Research Article

Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots

Zhizhen Liu,1 Hong Chen,1 Yan Li,2 and Qi Zhang1

1School of Highway, Chang’an University, Xi’an 710000, China
2Shenzhen Urban Transport Planning Centre, Shenzhen 518000, China

Correspondence should be addressed to Hong Chen; glch@chd.edu.cn

Received 18 November 2019; Revised 22 June 2020; Accepted 29 June 2020; Published 21 July 2020

Copyright © 2020 Zhizhen Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Accurate taxi demand prediction can solve the congestion problem caused by the supply-demand imbalance. However, most taxi demand studies are based on historical taxi trajectory data. In this study, we detected hotspots and proposed three methods to predict the taxi demand in hotspots. Next, we compared the predictive effect of the random forest model (RFM), ridge regression model (RRM), and combination forecasting model (CFM). Thereafter, we considered environmental and meteorological factors to predict the taxi demand in hotspots. Finally, the importance of indicators was analyzed, and the essential elements were the time, temperature, and weather factors. The results indicate that the prediction effect of CFM is better than those of RFM and RRM. The experiment obtains the relationship between taxi demand and environment and is helpful for taxi dispatching by considering additional factors, such as temperature and weather.

1. Introduction

Taxi is an essential part of urban public transportation, and taxi demand is different from others because of its stochastic trajectory and dependence of spatial location [1, 2]. However, the imbalance between the supply and demand of taxis is particularly severe due to the uneven information distribution between drivers and passengers [3]. Taxi drivers’ customer-searching behavior relies on historical experience, and passengers’ trips are random. The information asymmetry of taxis and passengers wastes limited public resources [4]. Thus, the taxi demand in the hotspots should be predicted [5].

Previous studies on taxi demand prediction are generally based on historical taxi trajectory data. Previous studies have shown the feasibility of obtaining predictions from historical taxi trajectory data [1, 5–23]. Methods of traffic demand prediction can be classified into three types: linear system theory (such as the autoregressive moving average model [24], Kalman filtering model, and time series model), nonlinear system theory (such as the neural network model, gray prediction model, and random forest model (RFM)), and combination forecasting model (CFM). The first application of the time series prediction model in traffic prediction research was modeling the univariate traffic flow data as seasonal autoregressive integrated moving average processes [25]. Shekhar used the Kalman filter model to study univariate traffic condition predictions [2]. Alvarez-Garcia et al. proposed a system based on the hidden Markov model to predict taxi trip destinations [26]. Chang et al. mined historical taxi trajectory data and predicted the time and spatial distributions of taxi demand [9]. Moreira-Matias et al. introduced a new method for using traffic flow data to predict the spatial distribution of taxi passengers in the short-term time. A CFM combining three time series prediction methods that can effectively determine the spatio-temporal distribution of taxi passenger demand was proposed [17]. Lv et al. proposed a traffic flow prediction method based on deep learning considering spatiotemporal correlation and used an autoencoder model to learn traffic flow characteristics [27]. Zhang et al. proposed an adaptive prediction method to predict a hotspot location and its heat [22]. Zhao et al. implemented and compared three predictors for predictive algorithms that determine maximum
predictability: Markov, Lempel–Ziv–Welch, and neural network predictors [13]. Davis used a time series model to predict taxi travel demand based on mobile app taxi services [28]. Zhao et al. proposed a new prediction model based on long short-term memory (LSTM) networks. The proposed LSTM network considered the spatiotemporal correlation in traffic systems [29]. Zhang et al. proposed a Dmodel based on the hidden Markov chain model for taxi prediction [21]. Yu et al. proposed a spatiotemporal recurrent convolutional network for traffic volume prediction based on the deep convolutional neutral network [30]. Ou et al. proposed a method of combining the bias-corrected random forest algorithm with the data-driven feature selection strategy for short-term urban traffic flow prediction to solve the problem of unreasonable feature selection [31]. Yao et al. proposed a deep multiview spatiotemporal network framework to simulate spatiotemporal relationships based on traffic prediction models [32]. Bao et al. considered the interaction between subways and taxis based on univariate traffic prediction and applied the residual neural network to predict the taxi demand in different regions [6]. Ishiguro et al. proposed a taxi demand prediction algorithm using real-time demographic data generated by cellular networks and used a stacked denoising autoencoder to assess the impact of real-time demographic data on taxi demand prediction accuracy [12]. Markou et al. considered the information provided by unstructured data while using taxi GPS data and used machine learning techniques to predict taxi demand [11]. Xu et al. believed that the occurrence of taxi request behavior is related to the historical traffic behaviors and proposed an LSTM model, which can predict taxi requests for each region of the city based on historical demand and other relevant information [19]. Past research has mostly focused on pickup points. Rodrigues et al. considered drop-off points and combined the time correlation with the spatial correlation to predict the taxi demand with an LSTM method [18]. Kuang et al. proposed two deep learning methods that combine unstructured textual information with historical taxi trip data for traffic demand prediction research [15]. Furthermore, Castro et al. conducted a review of studies on traffic GPS data and proposed a new direction based on GPS data [33].

Previous works have focused on mining the regularity of trajectory data to predict the traffic demand, but environmental data have been ignored. Furthermore, the method that combines linear and nonlinear system theory has been rarely proposed. This study aims to explore the prediction method combining RRM and RRM for predicting taxi demand in hotspots. Moreover, environmental data are considered. First, the method identifies the taxi demand hotspots in the city. Then, we predict taxi demand at various time periods using the RFM and RRM [34]. Next, we propose a CFR model that combines the RFM and RRM. The forecasting method considers environmental and historical taxi trajectory data. This study is beneficial for traffic management rebalancing taxis.

The paper consists of four sections: Section 1 describes the importance of taxi demand prediction and focuses on related research about taxi demand prediction; Section 2 describes the data and method we used in this study; Section 3 describes the results of the experiment; discussion and future research are included in Section 4; and Section 5 describes the conclusion.

2. Data

2.1. GPS Data. GPS data are from the Xi’an Taxi Management Office and consist of vehicle location data that are recorded every 5 s for 30 days. The dataset consists of 40 million track points. The GPS data have undergone extensive cleaning, and only error-free trip strings are used in this research (Figure 1).

2.2. Environmental Data. The purpose of this study is to accurately predict the demand for taxis in hotspots by constructing a set of affecting factors of the taxi demand. Therefore, the impacts of air quality, weather, wind speed, and temperature on demand for taxis are considered. In this study, the influencing factors of taxi demand are constructed on the basis of two types of data: air quality and meteorological data.

The air quality data are derived from the official website of Green Breathing. The detection indicators include various pollutant data, including PM2.5 and PM10, and the air quality level of the day can be defined according to the AQI. The meteorological data are from the National Meteorological Information Center. This study selects the hourly data of Xi’an, including hourly observations of temperature, pressure, humidity, wind speed, and precipitation. The air quality data used in this study have seven dimensions, and the meteorological data have five dimensions (Table 1).

3. Methods

3.1. Random Forest Model. RFM is an ensemble learning algorithm and an extension of bagging [35]. At each node of each decision tree, a subset of k feature attributes is randomly selected from the feature attribute set of the node; then, the best feature attribute is selected from the subset for division (Figure 2).

3.2. Ridge Regression Model. RRM is a partial estimation method designed for collinear data analysis and is an improved least-square estimation method. The regression coefficient becomes realistic and reliable by abandoning the unbiasedness of the least-square estimation and losing part of the information. An RRM fits the ill-conditioned data more accurately than the least-square estimation.

Given a dataset $D = \{x_1, y_1, x_2, y_2, \ldots, x_m, y_m\}$, where $x \in \mathbb{R}^n$ and $y \in \mathbb{R}$. The simplest linear regression model defines the loss function as the square of the residual. Then, the optimization objective is expressed as follows:

$$L = \|X\theta - y\|^2$$

where $\theta = \theta_1; \theta_2; \ldots; \theta_m$ is a regression coefficient. $X = x_1; x_2; \ldots; x_m$ and $y$ are predicted values. The abovementioned formula would easily overfit when the sample has many
Input parameters and training set $S$

\[ i = 1 \]

Extract the $i$th Bootstrap training sample $S_i$

Extract max_features variable subset in equal probability

Optimal variable

Split nodes

End training of $S_i$

Average the combination of $n_{\text{estimators}}$ regression trees to get the RF model $f(x)$

Table 1: Environmental data structure description.

<table>
<thead>
<tr>
<th>Indicators</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AQI</td>
<td>Air quality index</td>
</tr>
<tr>
<td>CO</td>
<td>Concentration of CO ($\mu$/m$^3$)</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>Concentration of NO$_2$ ($\mu$/m$^3$)</td>
</tr>
<tr>
<td>O$_3$</td>
<td>Concentration of O$_3$ ($\mu$/m$^3$)</td>
</tr>
<tr>
<td>PM2.5</td>
<td>Concentration of PM2.5 ($\mu$/m$^3$)</td>
</tr>
<tr>
<td>PM10</td>
<td>Concentration of PM10 ($\mu$/m$^3$)</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>Concentration of SO$_2$ ($\mu$/m$^3$)</td>
</tr>
<tr>
<td>Air quality</td>
<td>1: excellent; 2: good; 3: mild pollution; 4: serious pollution</td>
</tr>
<tr>
<td>Weather</td>
<td>1: sunny; 2: cloudy; 3: raining; 4: haze</td>
</tr>
<tr>
<td>Wind speed</td>
<td>Wind speed (m/s)</td>
</tr>
<tr>
<td>TEM</td>
<td>Temperature (°C)</td>
</tr>
<tr>
<td>RHU</td>
<td>Humidity (%)</td>
</tr>
<tr>
<td>PRE</td>
<td>Precipitation (mm)</td>
</tr>
</tbody>
</table>

Figure 1: GPS data representation.

Figure 2: The flow diagram of an RFM.
features, and the number of samples is relatively small. Regularization terms can be used in the aforementioned formula. The $L_1$ norm regularization is introduced into the RRM as follows:

$$L = \|X\theta - y\|^2 + \|\Gamma\theta\|^2.$$  \hfill (2)

We define $\Gamma = \alpha I$, where $I$ is the identity matrix, and is shown as

$$L(\alpha) = (X^TX + \alpha I)^{-1}X^Ty.$$  \hfill (3)

As $\alpha$ increases, the absolute values of the elements in $L(\alpha)$ tend to decrease, and the deviation of correct value $\theta_i$ increases. When $\alpha$ tends to infinity, $L(\alpha)$ tends to 0. The trajectory of $L(\alpha)$ that changes with $\alpha$ is called the ridge. When the ridge is stable, $\alpha$ is the optimal value. In general, the $R^2$ value of the ridge regression equation will be slightly low, but the significance of the regression coefficient is usually significantly high.

### 3.3. Combination Forecasting Model

CFM can solve special prediction problems in research by combining the characteristics of different models. The calculation can be expressed as

$$\tilde{y}_{\text{CFM},i} = \lambda_1\tilde{y}_{\text{RRM},i} + \lambda_2\tilde{y}_{\text{RFM},i},$$  \hfill (4)

where $\tilde{y}_{\text{CFM},i}$ is the predicted value of the CFM, $\tilde{y}_{\text{RRM},i}$ is the predicted value of the RRM, $\tilde{y}_{\text{RFM},i}$ is the predicted value of the RRM and CFM, respectively.

The core of the CFM is the determination of the weight coefficients $\lambda_1$ and $\lambda_2$. Inverse-variance weighting method is used to determine the weight coefficient of the CFM. The calculation equations are expressed as follows:

$$\lambda_1 = \frac{e_{\text{RRM}}(t)^{-1}}{e_{\text{RRM}}(t)^{-1} + e_{\text{RFM}}(t)^{-1}},$$  \hfill (5)

$$\lambda_2 = 1 - \lambda_1.$$  \hfill (6)

The squared error sum of the RRM is expressed as equation (7), and the squared sum of the RRM is expressed as equation (8):

$$e_{\text{RRM}}(i) = \sum_{i=1}^{276} (y_i - \tilde{y}_{\text{RRM},i})^2,$$  \hfill (7)

$$e_{\text{RFM}}(i) = \sum_{i=1}^{276} (y_i - \tilde{y}_{\text{RFM},i})^2,$$  \hfill (8)

where $e_{\text{RRM}}(i)$ represents the sum of squared errors of the RRM, $e_{\text{RFM}}(i)$ represents the sum of squared errors of the RFM, $y_i$ represents the true value, $\tilde{y}_{\text{RRM},i}$ represents the fitted value of RRM, and $\tilde{y}_{\text{RFM},i}$ represents the fitted value of the RFM.

### 4. Data Processing

#### 4.1. GPS Data Processing

The “STAT” attribute in taxi GPS data is the record of the taxi driving state, in which “4” represents the passenger and “5” represents empty driving. A change from “4” to “5” indicates that the passenger exits the vehicle. This record is recorded as point D. A change from “5” to “4” indicates that the passenger enters the vehicle. This record is recorded as point O.

#### 4.2. Feature Selection

Ensuring that the features are independent of one another is difficult because of their large number in the experiment. In the modeling process, two features with a strong correlation tend to exhibit multiple collinearities in the data. Therefore, the correlation of the experimental data features should be tested. The method chosen in this study is the Pearson correlation analysis, which can measure the linear relationship between variables. The calculation is expressed as follows:

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$  \hfill (9)

where $\text{cov}(X,Y)$ represents the covariance between the variables $X$ and $Y$, $\sigma_X$ and $\sigma_Y$ represent the standard deviations of the variables $X$ and $Y$, and $\rho_{XY}$ represents the correlation coefficient of two continuous variables; the value of $\rho_{XY}$ is between -1 and 1. If $\rho_{XY} > 0$, then the two variables are positively correlated; if $\rho_{XY} < 0$, then the two variables are negatively correlated. A large absolute value of $\rho_{XY}$ corresponds to a strong correlation. The corr function of the pandas library in Python is applied to obtain the correlation coefficient matrix (Figure 3).

Figure 3 shows that the correlation among PM2.5, PM10, and AQI is strong. A slight multicollinearity is observed in the correlation between $O_3$ and TEM (temperature); therefore, a correlation exists between RHU and TEM. Indicators with severe multicollinearity are excluded. Thus, indicators PM2.5 and PM10 are eliminated.

Four indicator variables of hour, wdy, week, and holiday are also added to explore the impact of time, week, weekday, and holiday factors on the taxi demand (Table 2).

#### 4.3. One-Hot Encoding

All data are encoded using the one-hot encoder function in the scikit-learn.preprocessing library. The week attribute is taken as an example (Figure 4).

After the one-hot encoding, the data dimension has expanded to 39. In the experiment, the sample size of the dataset is small, and the verification and test sets can be combined when dividing the dataset. The first 23 days of April 2017 are taken as the training set, with the other 7 days as the test set.

### 5. Results and Discussion

#### 5.1. Extract Hotspots

The ArcGIS 10.2 kernel density analysis tool is used to analyze the kernel density of the residents’ pickup and get-off positions in the three time periods of the working and rest days (Figure 5).

As shown in Figure 5, the taxi demand on weekdays and nonworking days are mainly distributed in the main roads of
Xi’an. The taxi demand at various peak hours is also distributed among the main roads of Xi’an. Xi’an taxi demand intensive areas are normalized and have no visible space-time character. The 30-day thermogram is superimposed (Figure 6).

Hotspots are distributed in areas such as Xi’anbei Railway Station, Bell Tower, Xiaozhai, Railway Station, and City Library. Xi’anbei Railway Station and Railway Station are transportation hubs. Xiaozhai, City Library, and Bell Tower are commercial areas. In this study, two representative areas, namely, Bell Tower and Xi’anbei Railway Station, are selected (Figure 7).

5.2. Random Forest Prediction. Using Python’s sklearn.ensemble library, we can use random forest regression (RFM) (Table 3).

The main influencing factor of RFM is “n_estimators.” We use the goodness of fit ($R^2$) to adjust the parameters of RFM. The calculation is expressed as follows:

$$R^2 = \frac{SSR}{SST} = \frac{1}{N} \left( \frac{\sum_{i=1}^{N} \left( y_i - \bar{y} \right)^2}{\sum_{i=1}^{N} \left( y_i - \bar{y} \right)^2} \right)$$

where $N$ is the sample size, $SST$ is the sum of squares, $SSR$ is the sum of squares of regression, $SSE$ is the sum of squared
residuals, $y_i$ is the value to be fitted, $\bar{y}$ is the mean of $y$, and $\tilde{y}_i$ is the fitted value.

Considering the number of samples and training speed of RRM, we choose $[1 – 200]$ as variable span. The relation between “$n_{\text{estimators}}$” and $R^2$ can be calculated (Figures 8 and 9).

The adjusted optimal parameters for Xi’anbei Railway Station and Bell Tower areas are shown in Tables 4 and 5.

Figure 5: Thermogram of workday (left) and nonworkday (right) in peak hours: (a) in the morning; (b) at noon; (c) in the evening.
The prediction results of RFM in Xi’anbei Railway Station and Bell Tower areas are shown in Figures 10 and 11. RFM can score the importance of feature attributes. In the RFM, evaluating the importance of feature attributes is based on the random replacement of the permutation principle. The reduction in the mean square residual and the prediction accuracy reflects the importance of characteristic variables. In this study, the calculation of the mean square residual reduction is used to evaluate the importance of the variables:

1. We assume $M$ regression trees in the random forest. OOB$_i$ represents the out-of-bag data of the $i$th tree. The out-of-bag mean square deviations of each tree are $\text{MSE}_{\text{OOB}_1}, \text{MSE}_{\text{OOB}_2}, \ldots, \text{MSE}_{\text{OOB}_M}$.

2. We assume that the total number of variables is $N$. For each input variable $X_p$, random replacement in $M$ out-of-bag data is conducted. $M$ new out-of-bag data OOB are obtained, and the mean square deviation of the new out-of-bag data is calculated. Then, an out-of-bag error matrix can be constructed as follows:

$$
\begin{bmatrix}
\text{MSE}_{1,\text{OOB}_1} & \text{MSE}_{1,\text{OOB}_2} & \cdots & \text{MSE}_{1,\text{OOB}_M} \\
\text{MSE}_{2,\text{OOB}_1} & \text{MSE}_{2,\text{OOB}_2} & \cdots & \text{MSE}_{2,\text{OOB}_M} \\
\vdots & \vdots & \ddots & \vdots \\
\text{MSE}_{N,\text{OOB}_1} & \text{MSE}_{N,\text{OOB}_2} & \cdots & \text{MSE}_{N,\text{OOB}_M}
\end{bmatrix}
$$
The out-of-bag error $MSE_{OOB_1}, MSE_{OOB_2}, \ldots, MSE_{OOB_M}$ before replacement is subtracted with the $i$th row of the out-of-bag error matrix. Then, the significance score of $X_i$ is the average of the abovementioned calculated results, as shown in the following equation:

$$VIM_i = \frac{1}{M} \sum_{j=1}^{M} (MSE_{OOB_j} - MSE_{OOB_i}) \quad 1 \leq i \leq N.$$  

A large value of $VIM_i$ corresponds to a great contribution of the variable. This study uses the `feature_importances_` function in RMM of the `scikit-learn` library to score the input variables (Figures 12 and 13).

### Table 4: Random forest parameters in the Xi’anbei Railway Station area.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Accuracy before adjustment</th>
<th>Accuracy after adjustment</th>
<th>Value</th>
<th>Increase in accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{\text{estimators}}$</td>
<td>0.857206</td>
<td>0.858649</td>
<td>80</td>
<td>0.001443</td>
</tr>
<tr>
<td>max_features</td>
<td>0.858649</td>
<td>0.884258</td>
<td>4</td>
<td>0.025609</td>
</tr>
<tr>
<td>max_depth</td>
<td>0.884258</td>
<td>0.884258</td>
<td>Default</td>
<td>0</td>
</tr>
<tr>
<td>min_samples_leaf</td>
<td>0.884258</td>
<td>0.8842589</td>
<td>Default</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 5: Random forest parameters in the Bell Tower area.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Accuracy before adjustment</th>
<th>Accuracy after adjustment</th>
<th>Value</th>
<th>Increase in accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{\text{estimators}}$</td>
<td>0.825876</td>
<td>0.836657</td>
<td>16</td>
<td>0.010781</td>
</tr>
<tr>
<td>max_features</td>
<td>0.836657</td>
<td>0.839204</td>
<td>28</td>
<td>0.002547</td>
</tr>
<tr>
<td>max_depth</td>
<td>0.839204</td>
<td>0.839204</td>
<td>Default</td>
<td>0</td>
</tr>
<tr>
<td>min_samples_leaf</td>
<td>0.839204</td>
<td>0.839204</td>
<td>Default</td>
<td>0</td>
</tr>
</tbody>
</table>

5.3. Ridge Regression Prediction. Using Python’s sklearn.ensemble library, we can find the implementation of ridge regression prediction models (Table 6).
The two most essential parameters in the RRM are the regularization intensity (alpha) and computational solver (solver) (Table 7).

After the RRM with the optimal parameters is constructed, the prediction results are shown in Figures 14 and 15. After the training of the RRM, the fitted model can be output. The standardization process is performed in advance. Thus, the model has no intercept term, and each index coefficient represents the importance of the index (Figures 16 and 17).

5.4. Combination Forecasting Model. The weight coefficients of two models in the CFM can be obtained by the sum of residuals of RFM and RRM on the training set. The weight coefficients of RFM and RRM are $\lambda_1 = 0.793067$ and $\lambda_2 = 0.206933$, respectively. The prediction results are shown in Figures 18 and 19.

We use mean square error, mean absolute error, and goodness of fit ($R^2$) to test the prediction effect of three models (Tables 8 and 9).

Figures 10, 11, 14, 15, 18, and 19 show the prediction results of taxi demand in the Xi’anbei Railway Station and Bell Tower areas through by RRM, RFM, and CFM. Then, Tables 8 and 9 analyze the forecast effect of three forecasting methods. The tables indicate that CFM has the highest accuracy among the three models.
Table 6: Model parameter descriptions of the RRM.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alpha</td>
<td>It indicates the regularization strength; it is the complexity parameter of</td>
</tr>
<tr>
<td></td>
<td>the control coefficient shrinkage; a large value of α corresponds to a</td>
</tr>
<tr>
<td></td>
<td>large shrinkage; thus, the coefficient is robust to collinearity</td>
</tr>
<tr>
<td>fit_intercept</td>
<td>It indicates whether to calculate the intercept of this model</td>
</tr>
<tr>
<td>max_iter</td>
<td>It is the maximum number of iterations of the conjugate gradient solver</td>
</tr>
<tr>
<td>Solver</td>
<td>It is the solution method for calculation</td>
</tr>
</tbody>
</table>

Table 7: Model parameter description of the RRM.

<table>
<thead>
<tr>
<th>Hotspots</th>
<th>Alpha</th>
<th>Solver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xi’anbei Railway Station</td>
<td>1.724102</td>
<td>‘Saga’</td>
</tr>
<tr>
<td>Bell Tower</td>
<td>5.050931</td>
<td>‘Saga’</td>
</tr>
</tbody>
</table>
Figure 16: Index importance results of the RRM in the Xi’anbei Railway Station area.

Figure 17: Index importance results of the RRM in the Bell Tower area.

Figure 18: Prediction result of the CFM in the Xi’anbei Railway Station area.

Figure 19: Prediction result of the CFM in the Bell Tower area.
As shown in Figures 12 and 13, the most crucial factor in taxi demand is hours in the Xi’anbei Station because the station is a transport hub. This finding illustrates that taxi demand in a transport hub has a strong correlation to the time factor. Figures 12 and 13 also show that O₃ is the main factor in the Bell Tower. Ozone concentration is related to temperature, and hot weather increases the taxi demand in the commercial area. However, Figures 16 and 17 imply that the main factors of RRM in two areas are time factor and O₃. Differences between the two areas of RRM are less than those of RFM.

6. Conclusions

In this study, we investigated the taxi demand prediction in hotspots and then proposed three prediction models, namely, RFM, RRM, and CFM. We extracted hotspots of taxi demand, and the taxi demand prediction model was constructed on the basis of taxi demand hotspots. The proposed models combined time, meteorological, and environmental characteristics to explain the generation of taxi demand. The prediction results show that CFM has better robustness and smaller error than FRM and RRM in the Xi’anbei Railway Station area and the Bell Tower area. The experiment also indicates that taxi demand prediction is mainly affected by the time period in the Xi’anbei Railway Station. In the Bell Tower area, the importance of ozone concentration and temperature to the model is relatively advanced. The study concludes that the proposed model can improve prediction accuracy. The most important influencing factor of the taxi demand prediction model is the time factor. Temperature and weather indicators are also relatively important.

Some limitations in the research on taxi demand prediction still need to be addressed. For example, the impact of other similar types of traffic demand is ignored in this study. If travel demand can be met by an online car-hailing service, then taxi demand will be greatly reduced. This study also ignores the impact of land use properties on taxi demand, which will be one of our future research directions. Part of environmental features is challenging to obtain. Thus, we will propose a method to predict environmental features for predicting taxi demand more precisely in the future.

### References


of 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, November 2018.


