Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3504704, 25 pages
https://doi.org/10.1155/2022/3504704

*Research Article*

# XGBoost-Based Travel Time Prediction between Bus Stations and Analysis of Influencing Factors

**Lingxiang Zhu**,[1] **Sisi Shu**,[2] **and Liang Zou**[2]

[1]*College of Mathematics and Informatics, South China Agricultural University, Guangzhou 510642, China*
[2]*College of Civil and Transportation Engineering, Shenzhen University, Shenzhen 518060, China*

Correspondence should be addressed to Liang Zou; zouliang@szu.edu.cn

Received 19 January 2022; Revised 13 June 2022; Accepted 21 June 2022; Published 27 July 2022

Academic Editor: Manuel Fernandez-Veiga



Real-time and accurate travel time information between bus stations is critical for passengers to make suitable travel plans to reduce waiting time at the stops. By mining and analyzing bus operational data, it can be obtained that factors such as the variation of vehicle speed in adjacent sections and the proportion of bus lanes between stations have affected the travel time between bus stations. Therefore, considering the temporal feature, spatial feature, and weather feature as the prediction model's input, travel time between bus stations prediction model based on eXtreme Gradient Boosting (XGBoost) was trained and established. The 28-day bus operation data of a certain bus line in Guangzhou was used for training and verification, and they were compared with the prediction models based on *K*-Nearest Neighbors (KNN), BP neural network, and Light Gradient Boosting Machine (LightGBM). In comparison with other models, the lowest MAPE of 11.96% was found for the XGBoost prediction model, which is 9.30% lower than other models on average. The sensitivity analysis of the proposed prediction model was further conducted: temporally, the accuracy of the prediction model was best during the flat peak hours; spatially, the MAPE of the model gradually decreased as the number of line units increased, and when the number of line units exceeded 18, the accuracy of the prediction model stabilized and was lower than 7%. The results confirm that the XGBoost model outperforms the KNN, BP, and LightGBM in terms of fitting, accuracy, and stability.

## 1. Introduction

For passengers, knowing arrival times of public vehicles in advance can reduce waiting times which is one of the most concerned questions. Compared with urban rail transit, the unreliability of conventional bus is mainly reflected in the uncertainty of arrival time, that is, the uncertainty of travel time between stations. Passengers may spend too much time waiting for the bus. A survey shows that more than 94% of passengers shave encountered a situation of waiting for a bus for too long. Zhang et al. found that the tolerable waiting time for passengers is actually 4.62 minutes. However, the 2020 Traffic Analysis Report of China's Major Cities jointly issued by AutoNavi Maps and other authoritative organizations shows that the average waiting time during peak hours in many cities such as Beijing, Shanghai, Guangzhou, and Shenzhen is no less than 9.8 minutes. Such overly long waiting time has greatly affected the travel experience, reduced the quality of bus service, and caused a certain waste of resources. With the continuous development of intelligent public transportation systems, people can now obtain real-time operating status information of buses through multiple channels such as electronic stop signs and smart phones. However, there are many negative comments emerging from the evaluation of relevant tools. For example, *Chelaile* and *MyBus* are two real-time bus query software with high downloads in the IOS system, and their negative comments takes 18.8% and 21.5%, respectively, of the total comments each. Among that, about 70% of these negative comments refer to the unpunctuality of bus arriving. Therefore, it can be shown that accurate prediction of travel time between bus stations remains a challenging issue. Accurately predicting the travel time between conventional bus stations can not only support bus operators to formulate scheduling plans, optimize operating routes, and allocate operating vehicles but also reduce unnecessary waiting time for

passengers, increase bus travel sharing ratio, thereby ease urban congestion and road traffic pollution, and promote green traffic development.

To address the problem of travel time prediction between bus stations, on the basis of existing research, this article proposes the key considerations from two aspects of time and space and uses parameters such as the maximum information coefficient (MIC) to deeply explore the correlation between influencing factors and travel time between stations, then further establishes a XGBoost-based travel time prediction model. Finally, using one-month bus operation data of a certain bus line in Guangzhou, China, for training and verification, the results show that the prediction model proposed in this paper outperforms the existing travel time prediction model between major bus stations in terms of prediction accuracy.

The main contributions of this paper are as follows:

Firstly, using examples of bus operation, the impact factors such as the state of the bus of the previous shift, the proportion of bus lanes between stations and other factors on the travel time between bus stations are analyzed in depth from the perspective of qualitative and quantitative. Secondly, combined with the above analysis, the feature construction for the influencing factors was carried out; the XGBoost-based bus station travel time prediction model was constructed and tested on the real data sets.

The rest of this paper is organized as follows: Section 2 discusses literature reviews, Section 3 explains the related concepts of maximum information coefficient and XGBoost, Section 4 describes data information and the analysis of influencing factors, and the experimental results are summarized in Sections 5 and 6.

## 2. Literature Review

*2.1. Research on Bus Travel Time Prediction.* With the rapid development of advanced bus systems, scholars from home and abroad have done a lot of researches on the prediction of travel time. The existing research methods can be divided into two categories: one is the prediction method based on statistics, and the other is based on machine learning. A review and the comparison of existing studies are provided below.

*2.1.1. Statistical Prediction Method.* The commonly used methods in statistical forecasting mainly include time series and regression analysis. The time series method analyzes the input in chronological order, researches, and predicts the output. Wang et al. used ARIMA to determine the influence weights of the travel time of the first few buses and improve the accuracy of the forecast of the travel time of the bus [1]. Tong et al. decomposed the nonstationary time series into several linear combinations of stationary time series and then used exponential smoothing model to predict the travel time between bus stations [2]. The regression analysis method is to analyze the relationship between multiple variables and establish a mathematical function formed by independent variables. According to the linear speed change shown by the relationship between bus speed and traffic density, Zhang et al. proposed a bus arrival time prediction model considering upstream signalized intersection and surrounding traffic flow [3]. Zhou et al. used the Kalman Filter algorithm to establish a prediction model of bus travel time [4]. By using the prediction model based on Kalman Filter, Ma et al. considered using road attributes and POI (Point of Interest) to divide similar sections to improve the prediction accuracy [5].

*2.1.2. Machine Learning Prediction Methods.* With the development of computer and communication technology, the acquisition and transmission of bus real-time operating status data have been basically achieved. Taking Shenzhen, China, as an example, more than 100 million vehicle GPS data and bus IC card swiping data are collected in one single day, and the data storage capacity exceeds 16G. In the face of these large and complex datasets, machine learning is an example of a data-driven method which is aimed at increasing efficiency and accuracy of the prediction.

Many scholars try to use different algorithms to improve the performance of bus $c$ time prediction model. For example, Huo et al. established a prediction model of bus arrival time based on the KNN algorithm [6]. Xie [7] and Han et al. [8] used a BP neural network which has better learning ability to construct bus travel time prediction models. He et al. proposed a prediction model based on LSTM (Long Short-Term Memory) that could solve the problem of long-term data dependence [9]. Yu et al. predicted the arrival time of buses based on SVM (Support Vector Machine) to avoid local optimal problems [10]. They also introduced a forgetting factor to assign different weights to time-varying feature data and used Grubbs test to filter out outliers in the data [11]. Jing et al. used a bus arrival time prediction model based on GBDT (Gradient Boosting Decision Tree), which has higher computational speed [12].

From the analysis results of the above studies, we can observe that time period, weather conditions, and distance between stations are the three most frequently used factors in the research of the factors affecting the travel time between bus stations. Meanwhile, whether it is a working day or a nonworking day, the number of signal lights and the historical travel time between stations are also more frequently used factors. Moreover, the travel time between stations in the previous section, traffic flow, and the number of passengers getting on and off are also investigated and discussed as impact factors that affect the travel time between stations.

*2.2. Research on XGBoost Model.* Compared with other prediction models, the XGBoost model has considerable advantages in terms of both generating predictions with higher accuracy and improving efficiency while reducing the occurrence of over-fitting. Therefore, scholars all around the world selected the XGBoost model in prediction researches in different fields:

In the prediction research of commodity sales, Xie et al. established the XGBoost prediction model for the application of housing rent prediction [13]. In the field of security

warning, Chen et al. used the XGBoost model to predict specific faults of power system [14]. In the field of medical treatment, Jia et al. constructed the XGBoost model based on clinical data to predict the prognosis quality score of fracture surgery [15]. In the prediction research of air quality, Zhang et al. established XGBoost model to predict haze concentration [16]. Such wide application of XGBoost model reflects its own superiority in prediction. And its analysis method can also be used for reference in the forecast research applied in the application domain of transportation. For example, Zhong et al. established the XGBoost prediction model to predict short-term traffic flow [17]. The application of XGBoost in traffic flow prediction opens up new ideas for improving the accuracy of travel time prediction between conventional bus stations. Zou et al. propose an ensemble tree method XGBoost to predict passenger flow of bus routes, taking the number of routes and the number of buses during the predicted interval into the model to improve the accuracy [18]. Dong et al. proposed a traffic flow prediction model combining wavelet decomposition and reconstruction with the eXtreme Gradient Boosting (XGBoost) algorithm [19]. Du et al. used the combined model of XGBoost and LSTM in the short-term traffic prediction of the base station [20]. Yun et al. built a local optimal fusion model based on LSTM, LightGBM, and dynamic regression device [21]. Wang et al. took Multivariable Linear Regression (MLR), $K$-Nearest Neighbor (KNN), XGBoost, and Gated Recurrent Unit (GRU) as four seed models to establish a regression integration model to accurately predict short-term passenger flows of urban public transport [22].

By summarizing the existing literature, it is found that scholars commonly take the travel time between bus stops of the previous bus work shift as the reference value to predict the current one; however, in-depth analysis on the correlation between the running state of the bus of the previous and the current bus work shift is of lack. Therefore, this paper makes a comparative analysis on the speed changes of the previous and current bus work shifts. Besides, bus lanes play a significant role in improving bus operation efficiency during peak hours. However, few people analyze the impact of bus lanes on bus travel time prediction which is discussed in this paper. And the real-time weather conditions frequently used as the input of the model have seldom been studied for its relativity, and this paper also focuses on its analysis.

At present, many first-tier cities have achieved the real-time acquisition function of public transport information, which generates massive amount of public transport data. Faced with huge and complex data, the XGBoost model outperforms other models in regression prediction. Therefore, this paper focuses on analyzing the respective correlations between the changes in the speed of the bus work shifts in the adjacent sections, the proportion of the bus lanes in the sections, the real-time weather, and the travel time between the bus stops. Based on this, feature engineering is constructed, and a XGBoost-based travel time prediction model between conventional bus stops is established.

## 3. Methodology

*3.1. Maximum Information Coefficient.* Maximum information coefficient is a method improved by Reshef et al. based on mutual information (MI) to measure the correlation between attributes [23], and it is an effective way to analyze the degree of correlation between variables. It can reflect the complex relationships between variables, such as linear relationships, nonlinear relationships, and nonfunctional relationships [24].

MI has some problems in feature selection, such as variables usually needed to be discretized, or cannot be normalized, and its calculations are not convenient enough. MIC overcomes these problems and can accurately calculate the degree of correlation between two variables even when the sample data is large. The universality of MIC shows that the functional relationship between variables can be found, whether it is linear or nonlinear; the fairness of MIC is shown as its ability to obtain the same results for the same level of noise existing in different forms of functions.

When using MIC analyzes the correlation between variables, first, grid the given scatter plot composed of variables $X$ and $Y$; second, calculate the corresponding maximum information value, and then normalize the obtained maximum information value with the range of [0,1]. Finally, the grid resolution that maximizes normalized mutual information is the MIC value. The greater the MIC value, the stronger the correlation of the variables. When the MIC value is 1, it indicates that the two variables have a strictly determined relationship and are not limited to the functional form. The smaller the MIC value, the weaker the correlation of variables. When MIC equals 0, it means that the two variables are completely independent.

*3.2. XGBoost.* XGBoost (eXtreme Gradient Boosting) is an improved learning algorithm based on the Gradient Boosting algorithm and Decision Tree. Its principle is to transform a large number of weak classifiers into strong classifiers by using the idea of iterative operation, so as to achieve accurate classification effect.

XGBoost is a highly efficient implementation of GBDT [25]. There are three main differences between XGBoost and GBDT. Firstly, GBDT only supports Decision Trees, while XGBoost also supports many other weak learners, such as gbtrees (General Balanced Trees), gblinear, and dart. Secondly, compared with GBDT, the target loss function of XGBoost increases the regular term. Third, GBDT's loss function only performs negative gradient (first-order Taylor) expansion on the error part, while XGBoost's loss function performs second-order Taylor expansion on the error part, which improves the accuracy of model prediction.

XGBoost is a classic method in Boosting. Boosting is aimed at building a strong classifier by integrating many weak classifiers, with XGBoost using CART (Classification and Regression Trees) [26]. The idea of XGBoost algorithm is to continuously add trees and continuously perform feature splitting to grow a tree. Through each addition, a new function can be learned to fit the residual of previous prediction. After training $K$ tree, when it is necessary to predict the

score of a sample, each tree will get the score of each child node according to the characteristics of the sample, and finally adding up the corresponding score of each tree is the predicted value of the sample.

## 4. Data and Influencing Factors

### 4.1. The Source of Data

*4.1.1. Bus Operation Data.* The bus operation data includes the information of all buses entering and leaving the station of a certain bus line in Guangzhou from April 24 to May 21, with a total of 7,532 bus work shifts and 291,800 entries and exits. The bus line consists of 25 bus stations: the origin stations in Baiyun District and the terminal stations in Haizhu District. The total length of the line is 18.3 kilometers. According to the information of bus pulling in and leaving the stations, the travel time of each bus in each work shift between stations can be calculated. There are a total of 300 station combinations and 1,604,144 data. Figure 1 shows the average travel speed distribution of the whole bus line.

It can be seen from the figure above that most of the average speed value gather around about 4 m/s, while some data deviated from it a lot. The reason for this phenomenon may be that the travel speed is significantly reduced as a consequence of the excessively heavy congestion during peak hours, or the volume of traffic flow and passenger flow on the road shrunk in the early morning or late night, so the travel speed between stations is accordingly higher.

*4.1.2. Bus Lanes.* The bus lane data is obtained from Baidu Street View map observation, which is the ratio of the length of the bus lane between two adjacent stations to the distance between them. The selected route has about 3.2 kilometers in total of special lanes, accounting for about 17.5% of the total bus route length. After calculating the proportion of the length of the bus lanes between two adjacent stations, the distribution of the proportions is shown in Figure 2 (the bus lanes of the selected line are distributed in the second half of the line).

In addition, the line signal light data is obtained from on-site observations, including the number of signal lights between stations and the corresponding turning information of the bus. The line has a total of 24 signal lights, containing 22 go straight, 2 right turns, and no left turns.

*4.1.3. Weather.* The weather data comes from the agricultural meteorological big data system-WheatA, which can provide time-by-time historical data of various meteorological indicators in various meteorological observatories across the country. The weather data specifically includes 672 pieces of information on the hourly temperature, rainfall, and wind speed in Guangzhou from April 24 to May 15.

Figure 3(a) shows the average temperature distribution in each hour of a month. As can be seen from the figure, the temperature is mainly distributed between 22°C and 31°C, which is basically in line with the weather and climate of Guangdong. In the morning and night, the temperature is low, while in the middle of the day, the temperature gets higher. Figure 3(b) shows the distribution probability of rainfall intensity per hour in a month. When the rainfall rate is less than 0.25 mm/h, the rainfall intensity is evaluated as sporadic light rain; when the rainfall rate is from 0.25 mm/h to 1.0 mm/h, the intensity would be light rain; when the rainfall rate is greater than 1.0 mm/h, the intensity would be moderate rain. The rainfall intensity in 28 days is only sporadic light rain or light rain, and most of them are sporadic light rain. Figure 3(c) shows the distribution of hourly wind speed in a month. When the wind speed is less than 0.2 m/s, the wind force is 0 scale, which is regarded as no wind. When the wind speed is greater than 0.3 m/s and less than 5.4 m/s, the wind-force is 1~3 scale; when the wind speed is more than 5.5 m/s and less than 10.7 m/s, the wind is 4~5 scale. Therefore, within 28 days, the wind is not over 5 scale, basically 1~3 scale.

### 4.2. Analysis of Influencing Factors.

The main reason why existing prediction models could not achieve excellent accuracy is the fact that the travel times are impacted by various factors such as the weather, road quality, traffic conditions, signal control and other complex and changeable factors. In this paper, any two bus stops on a bus line are selected as a road segment for analysis, focusing, respectively, on the correlations between the changes in the speed of the previous and current bus work shift in the adjacent road segment, the proportion of bus lanes, real-time weather, and the travel time between bus stops. In order to reflect the influence of various factors on bus operation more directly, this paper uses bus travel speed to analyze the impact factors.

*4.2.1. Changes in Speed of Adjacent Shifts on Adjacent Segment.* Traffic volume has both temporal and spatial distribution characteristics. Many scholars have explored and considered using factors the same as the predicted bus vehicles in space and time are closest to them for prediction. That is, the travel time between stations of the previous bus work shift is used to predict the current travel time between stations.

*(1) Verify the Correlation of Travel Time of Previous and Current Shifts.* The operational data of a bus line in Guangzhou were selected to obtain the entire travel speed of vehicles in different shifts. Divide the time $c$ into orderly intervals of 30 minutes to obtain the travel speed of the bus in different time periods. The variation of travel speed in each time period follows the pattern shown in Figure 4.

As can be seen from the figure, the variation of bus travel speed in adjacent periods shows a certain degree of continuity with little fluctuation. In the morning and evening rush hours, the travel speed of buses is obviously lower than other times, which is consistent with the real situation. The correlation coefficient of travel speed in adjacent time periods was calculated. The MIC value of travel speed in the former and latter time period was 0.661; that is, there was a certain correlation between the travel speed of vehicles in the former and latter time period. According to the actual bus scheduling, the vehicles in the time period can be approximately regarded as the buses in the different work shifts. Under

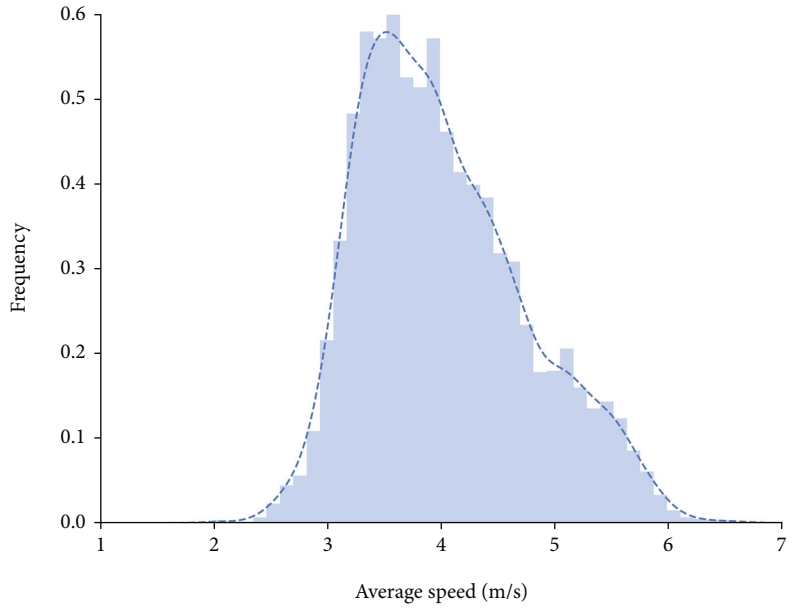



Figure 1: Average speed of the entire bus route.

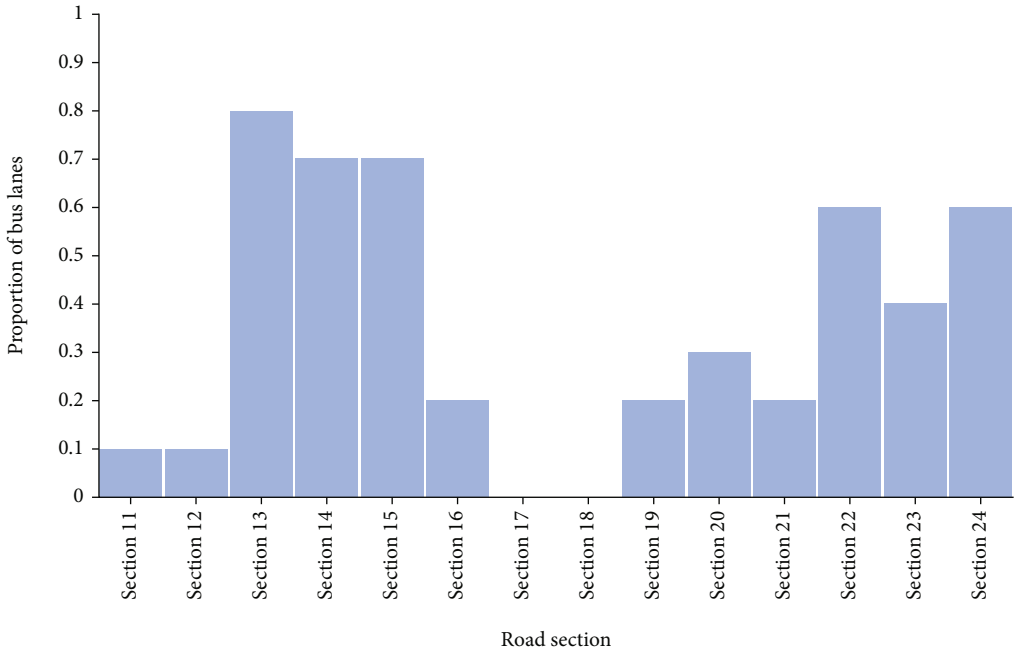


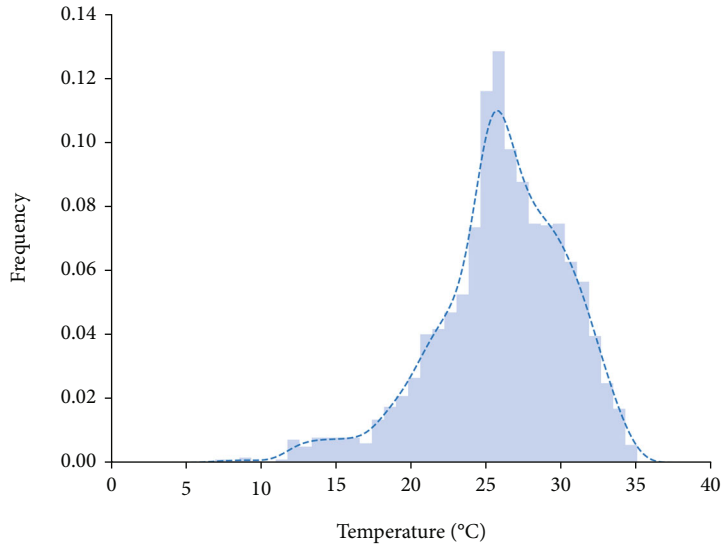

Figure 2: Distribution of bus lanes.

the condition that the road segments are the same and the road length is known, the operational data of vehicles in previous work shifts can be considered for prediction.

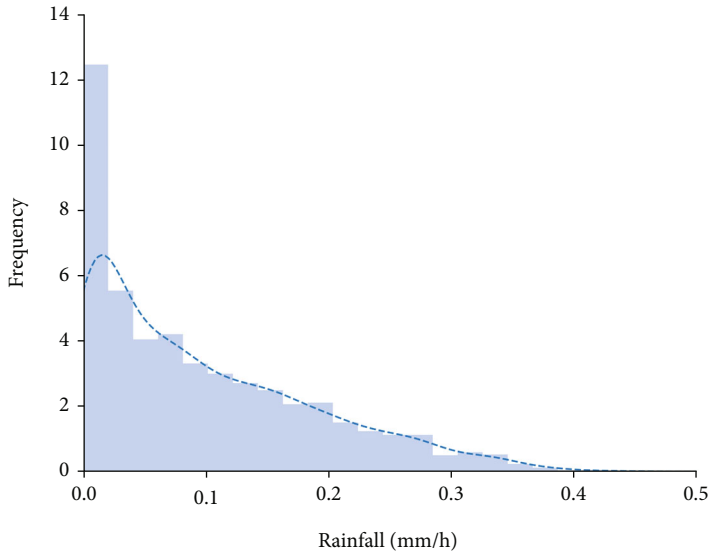
There is still room for improvement in real-time data of existing prediction research. Road traffic conditions may change about every ten minutes. The operational data of previous shifts of buses on the same road segment are not exactly the latest data available. In fact, that should be the travel time between stations of the current bus in the previous segment.

*(2) Correlation of Shift Speed Variations in Adjacent Road Segments*. There are differences between road conditions of the same shift in the adjacent segments; therefore, the speed of the adjacent segments cannot be directly used for correlation analysis.

According to the correlation of vehicle travel speed in the former and latter time periods, the speed changes of the two shifts in the adjacent road segments are compared. Ideally, the change in vehicle speed of different work shifts

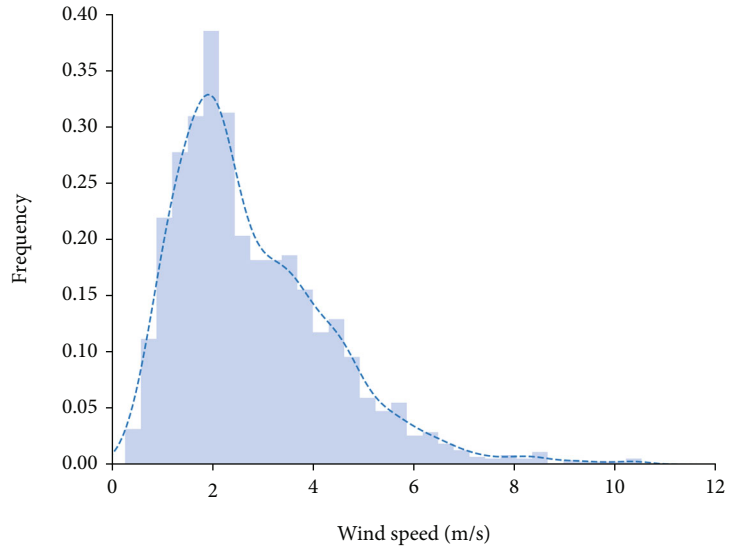



(a)




(b)

Figure 3: Continued.




(c)

Figure 3: Weather data distribution.

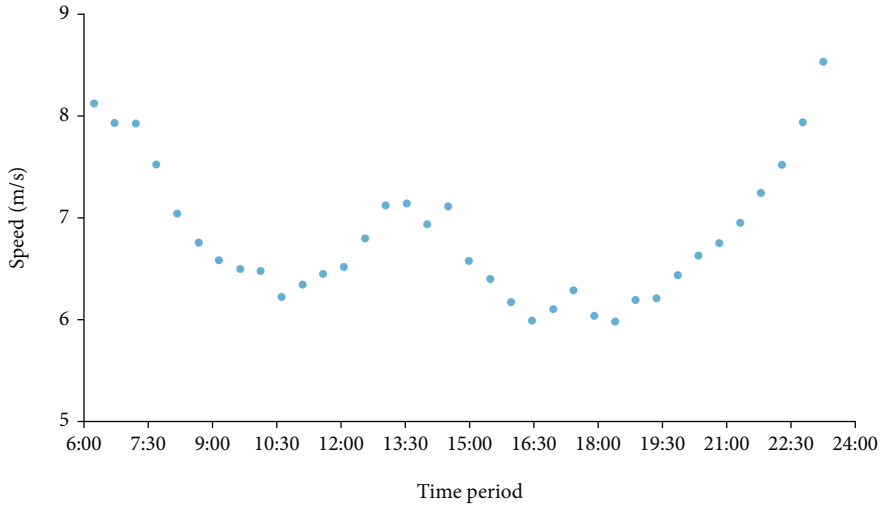



Figure 4: The full travel speed of the bus at different time periods.

on adjacent road segments should be the same. However, being affected by various factors, they are not the same. Here, SMAPE (Symmetric Mean Absolute Percentage Error) is used to measure the change of shift speed in the adjacent road segments. The expression of SMAPE is as follows:

$$\text{SMAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|x-y|}{(|x|+|y|)/2}, \tag{1}$$

where $x$ and $y$, respectively, represent the change of vehicle speed in the previous segment and current segment. The SMAPE of the change in the speed of bus work shift in the adjacent road segments in the same time period on different days basically does not exceed 25%; that is, the change of the speed of bus work shift in the adjacent road segments is generally similar.

In this paper, there are 25 stops in the whole bus route. The segment from the first stop to the 13th stop is selected as the former section, and the segment from the 13th stop to the last one is selected as the latter section. Based on that, Figure 5 shows the SMAPE distribution results of the changes in the speed of two bus work shifts in the two sections within 28 days.

In the figure, the abscissa is the SMAPE interval of the variation of the speed of the shift before and after in the adjacent road segments, and the ordinate is the probability of SMAPE in different intervals. It can be seen from the figure that the SMAPE of the speed change is basically between 10% and 20%, and the SMAPE of a certain time period will exceed 20%; that is, the changes of speed of work shift before and after in the adjacent segments are generally similar. In other words, travel times between bus stops can be predicted using variations in speed of shifts before and after in adjacent road segments.

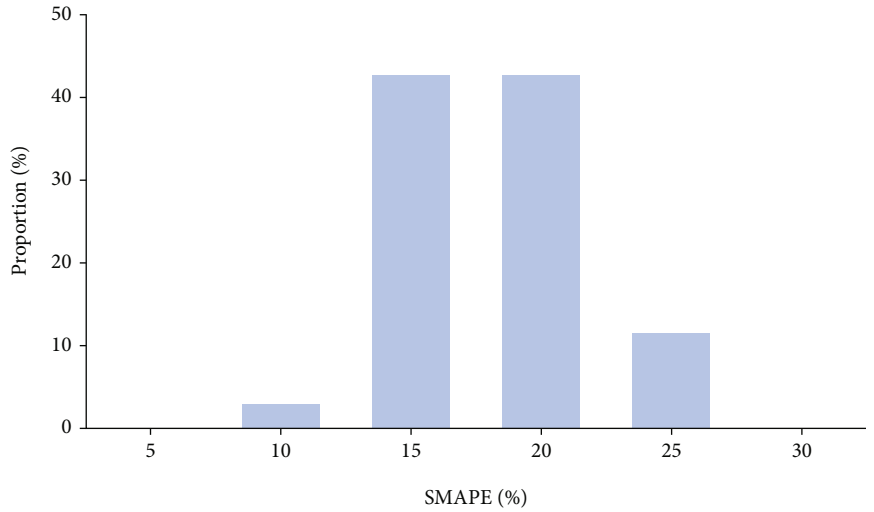



Figure 5: SMAPE for changes in travel speed of two work shifts on adjacent sections.

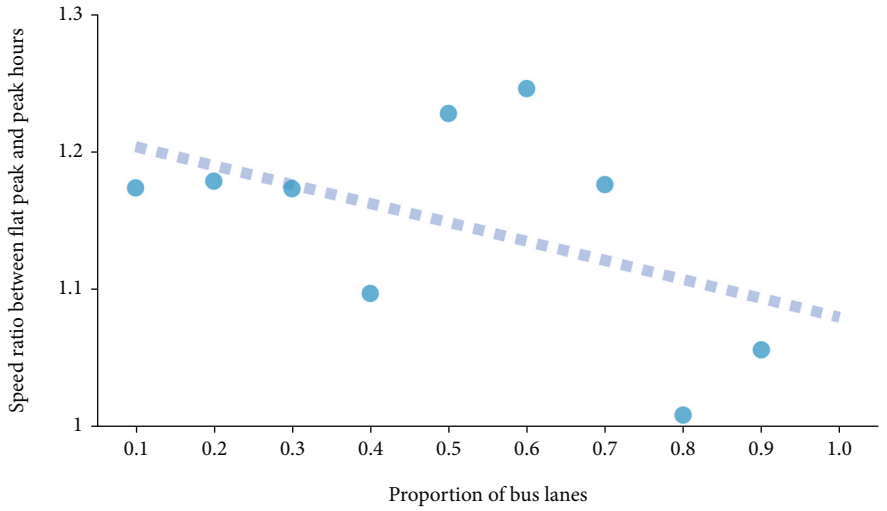



Figure 6: The ratio of the average bus travel speed between the flat peak and peak hours of different proportion of bus lanes.

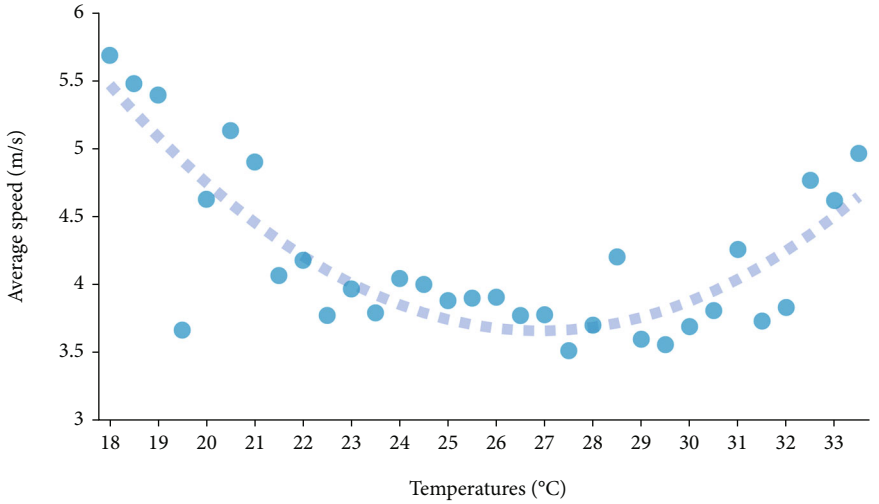



Figure 7: Average travel speed of buses under different temperatures.

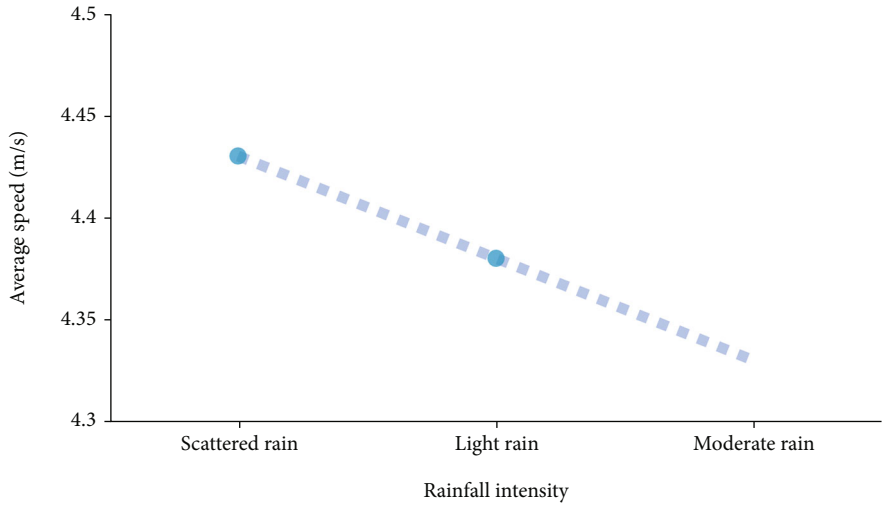



Figure 8: Average bus travel speed under different rainfall intensities.

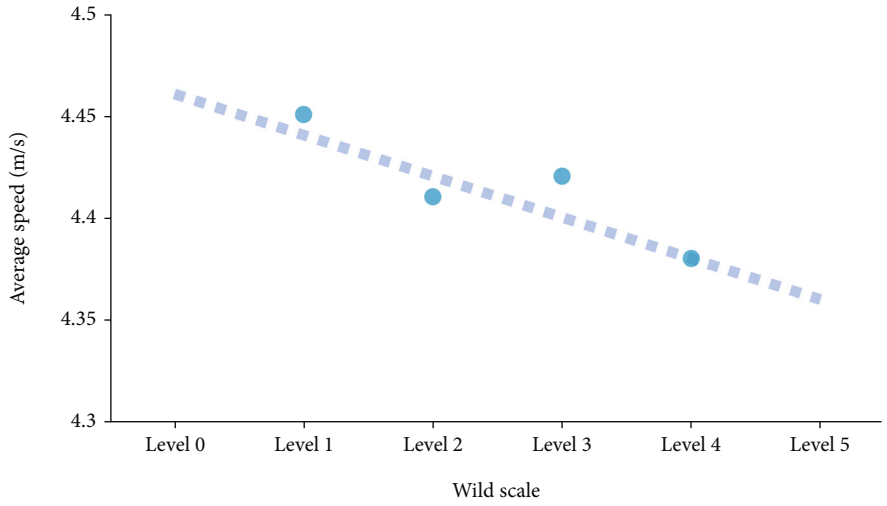



Figure 9: Average bus travel speed under different wind scales.

*4.2.2. Proportion of Bus Lanes.* In order to implement the development strategy of public transit priority, all provinces and cities in China have accelerated the construction of bus lanes in recent years. The length of bus lanes increased by 2101.5 kilometers in 2019 and 1599.9 kilometers in 2020. After the completion of the construction of bus lanes, the average speed of buses in Guangzhou during the morning peak hours increased by 13.91%. As an essential measure to ensure the priority of public transportation, using bus lanes to predict the road travel time can improve the accuracy of prediction to a higher degree.

The bus lanes in Guangzhou runs during two identified peak-hour periods: a morning peak from7 a.m. to 9 a.m. and an evening peak from 5 p.m. to 7 p.m. Those two periods are regarded as the traffic peak periods, and the rest periods of the day are the traffic off-peak periods. After processing the operational data of a selected bus lane, the average travel speed between stations can be clearly revealed. At the same time, the proportion of bus lanes in each segment is obtained from the map. Then, calculate the average travel speed of bus vehicles on each road segment at off-peak period and peak period, and get the ratio of average travel speed on each road segment at off-peak period and peak period with different proportion of bus lanes, which is depicted in Figure 6.

Table 1: Calculate the MIC of main influencing factor.

| Influencing factors | Variables | MIC |
|---|---|---|
| Proportion of bus lanes | Off-peak period | 0.236 |
| | Peak period | 0.612 |
| Real-time weather | Temperature | 0.407 |
| | Rainfall | 0.330 |
| | Wind | 0.245 |

In the figure, the abscissa is the proportion of bus lanes, and the ordinate is the ratio of average travel speed of vehicles between the flat peak and peak hours under the same

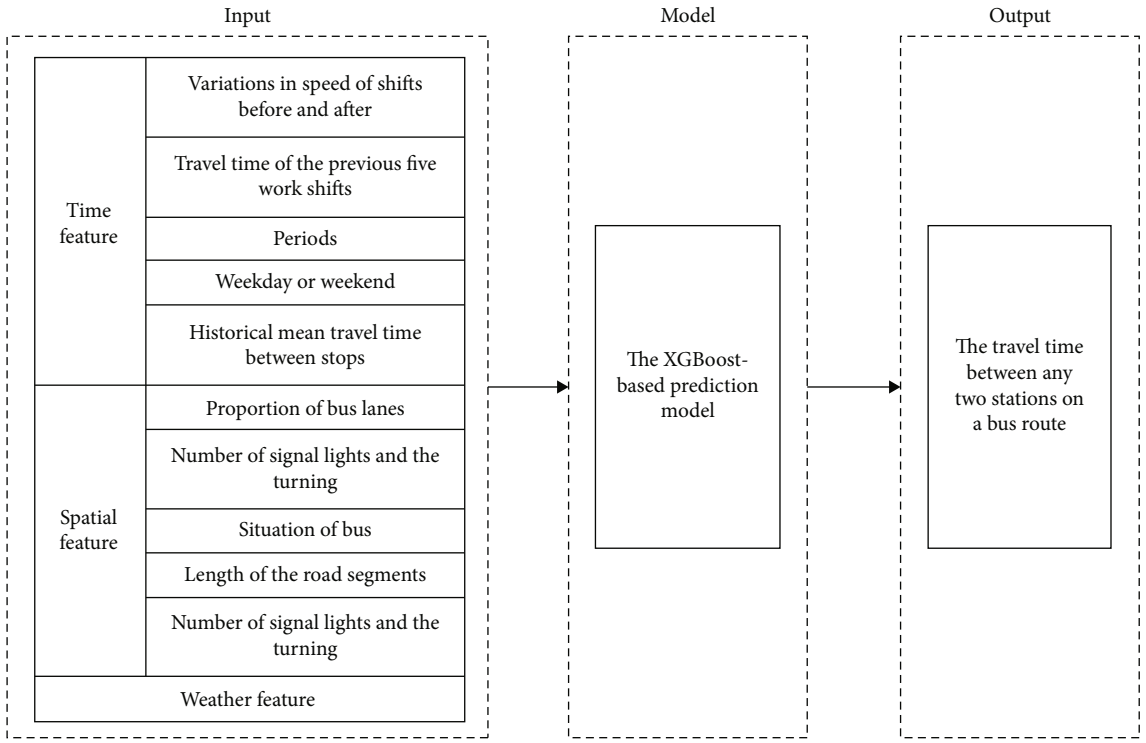



Figure 10: The structure plots of the XGBoost-based prediction model.

Table 2: Input of the predict model.

| Feature attributes | Influencing factors |
|---|---|
| Time | Variations in speed of shifts before and after |
| | Travel time of the previous five work shifts |
| | Periods |
| | Weekday or weekend |
| | Historical mean travel time between stops |
| Spatial | Proportion of bus lanes |
| | Number of signal lights and the turning situation of bus |
| | Length of the road segments |
| | Starting and ending bus station number |
| Weather | Weather information including temperature, rainfall intensity and the scale of wind force |

proportion. As the proportion of bus lanes increases, the ratio gradually approaches 1; that is, the difference of vehicle speed between flat peak and peak periods gradually decreases. It can be seen that bus lanes play a positive role in maintaining bus speed stability during peak hours, and the larger the proportion of bus lanes is, the more obvious the effect is. At the same time, the MIC between the ratio of average bus travel speed during off-peak time to average speed during peak hours and the proportion of bus lane was 0.236, and the MIC between the average bus travel speed in peak hours and the proportion of bus lane was 0.612. Hence, we can see that the proportion of bus lanes in the segment will affect the travel speed of bus vehicles. The larger the proportion, the higher the average travel speed, and also the shorter the travel time between stations. Therefore, the proportion of bus lanes can be selected as an input

Table 3: The best value of each parameter of XGBoost.

| Parameters | Value |
|---|---|
| learning_rate | 0.01 |
| n_estimators | 350 |
| max_depth | 15 |
| min_child_weight | 4 |
| subsample | 0.5 |
| subsample_bytree | 0.8 |
| gamma | 0.1 |
| reg_alpha | 1 |
| reg_lambda | 1 |

feature to establish a conventional prediction model of bus station travel time.

*4.2.3. Real-Time Weather.* Buses, as an important role in public transportation which is an activity has to be exposed outdoor, are susceptible to many factors, among which weather conditions is significant for predicting the travel time between bus stations. For example, when the road gets wet and slippery in rainy or snowy days, the drivers will slow down accordingly to ensure driving safety. Therefore, this paper analyzes the influence of temperature, rainfall intensity, and the level of wind force on vehicle speed.

For the element of temperature, calculate the average travel speed of buses at different temperatures, as shown in Figure 7. In the figure, the abscissa is the temperature, and the ordinate is the mean travel speed under the corresponding air temperature. As can be seen from the figure, when the temperature gets higher or lower, the vehicle speeds faster. The reason for this phenomenon is that the temperature is still low in the early morning and late night, when people stay at home and the travel demand is yet small, while the temperature reaches higher at noon, people show less motivation to travel, so the traffic flow reduces significantly, which is in line with the actual situation. The MIC of the temperature and average travel speed is 0.407.

For the element of rainfall, calculate the average travel speed of buses under different rainfall intensity, as shown in Figure 8. In the figure, the abscissa is rainfall intensity, and the ordinate is the mean travel speed under the corresponding rainfall intensity. It can be seen that the element of rainfall has a certain impact on the driving state of buses, which is presented as when the intensity of rainfall increases, the average travel speed of public transport vehicles will decrease. The correlation between rainfall intensity and average bus speed was calculated, and the MIC was 0.33.

For the element of scale of wind force, calculate the average travel speed of buses in different scales of wind force, as shown in Figure 9. In the figure, the abscissa is the level of wind force, and the ordinate is the mean of speed under the corresponding scales of wind force. It can be seen from the figure that in a certain range, the average travel speed of buses decreases with the increase of scale. The higher the scale of wind force, the greater the wind resistance of the vehicle and the lower the speed. The MIC value of the scale of wind force and average travel speed within 28 days was calculated too, and it was 0.245.

Therefore, temperature, rainfall rate, and the scale of wind force are selected as the input features of the travel time prediction model in this paper. In summary, the variation of speed between shifts before and after in the adjacent road segments, the proportion of bus lanes, and real-time weather are three discussed factors affecting the travel time between bus stations, which can be used as the input of the prediction model.

According to the previous analysis of influencing factors, the correlation coefficients of each influencing factor are shown in Table 1. Changes in speed of adjacent shifts on adjacent segment were assessed using SMAPE, and the changes are generally similar.

Table 4: The best value of each parameter of LightGBM model.

| Parameters | Value |
|---|---|
| *max_depth* | 5 |
| *num_leaves* | 10 |
| *learning_rate* | 0.01 |
| *random seed* | Fixed value |
| *n_estimators* | 1000 |

## 5. Result

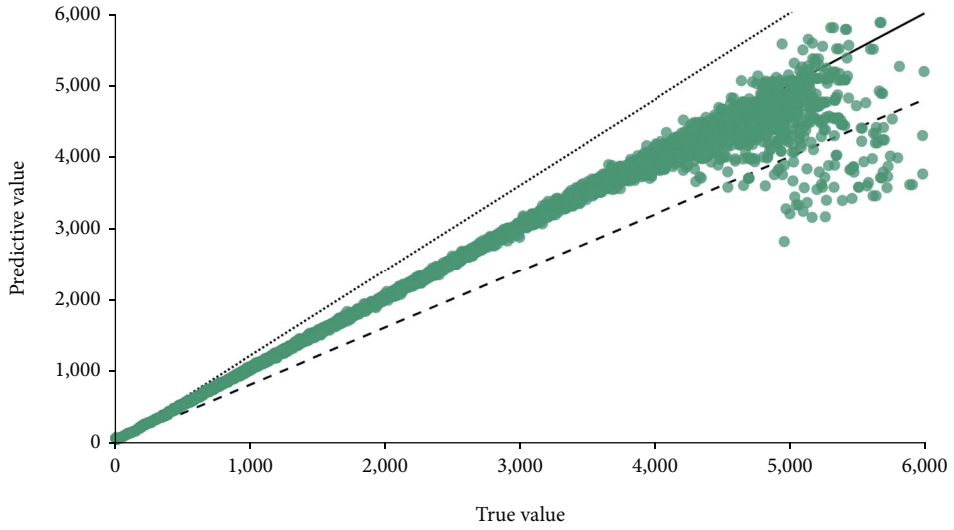
*5.1. Model Construction and Optimization.* According to the analysis of influencing factors in the previous section, combined with the characteristics used high-frequently in existing studies, the prediction model is constructed and the model parameters are tuned. The structure plots of the XGBoost-based prediction model constructed in this paper are shown in Figure 10.

The process of model optimization is as follows:

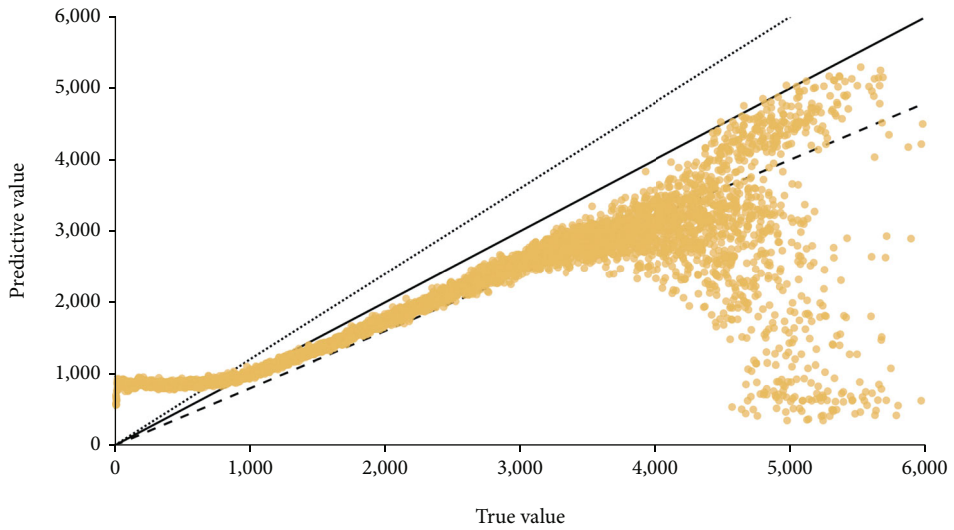
(i) Input: it is time feature that contains the speed variation, spatial feature that contains the proportion of bus lanes, and weather feature in Table 2.

(ii) Output: output is the travel time between any two stations on a bus route.

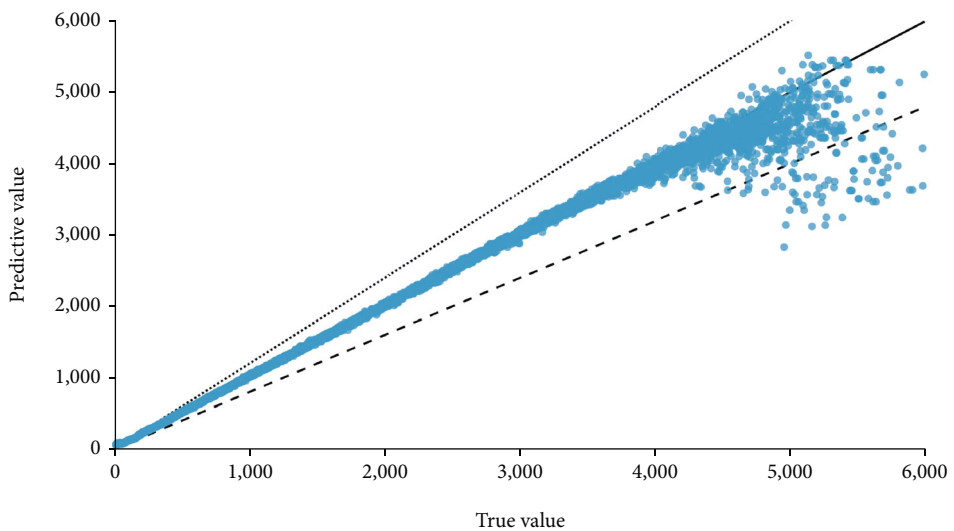
(iii) Parameter tuning process: use the network search cross-validation method to adjust the parameters. The network search cross-validation method is the "GridSearchCV" in the "Scikit-Learn" library. It returns the evaluation index score under all parameter combinations by means of cross-validation by traversing all permutations and combinations of incoming parameters. First, initialize the lift parameters and then adjust them. Taking parameter *n_estimators* as an example, the process is as follows: the candidate values of parameter *n_estimators* were determined to be 200, 300, 400, 500, and 600, and the value of parameter *n_estimators* was changed while the other parameters remained unchanged. Cross-validation and evaluation index RMSE were used to measure the performance of the model. Candidate values can be further divided on the basis of candidate values themselves. For example, the best candidate value is 350. According to the above steps, choose below in turn. The best values for column parameters are as follows: *learning_rate*, *max_depth*, *min_child_weight*, *subsample*, *subsample_bytree*, *gamma*, *reg_alpha*, and *reg_lambda*; the best values are shown in Table 3. When all the parameters finish determining the optimal parameter values, the construction of the prediction model of travel time between bus stations based on XGBoost is completed.




(a) KNN regression model

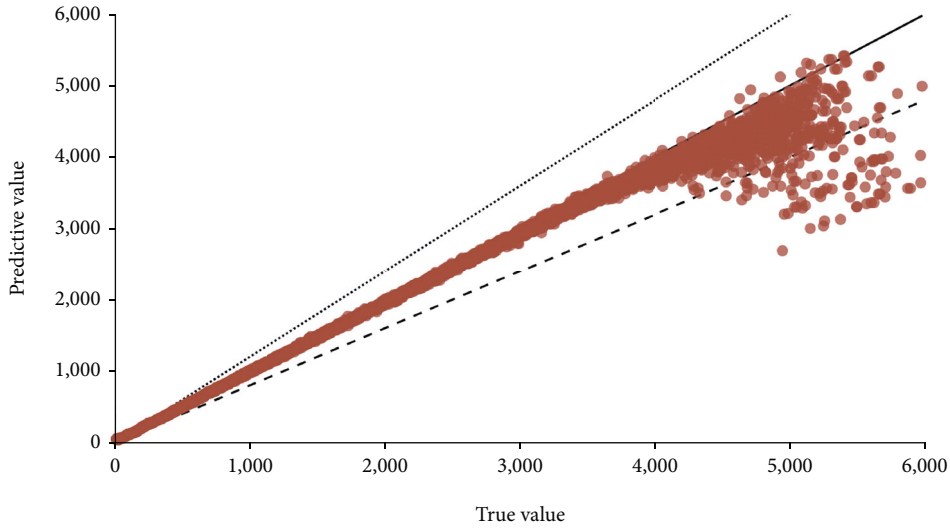



(b) BP neural network model




(c) LightGBM model

Figure 11: Continued.




(d) XGBoost model

Figure 11: Comparison of the predicted value of each model with the true value.

*5.2. Comparison with Traditional Model.* In this paper, the KNN regression model, BP neural network model, and LightGBM model are selected to be compared with XGBoost prediction model. In order to better understand and examine the effectiveness and accuracy of the bus travel time prediction model based on XGBoost established in this paper, the KNN regression prediction model, BP neural network model, LightGBM prediction model, and XGBoost prediction model are also established for comparison.

*5.2.1. KNN Regression Model.* The *KNeighborsRegressor* function in the Python module *sklearn.neighbors* is used to build a KNN-based regression prediction model. The key parameters are *n_neighbor* and *weight*, which represent the number of nearest neighbor data points and the weight of each neighbor data point, respectively. The optimal value of the parameter *n_neighbor* is determined to be 14 and the optimal value of the parameter weight is determined to be uniform by training the tuning parameters through the enumeration method.

*5.2.2. BP Neural Network Model.* In this paper, we use the Python module *Keras* to build the prediction model based on BP neural network. After repeated experiments, the number of nodes in the input layer, hidden layer, and output layer of the BP neural network model is determined, and the model structure is 24-14-1. The *relu* function is selected for the excitation function of the hidden layer and the output layer, and the number of iterations of the BP neural network is set to 200, and the learning rate is 0.01.

*5.2.3. LightGBM Model.* In this paper, we use the python module LightGBM to build a prediction model based on LightGBM. The parameters of the LightGBM model are determined by Grid Search. The best values of its parameters are shown in Table 4.

Table 5: Comparison of the average value of the evaluation indicators.

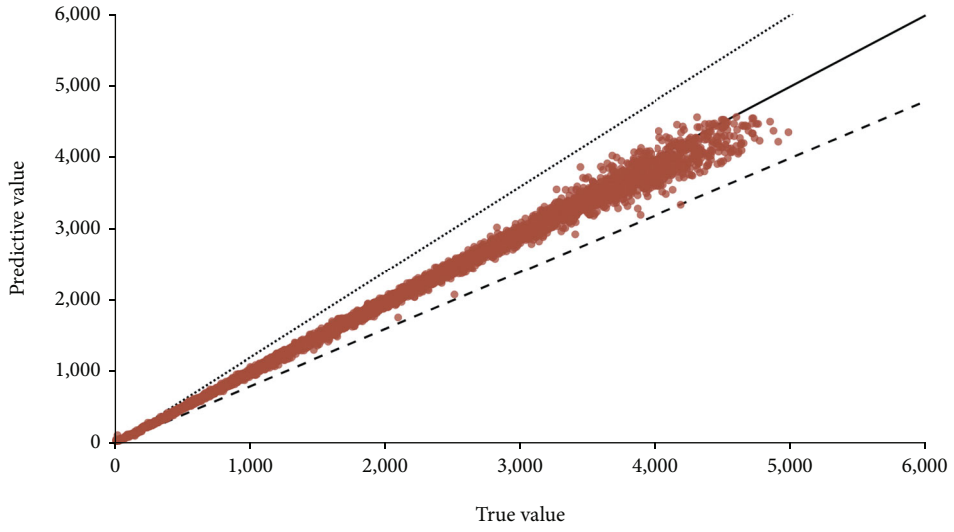
| Parameters | MAE | RMSE | MAPE (%) |
|---|---|---|---|
| XGBoost model | 137.01 | 247.92 | 11.96 |
| KNN regression model | 162.51 | 278.11 | 15.99 |
| BP neural network model | 459.27 | 673.26 | 30.75 |
| LightGBM model | 151.57 | 255.45 | 17.03 |

Table 6: Average evaluation value of XGBoost model in different time periods.

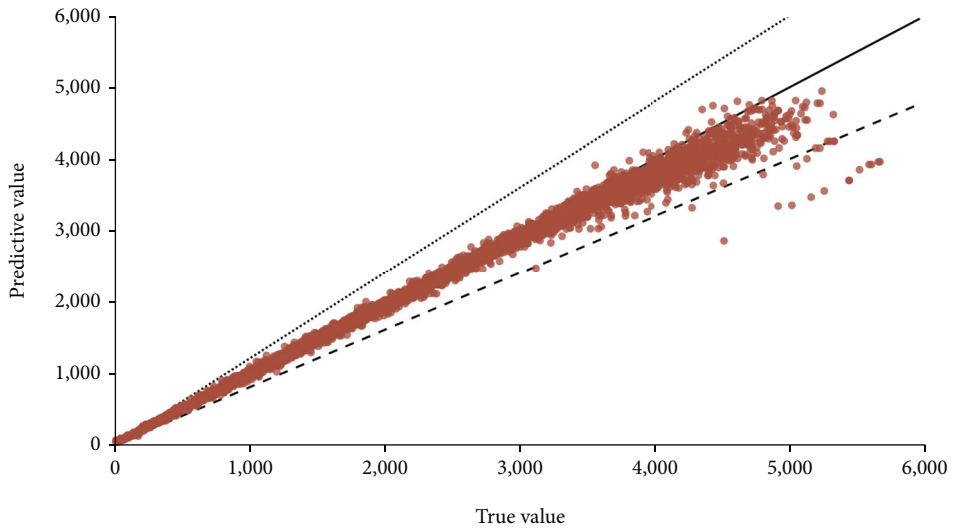
| Time period | MAE | RMSE | MAPE |
|---|---|---|---|
| Morning peak | 179.51 | 263.78 | 13.63% |
| Evening peak | 160.40 | 235.19 | 12.78% |
| Flat peak | 136.60 | 279.71 | 11.95% |

The selected influencing factors are used for feature engineering and as model input. And the KNN regression model, BP neural network model, LightGBM model, and XGBoost model are used for prediction. The comparison results between the predicted values of different prediction models and the real values are shown in Figure 11.

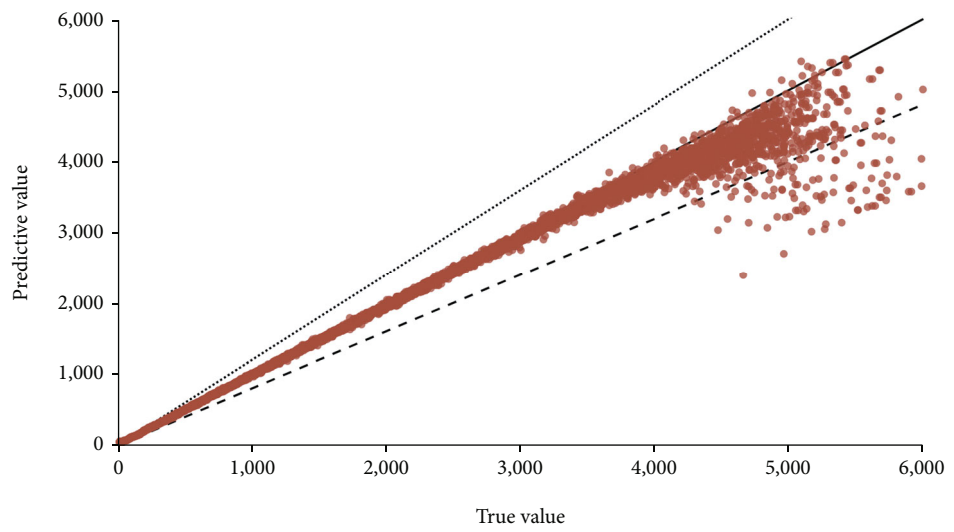
From the above figure, it can be seen that the XGBoost model has the best prediction. The evaluation index values of each prediction model are shown in Table 5. The values of the three-evaluation metrics of the XGBoost prediction model are lower than those of the other three models. Among them, the MAPE value of the XGBoost prediction model is 11.96%, which is 4.03%, 18.79%, and 5.07% lower than the KNN regression prediction model, BP neural network prediction model, and LightGBM prediction model, respectively, and its mean value is 9.30%. It is concluded that the accuracy of the XGBoost-based prediction model




(a) Morning peak




(b) Evening peak




(c) Flat peak

Figure 12: Comparison of the predicted value and the true value of the XGBoost model in different periods.

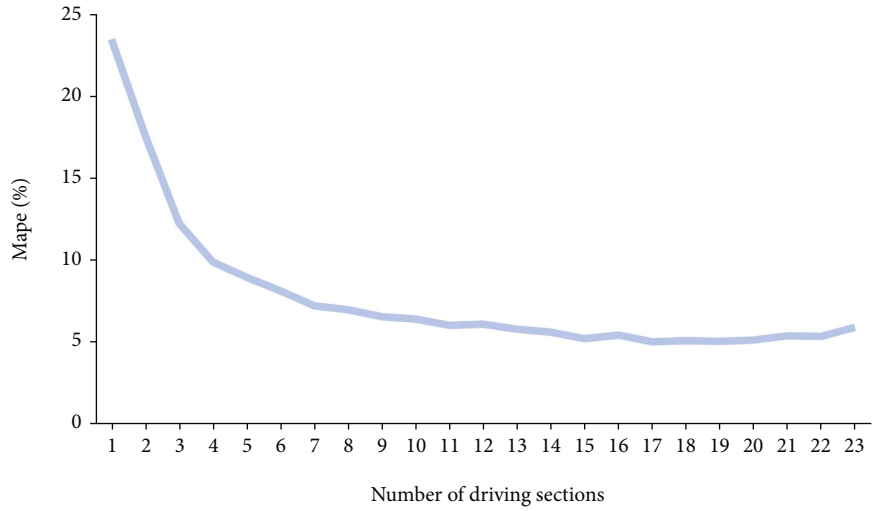



Figure 13: Average value of MAPE of XGBoost model under different numbers of driving section.

proposed in this paper is better when performing conventional bus stop-to-stop travel time prediction.

*5.3. Robustness of the Model.* To evaluate the robustness, we compare the performance of the proposed model.

*5.3.1. Analysis by Time Period.* This paper compares the performance of XGBoost model in different time periods (morning, evening peak and peak), and the results are shown in Table 6.

It can be seen that the MAPE of the XGBoost model in the morning peak period is significantly lower than that in the evening peak period and flat peak period, which indicates the prediction accuracy of the model in the morning peak period is higher, followed by the evening peak and flat peak period, and the flat peak period has the worst performance. The comparison between the predicted value and the real value of XGBoost model in different periods is shown in Figure 12.

The prediction result performs the best in the morning peak period, and the dots in the figure fall within the limit boundaries; the second is the evening peak period; only a few dots in the figure fall outside the 20% limit boundaries; and the flat peak period has the worst performance. The potential reason behind this could be that the overall road traffic flow volume during the morning rush hour is greater than that during the evening rush hour, and the residents' travel purposes during the morning rush hour are more unified, the travel time between stations is less volatile, and the various other factors show less significant impacts. Therefore, the prediction accuracy during the morning peak hours is higher, and the error is smaller. People travel for various purposes during peak hours, so the travel time between stations fluctuates greatly. Moreover, the bus company will increase the frequency of bus departures during peak hours. The more research samples, the higher the accuracy. All the above reasons contribute to prediction in peak hours which gets higher accuracy than it gets in nonpeak hours.

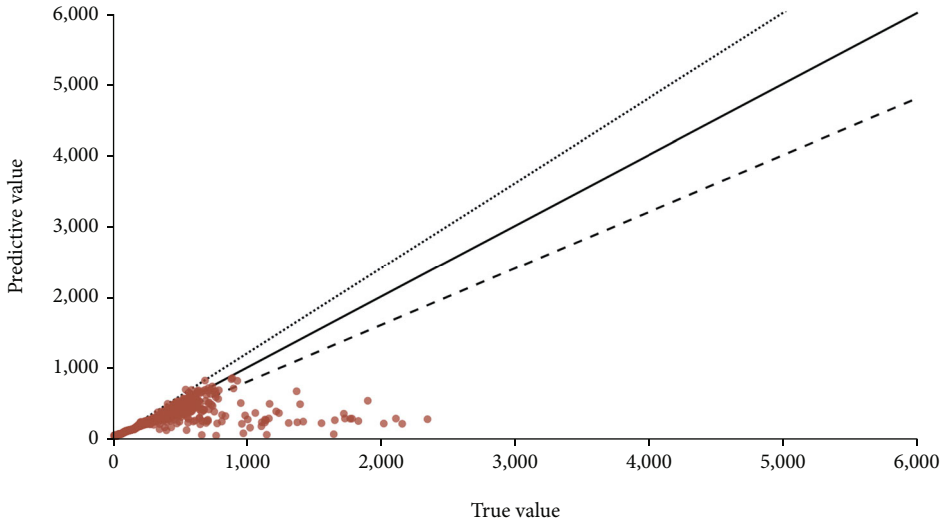
In the end, the calculated MAPE of the XGBoost model during peak hours is 10.10%, and the MAPE during normal peak hours is 10.63%. Therefore, the constructed forecasting model has good prediction result during peak hours and better performance during morning peak hours.

*5.3.2. Analysis by the Number of Driving Sections.* This paper regards the road segment between adjacent stations as a driving section. Compare the error between the model prediction result and the true value under different number of driving sections, as shown in Figure 13. The abscissa is the number of driving sections in the predicted road segment, and the ordinate represents the average absolute percentage error between the predicted value and the true value.

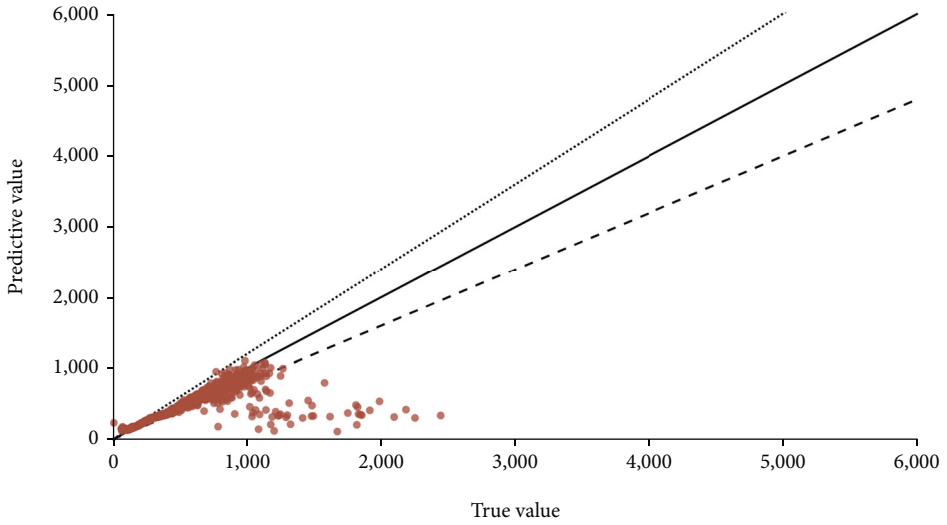
Obviously, with the increase of the number of driving sections, the average value of MAPE of XGBoost prediction model gradually decreases; that is, the prediction accuracy of XGBoost model gradually improves. The comparison between the predicted value and the real value of XGBoost model under different number of driving sections is shown in Figure 14. The number of driving sections corresponding to $(a) \sim (x)$ ranges from 1 to 24, respectively.

When the number of driving sections is small, more dots will fall outside the 20% limit. However, as the number of driving sections increases, the number of dots outside the limit boundaries decreases, which further verifies that as the number of driving sections increases, the prediction accuracy of the model improves. This may be because there are certain restrictions on the travel speed of public transport vehicles on urban roads. The greater the number of driving sections, the longer the length of the road section, the smaller the effect of various factors on driving vehicles will share, the more stably the average vehicle speed fluctuates, and the travel time between stations is gradually less affected by various factors, so the accuracy of the forecasting model is also improved.

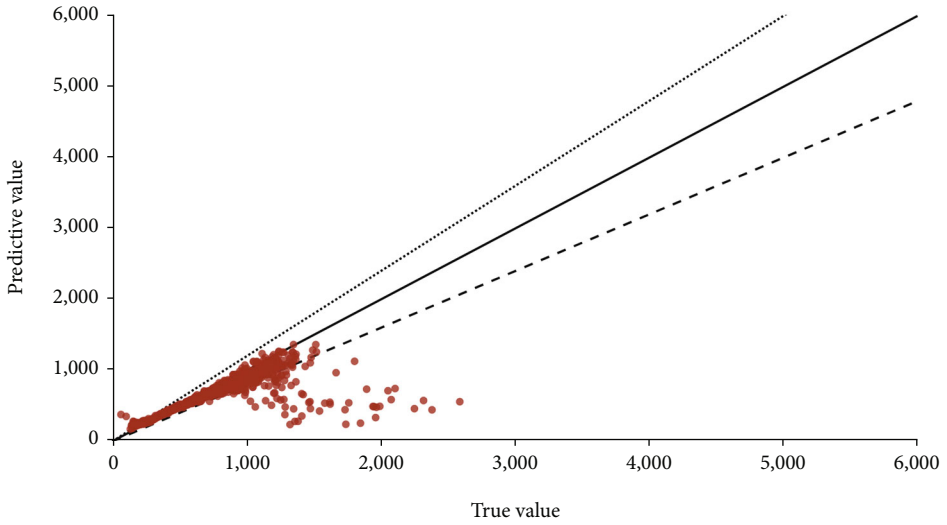
*5.4. Influence Degree of Variables.* In this paper, a prediction model based on XGBoost is established to predict the travel




(a)

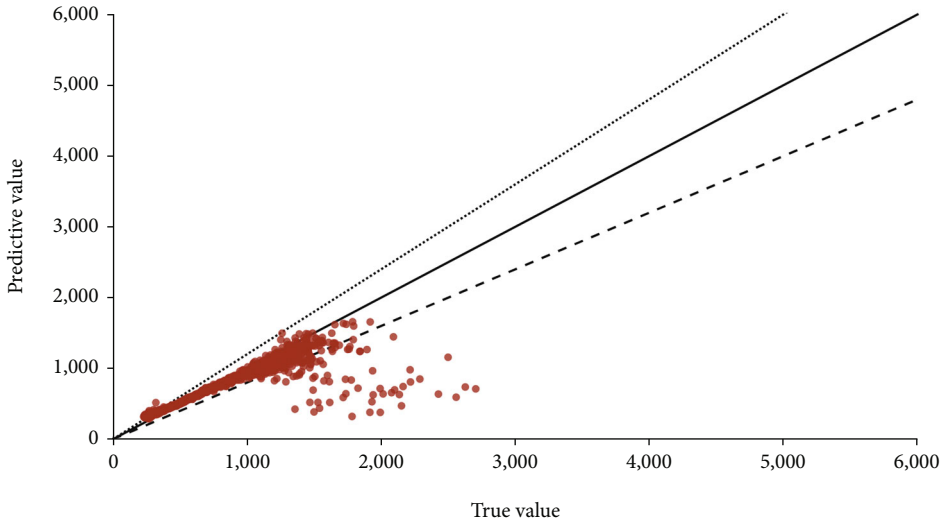



(b)

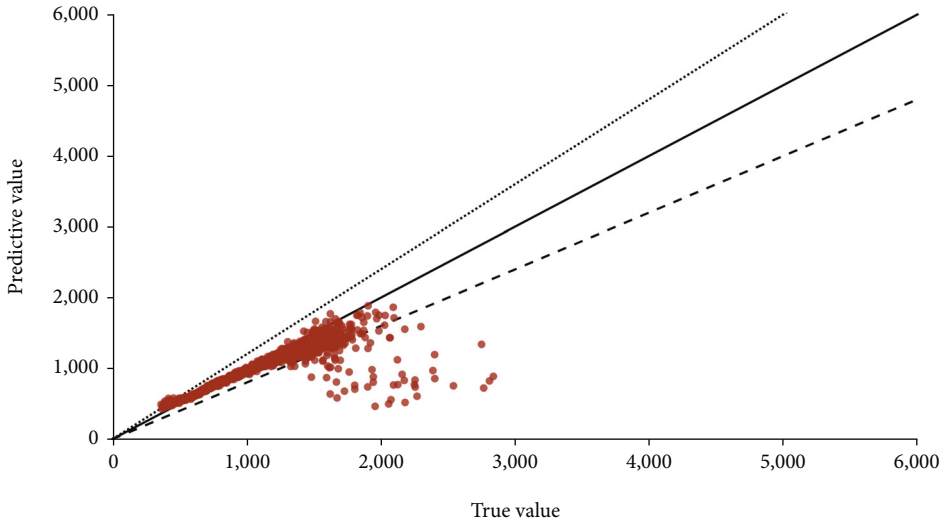



(c)

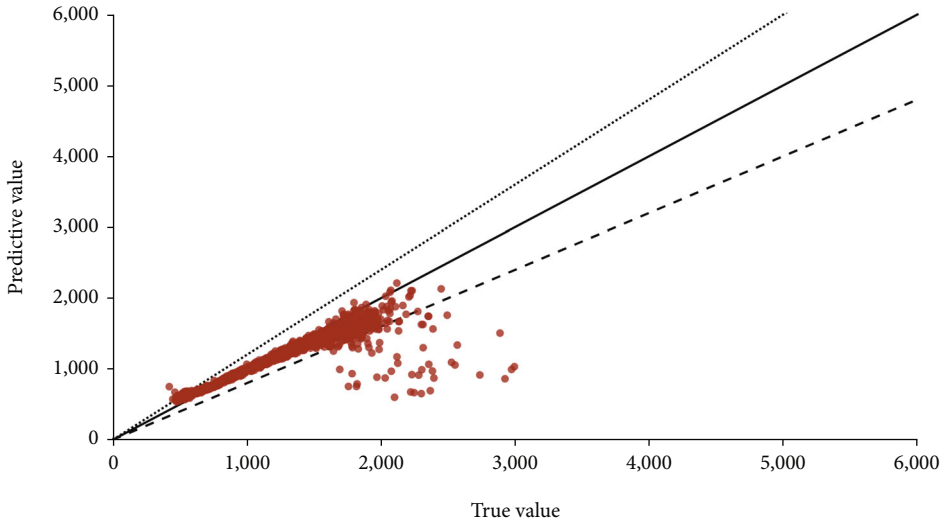
Figure 14: Continued.




(d)

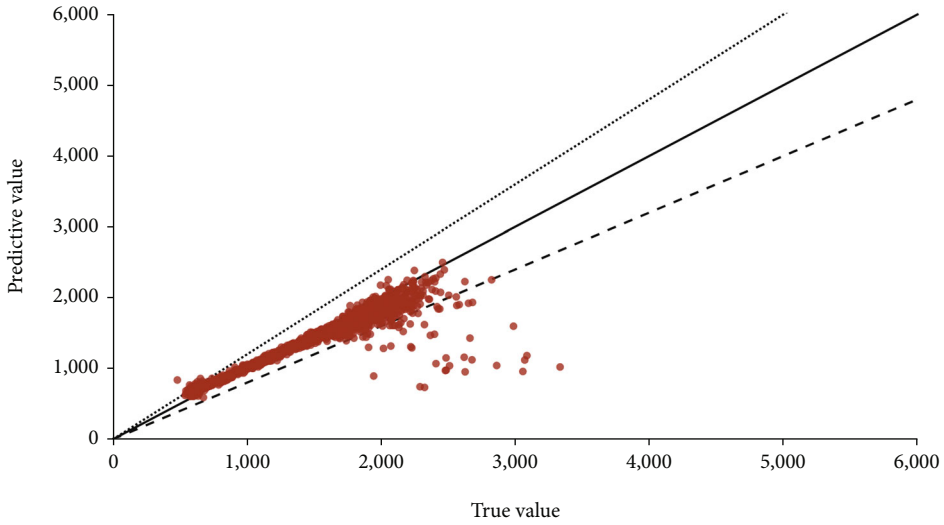



(e)

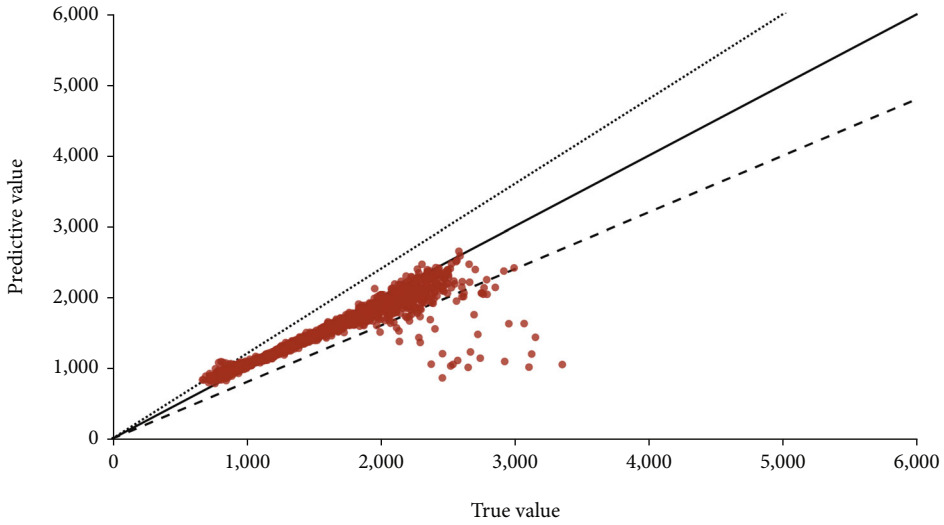



(f)

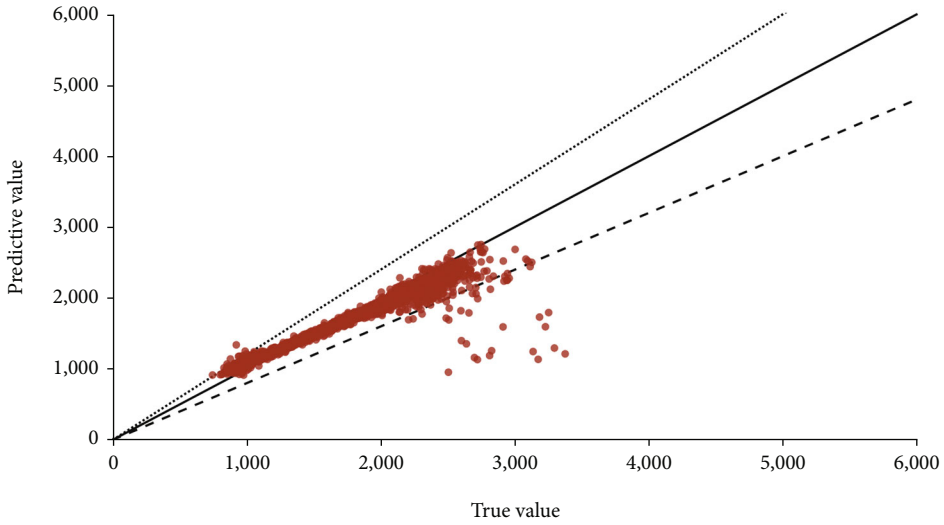
Figure 14: Continued.




(g)

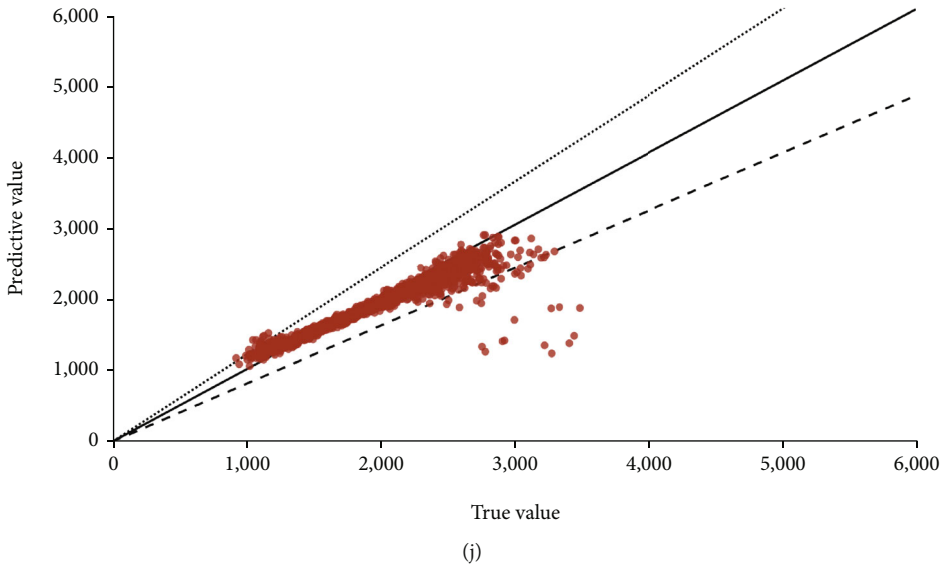



(h)

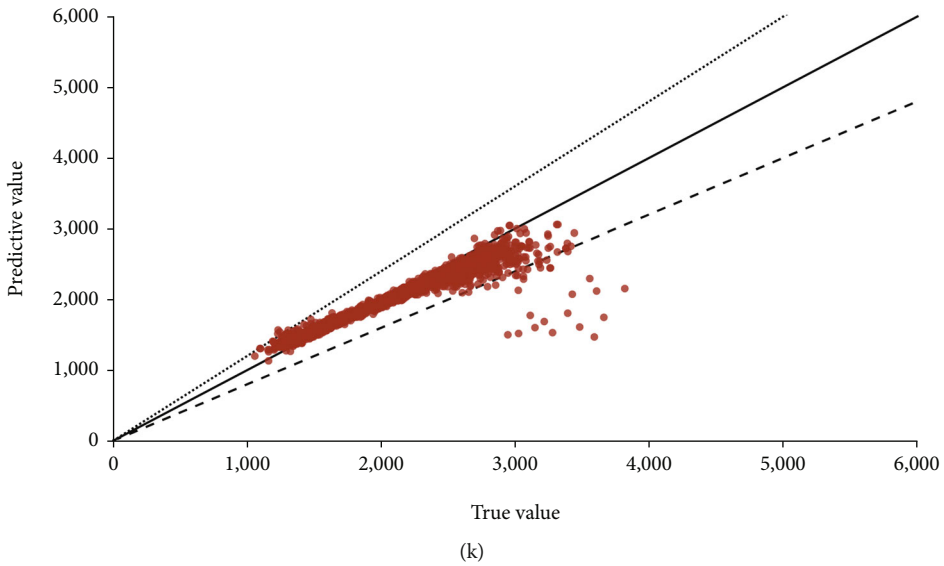



(i)

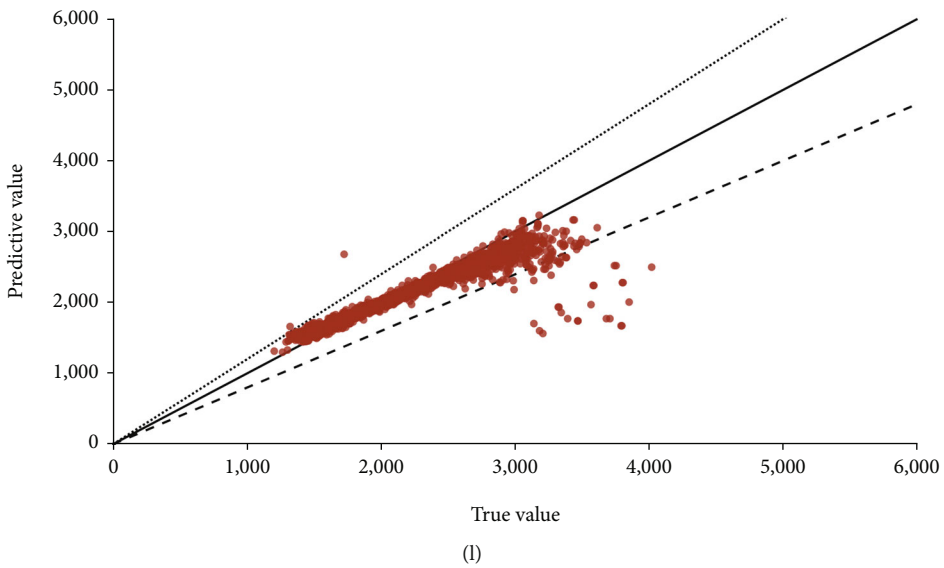
Figure 14: Continued.




(j)

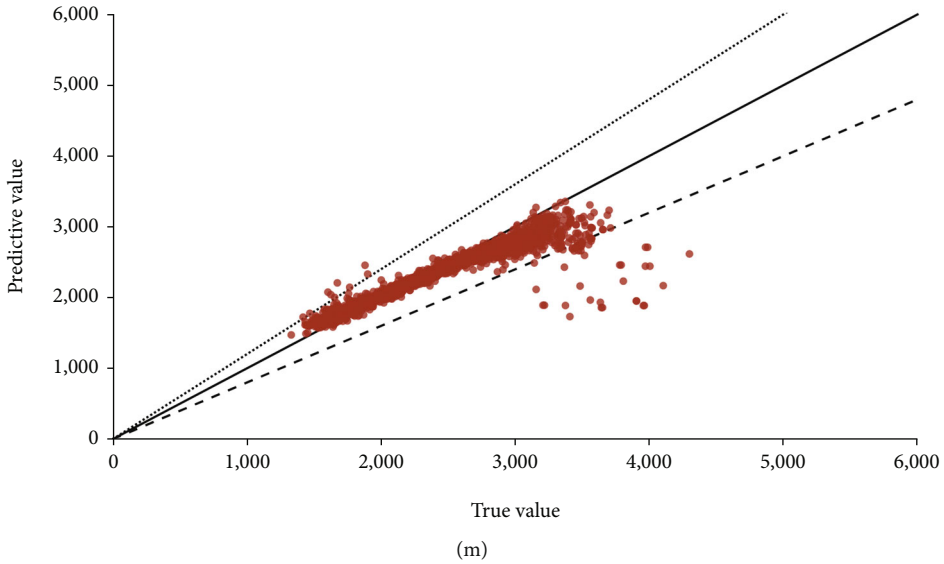



(k)

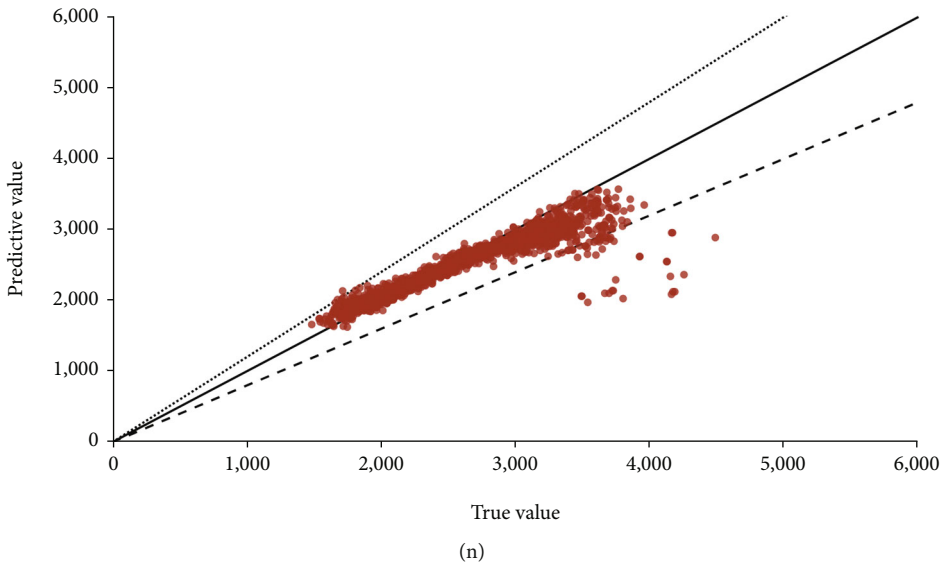



(l)

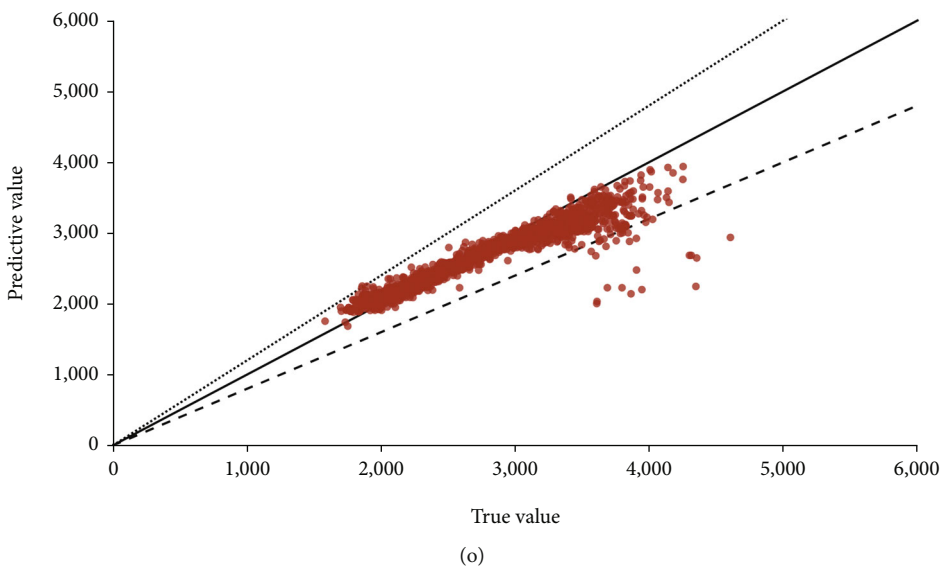
Figure 14: Continued.




(m)

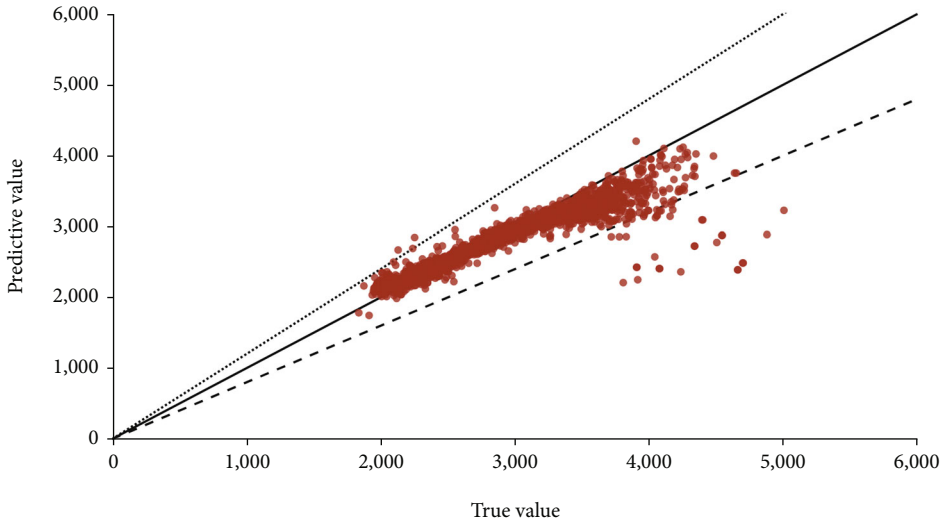



(n)

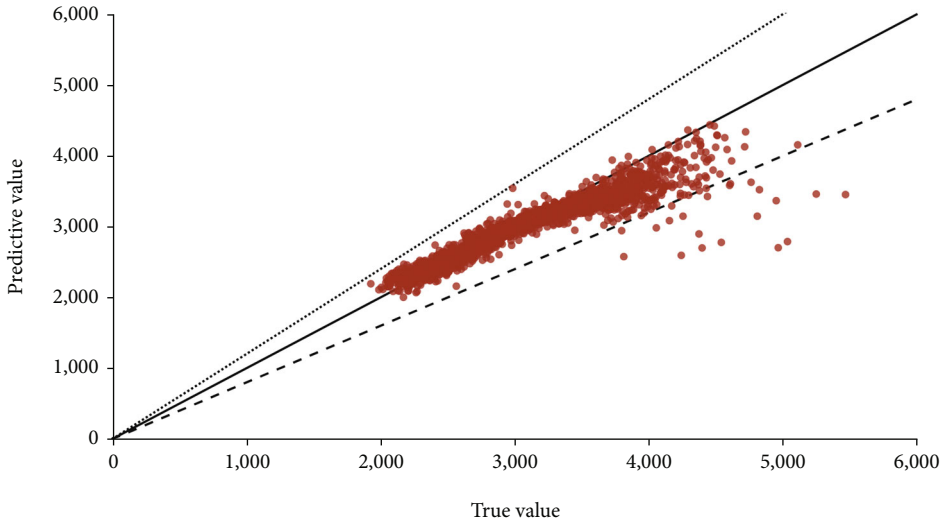



(o)

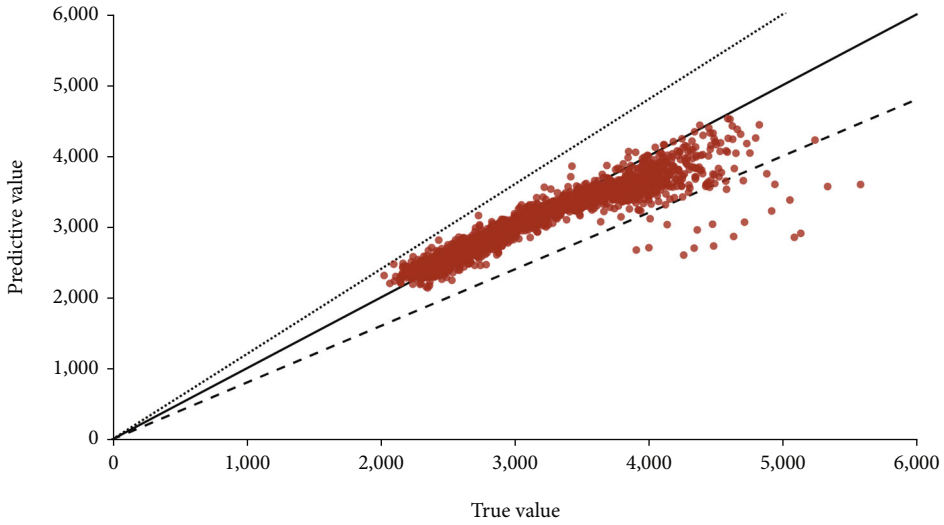
Figure 14: Continued.




(p)

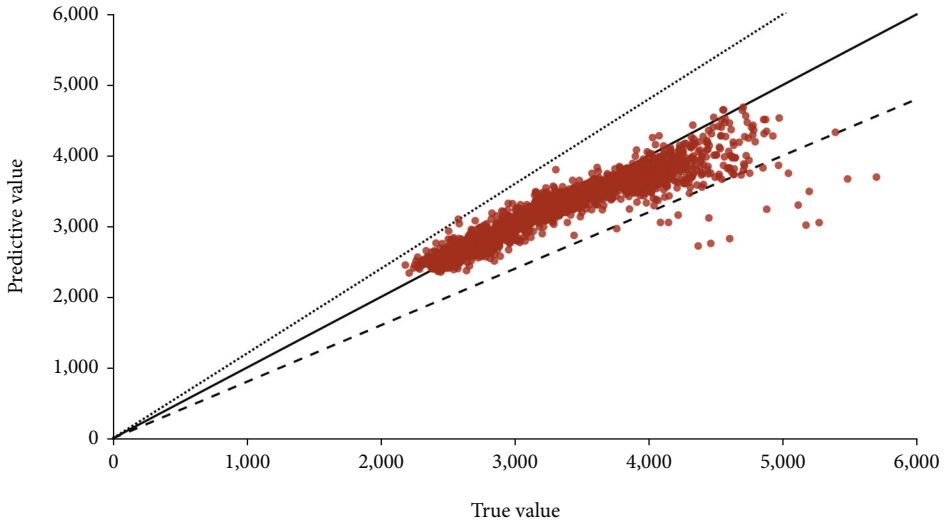



(q)

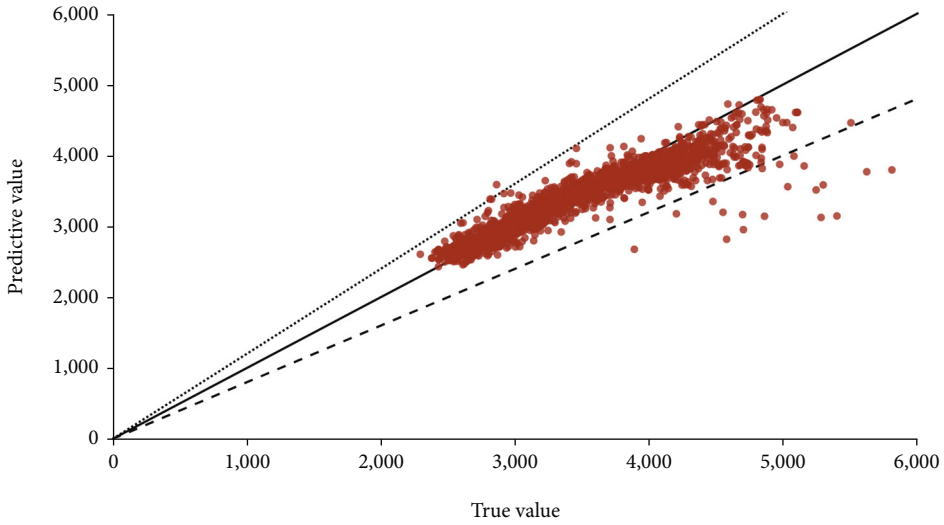



(r)

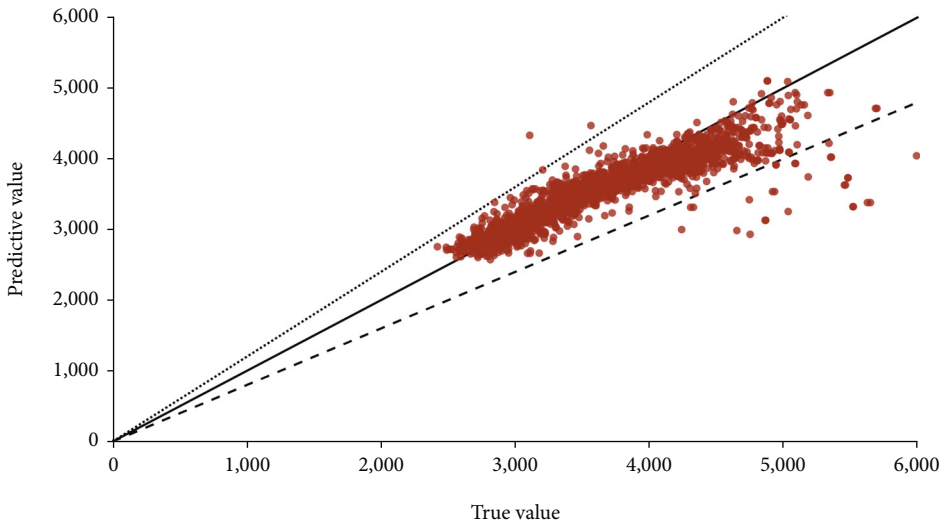
Figure 14: Continued.




(s)




(t)




(u)

Figure 14: Continued.

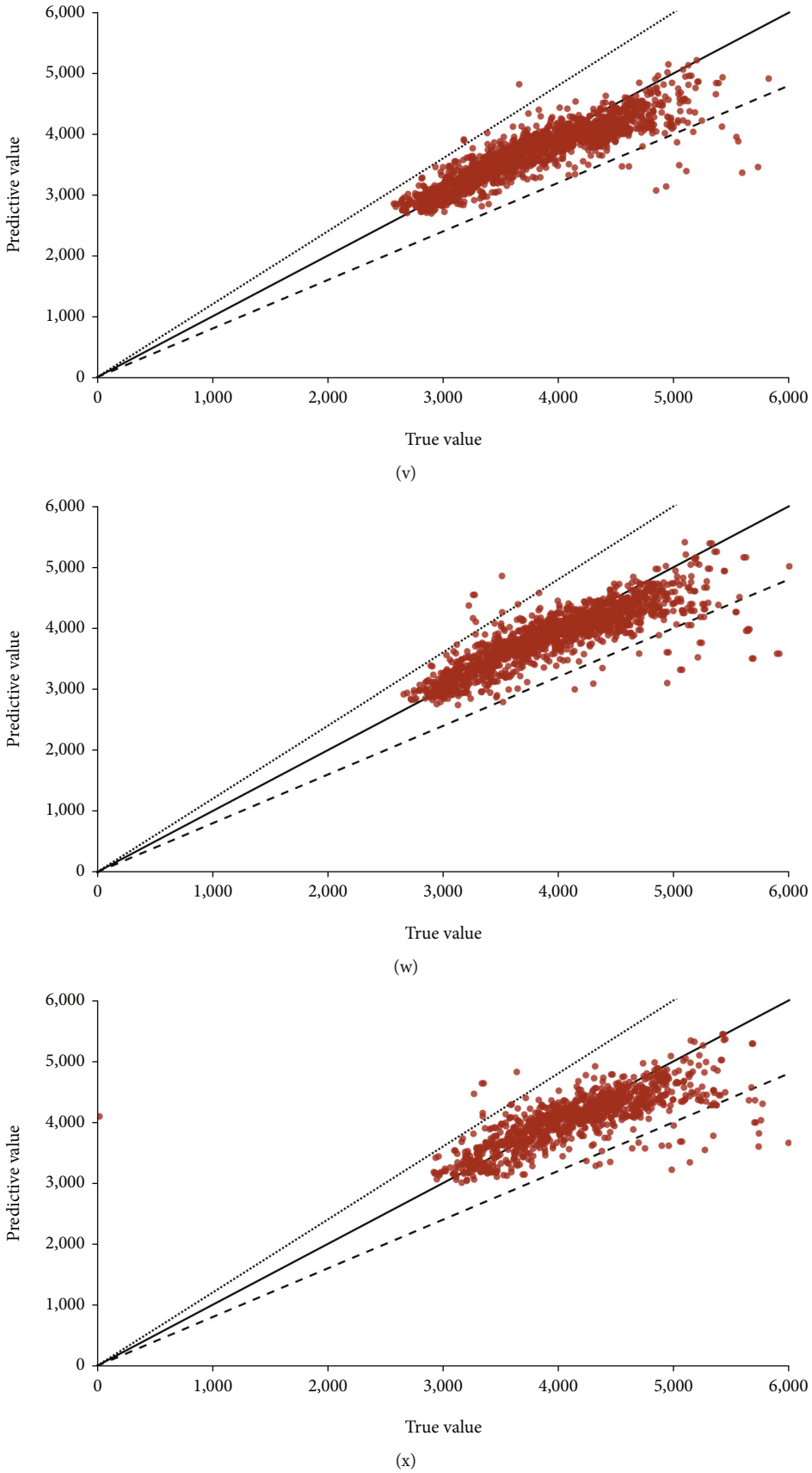



Figure 14: Comparison of the predicted value and the true value under different driving sections.

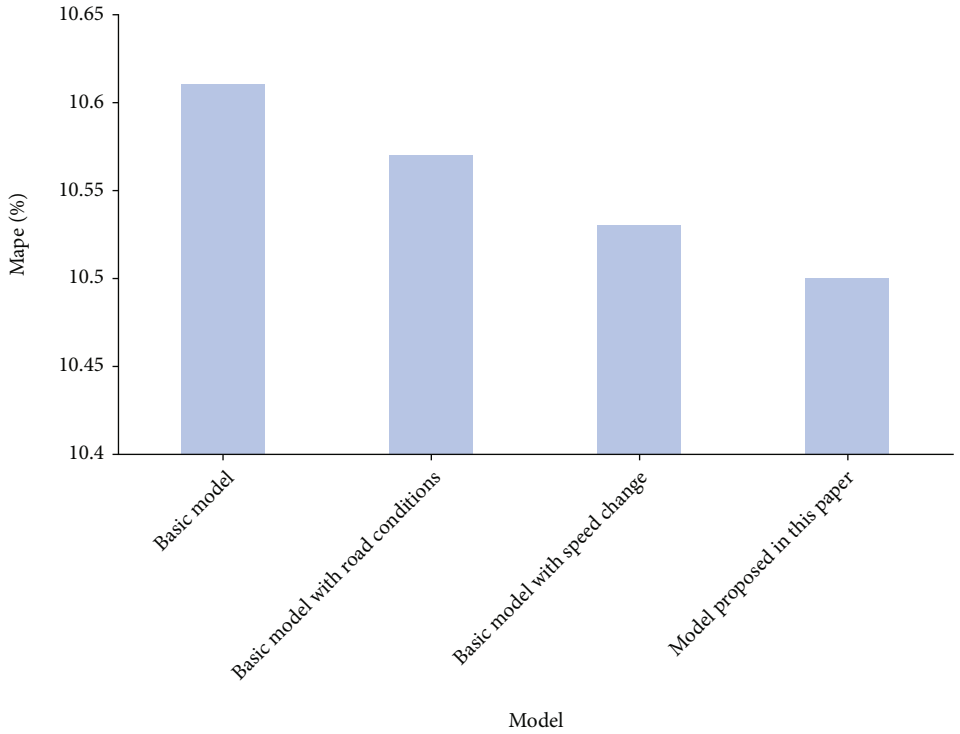



Figure 15: Contribution of proposed features.

time between bus stations by comprehensively considering the influence of temporal and spatialfeatures. To compare the different features of the building model prediction accuracy of contribution, the XGBoost prediction model, which does not include the two characteristics of the change of the shift speed before and after the adjacent road section and the road conditions, is used as the basic model. The road conditions include three influencing factors: the proportion of bus lanes, the number of traffic lights, and the turning situation of bus. Add these two features, respectively, or in combination into the feature project to predict, and get the performance of the model under different conditions, as shown in Figure 15. The MAPE of the benchmark model was 10.61%.

On the basis of the benchmark model, if any of the features including the change of shift speed before and after adjacent sections and road conditions are added, the model prediction will have a positive impact, that is, to improve the prediction accuracy of the model. The positive improvement of the model is the largest if the change of shift speed before and after the adjacent road section and the road condition are included simultaneously. This shows that the conventional prediction model of bus station travel time based on the proposed influencing factors is feasible.

## 6. Conclusion

This paper presents an algorithm for travel time between bus stations prediction based on XGBoost model, which utilizes variations of the flight speed in adjacent sections before and after, bus lane proportion, real-time weather influence on bus travel speed, etc. as characteristics. Compared with other prediction models, the XGBoost prediction model has the lowest MAPE value of 11.96%, which is 9.30% lower than the other prediction models on average. It is proved to have higher accuracy and stronger reliability. In addition, comparing the prediction results of the model in different time periods and different number of driving sections, extensive experiments demonstrate that the prediction accuracy of the model is high in the nonpeak hours, and the more the number of driving sections, the more stable and reliable the performance of the model, when the number of line units exceeds a certain value, the accuracy of the prediction model tends to stabilize and the prediction error is basically below 7%. Furthermore, the improvement effect of the proposed factors on the model is analyzed.

This paper has achieved the expected goal to a certain extent, but there are still some deficiencies. This paper only selects the bus operation data of one single line in Guangzhou for analysis and prediction, without considering other bus lines to firmly bridge the gaps between the theoretical research and the application of the developed model. The variation of the stopping time and boarding passenger flow of bus vehicles in the process of running is also an important factor affecting the travel time between conventional bus stations. Further research is needed to improve the prediction accuracy.

## Data Availability

Bus-related data used during the study were provided by a third party (Guangzhou Yangchengtong Company). Road and signal data during the study are available in a repository or online in accordance with funder data retention policies

(Baidu Map). Weather data used during the study are available in a repository or online in accordance with funder data retention policies (the agricultural meteorological big data system-WheatA).

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This paper is supported by Shenzhen Municipal Science and Technology Innovation Committee (Grant No. JCYJ2017 0818142947240), Shenzhen University, case teaching course construction project for professional degree postgraduates (Operational Research).

## References

[1] H. Wang and C. Xu, "Bus travel time prediction model based on GPS," *Computer Measurement & Control*, vol. 20, no. 8, pp. 2204–2206, 2012.

[2] X. Tong, D. Lu, and T. Zhang, "Research on prediction model of travel time between bus stations based on time series method——taking Suzhou no. 1 bus as an example," *Journal of Transportation Engineering and Information*, vol. 15, no. 4, pp. 114–119, 2017.

[3] H. Zhang, S. Liang, Y. Han, M. Ma, and R. Leng, "A prediction model for bus arrival time at bus stop considering signal control and surrounding traffic flow," *Access*, vol. 8, pp. 127672–127681, 2020.

[4] W. Zhou, J. Xu, and Z. Liu, "Prediction of travel time of public transport vehicles based on Kalman filter algorithm," *Communications Standardization*, vol. Z1, pp. 174–177, 2007.

[5] L. Ma, T. Chen, and M. Hao, "Bus travel time prediction algorithm based on multi-line information fusion," *Computer Science*, vol. 46, no. 11, pp. 222–227, 2019.

[6] H. Huo, J. Shen, and C. Zheng, "Prediction of bus arrival time based on KNN algorithm," *Journal of Transportation Engineering and Information*, vol. 18, no. 4, pp. 76–82, 2020.

[7] W. Xie, "Prediction of bus arrival time based on BP neural network," *Journal of Shengli College China University of Petroleum*, vol. 30, no. 4, pp. 38–40, 2016.

[8] Y. Han, L. Zhou, and P. Gao, "Research on dynamic travel time prediction method of public transport based on BP neural network," *Periodical of Ocean University of China*, vol. 50, no. 2, pp. 142–154, 2020.

[9] P. He, G. Jiang, S. K. Lam, and D. Tang, "Travel-time prediction of bus journey with multiple bus trips," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 11, pp. 4192–4205, 2019.

[10] B. Yu, Z. Yang, and J. Lin, "Application of support vector machine to predict bus running time," *System Engineering Theory and Practice*, vol. 4, pp. 160–164, 2007.

[11] Y. Bin, Y. Ting, T. Xiao-Mei, G. B. Ning, and S. Q. Zhong, "Bus travel-time prediction with a forgetting factor," *Journal of Computing in Civil Engineering*, vol. 28, no. 3, 2012.

[12] L. Jing, C. Xie, and A. Wang, "Research on prediction model of bus arrival time based on integrated learning," *Journal of Chongqing University of Technology (Natural Science)*, vol. 33, no. 10, pp. 47–53, 2019.

[13] Y. Xie, W. Xiang, and M. Ji, "Application analysis of forecasting monthly housing rent based on Xgboost and LightGBM algorithms," *Computer Applications and Software*, vol. 36, no. 9, pp. 151–155, 2019.

[14] M. Chen, Q. Liu, and J. Zhang, "Power system transient stability prediction method based on XGBoost," *Power System Technology*, vol. 44, no. 3, pp. 1026–1034, 2020.

[15] W. Jia, L. Sun, and Y. Jing, "Prediction of prognostic quality score of femoral neck fracture surgery based on XGBoost model," *Journal of Taiyuan University of Technology (Social Sciences Edition)*, vol. 49, no. 1, pp. 174–178, 2018.

[16] Y. Zhang, H. Chen, and Y. Zhang, "Haze prediction method based on XGBoost," *Computer Engineering and Design*, vol. 40, no. 12, pp. 3631–3638, 2019.

[17] Y. Zhong, Y. Shao, and W. Wu, "Short-term traffic flow prediction model based on XGBoost," *Science Technology and Engineering*, vol. 19, no. 30, pp. 337–342, 2019.

[18] L. Zou, S. Shu, X. Lin, K. Lin, J. Zhu, and L. Li, "Passenger flow prediction using smart card data from connected bus system based on interpretable XGBoost," *Wireless Communications and Mobile Computing*, vol. 2022, 13 pages, 2022.

[19] X. Dong, T. Lei, S. Jin, and Z. Hou, "Short-term traffic flow prediction based on XGBoost," in *2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 854–859, 2018.

[20] Q. Du, F. Yin, and Z. Li, "Base station traffic prediction using XGBoost-LSTM with feature enhancement," *IET Networks*, vol. 9, no. 1, pp. 29–37, 2020.

[21] Y. Jing, H. Hongtao, S. Guo, X. Wang, and F. Chen, "Dynamic differential pricing of high-speed railway based on improved GBDT train classification and bootstrap time node determination," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, pp. 1–13, 2021.

[22] X. Wang, L. Huang, H. Huang, B. Li, Z. Xia, and J. Li, "An ensemble learning model for short-term passenger flow prediction," *Complexity*, vol. 2020, Article ID 6694186, 13 pages, 2020.

[23] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of Max-dependency, Max-relevance, and Min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[24] D. N. Reshef, Y. A. Reshef, H. K. Finucane et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[25] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System: KDD '16," in *The 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2016.

[26] S. Thongsuwan, S. Jaiyen, A. Padcharoen, and P. Agarwal, "ConvXGB: a new deep learning model for classification problems based on CNN and XGBoost," *Nuclear Engineering and Technology*, vol. 53, no. 2, pp. 522–531, 2021.