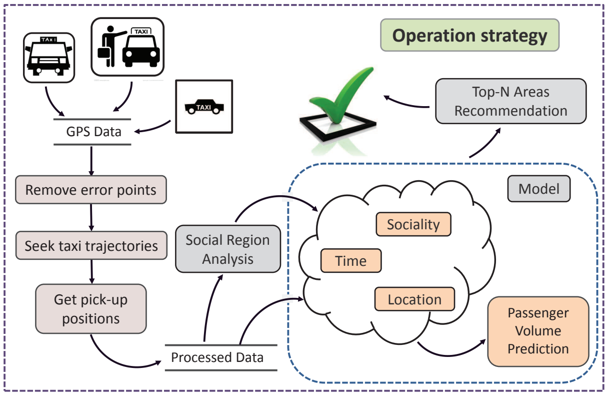


Fig. 1. Overview of the strategy.

TABLE I
RULES OF THE DEMANDS FOR TAXIS.

Social functional areas	Rush hour
Railway stations	Almost the whole day
Highway passenger station	Almost the whole day
Areas of historic interests	9:00-17:00 at weekend
Commercial and entertainment areas	16:00-18:00 on weekday
	10:00-20:00 at weekend
Government organizations/Public institutions	8:00-16:00 on weekday
Residential regions	7:00-11:00 on weekday
	17:00-20:00 on weekday
	10:00-13:00 at weekend
	17:00-20:00 at weekend

III. Operation Strategy

A. Overview

As shown in Fig. 1, we preprocess the data by removing the errors and duplicate records, seeking taxi trajectories and getting all the pick-up positions in different areas firstly. Then we analyze the social features of different areas based on the processed data which is of crucial importance in estimating passenger distributions of the following introduced model. Afterwards, we integrate the three dimensional properties, including time, location and sociality, into our model, which is responsible for predicting the passenger volume for different functional areas. Based on the prediction outcomes, we recommend Top-N areas to drivers at last, which contain more potential passengers in drivers' immediate areas. In following subsections, detail description for each part is presented.

B. Social Region Analysis

With the development of society, more and more functional buildings have been constructed. Around these landmark buildings, social functional regions have appeared one after another, and each of them has its own characteristic. As the improvement of people's living standards, these regions play a critical role in human's daily life and people rely more and more on them than before.

To some extent, people share the similar time rule and movement path in the same functional area. For example, for most staffs living in a residential area, they usually go to work around 8:00 o'clock, and their workplaces may be not far from each other. Especially, as more and more regions with the exclusive function appear, people's time rule and the way to travel form their own patterns, which lead to the variation of traffic conditions and demands for taxis. For instance, people mainly demanded taxis during commuting time before, while the demand for taxis has not been limited in rush hours because of the specific social properties of functional regions. Besides, people went to restaurants and cinemas independently which were not correlated in the past. However, they prefer to set relation between these two sites with the increasement of per capita income nowadays and the improvement of city modernization. So we can infer some rules from the social attributes of functional areas.

Hereby, we divide a big region into smaller ones based on social functions. We study some special functional areas like railway stations, highway passenger station, areas of historic interests, commercial and entertainment areas, government organizations / public institutions and residential regions [18] [26]. We analyze the demands for taxis in the different social areas encompassing the landmarks based on real datasets, and list some rules in Table I. The analysis in detail is shown in Section IV-B. We can conclude that each social functional area has its regular rules of demands for taxis, and the number of passengers in each area is not time-varying, which we can adopt to help drivers find more potential passengers.

C. Time-Location-Sociality Model

A lot of researchers [20] have divided a region into different social function areas based on human mobility and POIs. Moreover, passengers in the same area have similar demands for taxis because of their own social characteristics. From the perspective of sociality, we propose a Time-Location-Sociality model to predict the variation of passenger volume in different social function areas.

1) Time: Taxi drivers do not know how long it will take to ship passengers to destinations in advance. However, drivers are aware of the time point when passengers need them most, such as rush hours in morning. In a word, the number of passengers changes over time.

We set ϕ hour(s) as an interval for each day as shown in equation (1):

$$t_k = [k\phi, (k+1)\phi), k = 0, 1, \ldots, 24/\phi - 1 \qquad (1)$$

where t_k is the kth time interval. In our model, we set $\phi = 2$, and then k is from 0 to 11.

2) Location: Crowds of passengers may appear in different places at the same time. For example, office staffs will go home around 18:00 and students will go home from school at the same period, such that passengers near work places and schools are more than others at that time. So the number of passengers changes over locations.

We partition an area into n equal ones named R_j $(j = 1, 2, \ldots, n)$. Then the number of passengers in region R_j during day i $(D_i,\ i = 1, 2, \ldots, m)$ is denoted by $P_{D_i}^{R_j}$. The calculation of $P_{D_i}^{R_j}$ is illustrated as follows:

$$P_{D_i}^{R_j} = (P_{t_0}^{i,j}, P_{t_1}^{i,j}, \cdots, P_{t_k}^{i,j})^T \tag{2}$$

where $P_{t_k}^{i,j}$ is the number of passengers in region j from $2k$ o'clock to $2(k+1)$ o'clock of day i.

Consequently, the matrix P contains distribution of pick-up places in different regions at different days as shown in equation (3).

$$\begin{aligned} P &= (P_1, P_2, \cdots, P_n) \\ &= \begin{pmatrix} P_{D_1}^{R_1} & P_{D_1}^{R_2} & \cdots & P_{D_1}^{R_n} \\ P_{D_2}^{R_1} & P_{D_2}^{R_2} & \cdots & P_{D_2}^{R_n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{D_m}^{R_1} & P_{D_m}^{R_2} & \cdots & P_{D_m}^{R_n} \end{pmatrix} \end{aligned} \tag{3}$$

3) Sociality: Around the landmark building, more and more social functional circles appear. Nowadays, we can see commercial areas, industrial zones and residential areas in a city. Different people will go to different areas to satisfy their needs at different time points. For example, people will go to residential areas from commercial areas around 18:00, and leave the residential areas for the commercial areas around 8:00. So the number of passengers changes over sociality.

According to [20], we use a $l \times n$ matrix F to present which social property the region belongs to, as follows.

$$\begin{aligned} F &= (F_1, F_2, \cdots, F_n) \\ &= \begin{pmatrix} \times & & & \cdots & \\ & & \times & \cdots & \\ & & & & \times \\ \vdots & & & \ddots & \vdots \\ & \times & & \cdots & \end{pmatrix} \end{aligned} \tag{4}$$

In equation (4), if region j is regarded as a functional area with property of p $(p = 1, 2, \ldots, l)$, the corresponding element is denoted as $\times$ which is equal to 1 and the others in the same column are filled with 0.

We integrate these three vital properties into equation (5), which can identify three dimensional properties of city dynamics so as to estimate the number of passengers for different social functional regions in the following prediction.

$$\begin{aligned} R_j' &= P_j \times F_j^T \\ &= \begin{pmatrix} P_{t_0}^{1,j} & P_{t_0}^{2,j} & \cdots & P_{t_0}^{m,j} \\ P_{t_1}^{1,j} & P_{t_1}^{2,j} & \cdots & P_{t_1}^{m,j} \\ \vdots & \vdots & \ddots & \vdots \\ P_{t_{11}}^{1,j} & P_{t_{11}}^{2,j} & \cdots & P_{t_{11}}^{m,j} \end{pmatrix} \end{aligned} \tag{5}$$

In equation (5), R_j' can be regarded as the amount of pick-ups in the region j during the whole day. Because there is only one column vector which is not full of 0 in the matrix R_j', we choose this column vector and expand it. Eventually, we acquire a final matrix shown in equation (5).

Then, Support Vector Regression (SVR) is used to predict the volume of passengers. SVR is a kind of Support Vector Machine (SVM) which is an algorithm from statistical learning theory proposed by Vapnik in AT&Bell Lab and can get relatively accurate results from sparse data [27]. The parameter selection in SVR satisfies the real-time and accurate requirement in prediction, which has been proved by Wang *et al.* [28]. Its implementation details can be found in LIBSVM, which is a open source library for SVMs [29]. These advantages contribute to achieving more accurate predictions without too many complex operations. Especially, we choose the nonlinear processing of ϵ-SVR to deal with our discrete and nonlinear data, and its functional form is described in equation (6).

$$f(x) = \sum_{j=1}^{N} (\alpha_j^* - \alpha_j) K(x_j, x) + b \tag{6}$$

In equation (6), α_j^* and α_j are Largrange multipliers and Radial Basis Function is worked as the Kernel function K. By using LIBSVM in experiment, we only need to find how many passengers get on taxis in different areas as the input, which is actually the column vector in matrix R_j', and then select proper factors to obtain taxi demand prediction of x day.

D. Top-N Areas Recommendation

Drivers do not know where passengers want to go before passengers get on the taxi, and they can not reject passengers because of the desolate destinations. After passengers arriving, drivers have to find new passengers near the get-off positions which could be anywhere in the city and they are not familiar with. Thus they have no choice but to go randomly to find a passenger. Sometimes drivers may find passengers soon, while they may not find any passengers for a long time if unluckily and at worst they have to go back to the city center or original places to search for passengers. This empirical strategy can not ensure their income every day.

To solve this problem, our strategy provide drivers with some regions which contain more potential passengers based on the results of our prediction model. In the paper, we do not divide the areas into different social functional regions in terms of roads like [20]. On one hand, we have problems in choosing major loads like highways and ring roads or urban arterial roads to divide the map. On the other hand, there exist different sizes of regions, which means some of them may cover a large area and a pair of regions may be far from each other. It take too much time for drivers to leave from one region for another. Considering the above-mentioned problems, we further divide social functional areas into smaller ones.

Veloso *et al.* [26] divide an area into small squares with the edge length of 500 meters, which is so rough to make accurate prediction. So we divide the area into intervals of 0.005 degrees in both latitude and longitude in equation (2). In this situation, we have more chances to get distributions of passengers accurately. Even though the drivers are in a strange

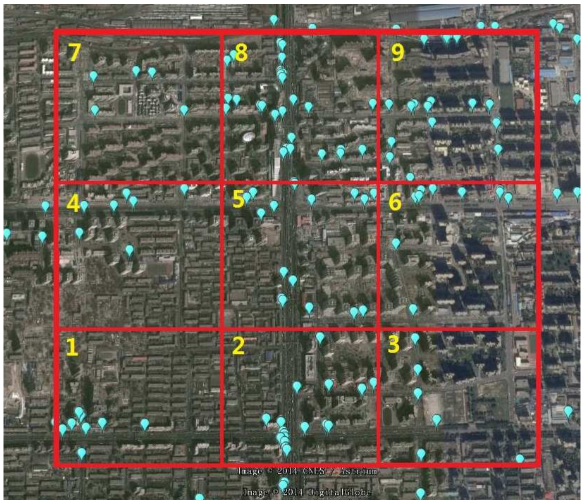


Fig. 2. Distribution of taxis.

place or run without any passengers for a long time, they can drive to some certain areas near them with more potential passengers according to the history and prediction. Thus, we can help the drivers to find more passengers effectively.

As shown in Fig. 2, we partition one region into 9 parts. All the points in the figure are the real get-on positions. If a passenger gets off at region 5, the driver locates at region 5. Then the driver tries to find some passengers near him. But we can know that it is not a smart choice for drivers waiting in this region according to the figure. Assume that if the driver goes to region 1, 4, 7, he may find some passengers in these regions. But we know that the driver can find more passengers in region 8 than other regions. So we recommend region 8 to the driver. In order to raise the chance of picking up passengers, our strategy recommend Top-N regions to the drivers with more potential passengers.

IV. Empirical Study

A. Dataset Description and Process

We use the GPS dataset collected from 12,000 taxis running in Beijing [30]. The size of the uncompressed dataset is 15GB and it contains hundreds of millions of records. Taxis uploaded their position information including time, latitude and longitude to the data base every day in November 2012. The format of the dataset is shown in Table II.

In the paper, we focus on GPS data of taxis in Dongcheng and Chaoyang Districts in Beijing. Dongcheng District locates in the east of Beijing that has established mutually beneficial partnerships with many metropolises all over the world. Meanwhile, Dongcheng Districts makes joint efforts with 45 cities in China. Similarly, Chaoyang District is an developed district in industry. These two districts have not only rich political, economical, cultural, educational, scientific and technological resources, but also high degree of city modernization and good infrastructures. Considering these factors, we select Dongcheng and Chaoyang districts to run the experiments.

TABLE II
Format of dataset

Name	Annotation
Car ID	Unique ID of the car
Trigger event	0:empty 1:nonempty
Running status	0:empty carrying 1:carrying 2, 3, 4: non-service
GPS time	yyyymmddhhmmss
GPS longitude	ddd.ddddddd
GPS latitude	ddd.ddddddd
GPS speed	ddd
GPS location	ddd
GPS status	0:invalid 1:valid

TABLE III
Dataset description.

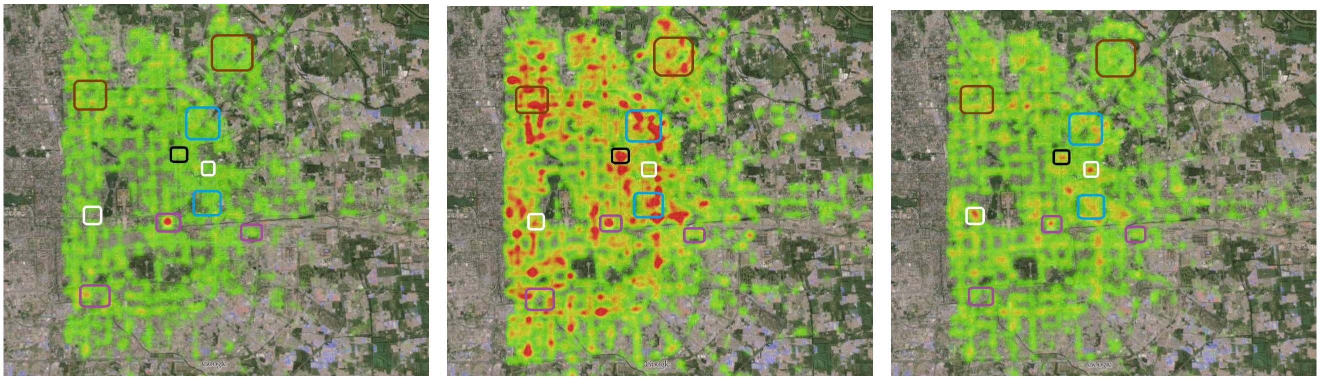
Dataset #	Latitude	Longitude	Time
Dataset 1	[39.820°N, 40.080°N]	[116.350°E, 116.600°E]	weekdays
Dataset 2	[39.820°N, 40.080°N]	[116.350°E, 116.600°E]	weekends
Dataset 3	[39.905°N, 39.960°N]	[116.390°E, 116.415°E]	weekdays
Dataset 4	[39.905°N, 39.960°N]	[116.390°E, 116.415°E]	weekends

The raw data uploaded by GPS devices lack accuracy, so we remove duplicate records, and the invalid records (GPS status = 0 in Table II) which may be caused by device errors or noises. Besides, we discard the records with the running status of 2, 3 and 4 (see Table II), which means the taxi is not in service. The percentage of these alterations on the dataset is about 0.92 %. Then we select pick-up records roughly in Dongcheng & Chaoyang district and Dongcheng district in terms of latitude and longitude, forming two preliminary datasets. Moreover, we separate each preliminary dataset into two parts, the records on weekdays and at weekends, which imply different human mobility patterns and can be of advantage to achieve more accurate results. After that, we acquire four datasets (as shown in Table III) for the following experiments.

When completing these steps, we obtain many taxi information in one file for one dataset, but it is difficult to find the track information of each taxi in a certain time interval. Thus, in each dataset, we further put the continuous taxi information which has the same ID into one file, and then divide each file based on the unit of 2 hours. At last, we get loads of files named by taxi ID and time. In this way, we can obtain the trace of a taxi in a certain period including not only the current position, but also where the passenger get on and get off.

B. Exploring Passenger Distributions

We map the pick-up locations from the processed data onto Google Map in each two hours. The heatmap Fig. 3 and Fig. 4 depict the distribution of pick-up positions in terms of 6:00-8:00, 8:00-10:00 and 16:00-18:00 of a weekday in Dataset 1 and a weekend in Dataset 2 respectively. Furthermore, in order

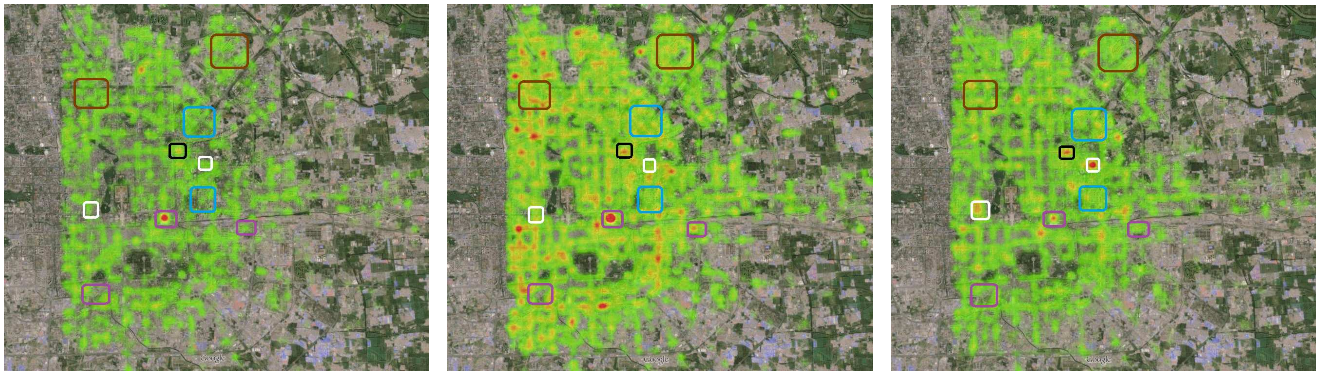


(a) 6:00-8:00

(b) 8:00-10:00

(c) 16:00-18:00

Fig. 3. Distribution of pick-up positions on a weekday.



(a) 6:00-8:00

(b) 8:00-10:00

(c) 16:00-18:00

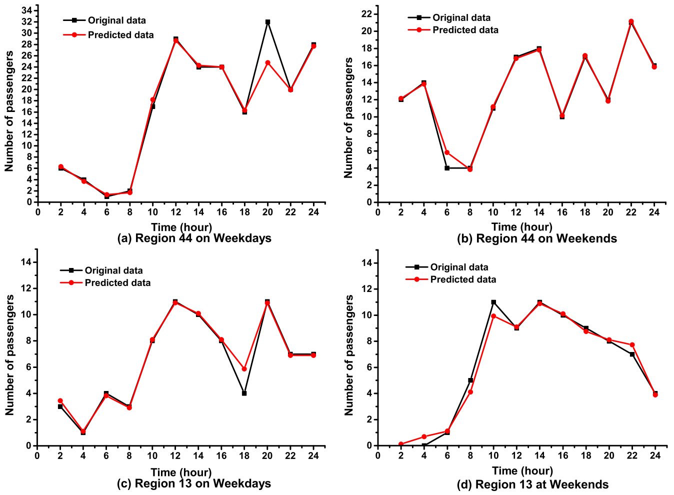
Fig. 4. Distribution of pick-up positions at a weekend.

to clarify the importance of social property, we circle several social regions in different colors in Figs. 3 and 4.

In Fig. 3, train stations representing by purple rectangles have a great demand for taxis in the whole day, especially the Beijing Railway Station in the middle of the map in the morning (see Figs. 3(a) and 3(b)). Similarly, the black one standing for a highway passenger station is also an all-day hotspot, and the peak appears in 8:00-10:00 as shown in Fig. 3(b). These stations provide twenty-four-hour services for travellers from all over China, and travellers prefer heading to their destinations by taking taxis. Residential regions in brown rectangles have more passengers than other social functional regions (except stations) in Fig. 3(a), and they are still filled with pick-ups in 8:00-10:00 as shown in Fig. 3(b). The phenomenon confirms the fact that residents rush to work through catching taxis. Diplomatic and embassy regions which contain government organizations and public institutions are located in the two blue rectangles of Fig. 3(b). It implies that people prefer to handle affairs on weekday morning as soon as possible, and then go to the next destination. According to Fig. 3(c), people in white rectangles (commercial and entertainment areas) finish shopping and try to go back home for dinner, which contributes to the rising of pick-ups in these area. Moreover, people have regular off hours in embassy regions, commercial areas and developed residential regions which contain various public infrastructures, such as schools, banks and emporiums, so they need more taxis in the period of 16:00-18:00 (see Fig. 3(c)).

Fig. 4 illustrates the decrease in the overall number of passengers at weekends in comparison to it on weekdays, while the reduction is not applicable to the train stations and highway passenger stations which have almost the same amount of passengers as usual. Especially, there are not so many people starting from residential regions in the morning (see Fig. 4(a) and Fig. 4(b)), which may arise from the consensus that people would rather stay at home or they are not so time-sensitive at weekends than they on weekdays. The inaccessibility of governmental agencies and public organizations in diplomatic and embassy regions directly brings about the dramatic reduction of passenger volume as the blue rectangles show in Figs. 3(b) and 4(b). On the contrary, the number of pick-ups significantly rise in commercial and entertainment areas in Fig. 4(c). Intuitively, more people are fond of relaxing themselves at weekends after a stressful week.

In brief, we can infer that the social properties of regions, along with time and location, play a vital part in the distribution of passengers according to Figs. 3 and 4. Based on the achievement in [20], we explore 8 kinds of social functional



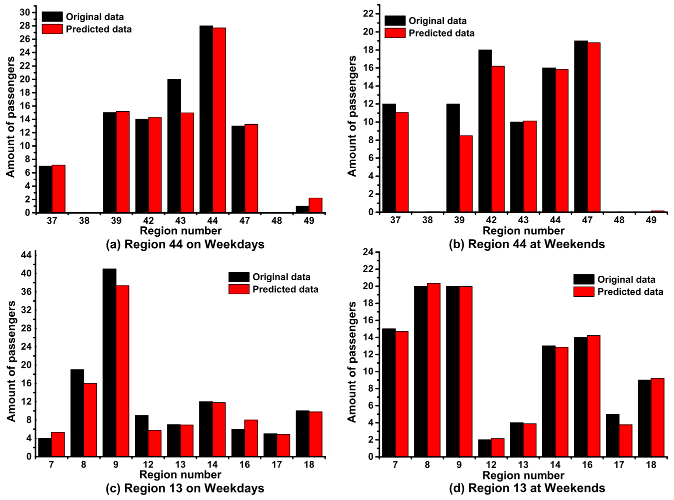
(a) Region 44 on Weekdays

(b) Region 44 on Weekends

(c) Region 13 on Weekdays

(d) Region 13 at Weekends

Fig. 5. Comparison between original data and predicted data.



(a) Region 44 on Weekdays

(b) Region 44 at Weekends

(c) Region 13 on Weekdays

(d) Region 13 at Weekends

Fig. 6. Distribution of passengers in near 9 regions

areas and each kind has its unique human mobility pattern, which can contribute to predicting passenger volume. To be specific, we set $l = 8$ in equation (4) and fill the matrix F depending on whether a area belongs to a kind of social region.

C. Predicting Passenger Volume

We conduct the following experiments in Datasets 3 and 4 respectively, which cover a typical social region with dense records. Later, we divide the whole region into 60 small ones denoted by Region 1, 2, ..., 60, and then apply our model to each of these 60 regions to predict the number of passengers. We use first 9 days' taxi data as a training set to predict the 10th day's passenger volume in Dataset 3, while for Dataset 4 the first 7 days' data constitute a training set to predict the last day's situation. Moreover, the widely range of passenger numbers would reduce the accuracy of prediction result, hence, we normalize these data into the interval of (0, 1), which achieves the best experiment results compared with others.

Due to space limitations, we only show the results of a typical residential region (Region 44, [39.945°N, 39.950°N] and [116.410°E, 116.415°E]), and an area of historic interests (Region 13, [39.9150°N, 39.9200°N] and [116.4050°E, 116.4100°E]). As shown in Fig. 5, we find that most of the prediction results are close to original data. By contrasting the data on residential areas and areas of historic interests, we find some valuable information: a) More taxis are demanded on weekdays in Region 44 (see Fig. 5(a)) than Region 13 (see Fig. 5(c)), this is due to more people live in residential areas than historic areas. b) For Region 44, people need more taxis on weekdays (203 pick-ups in total from Fig. 5(a)) than weekends (156 from Fig. 5(b)), however, the number of passengers in the Region 13 has not significantly changed from workdays (77 from Fig. 5(c)) to weekends (75 from Fig. 5(d)). c) According to Fig. 5(a) and Fig. 5(b), people prefer not to take taxis on weekends and they may stay at home for a rest in the morning, while after 8 p.m., many people prefer to go out by taxis. d) Region 13 is labelled as an area of historic interests, so the passengers there contain lots of tourists. As shown in Fig. 5(c), the number of passengers declines from 12 o'clock and reaches the trough at 18:00. However, we do not see this situation in Fig. 5(d) because tourists are more willing to visit places of interests at weekends.

By applying the proposed model to Datasets 3 and 4, we find average prediction accuracies of all these 60 regions are 90.14% on weekdays and 86.37% at weekends, which demonstrate our method is practicable to predict the distribution of passengers. And the results also imply that people have a more regular life on weekdays than weekends. For example, students go to school and employees go work at fixed times on weekdays, however, people have more options at weekends because of ample time. Due to the utilization of social properties which have long-term attributes, it is easier for our model to predict the accurate number of passengers on weekdays. Besides, we also observe the prediction accuracies are lower at hutong districts. In such case, the road situation is more complex and it is difficult to capture social attributes, which degrades the performance of our strategy.

D. Recommending Top-N Areas

We obtain the accurate distributions of passengers with our model, thereby our strategy can recommend the areas with more passengers to drivers in the near future. Fig. 6(a) and Fig. 6(b) depict the predicted results and the original results of passengers in 9 regions near Region 44 from 22:00 to 24:00 on weekdays and weekends respectively. Similar to Figs. 6(a) and 6(b), Fig. 6(c) and 6(d) contain the information of 9 regions around Region 13. These four sub-figures also imply that our proposed strategy achieves high accuracy comparing to original data. Taking Fig. 6(c) as an example, if a driver is located in Region 13, we can recommend Region 9 to the driver, which is the southeast area next to the current region and contains more profitable opportunities to pick up passengers. Therefore, drivers can improve their incomes, avoid wasting time and reduce energy consumption.

V. Conclusion

In this paper, we go one step further by exploring social attributes of functional regions upon big traffic data in Beijing and apply the knowledge to maximize drivers' profit. Therefore, a Time-Location-Sociality model is introduced in order to identify three-dimensional properties of city dynamics, which can effectively predict the distribution of passengers for different social functional regions. According to the prediction outcomes of the model, we recommend Top-N profitable areas near to the driver's real-time position. The extensive experiments on the real dataset show that we achieves prediction accuracies of 90.14% on weekdays and 86.37% at weekends respectively, which demonstrates the validity of the proposed taxi operation strategy.

As the first step on predicting passenger distributions in different social functional areas, we are short of providing experimental evidence to study the applicability of our strategy. Therefore, we consider to apply our strategy to other countries and cities with the most current datasets in the future. In addition, we will continue to improve the efficiency with other machine learning approaches and area division algorithms.

Acknowledgment

This work was partially supported by the Natural Science Foundation of China under Grant No. 61174174 and No. 61203165, the Foundation of Key Laboratory of System Control and Information Processing, Ministry of Education, China, under Grant No. SCIP2012001, Prospective Research Project on Future Networks, Jiangsu Future Networks Innovation Institute, China, under Grant No.BY2013095-3-01 and No.BY2013095-3-03, and the Fundamental Research Funds for the Central Universities under Grant DUT15YQ112 and No.DUT15QY06.

References

[1] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, "An information framework for creating a smart city through internet of things," *IEEE Internet of Things Journal*, 2014, DOI: 10.1109/JIOT.2013.2296516.

[2] http://www.fastcoexist.com/1679127/the-top-10-smart-cities-on-the-planet.

[3] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Dmodel: Online taxicab demand model from big sensor data in a roving sensor network," in *2014 IEEE International Congress on Big Data (BigData Congress)*, Anchorage, AK, USA, Jun. 2014, pp. 152–159.

[4] K. Su, J. Li, and H. Fu, "Smart city and the applications," in *2011 International Conference on Electronics, Communications and Control (ICECC)*, Zhejiang, China, Sep. 2011, pp. 1028–1031.

[5] http://zhengwu.beijing.gov.cn/ghxx/sewgh/t1237237.htm.

[6] D. Zhang, T. He, S. Lin, S. Munir, and J. Stankovic, "pcruise: Online cruising mile reduction for large-scale taxicab networks," *IEEE Transactions on Parallel and Distributed Systems*, 2014, DOI:10.1109/TPDS.2014.2364024.

[7] 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, Oct. 2013.

[8] D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, "Understanding taxi service strategies from taxi gps traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 123–135, Feb. 2015.

[9] N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng, and H. Xiong, "Discovering urban functional zones using latent activity trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 3, pp. 712–725, Mar. 2015.

[10] C. Zhong, X. Huang, S. M. Arisona, G. Schmitt, and M. Batty, "Inferring building functions from a probabilistic model using public transportation data," *Computers, Environment and Urban Systems*, vol. 48, no. 0, pp. 124–137, Nov. 2014.

[11] G. Pan, G. Qi, Z. Wu, D. Zhang, and S. Li, "Land-use classification using taxi gps traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 113–123, Mar. 2013.

[12] F. Xia, A. Ahmed, L. Yang, and Z. Luo, "Community-based event dissemination with optimal load balancing," *IEEE Transactions on Computers*, 2014, DOI:10.1109/TC.2014.2345409.

[13] O. Söderström, T. Paasche, and F. Klauser, "Smart cities as corporate storytelling," *City: analysis of urban trends, culture, theory, policy, action*, vol. 18, no. 3, pp. 307–320, Jun. 2014.

[14] N. Walravens, "Mobile city applications for brussels citizens: Smart city trends, challenges and a reality check," *Telematics and Informatics*, vol. 32, no. 2, pp. 282–299, May 2015.

[15] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou, "A new insight into land use classification based on aggregated mobile phone data," *International Journal of Geographical Information Science*, vol. 28, no. 9, pp. 1988–2007, May. 2014.

[16] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, Beijing, China, Aug. 2012, pp. 1–8.

[17] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi gps traces to social and community dynamics: A survey," *ACM Computing Surveys*, vol. 46, no. 2, pp. 17:1–17:34, Dec. 2013.

[18] G. Qi, X. Li, S. Li, G. Pan, Z. Wang, and D. Zhang, "Measuring social functions of city regions from large-scale taxi behaviors," in *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Seattle, USA, Mar. 2011, pp. 384–388.

[19] Y. Liu, F. Wang, Y. Xiao, and S. Gao, "Urban land uses and traffic 'source-sink areas': Evidence from gps-enabled taxi data in shanghai," *Landscape and Urban Planning*, vol. 106, no. 1, pp. 73–87, May. 2012.

[20] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Beijing, China, Aug. 2012, pp. 186–194.

[21] 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 *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, Seattle, USA, Mar. 2011, pp. 63–68.

[22] X. Li, G. Pan, Z. Wu, G. Qi, S. Li, D. Zhang, W. Zhang, and Z. Wang, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers of Computer Science*, vol. 6, no. 1, pp. 111–121, Feb. 2012.

[23] L. Moreira-Matias, J. Gama, M. Ferreira, J. Mendes-Moreira, and L. Damas, "Predicting taxi–passenger demand using streaming data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1393–1402, Sep. 2013.

[24] 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*, Washington, USA, Jul. 2010, pp. 899–908.

[25] Y. Ding, S. Liu, J. Pu, and L. M. Ni, "Hunts: A trajectory recommendation system for effective and efficient hunting of taxi passengers," in *2013 IEEE 14th International Conference on Mobile Data Management (MDM)*, vol. 1, Milan, Italy, Jun. 2013, pp. 107–116.

[26] M. Veloso, S. Phithakkitnukoon, and C. Bento, "Urban mobility study using taxi traces," in *Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis*, Beijing, China, Mar. 2011, pp. 23–30.

[27] G. W. Flake and S. Lawrence, "Efficient svm regression training with smo," *Machine Learning*, vol. 46, no. 1-3, pp. 271–290, Jan. 2002.

[28] F. Wang, G. Tan, C. Deng, and Z. Tian, "Real-time traffic flow forecasting model and parameter selection based on ε-svr," in *7th World Congress on Intelligent Control and Automation*, Chongqing, China, Jun. 2008, pp. 2870–2875.

[29] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[30] http://www.datatang.com/data/44502.