Time-Location-Relationship Combined Service Recommendation Based on Taxi Trajectory Data

Xiangjie Kong, Feng Xia, Jinzhong Wang, Azizur Rahim, and Sajal K. Das, *Fellow, IEEE*

***Abstract*—Recently, urban traffic management has encountered a paradoxical situation which is the empty carrying phenomenon for taxi drivers and the difficulty of taking a taxi for passengers. In this paper, through analyzing the quantitative relationship between passengers' getting on and off taxis, we propose a Time-Location-Relationship combined taxi service recommendation model (TLR) to improve taxi drivers' profits, uncover the knowledge of human mobility patterns, and enhance passengers' travel experience. Moreover, TLR model uses Gaussian Process Regression and statistical approaches to acquire passenger volume, mean trip distance, and average trip time in functional regions during every period on weekdays and weekends, and allow drivers to pick up more passengers within a short time frame. Finally, we compare our proposed model with Auto Regressive Integrated Moving Average model (ARIMA), Back-Propagation Neural Network model (BPNN), Support Vector Machine model (SVM), and Gradient Boost Decision Tree model (GBDT) by using the real taxi GPS data in Beijing. The experimental results show that our optimizing taxi service recommendation can predict more accurately than others by considering the three dimensional properties.**

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

I. Introduction

In recent years, leveraging the widely used mobile computing and sensing technologies, we acquire a variety of big data more easily including traffic data, mobile phone data, and noise data. These big data contribute to people's life improvement, economic development, and environmental sustainability through sensing, integration, and analysis [1].

However, with the increasing scale and population of modern cities (such as New York, London and Beijing), there exists a paradoxical situation in urban traffic control and management which is the inconvenience of taking a taxi for passengers and the empty carrying phenomenon for taxi drivers. Based on the statistics from Beijing traffic development and construction report during the 12th five-year plan [2], each taxi travels about 400 kilometers everyday in Beijing, while the average empty carrying ratio of taxis is about 40%. It implies that taxis run with low efficiency in nearly half time, which leads to more transportation resources consumption and environmental pollution. However, although taxis are mostly in the vacant scenario, there are still many citizens annoyed by the difficulty of taking a taxicab.

To our knowledge, there are a great number of methods 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 [3]–[5]. These literatures propose experimental proofs on the relationship between the spatio-temporal property of trajectory and the efficiency of passenger-finding algorithms. However, the relationship between passengers' getting on and off has not been explored deeply. Thus, further research is needed to improve the prediction accuracy of pick-up passengers. On the other hand, modern cities are made up of diverse functional areas, such as commercial areas, residential areas, and entertainment areas. These functional areas are historically associated with city planning, but are mostly shaped by people's actual needs for social activities over a long period [6] [7]. For example, entertainment areas are generally visited by people in order to relax on weekday evenings and weekends. Furthermore, people commute between these functional areas to engage in a sequence of related social activities. Intuitively, people usually go to entertainment areas from their offices on weekdays, whereas from residential areas on weekends. But these literatures have not exactly illustrated how to improve drivers' profits by leveraging the functional regions' characteristics.

In the paper, we aim to propose a data-driven taxi service recommendation using Time-Location-Relationship model (TLR) to tackle the tough issue above. By integrating, processing and analyzing the taxi trajectory data, we identify the quantitative relationship between passengers' getting on and off taxis in urban functional regions, and acquire people's mobility patterns efficiently. For example, we discover that people often go to diplomacy areas for handling affairs in the morning and then prefer to have lunch rather than leave immediately, which contributes to lots of cafe and tea restaurants that serve fairly decent food. In addition, we infer that most people would like to spend about 3 hours in visiting places of historical interest between 11 o'clock and 14 o'clock. The evolution law of passengers' going up and down in social functional areas is relatively stable and exclusive compared to individual mobility [8] [9], and can be used as a service recommendation for the taxi drivers.

To start with, we utilize the mobility patterns in TLR model,

This work was partially supported by the National Natural Science Foundation of China (61572106), the Natural Science Foundation of Liaoning Province, China (201602154), the Dalian Science and Technology Planning Project (2015A11GX015, 2015R054), and the US National Science Foundation Grants (CNS-1545037, CNS-1545050, CCF-1533918, CBET-1609642).

X. Kong, F. Xia, J. Wang, and A. Rahim are with the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China. J. Wang is also with Shenyang Sport University, Shenyang 110102, China.

S. K. Das is with Department of Computer Science, Missouri University of Science and Technology, USA.

The corresponding author of this paper is Feng Xia, Email: f.xia@ieee.org.

which considers three dimensional properties (time, location and relationship) of city dynamics. Then we can predict the number of passengers' getting on taxis with the help of the amount of passengers' getting off taxis in different social functional regions using the introduced model. Subsequently, we calculate mean trip distance and mean trip time between any two functional regions. Furthermore, we recommend Top-N areas for drivers according to our proposed model, which enables drivers to pick up more passengers for energy saving, and lightens the transportation pressure for a city.

A previous version of our work has been proposed in [10]. This paper differs from our previous work in that 1) TLR model mainly utilizes the quantitative relationship between passengers' getting on and off taxis, whereas the original model only leverages the number of passengers' going up in different functional regions; 2) in this paper, we expand simulation experiments and adopt more evaluation metrics to examine the performance of TLR compared with ARIMA model, BPNN model, SVM model, and GBDT model which leads to the improvement of prediction accuracy.

Our major contributions can be summarized as follows:

- We propose some important human mobility patterns of functional regions through analyzing the quantitative relationship between passengers' getting on and off taxis in every period.
- We present TLR 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, mean trip distance, and average trip time. Then they can decide where to pick up passengers to maximize their profits. The results achieve prediction accuracies of 90.9% on weekdays and 80.4% on weekends respectively.
- We evaluate and compare the performance of our proposed model, ARIMA model, BPNN model, SVM model, and GBDT model by utilizing the following metrics including Correlation Coefficient (CC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Normalized Mean Absolute Error (NMAE), which determines the effectiveness and stability of our proposed model.

The remainder of the paper is organized as follows. In Section II, we give a brief review on the related work. Section III mainly illustrates service recommendation for taxi drivers based on TLR model. Then we describe the dataset and the experiments we conduct to evaluate our proposed model in Section IV. In Section V, we compare our model with the other four methods based on four metrics. The work is concluded in Section VI.

II. Related Work

Mining taxi trajectory data has been a research hotspot in the smart city [1], intelligent transportation management [11], data mining and vehicular ad-hoc network [12]–[14]. Through analyzing some relevant literatures, functional regions' characteristics and taxi drivers' service strategies usually play an important role in increasing their revenues [15]–[23]. In this section, we mainly organize the related work in the following two subsections.

A. Functional Region Analysis

With the rapid economic development in urban areas, different functional regions have been generated. The characteristics of functional areas show the urban human mobility patterns, and then contribute to sensing and understanding urban dynamic problems. In [15] [16], authors use mobile phone position dataset to categorize land uses, whereas [13] utilize taxi GPS trajectories to assess the region function in cities. Tang *et al.* [18] firstly make a deep analysis of travel demand distributions through clustering get-on and get-off locations, and study human mobility patterns based on the three properties such as travel time, distance and average speed. Likewise, Liu *et al.* [19] propose that different functional regions show different temporal patterns of getting on and off, which is very useful to intelligent traffic management. Moreover, Qin *et al.* [24] explore abnormal traffic flows using independent component analysis with more than 90 percent accuracy. In [25], Csáji *et al.* use principal component analysis to uncover the relation between human behavior characteristics and their spatial locations from mobile phone dataset. Ahmadi *et al.* [26] propose an improved Non-negative Matrix Factorization framework and discover accurately the traffic motion patterns.

Additionally, a topic-modeling-based approach is proposed to analyze region functions by introducing points of interests (POIs) in taxi GPS trajectories [27], and [6] further gives a summary of the problem and utilizes location information and mobility semantics hidden in their time-stamped trajectories to enhance the framework performance. Zhong *et al.* [7] use multi-source data got from surveys and smart card systems to infer building functions and understand social activities. Mazimpaka *et al.* [17] consider the importance of non-taxi users' information and the inflexibility of predetermined POIs to combine taxi GPS and flickr data for discovering functional regions. Zhou *et al.* [28] extract potential functionally critical network locations from people's moving trajectories, uncover the space-time traveling patterns of a particular population and study the relationship between urban functional structures and people's activities. In sum, the above mentioned methods have made insights into the use of urban space and explored the functional regions' characteristics through mining different traffic data, which may facilitate some kinds of valuable application in our daily life. But it has not been elaborated how to utilize the acquired knowledge.

B. Taxi Service Strategies

Taxi service strategies have a direct impact on drivers' revenues, environmental pollution and inhabitant travel efficiency. There are several passenger-finding strategies to connect passengers to vacant taxis, which mainly focus on identifying and recommending some popular positions to drivers. Li *et al.* [20] adopt L1-Norm SVM to discover both efficient and inefficient passenger-finding strategies, then provide correct driving strategies to taxi drivers according to time and location.

Li *et al.* [21] propose an ARIMA based prediction method to predict the spatial-temporal variation of picking up passengers in a popular region and help taxi drivers to find the next passenger. Ge *et al.* [22] firstly formulate a mobile sequential recommendation problem and then provide a potential travel distance function and a recommendation algorithm to help drivers to get a sequence of potential pick-up positions. In [23], the authors introduce three time-series forecasting techniques to predict the spatial distribution of taxi-passengers for a short-term time horizon using streaming data. T-Finder [4] recommends drivers with locations and the routes to these locations, which also makes the passengers easily find vacant taxis in the locations. Then a probabilistic model is proposed to estimate profits of the candidate locations for a particular driver.

Furthermore, it is also a research hotspot on how to cruise in the process of finding the next passengers. Zhang *et al.* [5] utilise a feature matrix to categorize the taxi service strategies and show the relationship between strategies and revenues, therefore analysing the efficiency of each strategy from three perspectives. Zhang *et al.* [3] design a cruising system named PCruise to make taxi drivers to obtain an efficient cruising route with the minimum length and at least one arriving passenger with the help of a weighted cruising graph.

However, the above-mentioned methods mainly focus on recommending popular pick-up positions, potential travel lengths, but do not consider the hidden quantitative relationship between getting on and off taxis in the taxi GPS trajectories. In this paper, our proposed model goes one step further to uncover the correlation between passengers' going up and down, which is applied to predict passengers' spatio-temporal distribution in different social functional regions, and improve taxi drivers' incomes. In conclusion, TLR model is proposed in order to identify three-dimensional properties of city dynamics, which can help taxi drivers to find the next potential passengers in a short-term time. Based on the outcomes of the proposed model, we recommend profitable areas adjacent to the current driver's location, which would be more effective in practice since drivers are not willing to follow a particular route or cruise a longer distance just for the best pick-up position.

III. Service Recommendation

A. Overview

As shown in Fig. 1, for the acquired taxi trajectory data, we firstly delete the unrelated and error records, find taxi trajectories and utilize a statistical analysis method to get all the pick-up and drop-off positions in different areas. Then we analyze the quantitative relationship between passengers' getting on and off taxis, shed light on the human mobility patterns of different areas, and then sense the spatio-temporal distribution of passengers. Moreover, leveraging the three dimensional properties such as time, location and relationship, TLR model utilizes Gaussian Process Regression (GPR) method to predict the passenger volume for different regions. Finally, we provide Top-N areas to taxi drivers, which have the most potential passengers meaning higher profits in the surrounding areas. We have an elaborate description for taxi service recommendation in the following subsections.

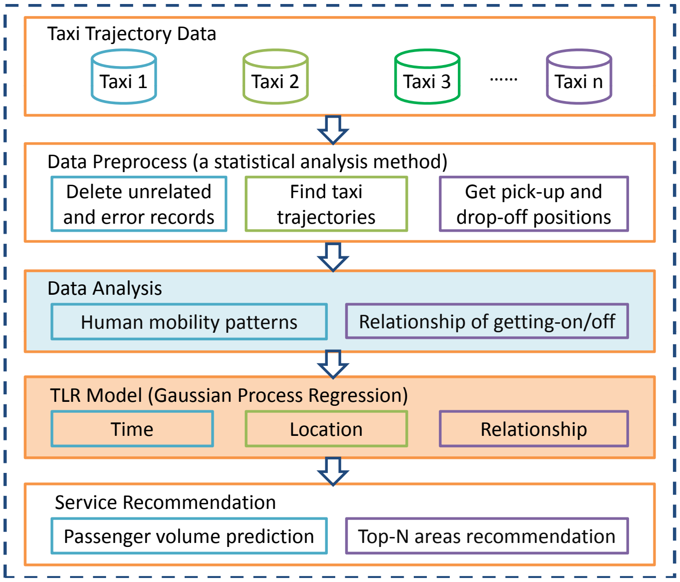

Taxi Trajectory Data
Taxi 1
Taxi 2
Taxi 3
......
Taxi n
Data Preprocess (a statistical analysis method)
Delete unrelated and error records
Find taxi trajectories
Get pick-up and drop-off positions
Data Analysis
Human mobility patterns
Relationship of getting-on/off
TLR Model (Gaussian Process Regression)
Time
Location
Relationship
Service Recommendation
Passenger volume prediction
Top-N areas recommendation

Fig. 1. Overview of taxi service recommendation.

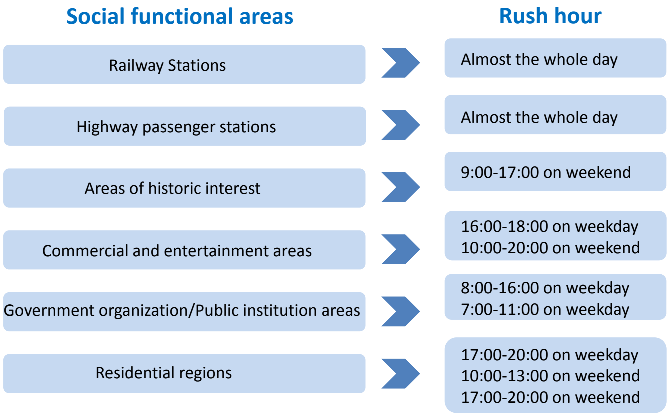


Fig. 2. Rules of the demands for taxis.

B. Quantitative Relationship Analysis

Nowadays, different functional regions have been formed along with the economic development, which facilitates human's daily life and impact on human travel patterns at the same time. What's more, get-on and get-off amount of social regions have a strong relationship, show their own characteristics with spatio-temporal variation, and reflect that people rely more and more on social functional regions than before.

According to our analysis, citizens usually follow the similar temporal movement trace in the same functional region. The distribution of functional regions has a crucial impact on the quantitative change of passengers' getting on and off taxis. For instance, in a residential area, people often go to work around 8 o'clock if their workplaces are not far apart, which leads to a great demand for taking a taxi. So get-on amount is much more than the get-off amount in rush hours, and it is an opposite situation in evening peak hours. Furthermore, in some other functional regions, people usually want to take a taxi for traveling not only in rush hours but also in the whole day because of the unique social characteristics, which contribute to different variation principles of passengers' getting on and off. Although people always keep the movement from one place to another, there are still some rules which we can infer from the amount of passengers' getting on and off in functional areas.

Besides, in different functional regions, the temporal variation of get-on/off amount demonstrates some social characteristics and depicts regional social dynamics. We discover that daily get-on/off amount in the city is closely associated with citizens' daily intensity of social activity. Moreover, the different get-on/off patterns in functional areas are very useful for the taxi drivers' service recommendation. In this paper, we divide a big region into smaller ones based on social functions. We study some special functional areas like railway stations, highway passenger stations, areas of historical interest, commercial and entertainment areas, government organization/public institution areas and residential areas, etc. [29]. Through a qualitative analysis, the patterns of different functional areas have their own characteristics on weekdays and weekends. In railway stations and highway passengers stations, the get-on/off number has a special temporal variation, which means citizens' travel is relevant to the days such as holidays. In contrast, we infer that the peak of get-on/off number is very different respectively in other functional areas, and shows the spatio-temporal distribution characteristics.

Therefore, we analyze the quantitative relationship between passengers' going up and down which reflects the demands for taxis in the different social areas encompassing the landmarks. Then we also summarise some rules as shown in Fig. 2. In Section IV-B, the detailed analysis is presented. We can conclude that each social functional region has its regular rules of passengers' getting on and off, and the get-on number is not independent of the get-off number, which can be used to help taxi drivers find more potential passengers in time.

C. TLR Model

Based on the three-dimensional properties (time, location, and relationship), we propose TLR model to capture the changing regulations of passenger volume in different social function areas. Particularly, TLR model analyzes the relationship between passengers' getting on and off during every period and adopts GPR to predict pick-up passengers' amount with the input of drop-off customers' number.

1) Time: Time is an important temporal factor in taxi services when taxi drivers drop off passengers and wait for their next customers. Especially, people usually represent different movement patterns during different specific time slots, such as in the morning or evening rush hours. So we separate a week into weekdays from Monday to Friday and weekends from Saturday to Sunday respectively, and set φ hour(s) as an interval for each day as shown in equation (1):

$$t_i = [i\varphi, (i+1)\varphi), i = 0, 1, \ldots, (24/\varphi) - 1 \tag{1}$$

where t_i is the ith time slot. We set $\varphi = 2$ and divide a day into 12 time slots. So i is from 0 to 11.

2) Location: Location is a crucial spatial factor in taxi services. It's well known that taxi drivers may find more passengers in popular regions than in rural areas with respect to the demand for a taxi. In addition, there are more passengers near work places and schools than other locations in the evening rush hours. So the choice of taxi cruising position plays a role in improving their revenues.

In this paper, we divide an area into n equal parts named R_j $(j = 1, 2, \ldots, n)$. Then the get-on and get-off number of passengers in Region R_j during day i $(D_i,\ i = 1, 2, \ldots, m)$ are denoted by $P_{D_i}^{R_j}$ and $Q_{D_i}^{R_j}$ respectively. The calculation of $P_{D_i}^{R_j}$ and $Q_{D_i}^{R_j}$ is illustrated as follows:

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

$$Q_{D_i}^{R_j} = (Q_{t_0}^{i,j}, Q_{t_1}^{i,j}, \cdots, Q_{t_k}^{i,j})^T \tag{3}$$

where $P_{t_k}^{i,j}$ and $Q_{t_k}^{i,j}$ are the get-on/off number of passengers in Region j from $2k$ o'clock to $2(k+1)$ o'clock of day i.

Consequently, as shown in equation (4) and (5), we utilize the matrix P to represent the distribution of pick-up places in different regions during a different period, whereas the matrix Q represents the distribution of drop-off places.

$$\begin{aligned} P &= (P_1, P_2, \cdots, P_n) \\ &= \begin{pmatrix} P_{D_1}^{R_1} & P_{D_1}^{R_2} & \cdots & P_{D_1}^{R_n} \\ P_{D_2}^{R_1} & P_{D_2}^{R_2} & \cdots & P_{D_2}^{R_n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{D_m}^{R_1} & P_{D_m}^{R_2} & \cdots & P_{D_m}^{R_n} \end{pmatrix} \end{aligned} \tag{4}$$

$$\begin{aligned} Q &= (Q_1, Q_2, \cdots, Q_n) \\ &= \begin{pmatrix} Q_{D_1}^{R_1} & Q_{D_1}^{R_2} & \cdots & Q_{D_1}^{R_n} \\ Q_{D_2}^{R_1} & Q_{D_2}^{R_2} & \cdots & Q_{D_2}^{R_n} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{D_m}^{R_1} & Q_{D_m}^{R_2} & \cdots & Q_{D_m}^{R_n} \end{pmatrix} \end{aligned} \tag{5}$$

3) Relationship: Relationship is the third important factor in taxi service recommendation. In our daily life, citizens usually follow the similar temporal rule and movement trace for their social activities, which impacts on the variation of passengers' getting on and off taxis in a certain way. Furthermore, we infer that the get-on number is closely connected with the get-off number in different functional regions. By utilizing the property, we can achieve more prediction accuracy of passenger volume in various social areas.

We introduce a $l \times n$ matrix F to classify social property of a region based on [27]. Furthermore, we explore 6 kinds of social functional areas and each kind has its unique human mobility pattern, which can contribute to predicting passenger volume. To be specific, we set $l = 6$ in equation (4) and (5), then fill the matrix P and matrix Q depending on whether an area belongs to a kind of social region.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2684163, IEEE Transactions on Industrial Informatics

$$\begin{aligned} F &= (F_1, F_2, \cdots, F_n) \\ &= \begin{pmatrix} \times & & & \cdots & \\ & & \times & \cdots & \\ & & & & \times \\ \vdots & & & \ddots & \vdots \\ & \times & & \cdots & \end{pmatrix} \end{aligned} \tag{6}$$

As shown in equation (6), the Region j is denoted by F_j. If F_j belongs to a functional region with property of p ($p = 1, 2, \ldots, l$), the element $\times$ is filled with 1 and the others in F_j are 0.

In equation (7) and (8), we leverage these three vital properties to identify three dimensional properties of city dynamics and predict the get-on number of passengers for different social functional regions in the following prediction.

$$\begin{aligned} R'_j &= P_j \times F_j^T \\ &= \begin{pmatrix} P_{t_0}^{1,j} & P_{t_0}^{2,j} & \cdots & P_{t_0}^{m,j} \\ P_{t_1}^{1,j} & P_{t_1}^{2,j} & \cdots & P_{t_1}^{m,j} \\ \vdots & \vdots & \ddots & \vdots \\ P_{t_{11}}^{1,j} & P_{t_{11}}^{2,j} & \cdots & P_{t_{11}}^{m,j} \end{pmatrix} \end{aligned} \tag{7}$$

$$\begin{aligned} S'_j &= Q_j \times F_j^T \\ &= \begin{pmatrix} Q_{t_0}^{1,j} & Q_{t_0}^{2,j} & \cdots & Q_{t_0}^{m,j} \\ Q_{t_1}^{1,j} & Q_{t_1}^{2,j} & \cdots & Q_{t_1}^{m,j} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{t_{11}}^{1,j} & Q_{t_{11}}^{2,j} & \cdots & Q_{t_{11}}^{m,j} \end{pmatrix} \end{aligned} \tag{8}$$

In equation (7) and (8), we use R'_j and S'_j to denote the pick-up amount and the drop-off amount respectively in the Region j during the whole day. In the matrix R'_j and S'_j, we select the one column vector which is not filled with 0 and expand it. At last, we extract the number of passengers getting on and off taxis in every functional area.

Then, GPR is used to predict the volume of passengers. Based on the context of Bayesian theory and statistical learning theory, GPR is a new machine learning method which can deal with the high-dimensional, small-sample and nonlinear regression problems [30] [31]. GPR can also achieve a better prediction and do not need too many complex operations at the same time. Particulary, we utilize the nonlinear processing to handle our discrete nonlinear data [32], and equation (9) illustrates the prediction process.

$$y = f(x) + \varepsilon, \varepsilon \sim N(0, \sigma^2) \tag{9}$$

In equation (9), (x, y) is a predictor-response pair measured for N subjects, *f(x)* is a regression function, and ε is a measurement error. In the experiment, we firstly acquire the number of passengers' getting off taxis in different regions which are represented by the column vector in matrix S'_j, and then input the data into the equation (9). Finally, we can predict passengers pick-up volume distributions with suitable factors for a day.

D. Top-N Areas Recommendation

Taxi service strategies have a direct impact on drivers' revenues, fuel consumption and carbon emission. According to the analysis of GPS traces data, we infer that service behaviors are very different from one driver to another. After dropping off passengers, some drivers may have no choice but to hunt the next passengers along the original route, whereas others may prefer to find new passengers in some popular regions. With respect to waiting time and cruising distance, these empirical strategies are sometimes inefficient and lead to lower profit margin for taxi drivers.

To address the above mentioned issue, our proposed TLR model recommends some suitable areas with more potential passengers based on get-on passenger volume, mean trip distance, and average trip time. In equation (10), N_{R_j} denotes the recommended value of Region j. $DI_{t_i}^{R_{i,j}}$ represents the mean trip distance from Region i to Region j during the period t_i, whereas $T_{t_i}^{R_{i,j}}$ denotes the average trip time accordingly. Our proposed model TLR gives taxi drivers an effective and efficient operation strategy based on the value of N_{R_j} in its adjacent regions.

$$N_{R_j} = \frac{P_{t_i}^{R_j}}{DI_{t_i}^{R_{i,j}} * T_{t_i}^{R_{i,j}}}, i, j = 1, 2, \ldots, 36. \tag{10}$$

In order to partition the functional regions, we do not employ the method which is used in [27] and mainly have the following two reasons. Firstly, it is difficult to choose arterial roads such as highways and ring roads to partition the map. Secondly, regions' segmentation size also play a role in taxi services, which means that drivers may take too much time to search for passengers from a subregion to another one within a large functional region. According to the above illustration, we decide to divide the social functional regions into equal smaller ones.

Moreover, the partition strategy employed by [29] is to divide an area into small squares whose edge length is 500 meters. But it is so rough that we do not make a more accurate prediction. As shown in Fig. 3, we partition the region with an interval of 0.015 degrees in both latitude and longitude. The region consists of 36 subregions named with numbers from 1 to 36. In the empirical study, taxi drivers can easily acquire an accurate spatio-temporal distribution of passengers volume. The fine-grained partition also helps drivers to drive quickly to some certain regions around them with more passengers and hunt customers in a short period.

As shown in Fig. 4, we enlarge 9 subregions and acquire the distribution of passengers' getting on. We can see the blue areas which denote real positions of pick-up passengers. The darker the colour, the more the get-on passengers. For example, if taxi drivers drop off passengers at Region 15, it is an important issue how to select the destination for the taxi. Waiting in this region is not obviously a smart choice. According to Equation (10), the N_{R_j} in Region 8 and Region 20 are much higher than the other regions, which means that taxi drivers will achieve high profits if they drive to the two regions. To improve the chances of picking up more passengers

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2017.2684163, IEEE Transactions on Industrial Informatics

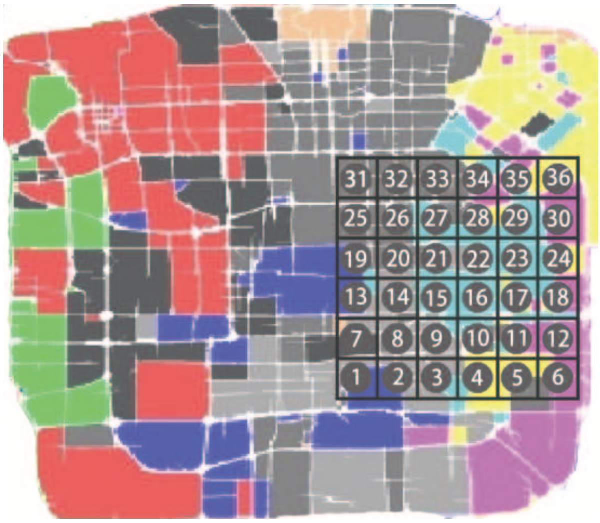


Fig. 3. Partition of the urban region.

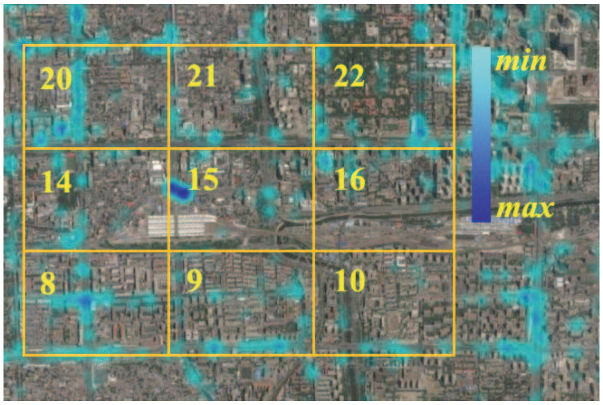


Fig. 4. Distribution of pick-up positions.

with lower energy consumption, the service recommendation provides Top-N areas to the vacant taxi drivers.

IV. Empirical Study

A. Dataset Description and Process

The GPS traces dataset is generated from 12,000 taxis running in Beijing [33]. Taxis equipped with sensor devices periodically transfer GPS data to a central server every day during November 2012, including time, latitude, longitude, and service status. Table I lists the format of the dataset.

TABLE I
Format of dataset.

Name	Annotation
Car ID	the taxi ID
Trigger event	0:getting off; 1:getting on
Running status	0:vacant; 1:carrying; 2, 3, 4: non-service
GPS time	yyyymmddhhmmss
GPS longitude	ddd.ddddddd
GPS latitude	ddd.ddddddd
GPS speed	ddd (000-255)
GPS location	ddd (000-360)
GPS status	0:invalid; 1:valid

TABLE II
Dataset description.

Dataset #	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Latitude	[39.875°N, 39.965°N]	[39.875°N, 39.965°N]	[39.875°N, 39.965°N]	[39.875°N, 39.965°N]
Longitude	[116.390°E, 116.480°E]	[116.390°E, 116.480°E]	[116.390°E, 116.480°E]	[116.390°E, 116.480°E]
Time	weekdays	weekdays	weekends	weekends
Get-on/off	on	off	on	off

Especially, we mainly analyze the GPS trajectory data in Dongcheng District in Beijing. Dongcheng District covers the eastern half of Beijing's urban core and an area of about 40.6 km^2. In addition, Dongcheng District has been as one of the city's four core zones and has developed partnerships with 45 cities in China, which has not only rich political, economical, cultural, educational, scientific and technological resources but also a high urban modernization and better infrastructure. Based on the factors, we focus on the district to do the experiments.

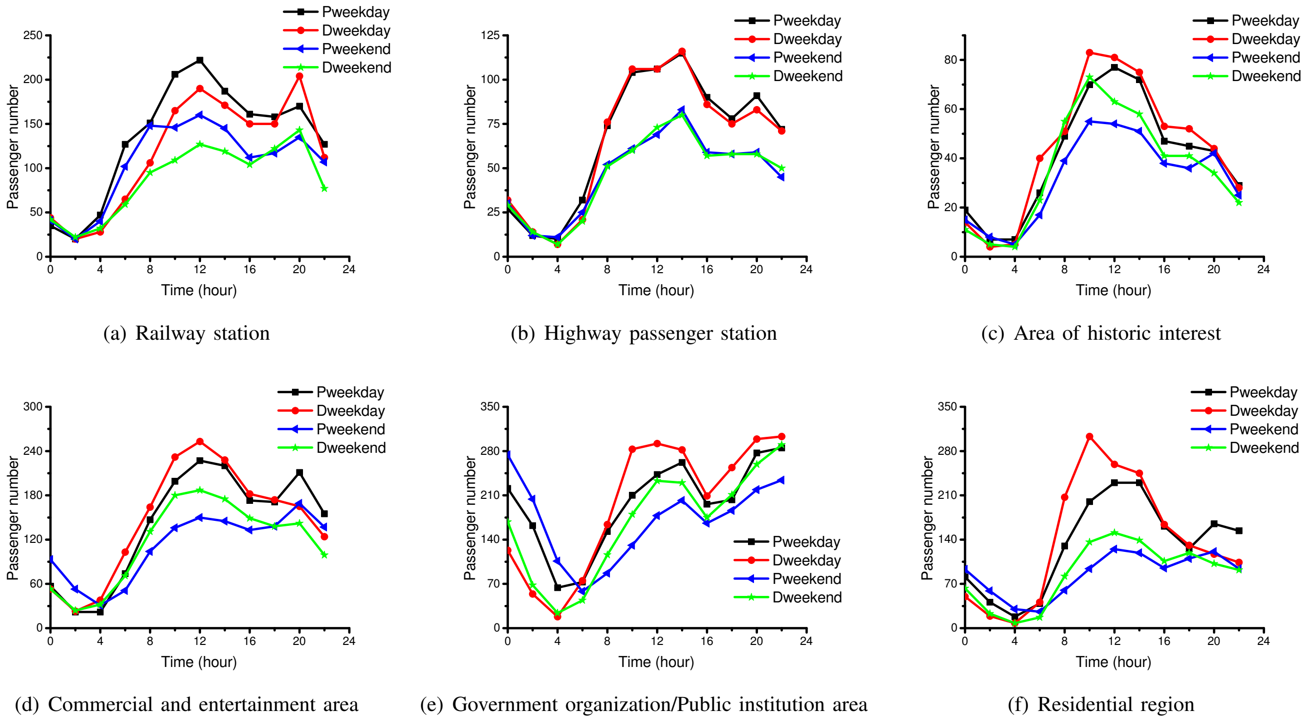
The raw dataset collected by GPS devices contain hundreds of millions of records. To improve prediction accuracy, we clean repeated data and invalid data which the status value is 0 as shown in Table I. In addition, we remove the logs with the status value of 2, 3 and 4, which represent the non-service taxis. The percentage of these error logs is about 0.85%. Subsequently, through utilizing a statistical analysis method, we extract pick-up and drop-off records in the Dongcheng District, and then acquire one initial dataset. Besides, we divide the initial dataset into two parts in terms of time periods such as weekdays and weekends, which contribute to a better prediction result with different human mobility. Then four datasets are formulated in Table II for the following evaluation.

However, even if the datasets contain a great deal of taxi information, we still have some difficulties in mining the trace of each taxi within a certain time interval. Therefore, we further extract the taxi service information with the same ID and save this data into one file. Finally, we categorize and name acquired files by taxi ID and time. In this way, we can sense each taxi's running status information at a certain time point such as current location and passengers' get-on/off positions.

B. Exploring the Get-on/off Relationship

As shown in Fig. 5, we show the quantitative relationship between passengers' getting on and off taxis in six regions on weekdays and weekends. Pweekday denotes the volume of get-on passengers on weekdays, while Dweekend denotes the volume of get-off passengers on weekends.

At the railway station, we discover that the passenger volume of taking a taxi on the weekends is lower than weekdays. As shown in Fig. 5(b), the peak value appears at 13



(a) Railway station

(b) Highway passenger station

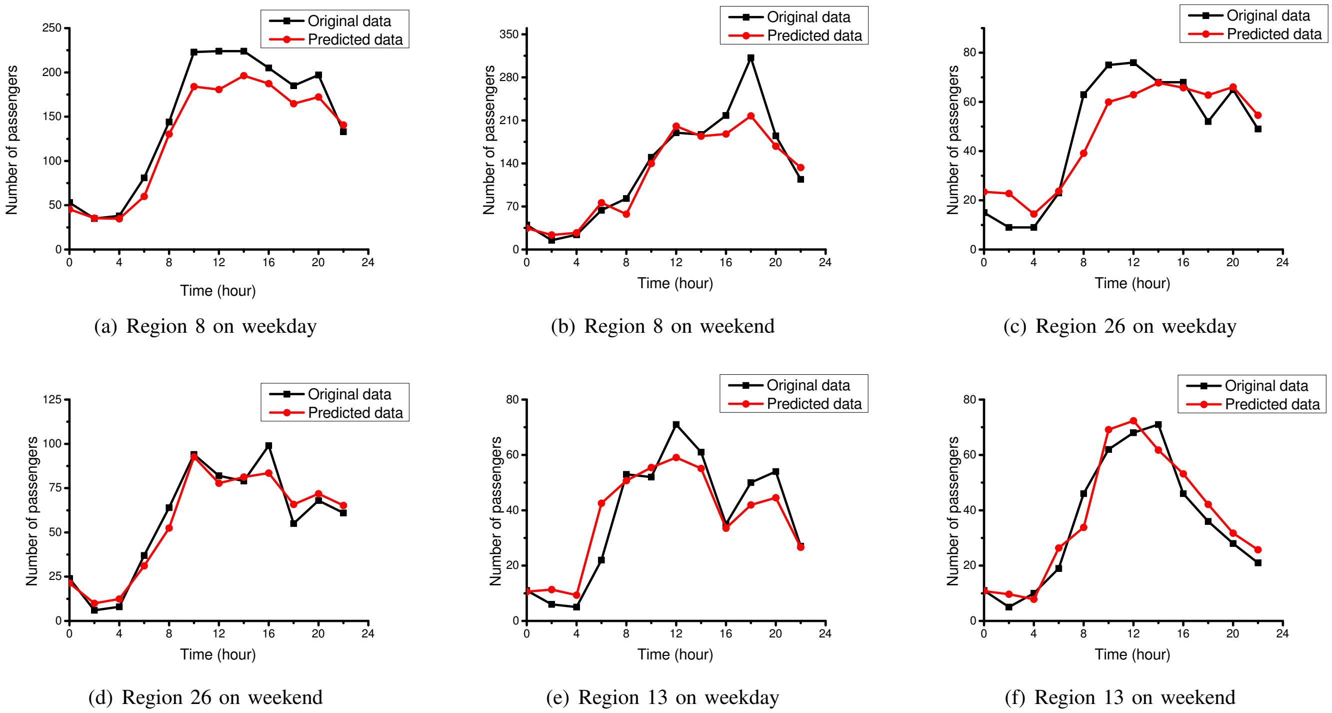
(c) Area of historic interest

(d) Commercial and entertainment area

(e) Government organization/Public institution area

(f) Residential region

Fig. 5. Distribution of pick-up and drop-off positions on weekday and weekend.



(a) Region 8 on weekday

(b) Region 8 on weekend

(c) Region 26 on weekday

(d) Region 26 on weekend

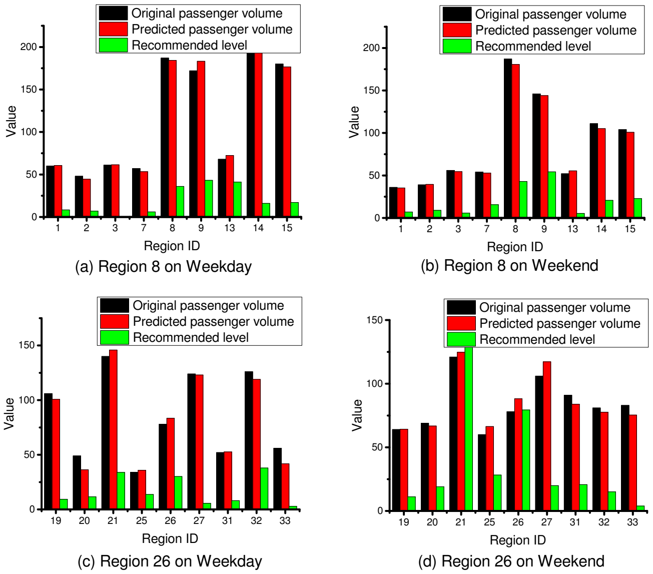
(e) Region 13 on weekday

(f) Region 13 on weekend

Fig. 6. Comparison between original data and predicted data on weekday and weekend.

o'clock which is useful for the taxi's service recommendation. In addition, as shown in Fig. 5(c), citizens usually go to visit areas of historical interest between 10 o'clock and 13 o'clock. So in areas of historic interest, taxi drivers will find some passengers and have the maximum profits around 13 o'clock.

In contrary, in emerging commercial entertainment areas which are convenient for leisure and shopping, we discover that the get-on/off peak is at 12 o'clock and the get-off amount is mostly higher than that of get-on. Moreover, as shown in Fig. 5(e), the get-off amount is much higher than the get-on amount from 6 o'clock to 18 o'clock, which is attributed to the regions' characteristics. According to Fig. 5(f), we infer that the get-on amount is higher than the get-off amount from 4 o'clock to 12 o'clock, which arises from people's going out



(a) Region 8 on Weekday

(b) Region 8 on Weekend

(c) Region 26 on Weekday

(d) Region 26 on Weekend

Fig. 7. Distribution of passengers and recommended value in near 9 regions.

for working.

Through the above-mentioned analysis, we infer that the evolution law of passengers' getting on and off is obviously different in some functional regions. With the respect of relationship between getting on and off, we easily sense human mobility patterns of every social region, which makes TLR model predict more accurately in various areas.

C. Predicting Passenger Volume

Based on our acquired taxi datasets, we carry out a series of experiments. At first, with respect to prediction accuracy, we partition the specific region into 6×6 small square ones which are named by a sequential number from 1 to 36. Then our TLR model is used to predict the amount of pick-up passengers in every subregion. Furthermore, we utilize different training set to predict the last day passenger number on weekday and weekend respectively. The former utilizes the first 19 weekdays data as a training set to predict the final weekday passenger volume in Dataset 1 and 2, whereas the latter leverages the first 7 weekends data as a training set to predict the last weekend in Dataset 3 and 4. To enhance prediction accuracy, this data is normalized into the interval of (0, 1), which contributes to better results with avoiding the influence by data fluctuation.

Through the empirical analysis, we show the results of some functional regions such as commercial entertainment areas (Region 8), residential areas (Region 26) and areas of historic interest (Region 13). As shown in Fig. 6, we infer that the predicted results are mostly consistent with the real data. According to the comparison analysis on the above mentioned 2 areas, we extract some important information as follows: a) There are more passengers taking a taxi on weekdays in Region 8 (see Fig. 6(a)) than Region 26 and Region 13 (see Fig. 6(c) and 6(e)), which makes more people go to the area for working and handling affairs than other areas. b) As shown in Fig. 6(a) and Fig. 6(b), the passenger volume on weekday is more than on weekend in Region 8. However, the passenger number in Region 13 has not too much fluctuation from weekdays (see Fig. 6(e)) to weekends (see Fig. 6(f)). c) On weekend morning, people prefer to stay at home enjoying themselves, whereas on weekend evening, citizen prefer to leave their own home for relaxation by taxi based on Fig. 6(c) and Fig. 6(d). d) As an area of historic interest, Region 13 represents its own regional characteristics. According to Fig. 6(e), the passenger volume increases firstly, then reduces gradually after reaching a peak value at 12 o'clock, finally falls into the trough at 20 o'clock. But in Fig. 6(f), the passenger volume declines gradually which implicates that people prefer to visit the region at weekends.

Through leveraging the proposed TLR model, the average prediction accuracy is 90.9% on weekdays and 80.4% on weekends respectively, which illustrates that the model is efficient and feasible to predict the passenger volume. Besides, the study also demonstrates that people have a more regular life on weekdays than weekends. On weekdays, people usually have the similar travel time and movement path, whereas on weekends, people often have more freedom to spend their leisure time. With the help of the relationship between getting on and off, it is more practicable to predict the accurate pick-up number on weekdays.

D. Analyzing Trip Distance and Trip Time

Except for passenger volume, we also introduce other two important factors: trip distance and trip time, which are calculated based on all taxis' historical trajectories. Trip distance is the inferior arc length by using great-circle from an origin to a destination on the surface of a sphere. Trip time is the elapsed time between an origin and a destination. According to our above-mentioned partition strategy, we acquire the average trip distance and trip time between any two adjacent regions.

To our knowledge, if a region has more pick-up passengers with a relatively long trip distance and time, it means more energy consumption and profit decline for taxi drivers. Furthermore, the volume of get-on passengers constantly changes. When taxi drivers spend some time to go to the recommended place, passengers are likely to leave for another place. Therefore, we consider the two impact parameters to provide a better service recommendation.

E. Recommending Top-N Areas

Through utilizing the proposed TLR model, we acquire the accurate volume of passengers, historical mean trip distance, and historical average trip time, and then we provide the the most potential regions to drivers who drop off customers for cruising. As shown in Fig. 7(a) and Fig. 7(b), we illustrate the passenger volume and recommended value of 9 regions around Region 8 by the histogram, when is at 14 o'clock on the weekday and at 12 o'clock on the weekend respectively. In addition, Fig. 7(c) and 7(d) represent the passenger volume and recommended value of 9 regions around Region 26. According to the four sub-figures, we also infer that TLR model has high accuracy and stability. For example, as shown in Fig. 7(c), if a taxi is hunting in Region 26, we can advise the driver to drive to Region 21 which is more suitable than any others. So it finally makes drivers achieve more profit, improves service efficiency and decreases fuel wasting.

Additionally, our proposed TLR model is only based on the local and adjacent regions. In other words, recommending results are just local optimal solution rather than global optimal method. So it does not result in being unable to take a taxi in some areas.

V. Performance Evaluation

Based on the acquired analysis data, we compare the performance of our proposed model, ARIMA model, BPNN model, SVM model, and GBDT model. The ARIMA model predicts future values of passengers' getting on by a linear combination of its past values and the time series; Whereas the BPNN model estimates future volume of pick-up passengers using back propagation algorithm; The SVM model mainly leverages kernel function to map the training samples into a high dimensional space for prediction; The GBDT model introduces gradient boosting method to improve the prediction accuracy. We introduce the following 4 metrics such as CC, RMSE, MAE, and NMAE to evaluate our proposed model.

- Correlation Coefficient (CC). It is used to determine the relationship between two properties. In this paper, we utilize the metric to evaluate whether or not our proposed model has the strongest correlation with the original data among the three methods,

$$r_{xy} = \frac{\sum\limits_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i=1}^{N}(y_i - \overline{y})^2}} \quad (11)$$

  where x_i and y_i denote predicted and original ratings, whereas $\overline{x}$ and $\overline{y}$ are an average value of prediction and real value.
- Root Mean Square Error (RMSE). It can measure the magnitude of the error between predicted and exact values, demonstrate the model precision accuracy, and can be formulated as:

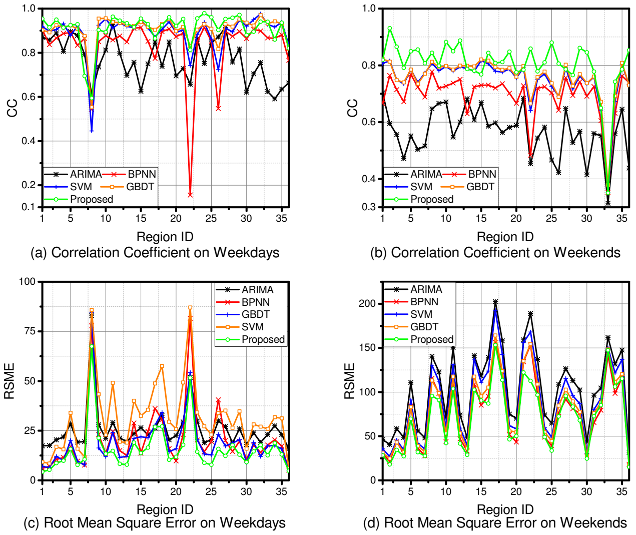
$$RMSE = \sqrt{\frac{\sum\limits_{i=1}^{n}|f_i - y_i|^2}{n}} \quad (12)$$

  where the predicted value and the real one is denoted by f_i and y_i respectively, and the number of measurements is defined as n.
- Mean Absolute Error (MAE). MAE is also widely used to measure how close predictions are to the real values. The smaller the value for MAE, the better the algorithm in performance. We define MAE as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (13)$$

- Normalized Mean Absolute Error (NMAE). NMAE is independent of the sample size and is widely used to compare the performance in several algorithms using different rating scale. It normalizes MAE measure and is defined as follows:

$$NMAE = \frac{MAE}{\sum\limits_{i=1}^{n} y_i} \quad (14)$$



(a) Correlation Coefficient on Weekdays

(b) Correlation Coefficient on Weekends

(c) Root Mean Square Error on Weekdays

(d) Root Mean Square Error on Weekends

Fig. 8. CC and RSME of 5 algorithms on weekdays and weekends.

TABLE III
Statistical results on Correlation Coefficient.

Dataset#	Time	ARIMA	BPNN	SVM	GBDT	Proposed
Dataset1&2	weekdays	0.7680	0.8332	0.8910	0.9065	0.9087
Dataset3&4	weekends	0.5635	0.6958	0.7498	0.7575	0.8041

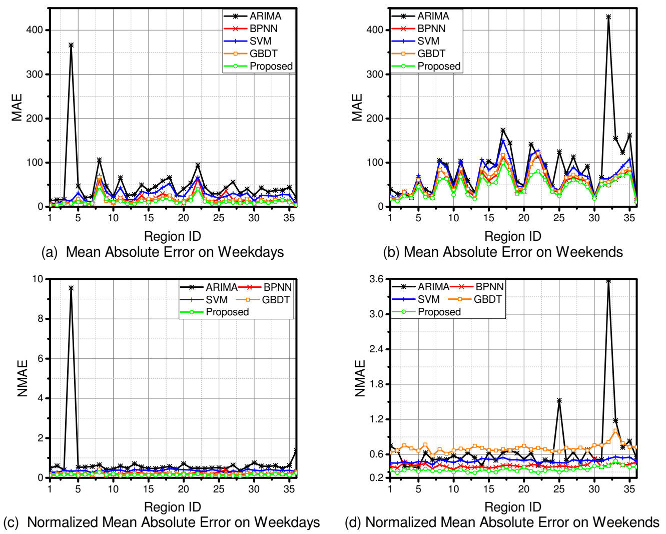
According to the horizontal comparison, the performance of five models on weekdays is a little better than that on weekends as shown in Fig. 8(a) and Fig. 8(b), which contributes to citizens' regular mobility patterns from Monday to Friday. Moreover, through our detailed quantitative analysis shown in Table III, CC mean value in GBDT model are 90.7% on weekdays and 75.8% on weekends accordingly. But for TLR model, it is higher than the other five models with 90.9% on

TABLE IV
Statistical results on Root Mean Square Error.

Dataset#	Time	ARIMA	BPNN	SVM	GBDT	Proposed
Dataset1&2	weekdays	25.683	20.916	31.630	19.335	15.881
Dataset3&4	weekends	103.16	76.013	90.067	81.015	71.849

TABLE V
Statistical results on Mean Absolute Error.

Dataset#	Time	ARIMA	BPNN	SVM	GBDT	Proposed
Dataset1&2	weekdays	49.643	16.871	27.313	15.214	11.775
Dataset3&4	weekends	90.107	54.854	67.493	58.297	46.240



(a) Mean Absolute Error on Weekdays

(b) Mean Absolute Error on Weekends

(c) Normalized Mean Absolute Error on Weekdays

(d) Normalized Mean Absolute Error on Weekends

Fig. 9. MAE and NMAE of 5 algorithms on weekdays and weekends.

TABLE VI
STATISTICAL RESULTS ON NORMALIZED MEAN ABSOLUTE ERROR.

Dataset#	Time	ARIMA	BPNN	SVM	GBDT	Proposed
Dataset1&2	weekdays	0.8128	0.2189	0.3616	0.2011	0.1576
Dataset3&4	weekends	0.6867	0.408	0.4921	0.7083	0.3435

weekdays and 80.4% on weekends respectively, which implies that our proposed model shows a strong correlation and works very well.

As shown in Fig. 8(c) and Fig. 8(d), the RMSE of our proposed model is the lowest compared with that of other prediction methods both on weekdays and weekends. As shown in Table IV, the RSME of BPNN model are 20.9 on weekdays and 76.0 on weekends, whereas they are 19.3 on weekdays and 81.0 on weekends respectively in GBDT model. However, the RSME of TLR model are 15.9% on weekdays and 71.8% on weekends correspondingly, which demonstrates that our proposed model has a good stability based on the analysis of the quantity relationship of get-on/off passengers.

As shown in Fig. 9(a) and Fig. 9(b), the MAE of our proposed model is the lowest among five prediction methods, which proves the high prediction accuracy and the effectiveness once again. Specifically, the MAE of TLR model are 11.775 on weekdays and 46.240 on weekends respectively as shown in Table V. In addition, our statistical results show that the performance of TLR model on weekdays is better than on weekends, which may be related to the diversity of urban human activities on weekends.

As shown in Fig. 9(c) and Fig. 9(d), the NMAE value generally keeps a small fluctuation except for ARIMA model, which contributes to normalized processing. Obviously, the error rate of TLR is the lowest with 15.8% on weekdays and 34.4% on weekends referring to Table VI. GBDT model ranks the second after our TLR model on weekdays, whereas BPNN is the second with the error rate of 40.8% on weekends.

In a nutshell, by applying the proposed TLR model, we find that our prediction results are more accurate than that by the other four models in all these 36 regions, which demonstrates that 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 with the Correlation Coefficient mean value of 90.9% on weekdays and 80.4% on weekends respectively. For example, students go to school and employees go to work at fixed times from Monday to Friday. However, people have more options on Saturday and Sunday because of ample time. Due to the utilization of the quantitative relationship between passengers of getting on and off, it is easier for our model to predict an accurate number of passengers on weekdays and weekends. And then 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.

VI. CONCLUSION

In this paper, we have proposed a taxi service recommendation model named TLR by analyzing the quantitative relationship between passengers' getting on and off taxis in different functional regions during each period. With the help of TLR model, we acquire some important human mobility patterns, predict the spatio-temporal distribution of passengers for different social functional regions efficiently, and calculate the mean trip distance and average trip time between any two functional regions. Based on the prediction outcomes of the model, we recommend Top-N profitable areas near the driver's real-time position, which leads to improve taxi drivers' profits and passengers' travel experience. Furthermore, we have conducted extensive simulations on TLR model and compared its performance against ARIMA model, BPNN model, SVM model, and GBDT model. The results have shown that TLR outperforms other four methods with higher prediction accuracies of 90.9% on weekdays and 80.4% on weekends respectively. Additionally, TLR model has the lowest prediction error rate with the NAME of 15.8% on weekdays and 34.4% on weekends. Finally, we measure the following metrics including CC, RMSE, MAE, and NMAE, which demonstrates the validity and stability of the proposed taxi service recommendation. At present, we continue to consider different social properties in one area. Besides, we have collaborated with Panda Travel co.Ltd to apply our proposed model into its bus service platforms.

In the future work, we will consider different social properties and multi-source datasets to improve our prediction accuracy. We also plan to quantitatively evaluate our proposed model using bus drivers' income data. Besides, we will focus on supply-demand matching and recommendation between passengers and taxis, which makes passengers find vacant taxis in less time.

REFERENCES

[1] 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.

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

[3] 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*, vol. 26, no. 11, pp. 3122–3135, Nov. 2015.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] F. Xia, A. Ahmed, L. Yang, and Z. Luo, "Community-based event dissemination with optimal load balancing," *IEEE Transactions on Computers*, vol. 64, no. 7, pp. 1857–1869, Jul. 2015.

[10] Q. Yang, Z. Gao, X. Kong, A. Rahim, J. Wang, and F. Xia, "Taxi operation optimization based on big traffic data," in *The 12th IEEE International Conference on Ubiquitous Intelligence and Computing (UIC 2015)*, Beijing, China, in press.

[11] X. Kong, Z. Xu, G. Shen, J. Wang, Q. Yang, and B. Zhang, "Urban traffic congestion estimation and prediction based on floating car trajectory data," *Future Generation Computer Systems*, vol. 61, pp. 97–107, 2016.

[12] 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.

[13] 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 (CSUR)*, vol. 46, no. 2, pp. 1–34, Dec. 2013.

[14] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[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] J. D. Mazimpaka and S. Timpf, "Exploring the potential of combining taxi gps and flickr data for discovering functional regions," in *AGILE 2015*. Switzerland: Springer International Publishing, 2015, pp. 3–18.

[18] J. Tang, F. Liu, Y. Wang, and H. Wang, "Uncovering urban human mobility from large scale taxi gps data," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 140–153, Nov. 2015.

[19] X. Liu, L. Gong, Y. Gong, and Y. Liu, "Revealing travel patterns and city structure with taxi trip data," *Journal of Transport Geography*, vol. 43, pp. 78–90, Feb. 2015.

[20] 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.

[21] 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.

[22] 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.

[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] T. Qin, X. Guan, W. Li, and P. Wang, "Monitoring abnormal traffic flows based on independent component analysis," in *Communications, 2009. ICC'09. IEEE International Conference on*, Dresden, Jun. 2009, pp. 1–5.

[25] B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, and V. D. Blondel, "Exploring the mobility of mobile phone users," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 6, pp. 1459–1473, 2013.

[26] P. Ahmadi, R. Kaviani, I. Gholampour, and M. Tabandeh, "Modeling traffic motion patterns via non-negative matrix factorization," in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuala Lumpur, Oct. 2015, pp. 214–219.

[27] 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.

[28] Y. Zhou, Z. Fang, J.-C. Thill, Q. Li, and Y. Li, "Functionally critical locations in an urban transportation network: Identification and space-time analysis using taxi trajectories," *Computers, Environment and Urban Systems*, vol. 52, pp. 34–47, Jul. 2015.

[29] 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.

[30] M. Belyaev, E. Burnaev, and Y. Kapushev, "Gaussian process regression for structured data sets," in *Statistical Learning and Data Sciences*, ser. Lecture Notes in Computer Science. Switzerland: Springer International Publishing, 2015, vol. 9047, pp. 106–115.

[31] J. R. Gattiker, M. S. Hamada, D. M. Higdon, M. Schonlau, and W. J. Welch, "Using a gaussian process as a nonparametric regression model," *Quality and Reliability Engineering International*, pp. 80–87, Feb. 2015.

[32] B. Wang and T. Chen, "Gaussian process regression with multiple response variables," *Chemometrics and Intelligent Laboratory Systems*, vol. 142, pp. 159–165, Mar. 2015.

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