Hindawi
Journal of Advanced Transportation
Volume 2020, Article ID 1302586, 13 pages
https://doi.org/10.1155/2020/1302586

Research Article

Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots

Zhizhen Liu,[1] Hong Chen,[1] Yan Li,[2] and Qi Zhang[1]

[1]*School of Highway, Chang'an University, Xi'an 710000, China*
[2]*Shenzhen 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

Academic Editor: Gonçalo Homem de Almeida Correia

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 spatiotemporal 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 RFM 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 CFM 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 estimate 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 $x \in R^d$ and $y \in R$. The simplest linear regression model defines the loss function as the square of the residual. Then, the optimization objective is expressed as follows:

$$L = \|X\theta - y\|^2, \quad (1)$$

$\theta = \theta_1; \theta_2; \ldots; \theta_m$ 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

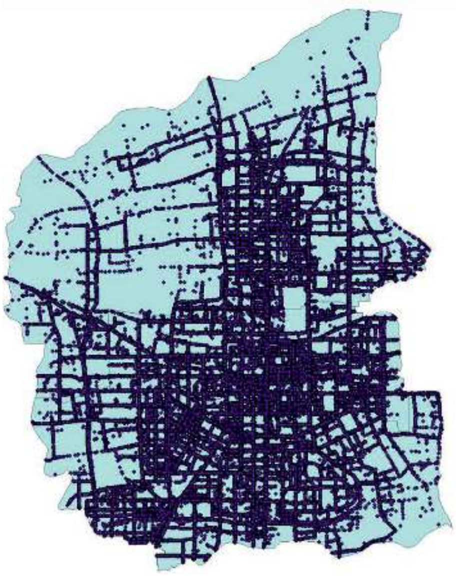


Figure 1: GPS data representation.

Table 1: Environmental data structure description.

Indicators	Description
AQI	Air quality index
CO	Concentration of CO (μg/m^3)
NO_2	Concentration of NO_2 (μg/m^3)
O_3	Concentration of O_3 (μg/m^3)
PM2.5	Concentration of PM2.5 (μg/m^3)
PM10	Concentration of PM10 (μg/m^3)
SO_2	Concentration of SO_2 (μg/m^3)
Air quality	1: excellent; 2: good; 3: mild pollution; 4: serious pollution
Weather	1: sunny; 2: cloudy; 3: raining; 4: haze
Wind speed	Wind speed (m/s)
TEM	Temperature (°C)
RHU	Humidity (%)
PRE	Precipitation (mm)

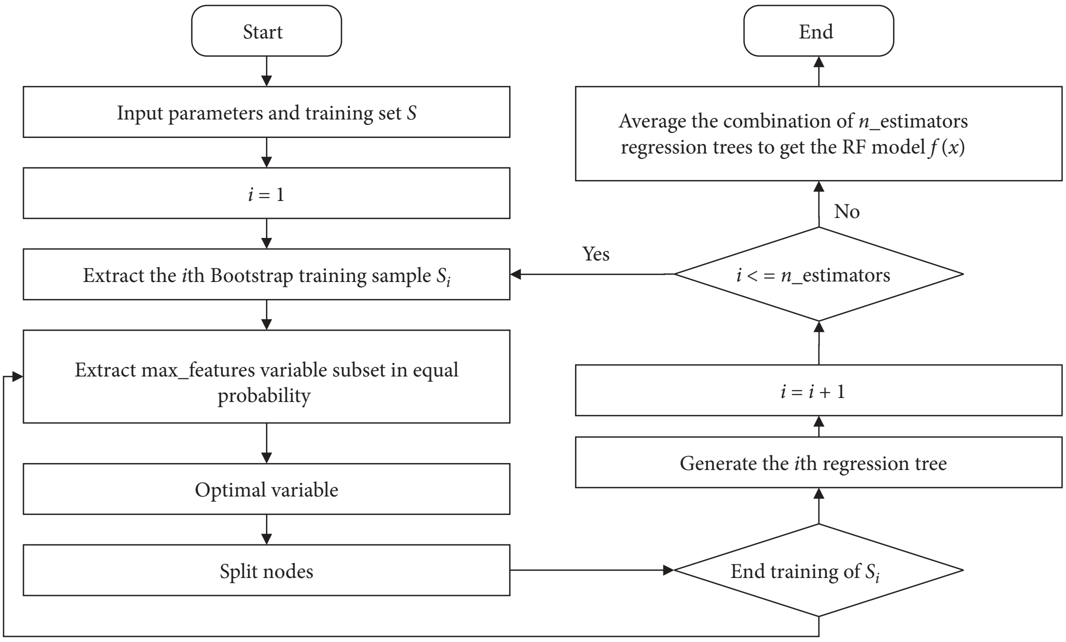

Start
Input parameters and training set S
$i = 1$
Extract the ith Bootstrap training sample S_i
Extract max_features variable subset in equal probability
Optimal variable
Split nodes
End training of S_i
Generate the ith regression tree
$i = i + 1$
$i < = n$_estimators
Yes
No
Average the combination of n_estimators regression trees to get the RF model $f(x)$
End

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_2 norm regularization is introduced into the RRM as follows:

$$L = \|X\theta - y\|^2 + \|\Gamma\theta\|^2. \tag{2}$$

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

$$L(\alpha) = \left(X^T X + \alpha I\right)^{-1} X^T y. \tag{3}$$

As α increases, the absolute values of the elements in $L(\alpha)$ tend to decrease, and the deviation of correct value θ_i increases. When α tends to infinity, $L(\alpha)$ tends to 0. The trajectory of $L(\alpha)$ that changes with α is called the ridge. When the ridge is stable, α 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

$$\widehat{y}_{\mathrm{CFM},i} = \lambda_1 \widehat{y}_{\mathrm{RRM},i} + \lambda_2 \widehat{y}_{\mathrm{RFM},i}, \tag{4}$$

where $\widehat{y}_{\mathrm{CFM},i}$ is the predicted value of the CFM, $\widehat{y}_{\mathrm{RRM},i}$ is the predicted value of the RRM, $\widehat{y}_{\mathrm{RFM},i}$ is the predicted value of the RFM, and λ_1 and λ_2 are the weight coefficients of RRM and RFM, respectively.

The core of the CFM is the determination of the weight coefficients λ_1 and λ_2. Inverse-variance weighting method is used to determine the weight coefficient of the CFM. The calculation equations are expressed as follows:

$$\lambda_1 = \frac{e_{\mathrm{RRM},k}(t)^{-1}}{e_{\mathrm{RMM},k}(t)^{-1} + e_{\mathrm{RFM},k}(t)^{-1}}, \tag{5}$$

$$\lambda_2 = 1 - \lambda_1. \tag{6}$$

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

$$e_{\mathrm{RRM}}(i) = \sum_{i=1}^{276} \left(y_i - \widehat{y}_{\mathrm{RRM},i}\right)^2, \tag{7}$$

$$e_{\mathrm{RFM}}(i) = \sum_{i=1}^{276} \left(y_i - \widehat{y}_{\mathrm{RFM},i}\right)^2, \tag{8}$$

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

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E\left[\left(X - \mu_X\right)\left(Y - \mu_X\right)\right]}{\sigma_X \sigma_Y}, \tag{9}$$

where $\operatorname{cov}(X,Y)$ represents the covariance between the variables X and Y, σ_X and σ_Y represent the standard deviations of the variables X and Y, and $\rho_{X,Y}$ represents the correlation coefficient of two continuous variables; the value of $\rho_{X,Y}$ is between -1 and 1. If $\rho_{X,Y} > 0$, then the two variables are positively correlated; if $\rho_{X,Y} < 0$, then the two variables are negatively correlated. A large absolute value of $\rho_{X,Y}$ 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

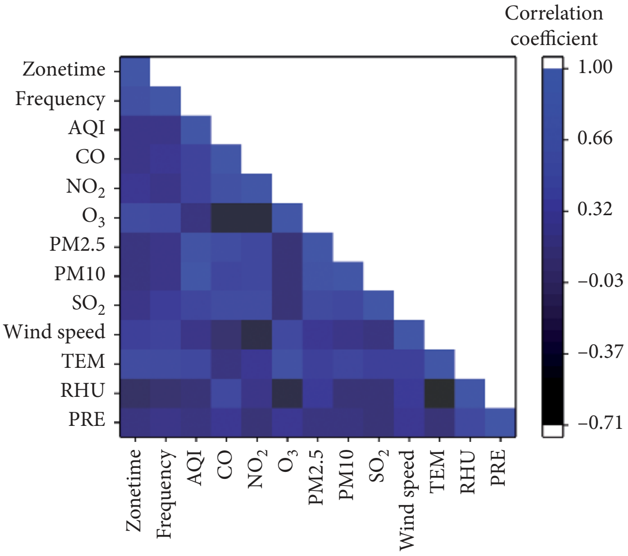
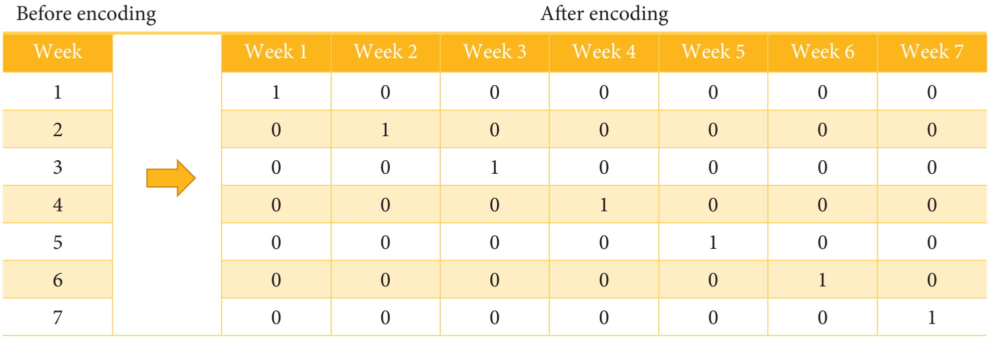


Figure 3: Correlation coefficient matrix.

Table 2: Added indicators description.

Indicator	Description
Hour	Represents the current time period {1~12 stands for 0:00~2:00, 2:00~4:00...22:00~24:00}
Wdy	Whether the day is a working day {"1" stands for weekdays, and "0" stands for nonworking days}
Week	The day of the representative is the day of the week {1~7 represents Monday to Sunday, respectively}
Holiday	Represents whether the day is a holiday (two holidays in April 2017: Ching Ming Festival and Labor day) {"0" stands for nonholiday, and "1" stands for holidays}


Before encoding

Week
1
2
3
4
5
6
7

After encoding

Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1

Figure 4: Example of the one-hot encoding process.

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).

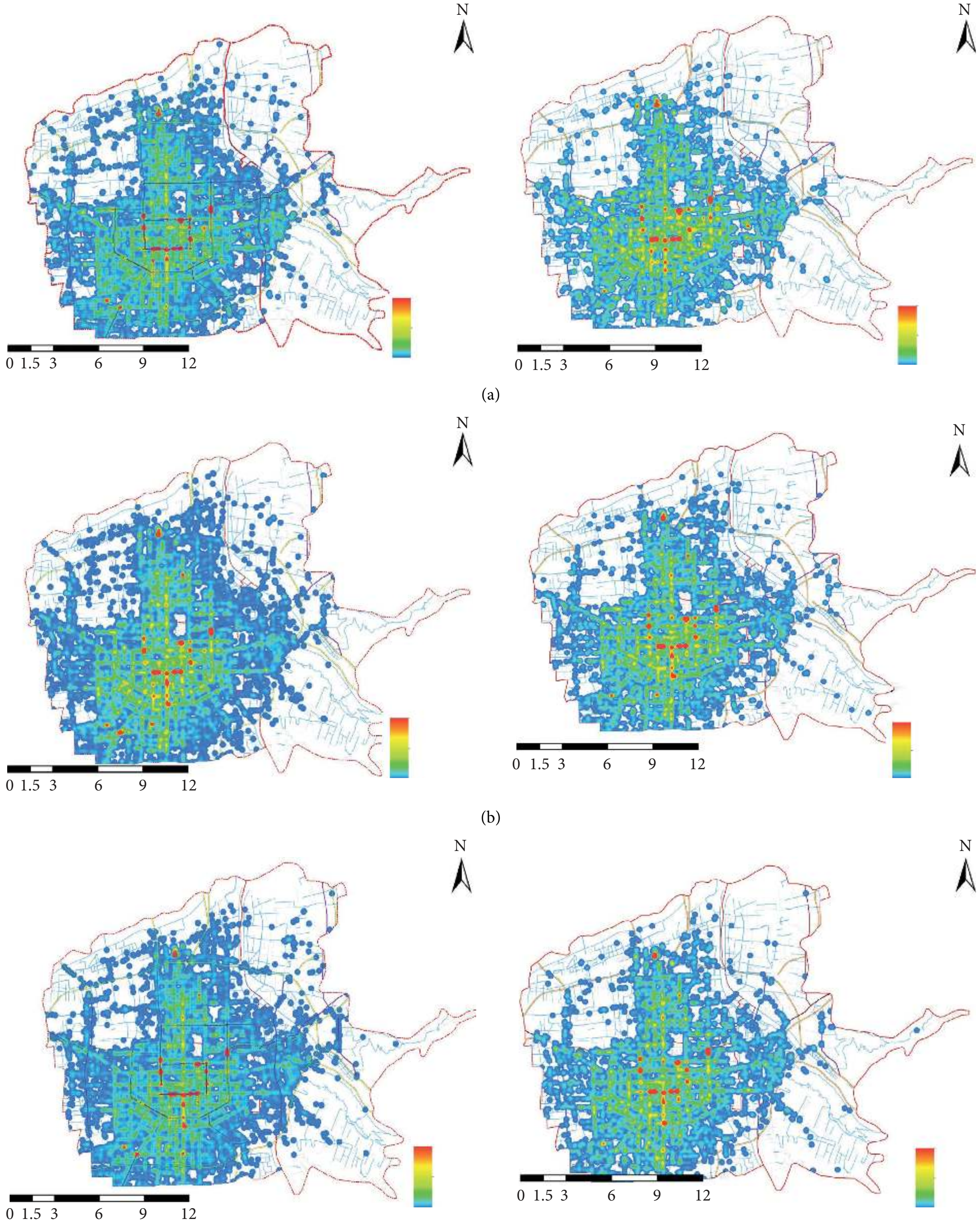
Hotspots are distributed in areas such as Xi'anbei Railway Station, Bell Tower, Xiaozhai, Railway Station, and City Library. Xi'anbei Railway Station and Railway Station are transportation hubs. Xiaozhai, City Library, and Bell Tower are commercial areas. In this study, two representative areas, namely, 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:

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} = 1 - \frac{\text{SSR}}{\text{SST}}, \tag{10}$$

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



(a)

(b)

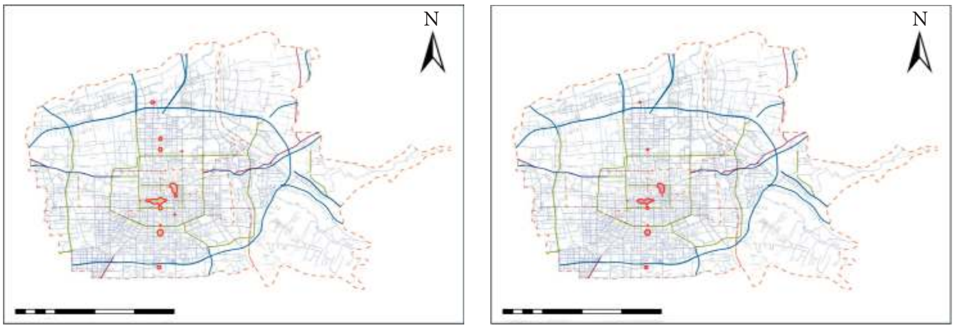
(c)

Figure 5: Thermogram of workday (left) and nonworkday (right) in peak hours: (a) in the morning; (b) at noon; (c) in the evening.

residuals, y_i is the value to be fitted, $\overline{y}$ is the mean of y, and $\widehat{y}_i$ is the fitted value.

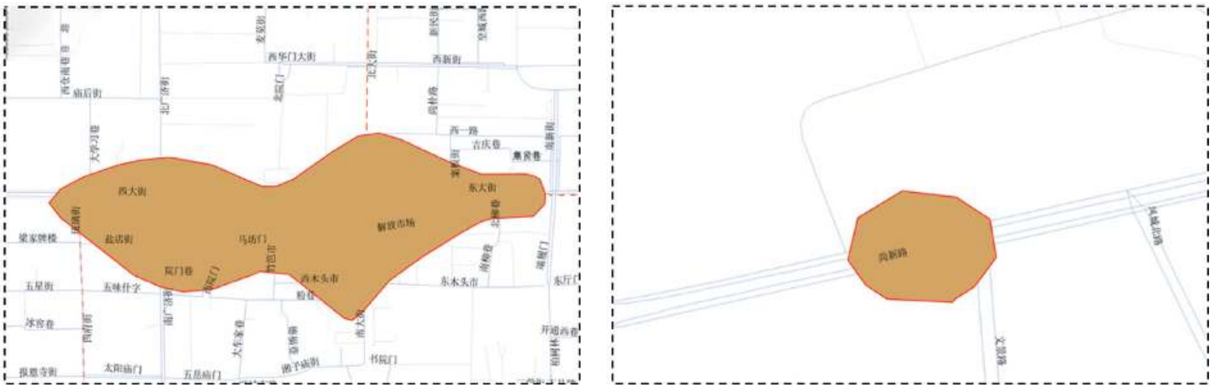
Considering the number of samples and training speed of RRM, we choose $[1 - 200]$ as variable span. The relation between "n_estimators" and R^2 can be calculated (Figures 8 and 9).

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



(a) (b)

Figure 6: Comparison of (a) workday and (b) nonworkday overlap thermogram.



(a) (b)

Figure 7: Hotspot selection: (a) description of hotspots in the Bell Tower; (b) description of hotspots in the Xi’anbei Railway Station.

Table 3: Random forest regression parameter description.

Parameters	Introduction
*n*_estimators	Number of submodules
Criterion	Method to judge whether the node continues to split
max_features	Maximum number of features involved in judging when node splits
max_depth	Maximum depth
min_samples_split	Minimum number of samples required for splitting
min_samples_leaf	Leaf node minimum sample number
min_weight_fraction_leaf	Minimum total sample weight of leaf nodes
max_leaf_nodes	Maximum number of leaf nodes

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. OBB_i represents the out-of-bag data of the ith tree. The out-of-bag mean square deviations of each tree are $\mathrm{MSE}_{\mathrm{OOB}_1}, \mathrm{MSE}_{\mathrm{OOB}_2}, \ldots, \mathrm{MSE}_{\mathrm{OOB}_\mathrm{M}}$.

(2) We assume that the total number of variables is N. For each input variable X_i, 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:

$$\begin{bmatrix} \mathrm{MSE}_{1,\mathrm{OOB}_1} & \mathrm{MSE}_{1,\mathrm{OOB}_2} & \cdots & \mathrm{MSE}_{1,\mathrm{OOB}_M} \\ \mathrm{MSE}_{2,\mathrm{OOB}_1} & \mathrm{MSE}_{2,\mathrm{OOB}_2} & \cdots & \mathrm{MSE}_{2,\mathrm{OOB}_M} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{MSE}_{N,\mathrm{OOB}_1} & \mathrm{MSE}_{N,\mathrm{OOB}_2} & \cdots & \mathrm{MSE}_{N,\mathrm{OOB}_M} \end{bmatrix}. \tag{11}$$

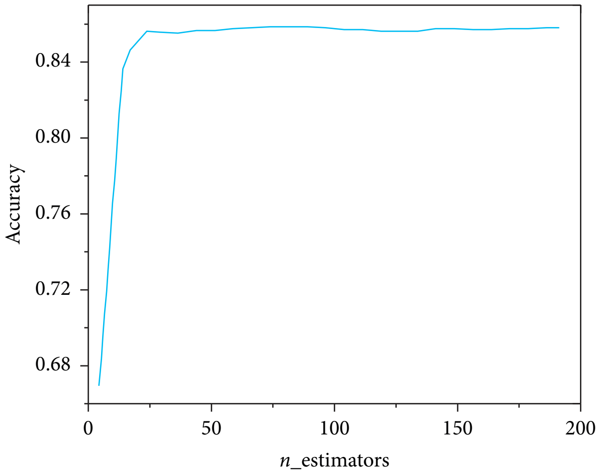


Figure 8: R^2 changes as "*n*_estimators" increases in the Xi'anbei Railway Station area.

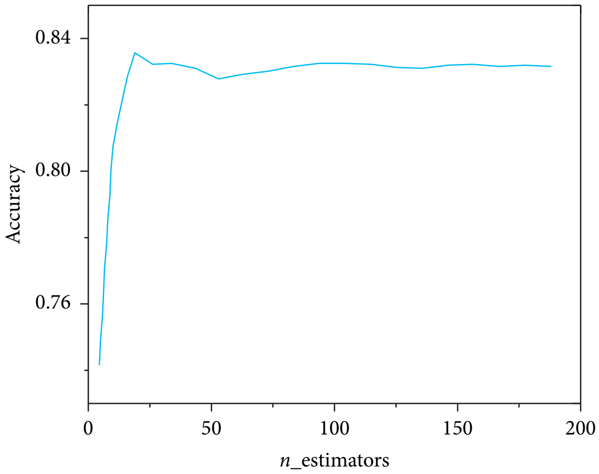


Figure 9: R^2 changes as "*n*_estimators" increases in the Bell Tower area.

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

Parameter	Accuracy before adjustment	Accuracy after adjustment	Value	Increase in accuracy
*n*_estimators	0.857206	0.858649	80	0.001443
max_features	0.858649	0.884258	4	0.025609
max_depth	0.884258	0.884258	Default	0
min_samples_leaf	0.884258	0.8842589	Default	0

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

Parameter	Accuracy before adjustment	Accuracy after adjustment	Value	Increase in accuracy
*n*_estimators	0.825876	0.836657	16	0.010781
max_features	0.836657	0.839204	28	0.002547
max_depth	0.839204	0.839204	Default	0
min_samples_leaf	0.839204	0.839204	Default	0

(3) The out-of-bag error $\mathrm{MSE}_{\mathrm{OOB}_1}, \mathrm{MSE}_{\mathrm{OOB}_2}, \ldots, \mathrm{MSE}_{\mathrm{OOB}_M}$ before replacement is subtracted with the ith 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:

$$\mathrm{VIM}_i = \frac{1}{M}\sum_{j=1}^{M}\left(\mathrm{MSE}_{\mathrm{OOB}_j} - \mathrm{MSE}_{i,\mathrm{OOB}_j}\right) \quad 1 \le i \le N. \tag{12}$$

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).

5.3. Ridge Regression Prediction. Using Python's sklearn.ensemble library, we can find the implementation of ridge regression prediction models (Table 6).

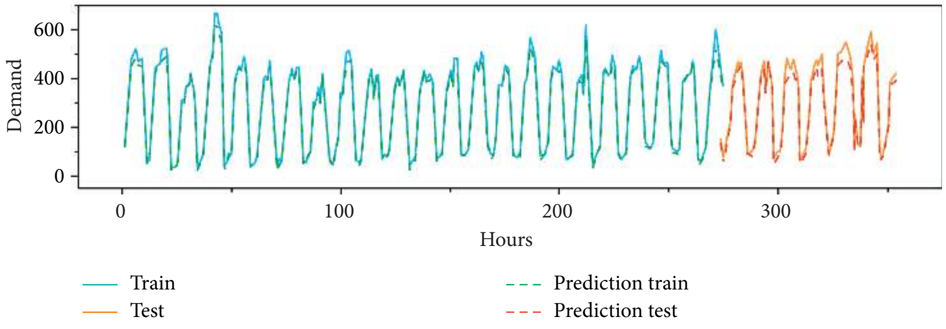


Figure 10: Prediction result of RFM in the Xi’anbei Railway Station area.

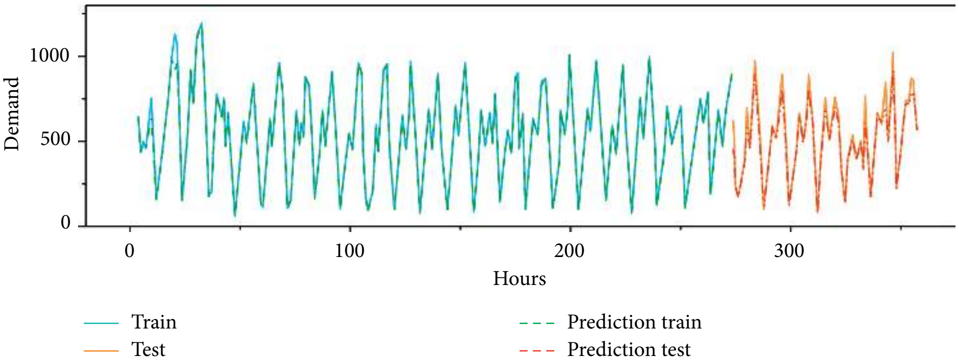


Figure 11: Prediction result of RFM in the Bell Tower area.

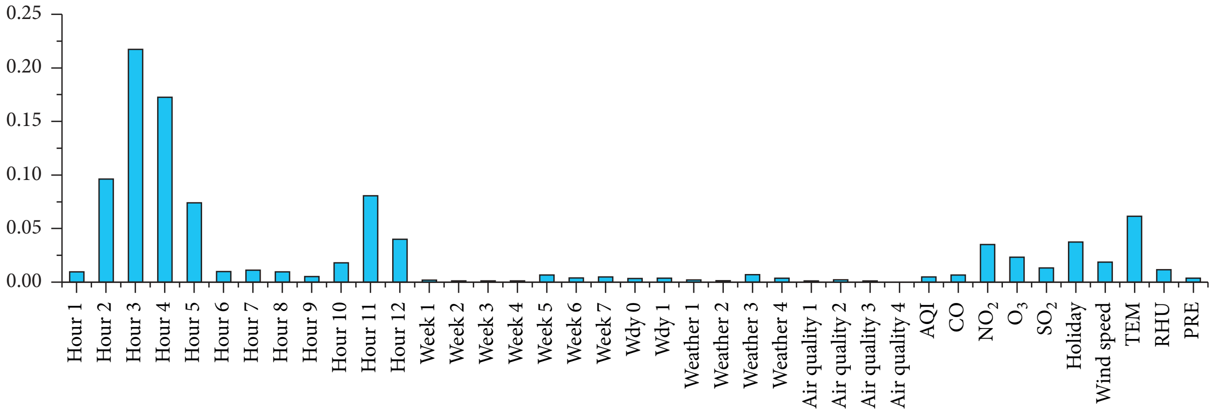


Figure 12: Index importance results of RFM in the Xi’anbei Railway Station area.

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 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.

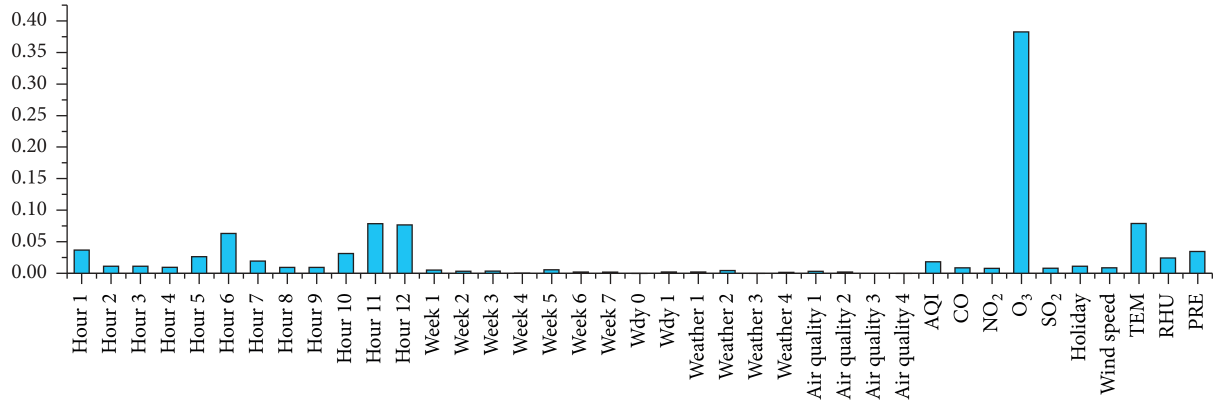


FIGURE 13: Index importance results of RFM in the Bell Tower area.

TABLE 6: Model parameter descriptions of the RRM.

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

TABLE 7: Model parameter description of the RRM.

Hotspots	Alpha	Solver
Xi'anbei Railway Station	1.724102	'Saga'
Bell Tower	5.050931	'Saga'

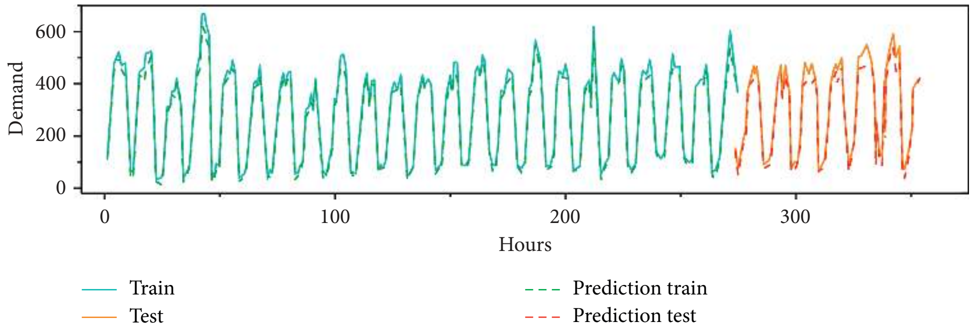


FIGURE 14: Prediction result of the RRM in the Xi'anbei Railway Station area.

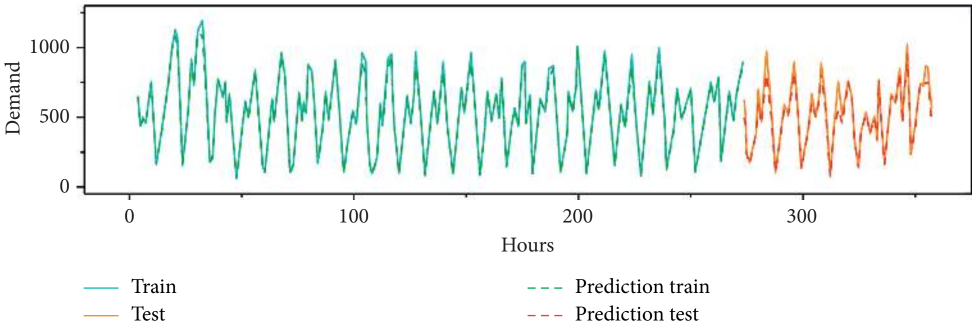


FIGURE 15: Prediction result of the RRM in the Bell Tower area.

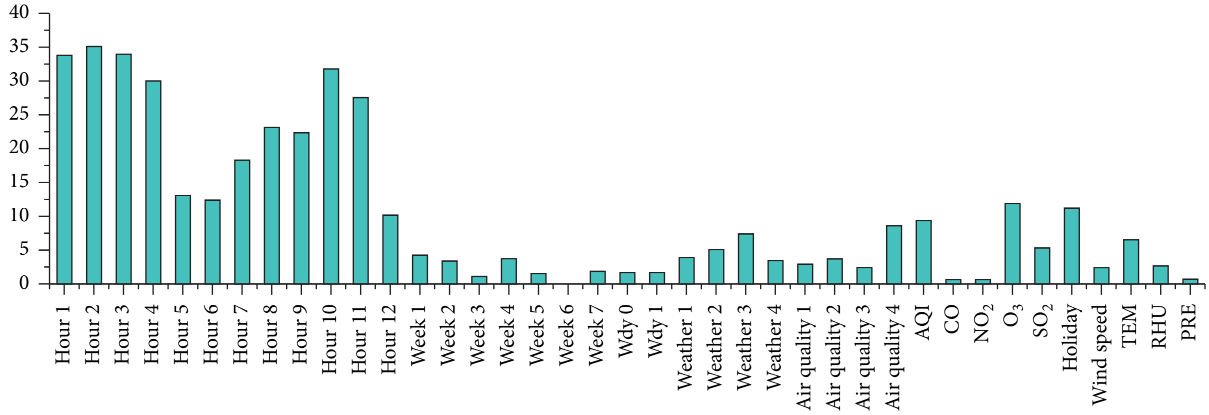


Figure 16: Index importance results of the RRM in the Xi'anbei Railway Station area.

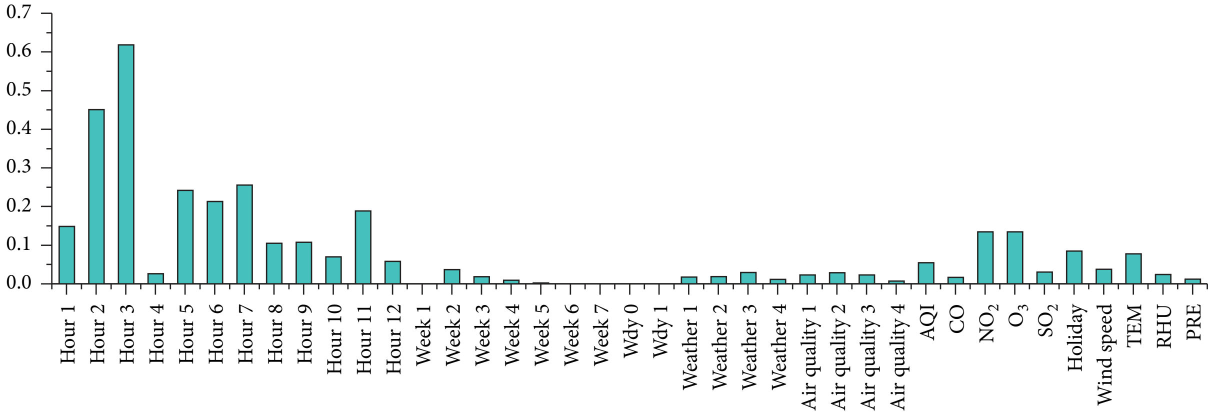


Figure 17: Index importance results of the RRM in the Bell Tower area.

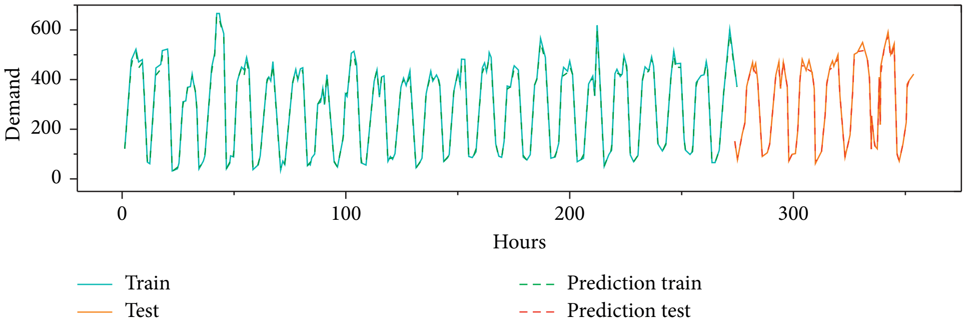


Figure 18: Prediction result of the CFM in the Xi'anbei Railway Station area.

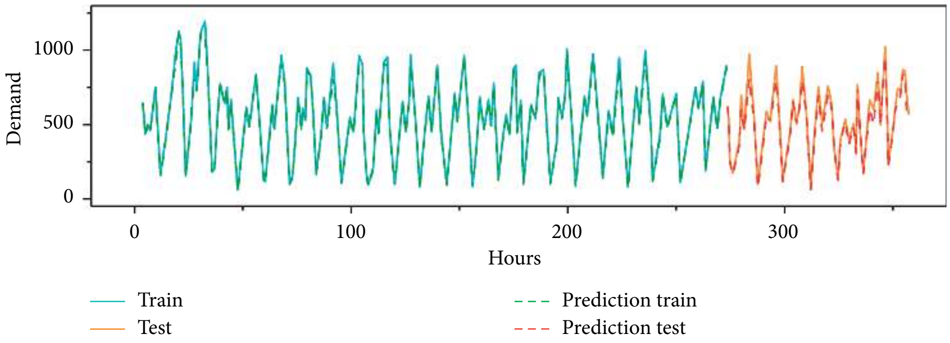


Figure 19: Prediction result of the CFM in the Bell Tower area.

Table 8: Comparison of three models in the Xi'anbei Railway Station.

Model	MSE	MAE	R^2
RRM	2448.86	35.34	0.854136
RFM	2205.06	30.59	0.864258
CFM	2025.67	28.33	0.885365

Table 9: Comparison of three models in the Bell Tower.

Model	MSE	MAE	R^2
RRM	9076.01	67.52	0.749921
RFM	5339.22	47.21	0.829204
CFM	5143.12	42.41	0.835642

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 O_3 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 O_3. 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.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This study was supported jointly by the Technology Project of Shaanxi Transportation Department (grant number 15-39R) and Special Fund for Basic Scientific Research of Central Colleges of Chang'an University (grant number 300102218409).

References

[1] H. Yang, K. I. Wong, and S. C. Wong, "Modeling urban taxi services in road networks: progress, problem and prospect," *Journal of Advanced Transportation*, vol. 35, no. 3, pp. 237–258, 2001.

[2] S. Wong and B. M. Williams, "Adaptive seasonal time series models for forecasting short-term traffic flow," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2024, no. 1, pp. 116–125, 2007.

[3] F. Miao, S. Han, S. Lin et al., "Taxi dispatch with real-time sensing data in metropolitan areas: a receding horizon control approach," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 463–478, 2016.

[4] S. Zhang, J. Tang, H. Wang, Y. Wang, and S. An, "Revealing intra-urban travel patterns and service ranges from taxi trajectories," *Journal of Transport Geography*, vol. 61, pp. 72–86, 2017.

[5] I. Markou, K. Kaiser, and F. C. Pereira, "Predicting taxi demand hotspots using automated internet search queries," *Transportation Research Part C: Emerging Technologies*, vol. 102, pp. 73–86, 2019.

[6] Y. Bao, Y. Sun, X. Bu et al., "How do metro station crowd flows influence the taxi demand based on deep spatial-temporal network?" in *Proceedings of 2018 14th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, Shenyang, China, December 2018.

[7] A. Gholami and A. S. Mohaymany, "Analogy of fixed route shared taxi (taxi khattee) and bus services under various demand density and economical conditions," *Journal of Advanced Transportation*, vol. 46, no. 2, pp. 177–187, 2012.

[8] J. Gui and Q. Wu, "Taxi efficiency measurements based on motorcade-sharing model: evidence from GPS-equipped taxi data in sanya," *Journal of Advanced Transportation*, vol. 2018, Article ID 4360516, 10 pages, 2018.

[9] H. W. Chang, Y. C. Tai, and J. Y. J. Hsu, "Context-aware taxi demand hotspots prediction," *International Journal of Business Intelligence and Data Mining*, vol. 5, no. 1, pp. 3–18, 2010.

[10] X. Hu, S. An, and J. Wang, "Taxi driver's operation behavior and passengers' demand analysis based on GPS data," *Journal of Advanced Transportation*, vol. 2018, Article ID 6197549, 11 pages, 2018.

[11] I. Markou, F. Rodrigues, and F. C. Pereira, "Real-time taxi demand prediction using data from the web," in *Proceedings*

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

[12] S. Ishiguro, S. Kawasaki, and Y. Fukazawa, "Taxi demand forecast using real-time population generated from cellular networks," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers-UbiComp'18*, Singapore, October 2018.

[13] K. Zhao, D. Khryashchev, J. Freire, C. T. Silva, and H. T. Vo, "Predicting taxi demand at high spatial resolution: approaching the limit of predictability," in *Proceedings of 2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, December 2016.

[14] L. Kattan, A. De Barros, and S. C. Wirasinghe, "Analysis of work trips made by taxi in canadian cities," *Journal of Advanced Transportation*, vol. 44, no. 1, pp. 11–18, 2010.

[15] L. Kuang, X. Yan, X. Tan, S. Li, and X. Yang, "Predicting taxi demand based on 3D convolutional neural network and multi-task learning," *Remote Sensing*, vol. 11, no. 11, p. 1265, 2019.

[16] X. Liu, L. Sun, Q. Sun, and G. Gao, "Spatial variation of taxi demand using GPS trajectories and POI data," *Journal of Advanced Transportation*, vol. 2020, Article ID 7621576, 20 pages, 2020.

[17] 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, 2013.

[18] F. Rodrigues, I. Markou, and F. C. Pereira, "Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach," *Information Fusion*, vol. 49, pp. 120–129, 2019.

[19] J. Xu, R. Rahmatizadeh, L. Boloni, and D. Turgut, "Real-time prediction of taxi demand using recurrent neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 8, pp. 2572–2581, 2018.

[20] Y. Yang, X. Wang, Y. Xu, and Q. Huang, "Multiagent reinforcement learning-based taxi predispatching model to balance taxi supply and demand," *Journal of Advanced Transportation*, vol. 2020, Article ID 8674512, 12 pages, 2020.

[21] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic, "Taxi-passenger-demand modeling based on big data from a roving sensor network," *IEEE Transactions on Big Data*, vol. 3, no. 3, pp. 362–374, 2017.

[22] K. Zhang, Z. Feng, S. Chen, K. Huang, and G. Wang, "A framework for passengers demand prediction and recommendation," in *Proceedings of 2016 IEEE International Conference on Services Computing (SCC)*, San Francisco, CA, USA, June 2016.

[23] W. Zhu, J. Lu, and Y. Yang, "A pick-up points recommendation system for ridesourcing service," *Sustainability*, vol. 11, no. 4, p. 1097, 2019.

[24] J. Klepsch, C. Klüppelberg, and T. Wei, "Prediction of functional ARMA processes with an application to traffic data," *Econometrics and Statistics*, vol. 1, pp. 128–149, 2017.

[25] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results," *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[26] J. A. Alvarez-Garcia, J. A. Ortega, L. Gonzalez-Abril, and F. Velasco, "Trip destination prediction based on past GPS log using a Hidden Markov model," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8166–8171, 2010.

[27] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2015.

[28] N. Davis, G. Raina, and K. Jagannathan, "A multi-level clustering approach for forecasting taxi travel demand," in *Proceedings of 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, Rio de Janeiro, Brazil, November 2016.

[29] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.

[30] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," *Sensors*, vol. 17, no. 7, p. 1501, 2017.

[31] J. Ou, J. Xia, Y.-J. Wu, and W. Rao, "Short-term traffic flow forecasting for urban roads using data-driven feature selection strategy and bias-corrected random forests," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2645, no. 1, pp. 157–167, 2017.

[32] H. Yao, F. Wu, J. Ke et al., "Deep multi-view spatial-temporal network for taxi demand prediction," in *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, LA, USA, February 2018.

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

[34] U. Grömping, "Variable importance assessment in regression: linear regression versus random forest," *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.

[35] M. Ristin, M. Guillaumin, J. Gall, and L. Van Gool, "Incremental learning of random forests for large-scale image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 490–503, 2016.