World Wide Web (2020) 23:1381–1405
https://doi.org/10.1007/s11280-019-00700-1

TBI2Flow: Travel behavioral inertia based long-term taxi passenger flow prediction

Xiangjie Kong[1] · Feng Xia[1] · Zhenhuan Fu[1] · Xiaoran Yan[2] · Amr Tolba[3,4] · Zafer Almakhadmeh[3]

Received: 13 December 2018 / Revised: 15 April 2019 / Accepted: 23 May 2019 /
Published online: 17 June 2019

Abstract
Taxis are one of the representative modes of traffic systems. However, with the emergence of shared cars led by DiDi and Uber in recent years, the traditional taxi companies are facing unprecedented competitions. Without personalized data collected from the mobile devices, passenger flow prediction based on vehicle GPS records presents a unique solution that can improve taxis' operating efficiency while preserving personal privacy. In this paper, we propose the Travel Behavioral Inertia (TBI) from taxi GPS records, which embodies Driver Inertia (DI) and Passenger Inertia (PI). Then we integrate TBI with other features to construct multi-dimensional features and predict taxi passenger flow based on a deep learning algorithm. We call the entire framework TBI2Flow. Extensive experiments demonstrate that TBI features has outstanding contribution to passenger flow prediction and TBI2Flow outperforms state-of-the-art methods including time series-based method and other deep learning-based methods on long-term taxi passenger flow prediction.

Keywords Travel behavior · Traffic flow prediction · Smart city · Taxi operation

This article belongs to the Topical Collection: *Special Issue on Smart Computing and Cyber Technology for Cyberization*
Guest Editors: Xiaokang Zhou, Flavia C. Delicato, Kevin Wang, and Runhe Huang

✉ Feng Xia
f.xia@ieee.org

[1] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian 116620, China

[2] Indiana University Network Science Institute, Indiana University Bloomington, Bloomington, IN 47408, USA

[3] Computer Science Department, Community College, King Saud University, Riyadh 11437, Saudi Arabia

[4] Faculty of Science, Mathematics and Computer Science Department, Menoufia University, Shebin-El-kom 32511, Egypt

1 Introduction

The prosperity of big data and mobile computing has brought many new opportunities for the development of scientific research, such as user network analysis [44], academic influence aware [45], and traffic prediction [4]. Taxis are ubiquitous in urban residents' life and participate in urban perception to reflect dynamic changes in traffic status. In New York City, there are nearly 13000 taxis in service every day. They make, on average, 500,000 trips each day, totaling over 170 million trips a year [7]. Recently, the Internet of Things and cognitive cyber-physical systems are developing rapidly [21, 46]. Because of their wide distribution and high mobility, taxis have great application value in the construction of smart city. For example, taxis can be used as the carrier of fog calculation as well as the mobile crowdsensing. Besides, taxis are valuable sensors of city life. Information hidden in taxi trips bring a great opportunity to understand city life in depth [35, 36]. These are likely to pave the way for big data-driven analysis of taxi operation status.

With the rapid development of global economy in recent 10 years, the number of urban vehicles has grown tremendously. It leads to increasingly severe traffic problems, including traffic congestion [15], environmental pollution, and frequent traffic accidents. For many of these problems, taxis are major contributors due to their operational characteristics. One of the most typical aspects is the unbalanced relationship between the passenger travel demand and the empty carrying phenomenon for taxi drivers. Take the taxi operation status of Beijing for example. We get from Beijing traffic development and construction report that each taxi travels about 400 kilometers everyday and the average empty carrying ratio of taxis is around 40% [16]. It indicates that taxis in cities operate with low efficiency, which leads to wasteful resources consumption and environmental pollution. On the other hand, it also increases the operating costs of the companies, and pulls down the level of passenger satisfaction.

Besides traditional taxi services, shared transportation based on the mobile/internet platform are revolutionizing people's travel choices, such as Mobike [25], DiDi [42], and Uber [8]. Among them, DiDi has been rapidly developed in Chinese market in recent years and provides convenient and quick travel service. As a result, it reduces the market of taxis further and has a powerful impact on the operation of taxis. In the age of internet, traditional taxi companies have fallen behind in competition due to outdated technology and lack of personalized data. Previously all taxi services in China are provided by taxi companies. The emergence of DiDi broke the monopoly of the industry. Everyone can provide online car-hailing as long as he or she has a car and a driver's license. Equipped with data collected directly from the mobile devices, DiDi is able to deliver better personalized services and gain market share [29]. However, frequent driver crimes recently have led people to question the safety of online car-hailing services [40]. Therefore, how to improve taxis' operating status has become an urgent problem. The key lies in how to resolve the unbalanced relationship between the passenger travel demand and the empty carrying phenomenon for taxi drivers, that is, how to inform taxi drivers where passengers locate, without personalized location data from the mobile devices. Taxi passenger flow prediction presents a good solution to this challenge.

Traffic flow prediction in general is of great significance in lots of applications, such as urban anomaly detection and city road network planning, and has sparked high attention of many scholars as an important means to solve urban traffic problems. However, due to the complexity of taxi trajectory data, which contains pick up and drop off locations and times,

taxi id, distance traveled, fare, tip amount, and so on, there are relatively few studies on taxi passenger flow prediction. Predicting taxi passenger flow combining the characteristics of current taxi operating status is even more rare. As mentioned above, taxis' operating status undergoes major changes recently.

According to current taxis' operating status, we first analyze taxi travel behavior based on massive taxi trajectory data. Then we refine inert travel behavioral from two types of subjects involved in taking taxi events, taxi drivers and taxi passengers. Utilizing a deep learning algorithm, we present a travel behavioral inertia-based approach TBI2Flow for long-term taxi passenger flow prediction.

The major contributions of this work can be summarized as follows:

- We propose the definition of Travel Behavioral Inertia (TBI) which contains Driver Inertia (DI) and Passenger Inertia (PI) and verify its effectiveness in long-term taxi passenger flow prediction.
- We present TBI2Flow that includes data modeling, inertia extracting, and flow prediction and uses multiple features (TBI, location, time, weather, and flow) to predict taxi passengers' distribution over different time intervals.
- We design a simulation experiment using a real taxi dataset. The results show that TBI2Flow has great value for taxi operation, as well as urban traffic planning in general.

We also conduct extensive comparative experiments from three aspects, which are with or without TBI feature, regression algorithms selection, and prediction approaches comparing, to demonstrate the effectiveness of TBI2Flow.

The remainder of this paper is structured as follows. In Section 2, we review the related work according to different traffic flow prediction methods categories. Section 3 mainly illustrates our proposed approach TBI2Flow in detail. Data description and experiment results are displayed in Section 4. Following that, we verify the effectiveness of TBI2Flow from three aspects in Section 5. Finally, we give the conclusions of our work and offer further discussion on open issues in Section 6.

2 Related work

Traffic flow prediction plays an essential role in tackling urban traffic problems and is valuable in a large quantity of applications, which embody smart driving [38], modeling spatiotemporal structure of taxi services [5, 20, 39], building passenger-finding strategies [18, 19] and other applications [16, 17, 23]. Prediction methods can be classified into short-term prediction and long-term prediction[26], parametric approaches and non-parametric approaches [22]from different perspectives.

2.1 Short-term prediction and long-term prediction

According to the time span of predictive objective, traffic flow prediction is roughly divided into short-term prediction and long-term prediction. The predictive period of short-term prediction is usually less than 30 minutes, and long-term prediction can span a day, a week, or even longer [26].

Faced with the fast changing traffic flow data, most of the previous literature has focused on short-term prediction. Zhang et al. [41] propose a deep learning-based prediction model for spatial-temporal data. Yang et al. [37] put forward a novel model, stacked

auto-encoder Levenberg-Marquardt model, which is a type of deep architecture of neural network approach aiming to improve forecasting accuracy. Zhao et al. [43] adopt the Gaussian Process Dynamical Model (GPDM) to a fourth-order GPDM which is rather suitable for traffic flow prediction.

The complexity and variety of traffic conditions cause great impediments to the research of long-term prediction. What's more, long-term prediction has crucial practical significance because short-time traffic flow prediction will fail under some conditions especially network delay. Unfortunately, there are few related research on long-term prediction. Most current long-term prediction methods are based on traffic flow time series analysis, which mines flow laws from historical data with the support of large-scale data. Considering traffic flow as a continuous time stochastic process, Fei et al. [26] propose an non-parametric regression-based long-term prediction model. They employ auto-correlation analysis to analyze the traffic flow and select the state vector. Besides, they further use the functional principal component analysis as distance function to compute proximity between different time series. Zhong et al. [13] propose the concepts of traffic flow pattern similarity and repeatability. They split traffic flow series into basis series and deviation series. Besides, He et al. [11] employ social media features for long-term traffic prediction. They find that social activity and traffic activity have a strong correlation. All the above methods, whatever time series-based or social media-based, only focus on extracting information contained in the statistical data while ignoring the behavioral characteristics of traffic participants. The main innovation of our work is to include taxis and passengers behavior into the prediction model and experimental results show that our method can greatly improve prediction accuracy.

2.2 Parametric techniques and non-parametric techniques

From the perspective of predictive techniques, traffic flow prediction can be classified into parametric techniques and non-parametric techniques. Parametric approaches are also known as model-based methods. Model parameters are often computed with empirical data and model structure is predetermined based on certain theoretical assumptions [22]. Common parametric techniques include AutoRegressive Integrated Moving Average model (ARIMA), exponential smoothing [32], and historical average [12]. Particularly, ARIMA has become one of the most common parametric prediction approaches since the 1970s. It is a classical time series prediction method proposed by Box and Jenkins [3]. ARIMA treats the data sequence formed by the predicted object over time as a random sequence, and uses a mathematical model to approximate the sequence. ARIMA includes moving average process (MA), autoregressive process (AR), autoregressive moving average process (ARMA), and autoregressive moving average mixing process (ARIMA) depending on whether the original sequence is stationary or not. ARIMA model has been widely used in short-term traffic predictions, such as traffic flow, travel time, speed, and occupancy [2, 10]. In addition, many existing studies proposed variants of the ARIMA model, such as the seasonal ARIMA [33].

For non-parametric techniques, model structure and parameters are not fixed. With the rise of data mining and science, many techniques have been widely adopted recently, such as Support Vector Machine for Regression (SVR) [34], Kalman filtering models [31], Neural Networks (NN) [27] and K-Nearest Neighbor (KNN) [9]. Among these techniques, SVR is widely used for traffic flow prediction, but it may has a bad performance when dealing with a large numbers of variables because of the choice of the appropriate kernel function for the practical problem [30]. The KNN algorithm is capable of

non-parametric predictions without prior knowledge given the k parameter. However, KNN requires long computation time if the historical data increases [24]. NN usually can obtain good results and it is suitable for solving complex non-linear problems without prior knowledge regarding the relationships between input and output variables. The most commonly used model in artificial neural networks methods is Multi-Layered Networks (MLNN) model. It has become popular as a traffic flow prediction model with its excellent prediction performance.

Existing methods for taxi passenger flow prediction only focus on the relevance of the data but neglect the subjective factors of the individual. The time span of the data used by these methods is usually short and can't reflect the preference of taxis and passengers. Differing from the existing researches, we take into account the inertia characteristics of taxis and passengers. We will propose the concept of TBI and utilize it to conduct passenger flow prediction. Combined with other features, TBI will ultimately improve the accuracy of traffic flow prediction.

3 Design of TBI2Flow

In order to build an accurate and effective taxi passenger flow prediction framework, we perform three steps as shown in Figure 1. In step 1 *data modeling*, we apply trajectory extraction, time division, and region division to the taxi GPS records for flow prediction tasks. Step 2 *inertia excavation* is the key innovation of our framework. We will explain in detail what TBI is, which integrates its motivation, definition, and quantitative methods, to pave the way for passenger flow prediction. In step 3 *flow prediction*, we combine TBI with four other features (location, weather, time, and flow) to predict taxi passenger flow based on a deep learning algorithm.

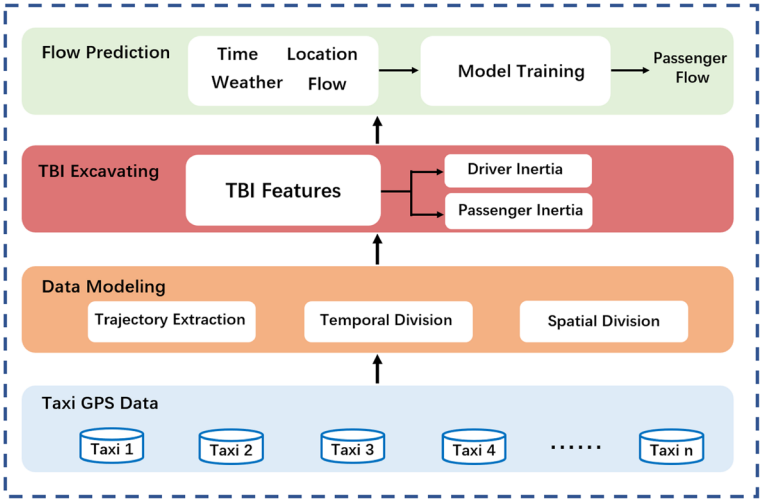


Figure 1 Framework of TBI2Flow

3.1 Data modeling

Before data modeling, we conduct general data preprocessing operations, which include data cleaning, data matching, and data organizing. Data cleaning is to remove errors and outliers and to filter out useless records of raw GPS data. Traffic data records are discrete sample points and may not match urban road networks. We utilize map matching to align a sequence of observed vehicle positions with the road network on a digital map.

3.1.1 Trajectory extraction

From the acquired taxi GPS records, we can get longitude, latitude, data, time, and status of each taxi, which indicate the taxi's geographical location and the taxi is occupied or vacant at different time points. Each taxi is equipped with an on-board GPS positioning equipment, which generates an unique ID number. The taxi carrying a GPS device sends a signal to the data center every 20 seconds. In order to achieve the goal of traffic flow prediction, we need to extract taxi travel trajectories from a huge quantity of discrete GPS records. We first aggregate GPS records according to unique taxi IDs and then arrange corresponding position information of taxis in sequence chronologically on the map. Consequently, a continuous series of points on the map represents a taxi trajectory. We distinguish occupied trajectories and vacant trajectories based on the status field of GPS records data. Such operation is indispensable for subsequent traffic flow prediction. The trajectory extraction process is displayed in Figure 2 and we use black and red to denote occupied trajectory and vacant trajectory respectively. As shown in Algorithm 1, we segment a trajectory by the change of taxi status. The pseudocode offers more details about trajectory extraction.

Algorithm 1 Taxi trajectory extraction.

input: GPS records
output: GetOn & GetOff

```
for ID in Database do
    for time i of ID, i++ do
        TimeSeries.add(i)
        sort(TimeSeries)
    end for
    T(j) ← Trajectory.add(ID)
end for
for j in T(j) do
    for time i of j, i++ do
        if state 1 turn into state 0 then
            GetOn.add(i)
        end if
        if state 0 turn into state 1 then
            GetOff.add(i)
        end if
    end for
end for
```

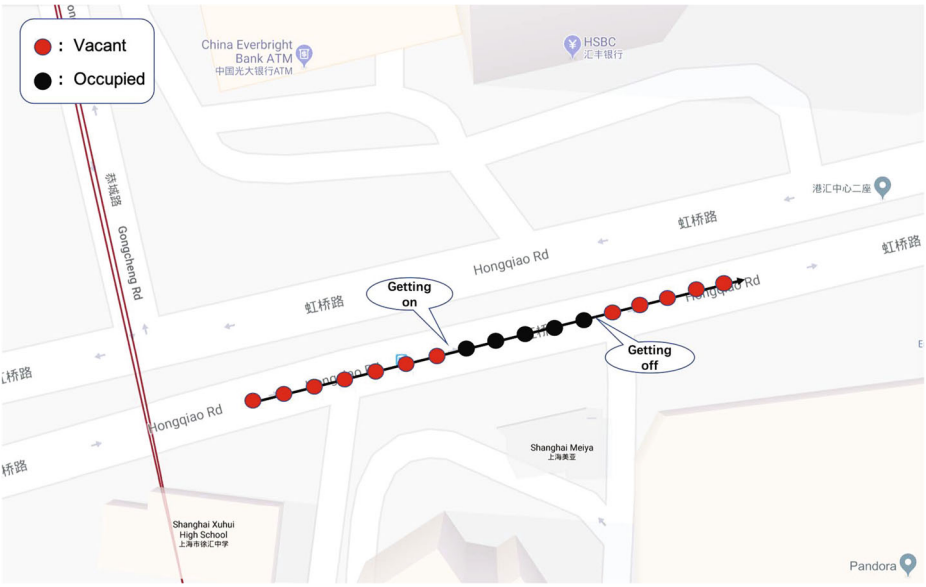


Figure 2 An example of taxi trajectory extraction

3.1.2 Temporal division

Time has a great impact on taxi passenger flow. Therefore, we need to divide the study time range into multiple time intervals. Time division is an essential operation in conventional traffic data processing. Common time division methods include regular time division, that is, to divide the study time range into time intervals with equal length, and irregular time division, which means to divide the study time range based on traffic flow distribution. Passenger flow is highly sensitive to time, and the right choice of time division methods dependents on analyzing goals. According to the objective of this work, passenger flow prediction, we choose regular time division to divide a day into regular time intervals. The granularity of time division is also critical. To keep the granularity of passenger flow prediction as fine as possible, we should make the length of a time interval as short as possible. Meanwhile, time division operation is supposed to follow the principle of keeping taxi trajectories as complete as possible. In our experiments, we choose an hour as a time interval and use (1) to do time division of taxi trajectory data,

$$Time_t = [t\theta, (t+1)\theta), t = 0, 1, \ldots, 23 \tag{1}$$

where t is time interval's number and θ is the duration of each time interval. We then analyze the periodicity of passenger flow changes. We first calculate the similarity of seven days based on passenger flow and divide them into several types. According to the results of similarity analysis, we divide a week into weekdays and weekends and model them separately.

3.1.3 Spatial division

Spatial division is to divide the whole study area into multiple small regions to facilitate fine-grained analysis. The most common method of spatial division is grid partitioning.[1] Urban traffic passenger flow prediction ought to be as fine-grained as possible in spatial division. Therefore, we first partition the whole study area into grids with a granularity of longitude and latitude of 0.1. Traffic flow trend varies over different region categories. For example, traffic flow trend of residential area is obviously different from that of workspace. In this work ,we consider a factor that is directly related to the target of region passenger flow prediction, that is, region historical passenger flow. We aggregate the results of above spatial division to multiple kinds of regions and one kind of regions contains one or more regions. In this way, we can perform a fine-grained passenger flow prediction on multiple region types.

3.2 TBI excavation

As mentioned earlier, taxi operating status undergoes considerable changes in recent years due to the impact of DiDi and Uber. In consequence, we consider combining characteristics of taxi operation to predict passenger flow. In real life, people usually has their own habits. For example, they like listening to a singer's songs, or visit several hotels frequently, or always go to a same store to buy something they like. Such kind of behavior is also reflected in travel behavior. For instance, people often take the taxi in fixed places, such as homes and companies. At the same time, taxi drivers usually pick passengers up in several fixed areas. In this paper, we try to extract these preference behavior which we called inertia features, and employ them to predict taxi passenger flow. In the composition of traffic flow, the proportion of inert travel is higher and much easier to predict when compared with random travel. Therefore, based on analysis of taxi travel behavior of the two types of participants, taxi drivers and passengers, we define Travel Behavioral Inertia as follows:

Definition Travel Behavioral Inertia (TBI) is a vector consists of Driver Inertia (DI) component and Passenger Inertia (PI) component, as shown in the following equation:

$$TBI_t = \{DI_t, PI_t\}, t = 0, 1, ..., 23 \tag{2}$$

where t denotes time interval's number.

Driver inertia DI stands for inert taxi drivers who tend to pick up passengers at certain places frequently. Drivers of taxis that meets the conditions in following equations and equation are regarded as inert drivers.

$$DI_{(u,t)} = \begin{cases} 1, Freq_{(u,t)} \geq \theta\delta - Aver_{(u,t)} \\ 0, Freq_{(u,t)} < \theta\delta - Aver_{(u,t)} \end{cases} \delta \in (0, 1] \tag{3}$$

$$Aver_{(u,t)} = \frac{\sum_{u \in U_{(t)}} Freq_{(u,t)}}{|U_{(t)}|} \tag{4}$$

[1]This method is not only simple, but also can meet the requirements of most traffic researches. Other methods embody Voronoi tessellation division based on particles and division based on urban road network framework. Researches with special spatial needs can consider the above two complex spatial division methods.

where u stands for a taxi and $U_{(t)}$ is the complete set of all taxis in tth time interval. Note that we employ unique taxi ID to distinguish each taxi. $Freq_{(u,t)}$ presents the frequency of the taxi u pick up passengers at a same place in tth time interval. θ is the duration of a time interval. δ should be adjusted according to concrete time intervals and during the adjustment process, we ought to ensure $Freq_{(u,t)} > 0$. The number of inert drivers in a region are regarded as DI_t feature, as shown in (5). For example, driver A has pick up passengers in a certain region for 5 times in tth time interval, and the threshold we set is 4, so driver A can be regarded as a inert driver. If there are K inert drivers in a region, the DI_t feature can be set to K. After given the definition of DI, we identify DI in each time interval using the pseudocode shown in Algorithm 2.

$$DI_t = \sum_{u \in U_{(t)}} DI_{(u,t)} \tag{5}$$

DI stands for inert taxi drivers who

Algorithm 2 DI identification.

input: θ, U
output: DI
1: **Calculate** $\theta\delta$, **when** $\delta = 0.25$
2: **for** t in θ **do**
3: **Calculate** $Aver_{(u,t)}$
4: **for** u in $U_{(t)}$ **do**
5: **if** u meet (3) **then**
6: $DI_{(u,t)}$.add(u)
7: **end if**
8: **end for**
9: $DI_t \leftarrow \sum_{u \in U_{(t)}} DI_{(u,t)}$
10: **end for**
11: **until** Identify DI for all tth time periods,$t = 0, 1, ..., 23$

Passenger inertia PI indicates inert passengers who tend to take taxies in a particular region, measuring the taxi demand of the region. Due to the lack of public transportation, the taxi demand in some regions is significantly higher than others. To determine the bound of the taxi demand, we analyse the historical data. The historical passenger flow is a typical time series data which fluctuates within a certain range, and has stable periodicity. Therefore, we employ the ARIMA model to predict the passenger flow of a region in a time interval. After that, we calculate PI based on the prediction results of ARIMA. PI considers three factors, the maximum, the minimum and the average of prediction results at a certain time interval which are denoted as $\max(t)$, $\min(t)$, and $aver(t)$ respectively to estimate passenger flow stability. As displayed in (6), PI is the linear combination of the above three factors and α, β, γ are the weights of $\max(t)$, $\min(t)$, and $aver(t)$ which are determined by grid-based search algorithm.

$$PI_t = [\max(t), \min(t), aver(t)]\,[\alpha, \beta, \gamma]^T \tag{6}$$

Based on the definition of TBI, we can excavate DI and PI features data from taxi trajectory for passenger flow prediction.

3.3 Passenger flow prediction

Apart from TBI features we proposed, other important features also have a great impact on passenger flow prediction and we extract them based on travel behavior analysis.

3.3.1 Time

Time plays an important role in passenger flow prediction. Taxi passenger flow is changing constantly because urban residents travel frequency is affected greatly by time. To be specific, passenger flow during the day is higher than that at night. Meanwhile, passenger flow on weekdays is more compared with that of weekends and holidays. So it's essential to take into account of temporal factor for passenger flow prediction. We divide the day into 24 time intervals and number them chronologically.

3.3.2 Location

As we know, the probability of traffic congestion occurring in the city center is generally higher. While remote areas are much less likely to suffer from serious traffic problem Obviously, passenger flow has a strong tie with geographical positions. Therefore, it's vital to consider location feature in passenger flow prediction. We build independent prediction models for each kind of regions based on spatial division.

3.3.3 Weather

In real life, we usually consider sunny day is suitable for going out to picnic and watching a movie at home is more comfortable when it's raining. Weather affects people's travel choices, and consequently their choice of transportation. For example, people may choose to ride a bicycle in clear, windless weather, and taxi becomes a much more compelling option under bad weather condition. Similar to most traffic flow prediction studies, we also consider weather factor in passenger flow prediction. We classify the weather into three main categories, which are sunny, cloudy, and rainy, and quantified them as 0.5, 0.3 and 0.1 respectively.

3.3.4 Flow

As introduced in the section of related work, researchers apply historical flow data to predict traffic flow. Furthermore, traditional parametric techniques rely solely on historical data to conduct flow prediction, such as ARIMA [10, 28] and Bayesian dynamic model [6]. In other words, historical flow data is the most relevant and straightforward factor related with future passenger flow. In this work, we use the flow data in the corresponding time interval in the previous day as the measure of historical flow data.

3.3.5 MLNN prediction

Multi-Layered Neural Network is an algorithm that roughly mimics the design of the human brain to identify patterns. It attempts to use multiple processing layers containing complex structures or multiple non-linear transformations to learn high-level abstraction of data. It has a strong performance in learning the relationship between features, which not only can be used for classification problems, but also widely applied in the studies of regression

problems, such as passenger flow prediction. The patterns that a neural network can recognize are numerical forms contained in vectors. Therefore, all real-world data including images, sounds, texts, time series, and so on, must be converted into numerical values. So do location, time, weather, TBI. In this work, we employ MLNN to learn multiple features embodying TBI feature to construct multi-feature passenger flow prediction model to predict urban taxi passenger flow.

The regression effect of MLNN is highly dependent of good parameter tuning, so we train the prediction model various times and make detailed adjustments. The number of hidden layers and the number of neurons in each layer are two most significant parameters and they have corresponding value ranges. We perform comparative experiments to optimize their values for the best prediction performance. In order to evaluate the generalization error of the algorithm, we employ the hold-out method to divide the dataset into training set and testing set, as well as validation set. The training set consists of 60 percent of the dataset. The rest data is divided into the testing set and validation set, each with 20 percent respectively. Based on grid search, we determine the number of hidden layers as 4 and the number of neurons in each layer as 30. The maximum number of iterations is 1000. To avoid over-fitting, we add L1 and L2 regular constraints in our model and the strength is set to 0.01. The optimizer used in our network is proximal adagrad optimizer and the activation function is Relu function. We use the batch gradient descent to updated parameters and the learning rate is 0.01. We utilize a machine learning visualization framework, TensorFlow, to visualize the tuning process [1].

4 Empirical study

In this section, we first describe experimental datasets in detail and present data modeling results from the aspects of data preprocessing, temporal division, and spatial division. Then we verify the strong relationship between TBI and taxi passenger flow. Finally we display the prediction results of our proposed approach TBI2Flow.

4.1 Dataset description and processing

We employ real-world taxi dataset to demonstrate the effectiveness of TBI2Flow and the datset we used in our experiments is generated by Qiangsheng Taxi Company in Shanghai. As shown in Table 1, there are 7 main fields in taxi dataset and it covers from 1st to 30th of April, 2015, which contains 21 weekdays and 9 weekend days. The dataset is generated by

Table 1 Description of taxi GPS records dataset

Field	Annotation	Example
TaxiId	Taxi Id	1901252167
Latitude	Coordinates	31.215545
Longtitude	Coordinates	121.404068
State	1:occupied, 0:vacant	1
Date	Date that GPS record wat sent	2015-04-15
Time	Time that GPS record wat sent	10:53:26
Speed	Taxi running speed	16.613991

Table 2 Statistics of travel time's duration

Duration (minute) \ Date	5th, April	6th, April	7th, April	10th, April
[0, 10)	37.4%	35.19%	34.95%	38.16%
[0, 20)	68.33%	66.24%	65.76%	67.63%
[0, 30)	80.72%	78.92%	78.78%	79.57%
[0, 40)	87.23%	85.56%	85.53%	85.96%
[0, 50)	90.81%	89.45%	89.49%	89.67%
[0, 60)	92.85%	91.88%	91.88%	91.97%

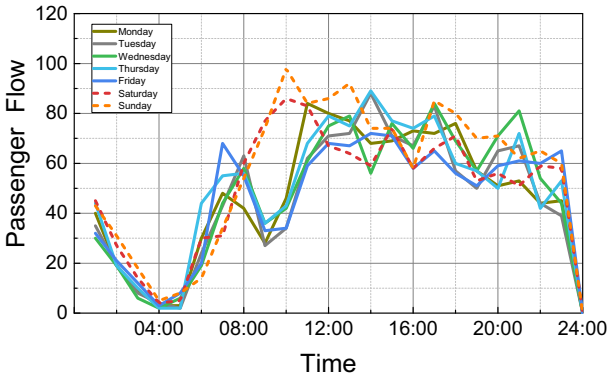
13,685 taxis which accounts for about 25% of the taxis in Shanghai. The dataset contains 34 billion GPS records with a total size of 619 GB.

Apart from cleaning errors of the dataset, we also measure the deviation between the data and the average and filter out the records which have the largest deviation. In order to determine an appropriate temporal division granularity, we first run a cumulative statistical analysis of the duration of taxi travel time and select the results of four days in Table 2. These four days include three weekdays and one weekends, they are Sunday, Monday, Tuesday, and Friday respectively.

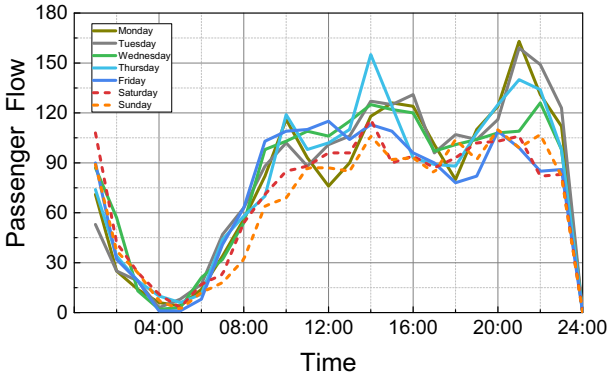
We can obtain from the table that for all days, the proportion of trajectories whose duration of travel time is less than one hour is more than 90% of the total number of trajectories. Therefore, to follow the principle of trajectory integrity and fine-grained division, we set the duration of a time interval in temporal division as one hour. Then we calculate the similarity of seven days of a week using passenger flow and display the results in Table 3. We can get from the table that similarity between weekdays is over 90%. So is the similarity between weekends. However, similarity between weekdays and weekends is relatively low, which provide the basis of days division. From Figure 3 we can intuitively observe the differences in passenger flow trend between weekdays and weekends. Especially in Figure 3a, there exists a flow trough at around ten o'clock of weekdays, while the corresponding time of the weekends is a flow peak. The differences of peaks between weekdays and weekends are accessible. People usually rise early to work on weekdays. Therefore, the flow peak of weekdays appear at around eight o'clock in general. Meanwhile, people tend to stay home at weekends. Thus the delay of flow peak of weekends is reasonable.

Table 3 Similarity of seven days in a week

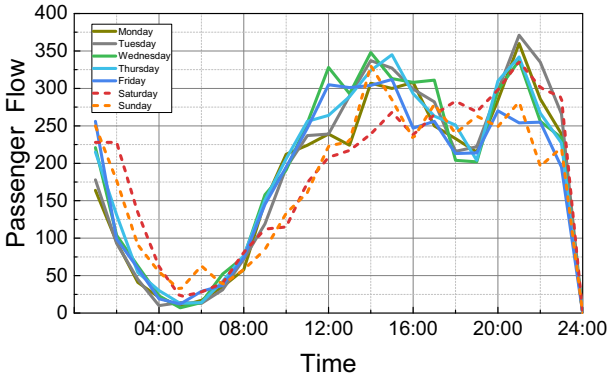
Week	Mon	Tue	Wed	Thur	Fri	Sat	Sun
Mon	1	0.985	0.9611	0.979	0.9245	0.868	0.881
Tue	0.985	1	0.964	0.977	0.935	0.876	0.891
Wed	0.961	0.964	1	0.979	0.973	0.825	0.899
Thur	0.979	0.977	0.979	1	0.967	0.866	0.908
Fri	0.924	0.935	0.973	0.967	1	0.808	0.894
Sat	0.868	0.876	0.825	0.866	0.808	1	0.908
Sun	0.881	0.891	0.899	0.908	0.894	0.908	1



(a) No.1 region



(b) No.2 region



(c) No.3 region

Figure 3 Passenger flow trend of seven days in a week at three typical regions

We select Xuhui District, one of the most prosperous areas in Shanghai, as the study area, which is located at $31.19°N$ to $31.26°N$ and $121.38°E$ to $121.45°E$. According to the method introduced in spatial division, we first divide the whole study area into 49 grids as displayed in Figure 4, in which a grid stands for a region, and then group them into

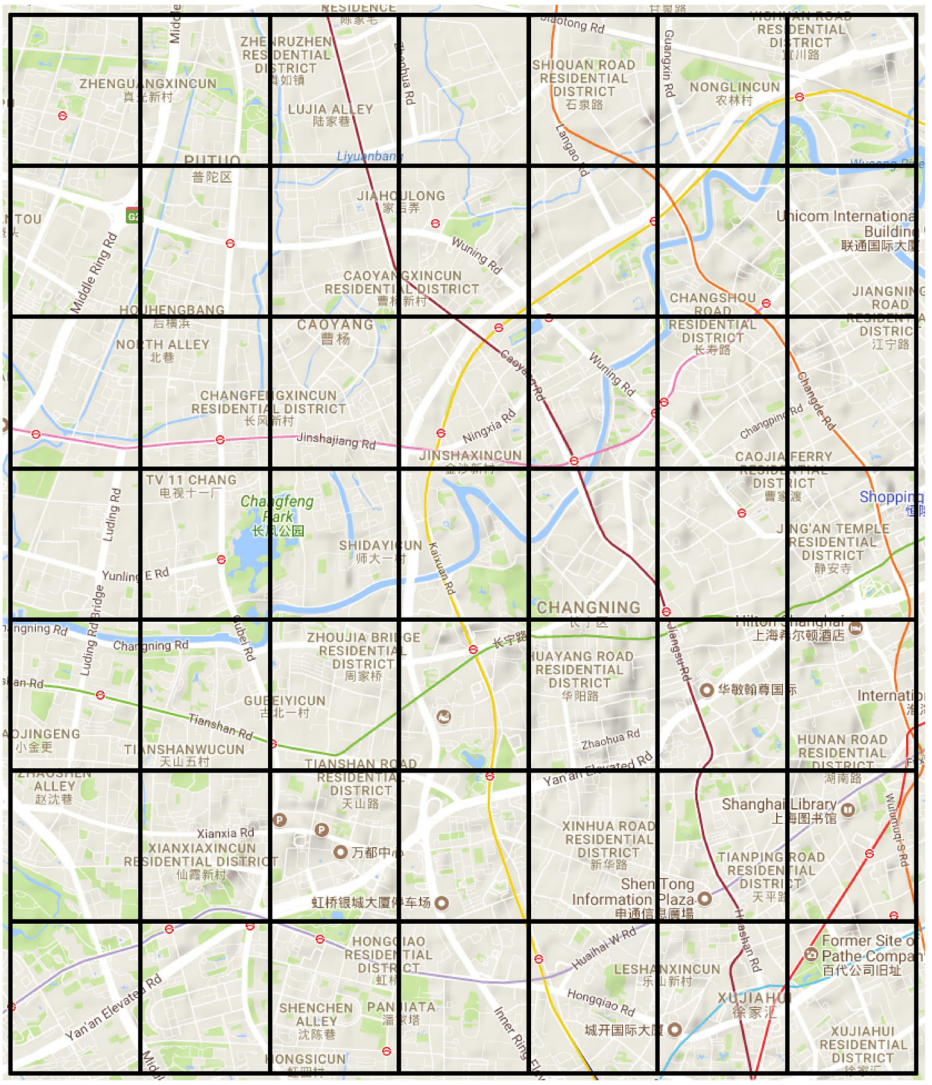


Figure 4 Grid-based spatial division of Xuhui district

ten categories based on average flow distribution of regions. Theses ten categories cover a variety of urban functional regions, such as Shanghai Railway Coach Station, Xuhui District Government, Jiuzhou Living Square, and can support comprehensive traffic travel behavior analysis.

4.2 Excavating inert travel behavior

Based on the analysis of taxi travel behavior, we excavate TBI from real-world taxi GPS data after data modeling. What is the relationship of TBI features and region passenger flow? Is

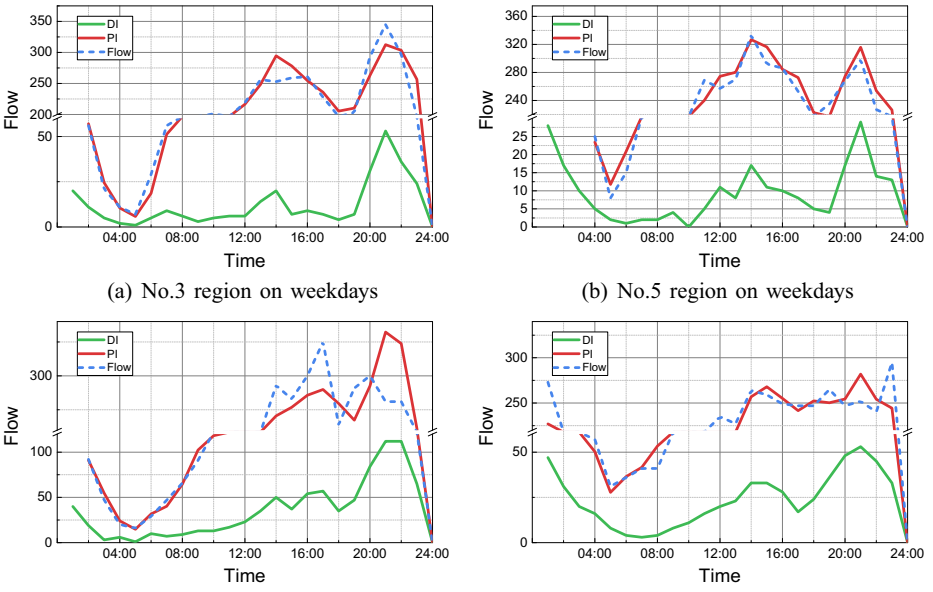
Table 4 Relevance analysis of TBI (DI, PI) and passenger flow

Feature	No.1 Region		No.2 Region		No.3 Region		No.4 Region		No.5 Region	
	weekday	weekend	weekday	weekend	weekday	weekend	weekday	weekend	weekday	weekend
DI-Flow	0.275	0.827	0.3	0.814	0.612	0.939	0.3	0.919	0.443	0.955
PI-Flow	0.915	0.52	0.969	0.526	0.959	0.78	0.946	0.925	0.969	0.802

TBI effective at taxi passenger flow prediction? To answer these questions, we conduct an in-depth analysis of the framework. We measure the relevance between DI and passenger flow, PI and passenger flow respectively, using Correlation Coefficient (CC) as the metric. Results are displayed in Table 4. We can gain from the table that the relevance between PI and passenger flow on weekdays can be up to 90% and the relevance is relatively low on weekends, while it still can reach 80% at some regions. However, the relevance of DI and passenger flow is exactly the opposite of that of PI, which can reach over 80% on weekends and on weekdays the relevance at certain regions is up 60%. In Figure 5, the curves of PI and passenger flow are close and DI shows a similar trend as passenger flow. We can conclude that not only does PI and DI features demonstrate strong correlations with passenger flow data, their contributions also complement each other perfectly.

4.3 Predicting passenger flow

In addition to TBI features, we also extract multi-dimensional features related to taxi travel behavior, including time, weather, and flow. To account for the spatial variations, we build independent prediction models for each region. Using the MLNN model described in



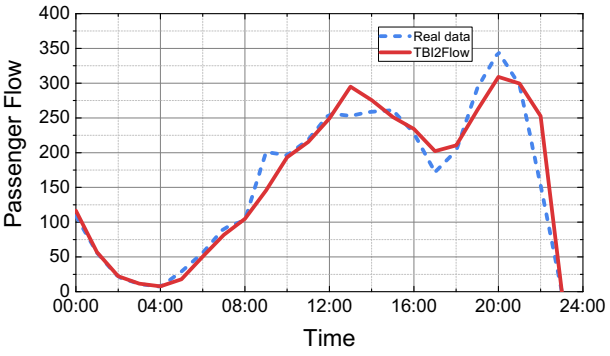
(a) No.3 region on weekdays

(b) No.5 region on weekdays

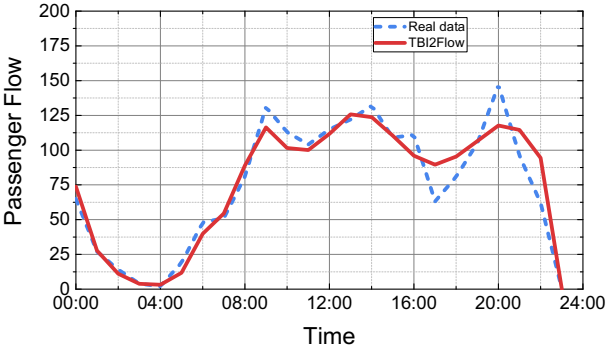
(c) No.3 region on weekends

(d) No.5 region on weekdays

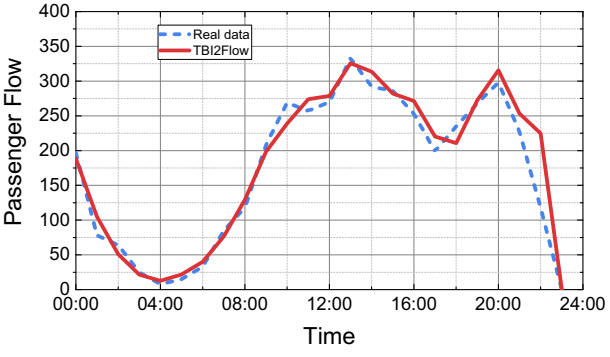
Figure 5 Comparison of TBI (DI, PI) with passenger flow on weekdays and weekends



(a) No.1 region



(b) No.2 region



(c) No.3 region

Figure 6 Prediction results of TBI2Flow at three typical regions

Section 3.3, we predict taxi passenger flow of the study area and the result of three representative regions are presented in Figure 6. By comparing passenger flow trend of predicted data with that of real data, the predictive power of the TBI2Flow framework is full display. For instance, in Figure 6a and especially Figure 6c, the passenger flow trend of real data has two flow peaks at 13 and 20, two troughs at 4 and 17, which are all accurately predicted by TBI2Flow.

5 Performance evaluation

In this section, we conduct extensive comparative contrast experiments to verify the effect of our proposed approach from three aspects, which embody passenger flow prediction with TBI and without TBI, comparison of regression algorithms, and comparison among TBI2Flow, ARIMA, and a Deep Learning Architecture (DLA) [14]. We utilize three indicators that are commonly used in traffic flow prediction to evaluate the effectiveness of algorithms, including CC, MAE, and RMSE.

– Correlation Coefficient (CC). CC can reflect the degree of correlation between variables. And as the degree of correlation increases, the value of CC gets closer to 1. In traffic flow prediction, CC is regarded as an important reference for accuracy by analyzing the linear correlation of predicted data and real data. We define CC in this paper as the following equation:

$$CC = \frac{\sum_{i=1}^{n}(p_i - \overline{p})(r_i - \overline{r})}{\sqrt{\sum_{i=1}^{n}(p_i - \overline{p})^2}\sqrt{\sum_{i=1}^{n}(r_i - \overline{r})^2}} \tag{7}$$

where p_i and r_i denote predicted values and real values respectively.

– Mean Absolute Error (MAE). MAE is the average of the absolute values of the deviation between all the individual predicted values and real ones. The smaller the value of MAE, the better the performance of the algorithm. The definition of MAE is given by:

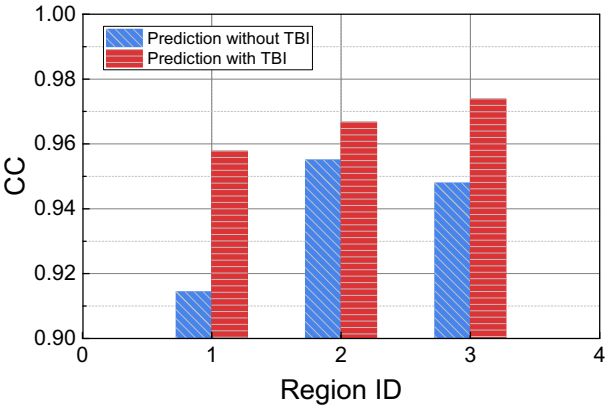
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \tag{8}$$

– Root Mean Square Error (RMSE). The deviation of predicted data and real data can alternatively be measured by RMSE. Its calculating formula is shown as follows:

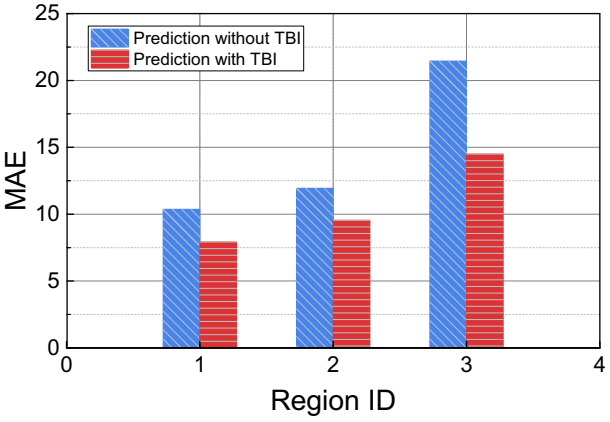
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}|p_i - r_i|^2}{n}} \tag{9}$$

5.1 TBI contributions

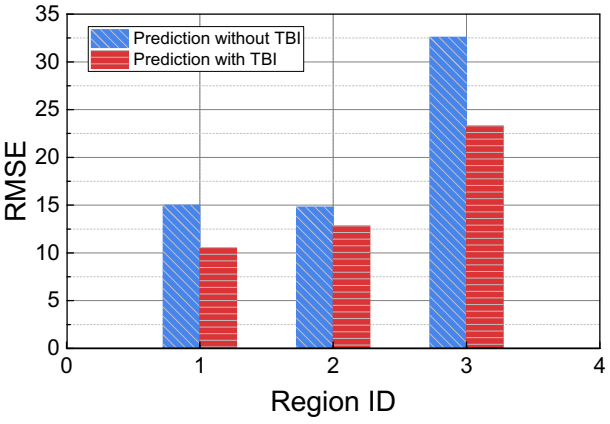
In empirical study, we illustrated the relevance between TBI features and taxi passenger flow and draw the conclusion that there exists a strong linear correlation of TBI and passenger flow. Then how does TBI perform in taxi passenger flow prediction experiments? In this section, we conduct prediction experiments with TBI features data and without TBI features data respectively. The result in Figure 7 demonstrates that adding TBI features to the input can significantly improve the prediction accuracy. Meanwhile, the values of MAE and RMSE also drop after adding TBI features. In Table 5, we offer a detailed and quantitative illustration of TBI's contribution in prediction accuracy and stability, where CC represents the percentage of improving prediction accuracy. MAE and RMSE represents the percentage of reducing predictive error. Even though the field of passenger flow prediction is relatively matured, prediction methods can achieve a high accuracy, the addition of TBI features can significantly improve the value of CC. What's more, the improvement in prediction deviation is even greater. We can conclude that TBI are



(a) CC



(b) MAE



(c) RMSE

Figure 7 TBI contributions analysis

Table 5 TBI contributions analysis of taxi passenger flow prediction

Region	CC		MAE		RMSE	
	weekday	weekend	weekday	weekend	weekday	weekend
No.1 region	4.34%	5.24%	23.6%	27.95%	29.74%	25.35%
No.2 region	1.17%	3.56%	20.04%	24.73%	13.32%	21.08%
No.3 region	2.59%	2.17%	32.42%	37.45%	28.47%	32.05%

indispensable features in taxi passenger flow prediction due to the high correlation with passenger flow and have outstanding contributions in predicting passenger flow. In particular, to verify that our proposed DI features can complement PI features constructed by ARIMA, we further verify the performance of the model without DI features, as shown in Figure 8. The result show strong evidence of DI's contribution to the accuracy of prediction results.

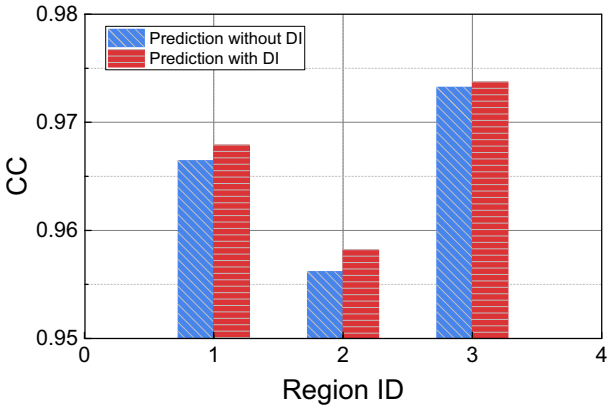
5.2 Comparison of regression algorithms

Having demonstrated the effectiveness of our proposed features, we now compare different learning algorithms. Like any machine learning problem in general, models and features determines the upper limit of the performance, but a good regression algorithm is also critical. While there is no universal algorithm for prediction. Different algorithms will differ in effectiveness for different data sets.

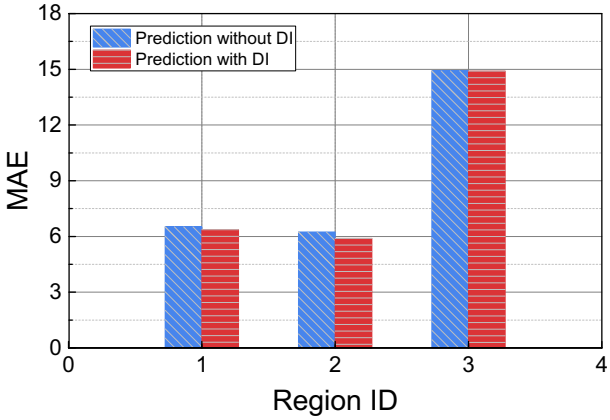
In this subsection, we select three common regression algorithms, including SVR, Back Propagation Neural Network (BPNN), and Gradient Boosting Decision Tree (GBDT) to compare with MLNN in prediction performance. Figure 9 illustrates the comparison results of four algorithms. GBDT has the worst performance, with lower accuracy and higher MAE and RMSE. Compared with GBDT, TBI-based SVR and BPNN performed much better. However, the prediction accuracy and stability of the above two algorithms are lower than that of MLNN. In summary, based on the research background of this paper, TBI-based MLNN has significant advantages over taxi passenger flow prediction compared with common regression models.

5.3 Comparison of prediction approaches

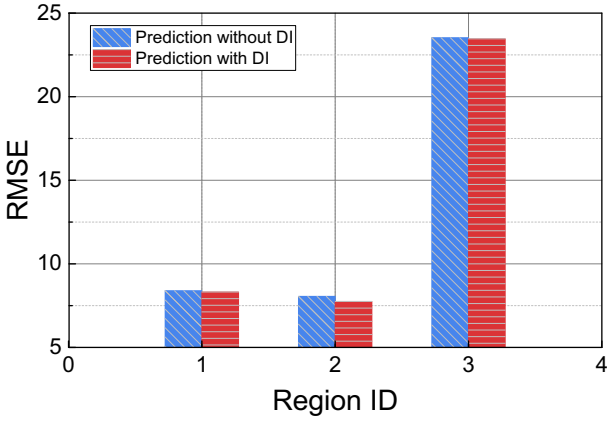
In the previous subsection, we conduct assessment of each part of our proposed approach TBI2Flow and their validity is verified. However, the sum of partial optimum isn't supposed to be the global optimum. Therefore, it's necessary to perform the assessment of the overall approach's effectiveness. Here we choose two traffic flow prediction methods for comparison experiments, which embody ARIMA and DLA. The thought of ARIMA is to learn the law of data changing over time from historical data and then utilize the law to predict future data, that is, ARIMA is used to predict time series. As a simple prediction model, ARIMA can predict future passenger flow based on historical data alone and is widely used in traffic flow prediction. Another algorithm, DLA, is a deep learning architecture proposed by Huang et al. for traffic flow prediction. The architecture has two steps. In the first step, multi-dimensional features are extracted from data by Deep Belief Network (DBN). The features are employed to predict traffic flow using SVR. The number of features in our experiments



(a) CC



(b) MAE



(c) RMSE

Figure 8 DI contributions analysis

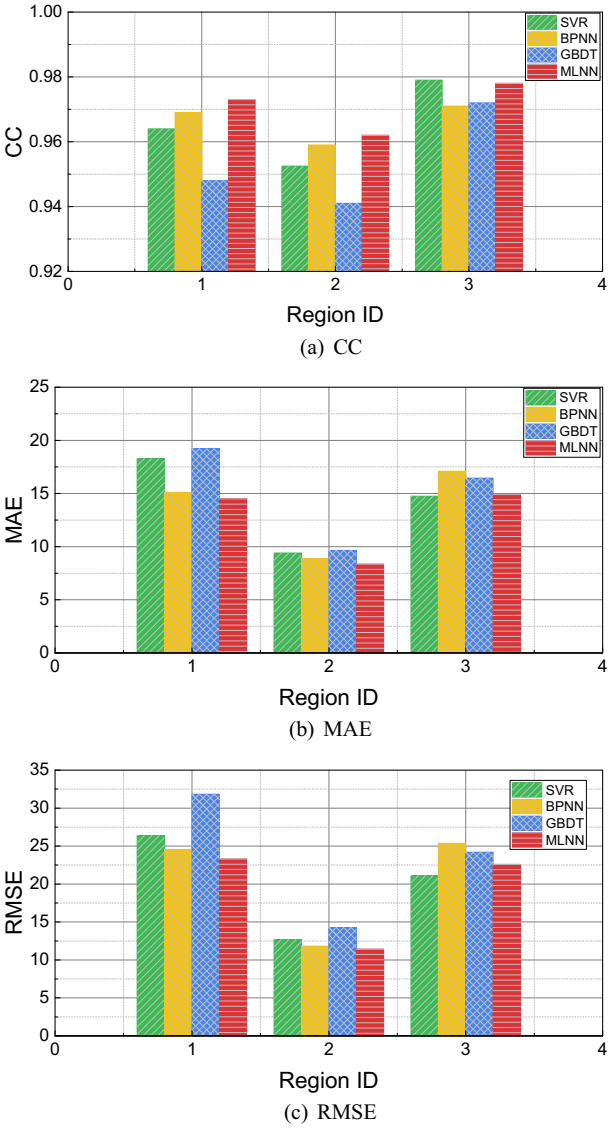
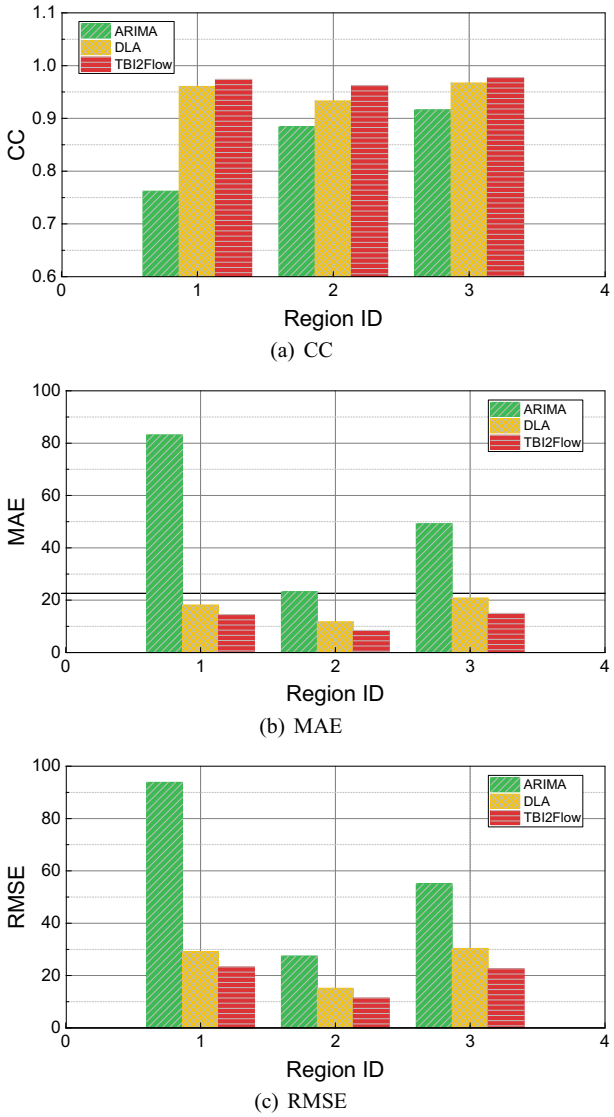


Figure 9 Comparison of four regression algorithms, including SVR, BPNN, GBDT and MLNN

is set as 5. This architecture is similar to ours because they are all based on deep learning algorithms.

Prediction experiments of the three algorithms is based on the same dataset, that is, the preprocessed Shanghai taxi trajectory data. Results of prediction evaluation are presented in Figure 10. Compared with prediction methods based on deep learning algorithms, ARIMA has far less prediction performance, which relies solely on historical data to simulate trends. At some regions its accuracy is even less than 80%, which is a quite poor result in taxi passenger flow prediction. What's worse, ARIMA's prediction results is extremely unstable.



(a) CC

(b) MAE

(c) RMSE

Figure 10 Comparison of TBI2Flow, ARIMA, and DLA

An potential explanation is that ARIMA model only considers a single factor, data trends, and neglects some other significant factors. In contrast, the prediction performance of DLA is respectful, but our proposed TBI2Flow is consistently better. For example, in terms of accuracy, the basic difference between DLA and TBI2Flow at each region is around one percentage point. DLA is a algorithm proposed for short-term passenger flow prediction and has poor expansibility and adaptability to long-term passenger flow prediction of our work. To sum up, the validity of TBI2Flow on long-term taxi passenger flow prediction is well demonstrated.

6 Conclusion and future work

In this paper, we put forward a taxi passenger flow prediction framework TBI2Flow. The approach is based on TBI features, containing DI feature and PI feature, which are excavated from taxi trajectory data according to travel behavior analysis. The strong relationship between TBI and the the passenger flow is verified by statistic analysis. We conduct extensive experiments and the results demonstrate that TBI2Flow is remarkably effective to predict long-term taxi passenger flow. Our TBI2Flow has considerably practical significance in traffic management, especially for taxi industry and smart city construction.

There are multiple venues for future work. Real-time is extremely essential for passenger flow prediction. Therefore, we will construct a real-time prediction architecture based on TBI2Flow and even build a real-time prediction system to provide effective suggestions for urban taxi drivers and passengers. We also will focus on further refinement of passenger flow prediction granularity.

Acknowledgments The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group No. RG- 1439-088. This work was partially supported by the National Natural Science Foundation of China (61572106), the Dalian Science and Technology Innovation Fund (2018J12GX048), and the Fundamental Research Funds for the Central Universities (DUT18JC09).

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12Th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283 (2016)
2. Ahmed, M.S., Cook, A.R.: Analysis of freeway traffic time-series data by using Box-Jenkins techniques. 722 (1979)
3. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time Series Analysis: Forecasting and Control. Wiley, New York (2015)
4. Chen, C., Jiao, S., Zhang, S., Liu, W., Feng, L., Wang, Y.: Tripimputor: real-time imputing taxi trip purpose leveraging multi-sourced urban data. IEEE Trans. Intell. Transp. Syst. (99):1–13 (2018)
5. Deng, Z., Ji, M.: Spatiotemporal structure of taxi services in Shanghai: Using exploratory spatial data analysis. In: International Conference on Geoinformatics, pp. 1–5 (2011)
6. Fei, X., Lu, C.C., Liu, K.: A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction. Transportation Research Part C Emerging Technologies **19**(6), 1306–1318 (2011)
7. Ferreira, N., Poco, J., Vo, H.T., Freire, J., Silva, C.T.: Visual exploration of big spatio-temporal urban data: a study of new york city taxi trips. IEEE Trans. Vis. Comput. Graph. **19**(12), 2149–2158 (2013)
8. Glöss, M., McGregor, M., Brown, B.: Designing for labour: uber and the on-demand mobile workforce. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 1632–1643. ACM (2016)
9. Guo, F., Krishnan, R., Polak, J.: A computationally efficient two-stage method for short-term traffic prediction on urban roads. Transp. Plan. Technol. **36**(1), 62–75 (2013)
10. Hamed, M.M., Al-Masaeid, H.R., Said, Z.M.B.: Short-term prediction of traffic volume in urban arterials. J. Transp. Eng. **121**(3), 249–254 (1995)
11. He, J., Shen, W., Divakaruni, P., Wynter, L., Lawrence, R.: Improving traffic prediction with tweet semantics. In: International Joint Conference on Artificial Intelligence, pp. 1387–1393 (2013)
12. Hobeika, A.G., Chang, K.K.: Traffic-flow-prediction systems based on upstream traffic. In: Vehicle Navigation and Information Systems Conference, 1994. Proceedings, pp. 345–350 (2002)
13. Hou, Z., Li, X.: Repeatability and similarity of freeway traffic flow and long-term prediction under big data. IEEE Trans. Intell. Transp. Syst. **17**(6), 1786–1796 (2016)
14. Huang, W., Song, G., Hong, H., Xie, K.: Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. IEEE Trans. Intell. Transp. Syst. **15**(5), 2191–2201 (2014)

15. Kong, X., Xu, Z., Shen, G., Wang, J., Yang, Q., Zhang, B.: Urban traffic congestion estimation and prediction based on floating car trajectory data. Futur. Gener. Comput. Syst. **61**(C), 97–107 (2016)
16. Kong, X., Xia, F., Wang, J., Rahim, A., Das, S.K.: Time-location-relationship combined service recommendation based on taxi trajectory data. IEEE Trans. Ind. Inf. **13**(3), 1202–1212 (2017)
17. Kong, X., Li, M., Li, J., Tian, K., Hu, X., Xia, F.: Copfun: an urban co-occurrence pattern mining scheme based on regional function discovery. World Wide Web, pp. 1–26 (2018)
18. Lee, J., Shin, I., Park, G.L.: Analysis of the passenger pick-up pattern for taxi location recommendation. In: International Conference on Networked Computing and Advanced Information Management, pp. 199–204 (2008)
19. Li, B., Zhang, D., Sun, L., Chen, C., Li, S., Qi, G., Yang, Q.: Hunting or waiting? discovering passenger-finding strategies from a large-scale real-world taxi dataset. In: IEEE International Conference on Pervasive Computing and Communications Workshops, pp. 63–68 (2011)
20. Liu, L., Andris, C., Biderman, A., Ratti, C.: Revealing taxi driver's mobility intelligence through his trace. IEEE Pervasive Comput **160**, 1–17 (2009)
21. Ma, J., Choo, K.K.R., Hsu, H.h., Jin, Q., Liu, W., Wang, K., Wang, Y., Zhou, X.: Perspectives on cyber science and technology for cyberization and cyber-enabled worlds. In: Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), 2016 IEEE 14th Intl C, pp. 1–9 (2016)
22. Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y.: Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies **54**, 187–197 (2015)
23. Ni, M., He, Q., Gao, J.: Forecasting the subway passenger flow under event occurrences with social media. IEEE Trans. Intell. Transp. Syst. **PP**(99), 1–10 (2017)
24. Oh, S.D., Kim, Y.J., Hong, J.S.: Urban traffic flow prediction system using a multifactor pattern recognition model. IEEE Trans. Intell. Transp. Syst. **16**(5), 2744–2755 (2015)
25. Ouyang, Y., Guo, B., Lu, X., Han, Q., Guo, T., Yu, Z.: Competitivebike: Competitive analysis and popularity prediction of bike-sharing apps using multi-source data. IEEE Trans. Mob. Comput. (2018)
26. Su, F., Dong, H., Jia, L., Qin, Y., Tian, Z.: Long-term forecasting oriented to urban expressway traffic situation. Adv. Mech. Eng. **8**(1). https://doi.org/10.1177/1687814016628397 (2016)
27. Vlahogianni, E.I., Golias, J.C., Karlaftis, M.G.: Short-term traffic forecasting: Overview of objectives and methods. Transp. Rev. **24**(5), 533–557 (2004)
28. Voort, M.V.D., Dougherty, M., Watson, S.: Combining kohonen maps with arima time series models to forecast traffic flow. Transportation Research Part C Emerging Technologies **4**(5), 307–318 (1996)
29. Wang, D., Cao, W., Li, J., Ye, J.: Deepsd: Supply-demand prediction for online car-hailing services using deep neural networks. In: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 243–254. IEEE (2017)
30. Wang, J., Shi, Q.: Short-term traffic speed forecasting hybrid model based on chaosĺcwavelet analysis-support vector machine theory. Transportation Research Part C Emerging Technologies **27**(2), 219–232 (2013)
31. Wang, Y., Papageorgiou, M., Messmer, A.: Real-time freeway traffic state estimation based on extended kalman filter: a case study. Transp. Sci. **41**(2), 167–181 (2007)
32. Williams, B., Durvasula, P., Brown, D.: Urban freeway traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models. Transp. Res. Rec. **1644**(1), 132–141 (1998)
33. Williams, B.M., Hoel, L.A.: Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. J. Transp. Eng. **129**(6), 664–672 (2003)
34. Wu, C.H., Ho, J.M., Lee, D.T.: Travel-time prediction with support vector regression. IEEE Trans. Intell. Transp. Syst. **5**(4), 276–281 (2004)
35. Xia, F., Rahim, A., Kong, X., Wang, M., Cai, Y., Wang, J.: Modeling and analysis of large-scale urban mobility for green transportation. IEEE Trans. Ind. Inf. **14**(4), 1469–1481 (2018)
36. Xia, F., Wang, J., Kong, X., Wang, Z., Li, J., Liu, C.: Exploring human mobility patterns in urban scenarios: a trajectory data perspective. IEEE Commun. Mag. **56**(3), 142–149 (2018)
37. Yang, H.F., Dillon, T.S., Chen, Y.P.: Optimized structure of the traffic flow forecasting model with a deep learning approach. IEEE Transactions on Neural Networks and Learning Systems **PP**(99), 1–11 (2016)
38. Yuan, J., Zheng, Y., Xie, X., Sun, G.: T-drive: Enhancing driving directions with taxi drivers. IEEE Trans. Knowl. Data Eng. **25**(1), 220–232 (2013)
39. Yue, Y., Zhuang, Y., Li, Q., Mao, Q.: Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In: International Conference on Geoinformatics, pp. 1–6 (2009)

40. Zhang, F., Zhang, L.: Regulation of car-hailing against the background of "Internet Plus" in China. In: 2017 2nd International Seminar on Education Innovation and Economic Management (SEIEM 2017), Atlantis Press (2017)
41. Zhang, J., Zheng, Y., Qi, D., Li, R., Yi, X.: Dnn-based prediction model for spatio-temporal data. In: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, p. 92. ACM (2016)
42. Zhang, L., Lu, J., Zhou, J., Zhu, J., Li, Y., Wan, Q.: Complexities' day-to-day dynamic evolution analysis and prediction for a didi taxi trip network based on complex network theory. Mod. Phys. Lett. B **32**(09), 1850,062 (2018)
43. Zhao, J., Sun, S.: High-order gaussian process dynamical models for traffic flow prediction. IEEE Trans. Intell. Transp. Syst. **17**(7), 2014–2019 (2016)
44. Zhou, X., Bo, W., Jin, Q.: Analysis of user network and correlation for community discovery based on topic-aware similarity and behavioral influence. IEEE Transactions on Human-Machine Systems **PP**(99), 1–13 (2017)
45. Zhou, X., Liang, W., Kevin, I., Wang, K., Huang, R., Jin, Q.: Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data. IEEE Trans. Emerg. Top. Comput. (2018)
46. Zhou, X., Zomaya, A.Y., Li, W., Ruchkin, I.: Cybermatics: Advanced strategy and technology for cyber-enabled systems and applications (2018)