Hindawi
Complexity
Volume 2020, Article ID 6694186, 13 pages
https://doi.org/10.1155/2020/6694186

Research Article

An Ensemble Learning Model for Short-Term Passenger Flow Prediction

Xiangping Wang, Lei Huang, Haifeng Huang, Baoyu Li, Ziyang Xia, and Jing Li

School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Jing Li; jingli@bjtu.edu.cn

Received 3 November 2020; Revised 16 November 2020; Accepted 8 December 2020; Published 19 December 2020

Academic Editor: Abd E.I.-Baset Hassanien

In recent years, with the continuous improvement of urban public transportation capacity, citizens' travel has become more and more convenient, but there are still some potential problems, such as morning and evening peak congestion, imbalance between the supply and demand of vehicles and passenger flow, emergencies, and social local passenger flow surged due to special circumstances such as activities and inclement weather. If you want to properly guide the local passenger flow and make a reasonable deployment of operating buses, it is necessary to grasp the changing law of public transportation short-term passenger flow. This paper builds a short-term passenger flow prediction model for urban public transportation based on the idea of integrated learning. The goal is to use the integrated model to accurately predict the short-term passenger flow of urban public transportation, using Multivariable Linear Regression (MLR), K-Nearest Neighbor (KNN), eXtreme Gradient Boosting (XGBoost), and Gated Recurrent Unit (GRU) as the four seed models, and then use regression algorithm to integrate the model and predict the passenger flow, station boarding and landing, and cross-sectional passenger flow data of the typical representative line 428 in the "Huitian Area" of Beijing from January 1, 2020, to May 31, 2020. Finally, the prediction results of the submodels are compared with those of the integrated model to verify the superiority of the integrated model. The research results of this paper can enrich the short-term passenger flow forecasting system of urban public transportation and provide effective data support and scientific basis for the passenger flow, vehicle management, and dispatch of urban public transportation.

1. Introduction

According to the annual report on Beijing's Transport Development in 2020, by the end of 2019, the Beijing Public Transport Group has 28,271 buses and 1,620 routes in operation. The annual passenger volume of electric buses reached 3.564 billion, with an average daily passenger volume of 9.7377 million, providing great convenience for Beijing residents to travel, and it is the main undertaker of Beijing's surface public transportation.

In recent years, the characteristics of public transport network operation have become increasingly obvious; also, some potential problems gradually emerged, such as traffic jams during rush hours, traffic supply and demand not matching, a large number of passengers commuting security hidden danger in passenger flow gathering in a certain space, and some large activities, bad weather, and bus fault under the special operating environment urgent need for rapid evacuation etc. At the same time, with the increasing development of urban public transportation informatization, the Advanced Public Transportation System (APTS) has become an indispensable part of the construction of a "Smart City" as a result of the accumulation of massive public transport IC card data assets. At present, Beijing has built a bus GPS data acquisition system, line network management system, and other basic data such as BUS IC card, BUS GPS, and bus network. Thanks to the rapid development of computer technology, the methods of machine learning and deep learning display advantages such as high computational efficiency and strong data processing ability. Applications based on big data prediction technology, comprehensive and accurate projections for short bus traffic, are to promote effective shuttle buses and other public traffic modes; to improve the utilization rate of public transport vehicles, optimization of vehicle dispatching, and the important measures to enhance the level of public transportation system management and operation; and also are the core of the realization of the intelligent transport system.

At present, the public transportation enterprises are in the process of actual operation, and the vehicle operation dispatching scheme formulation depends largely on historical experience. The forecasting ability for short-time bus passenger flow of each station, line and period is insufficient. It will inevitably lead to public transport vehicles not being able to get reasonable scheduling, affect the passengers, and impact on the effective running of the bus system. Therefore, it is of great importance to use big data situation analysis technology to accurately predict short-time bus passenger flow based on traffic IC card data and external weather data to analyze and master the transport demand and passenger flow rule of public transportation. Forecasting traffic demand is a core issue in any transportation system organization, and the future demand provided by predictive algorithms means that a reasonable supply can be planned in advance. Bus passenger flow related indicators reflect the passenger travel demand and regularity; can, for the operators in time according to the current system resource, adjust operation plans such as temporary or reduce extra trains and other transportation emergency cases combined effective disposal; and provide a scientific basis for narrowing the scope of the influence of the incident. As a result, it is necessary for public transportation to study the short-term passenger flow forecast, build higher prediction accuracy of the model, and obtain more reliable short-term passenger flow distribution, so as to solve the above problems effectively.

In this paper, the integrated learning method is introduced into the model of short-term bus passenger flow prediction, which significantly improves the accuracy of bus passenger flow prediction and provides a new modeling method for the quantitative research of public transportation, which has the dual significance of theoretical guidance and method innovation.

2. Literature Review

Short-term passenger flow prediction is an important part of the intelligent transportation system, which can be used to assist the adjustment of travel behavior, reduce passenger flow congestion, and improve the service quality of the transportation system. The evolution of the passenger flow prediction method is a process of continuous development and expansion, from the initial linear estimation model to the current model of machine learning and deep learning, gradually towards maturity. Generally speaking, short-term passenger flow prediction methods can be divided into two categories: parametric method and nonparametric method. The main difference between these two types of methods lies in the assumed functional dependence between independent variables and dependent variables [1].

In the traditional parametric methods, there are mainly Autoregressive Models (AR), Exponential Smoothing (ES) [2], Autoregressive Integrated Moving Average (ARIMA) model [3], and so on. ARIMA model is a linear combination of time-delay variables and error terms. Since the 1970s, the ARIMA model has become one of the commonly used parameter prediction methods and has been widely applied to the prediction of short-term traffic data such as traffic flow, travel time, and speed. Based on the historical passenger flow data collected by the urban rail transit automatic ticketing system, Cai et al. [4] used the ARIMA model to predict the passenger flow of Guangzhou metro. In addition, due to the seasonal and trend characteristics of passenger flow time series data, some researchers have applied Seasonal Autoregressive Integrated Moving Average (SARIMA) model to predict passenger flow. In order to deal with the strong seasonal autocorrelation of the time series of passenger flow of Serbian railway, Milenković et al. [5] used the SARIMA model to predict the passenger flow of Serbian railway, which shows good prediction performance. Wang et al. [6] analyzed the rule of passenger flow in and out of Beijing subway station with time change, and the SARIMA model is used for modeling. The results show that the predicted results can accurately reflect the time change rule of passenger flow in and out of Beijing subway station. Because these parametric models assume linear relationships between variables with time delay, it is difficult to capture nonlinear relationships between variables, so the use of traditional parametric methods is limited [7, 8].

In order to better deal with the nonlinear characteristics of passenger flow data, the nonparametric method is introduced. Different from the parametric method, the nonparametric method is to establish the nonlinear relationship between input variables and output variables without prior knowledge. Therefore, it is more flexible and widely used in passenger flow prediction. Guo et al. [9] used 15 minutes of time interval summary of real traffic flow data compared, and the experiment shows that the adaptive Kalman filtering method can get a feasible prediction accuracy, especially under the condition of traffic high volatility, shows how to improve the adaptability of this method and, finally, puts forward the suggestions to improve the short-term traffic flow prediction performance. According to the characteristics of bus passenger flow and the law of changing with time, Deng et al. [10] proposed a prediction model of multicore least-squares support vector machine. The model fully considers the influence of historical data on bus passenger flow. Zhao et al. studied the passenger flow distribution in each period of the bus line by using the method of combining wavelet analysis and neural network and predicted the passenger flow of the short-time bus line, so as to realize the dynamic control and reasonable scheduling of the bus. Zhang and Yang [11] combined the main factors affecting passenger flow with the neural network self-learning method and established a subway passenger flow prediction model based on the neural network of spline weight function. Wang et al. [12] used a correlation analysis method to analyze the relationship between pedestrian flow and its influencing factors, extracted 11 important influencing factors, and established a prediction model of pedestrian flow using the modular neural network. Among these nonparametric methods, neural networks are widely used because of their good adaptability, nonlinearity, and ability to map arbitrary functions [13, 14].

In the era of big data, the data processing capacity and prediction accuracy of the model have higher requirements. Researchers have made efforts to increase network density, and Hinton et al. [15] first proposed the concept of deep learning in 2006. Compared with the traditional neural network and other shallow learning models, deep learning is equivalent to a deeper neural network; that is, there are more hidden layers, which enable it to express more abstract and higher-level nonlinear features and more accurately capture the "deep" features of short-term passenger flow. Bai et al. [16], aiming at the short-term prediction of bus passenger flow, used the Deep Belief Network (DBN) to establish a prediction model. Compared with the classical parametric method and nonparametric method, this model shows a good predictive advantage. Li Bangpeng, respectively, used the convolutional neural network and the time-length neural network prediction model in deep learning to predict the future indoor spatial and temporal passenger flow distribution based on the real spatial and temporal passenger flow data and made a model comparison.

At present, integrated learning is a widely used method in machine learning, which integrates different learner sets so as to improve the accuracy of prediction [17]. In order to facilitate the collection, the mainstream of the current research is the design algorithm which promotes the weak learner to the strong learner and integrates multiple learners generated by the same algorithm. Freund and Schapire [18] proposed the Adaptive Boosting (AdaBoost) algorithm, which uses sequence sampling and has high operational efficiency and practical application value. The Bagging algorithm proposed by Breiman [19], which uses self-sampling to combine the base learner, was subsequently improved into Random Forest (RF) in 2001 [20] and had become the most classic algorithm in Bagging integration. In 1992, Wolpert [21] proposed the stacked generalization (stacked generalization) model, but the stacking algorithm only provides the integrated idea, for its selection of learning has certain subjectivity and then the selection of some scholars to study the certain research, such as Ledezma et al. [22] and Xu Huili, to use the genetic algorithm in the metamodel and the selection of the base model is optimized. The stacking algorithm has difficulty in obtaining the correct base learner assembly. Integrated learning, due to the combination of multiple learners, greatly improves the prediction accuracy and generally performs better than each component model, which benefits from the diversity among models, reduces the risk of using isolated models, and compensates for the shortcomings of each model [23, 24]. In addition, its models can solve many problems that a single model cannot solve. The passenger flow of urban public transport is dynamic and random, so it is difficult for a single model to fit its trend well, and integrated learning can better make up for this deficiency.

In conclusion, due to the complexity and randomness of bus passenger flow, as well as the higher requirements of big data on the data processing capacity and prediction accuracy of the prediction model, the use of traditional parameter methods and shallow neural network methods is limited. The application of deep learning, integrated learning, and other methods provides a new opportunity for accurately capturing the nonlinear characteristics of STW passenger flow and processing large quantities of multisource data.

3. Materials and Methods

3.1. Data Selection and Processing. This paper selects the card-swiping passenger volume, station boarding and landing volume, and section passenger volume data of the typical representative bus line 428 in the "Huitian area" from January 1, 2020, to May 31, 2020, for the key index prediction. The data source is the IC card data of Beijing Public Transport Group, with a total amount of about 107,000 pieces of data. Based on the basic analysis of the card-swiping data, it can be known that most of the bus operation time period is from 05:00 to 24:00, and the number of card-swiping times within 15 minutes during this time period is counted; that is, each indicator should get 76 data based on the granularity of 15 minutes a day. The processing of time series data first needs to be converted into a supervised sequence according to the set time step; that is, for certain data, it is considered that the data of its previous time step bar has obtained this data (time step is the number of time steps). In this process, the daily supervised sequence length is the original daily time series length minus the time step.

3.2. Analysis of the Key Indexes of Urban Bus Network Monitoring in "Huitian Area". This part monitors and analyzes the three key indicators related to the passenger volume of route 428, namely, the card-swiping passenger volume, station boarding and landing volume, and section passenger volume. The time frame is from January 1, 2020, to May 31, 2020.

Route 428 is metro Longze Station-Tiantong Beiyuan Station, including 32 stations. The operating mileage of the line is 13.9 km, the average one-way running time is 47.73 minutes, and the average running speed is 17.74 km/h. There are 20 vehicles in operation. There are 100 trains per day and 19 in peak hours. The average daily passenger throughput is 3,474.

3.2.1. Card-Swiping Passenger Volume. As shown in Figure 1, due to the impact of the epidemic, the passenger volume of card swiping during the Spring Festival and the epidemic prevention and control period after the festival was significantly lower than the normal situation before the festival, while the passenger volume of card swiping during the epidemic prevention and control period after the festival was generally low and slowly picked up, with a weekly increase.

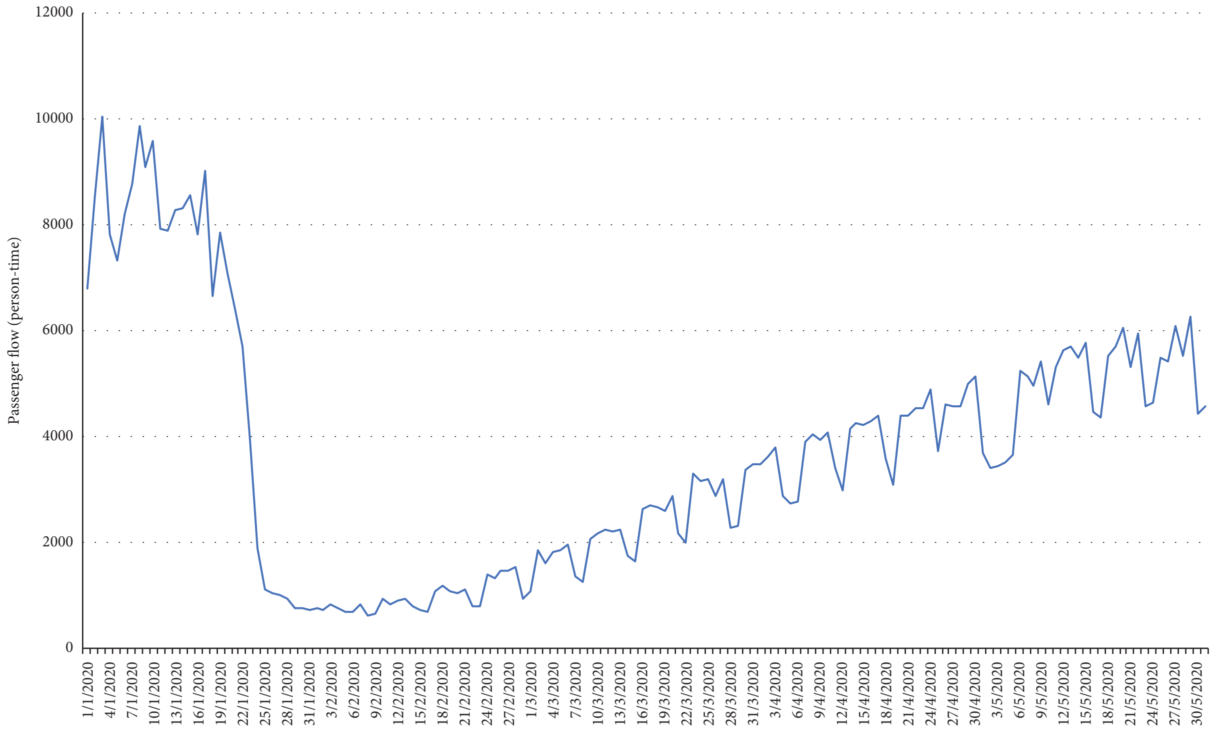

Figure 1: Passenger volume of the No. 428 bus.

3.2.2. Boarding and Landing Volume. The boarding and landing volume of bus number 428 is shown in Figure 2, in the direction of Metro Longze Station to Tiantong Beiyuan Station, the average daily volume for the largest station is 1036 (Banjieta Village North Station), and the average daily volume for the smallest station is 14 (the north gate of District 1, Harmony Garden). In the direction of Tiantong Beiyuan Station to Metro Longze Station, the average daily volume for the largest station is (Longjinyuan Area 4) and the minimum is 33 (Longxiyuan 3rd District Intersection West).

3.2.3. Sectional Passenger Volume. The average daily section passenger volume of bus number 428 is shown in Figure 3. The stations with the largest passenger volume in the direction of Longze Station and Tiantong Beiyuan Station are Banjieta Village North Station and Banjieta Village East Station. Tiantong Beiyuan Station–Metro Longze Station direction section passenger volume is the largest station for Xiaoxinzhuang East Station.

3.3. Model Selection. Bus passenger volume is affected by more external environment, and it is difficult for a single model to learn its complicated rules. Short-term prediction is essentially a question of time sequence, to the problem of the prediction which is usually not a model that can be applied to all scenarios, and integrated thinking is through a combination of several single models to reduce the risk of the error model, by giving full play to the information of the prediction results of each submodel to make up for the shortcoming of single model that the prediction error is large due to the influence of random factors, thus improving the prediction performance. This paper constructs four seed models of Multivariable Linear Regression (MLR), K-Nearest Neighbor (KNN), eXtreme Gradient Boosting (XGBoost), and Gated Recurrent Unit (GRU) and also constructs the regression integration model.

3.3.1. MLR. In this paper, we study the influence of many factors and so the selection of the most commonly used multiple linear regression, the simple model principle as shown in Figure 4.

3.3.2. KNN. KNN is a model based on distance. Figure 5 shows the algorithm principles of the classification model, according to the K value selection near the element, the element near the largest number of categories.

3.3.3. XGBoost. XGBoost is a boosting tree model based on ensemble learning boosting, which is based on regression tree. Once proposed, this method has been widely used in much research and many enterprises because of its high efficiency and accuracy. Some studies have shown that the prediction accuracy of this method can be comparable to the neural network and deep learning in dealing with time series problems.

3.3.4. GRU. GRU combines the forget gate and the input gate into one and mixes the cell state C and the hidden state. The final model is simpler than the standard LSTM, as shown in Figure 6.

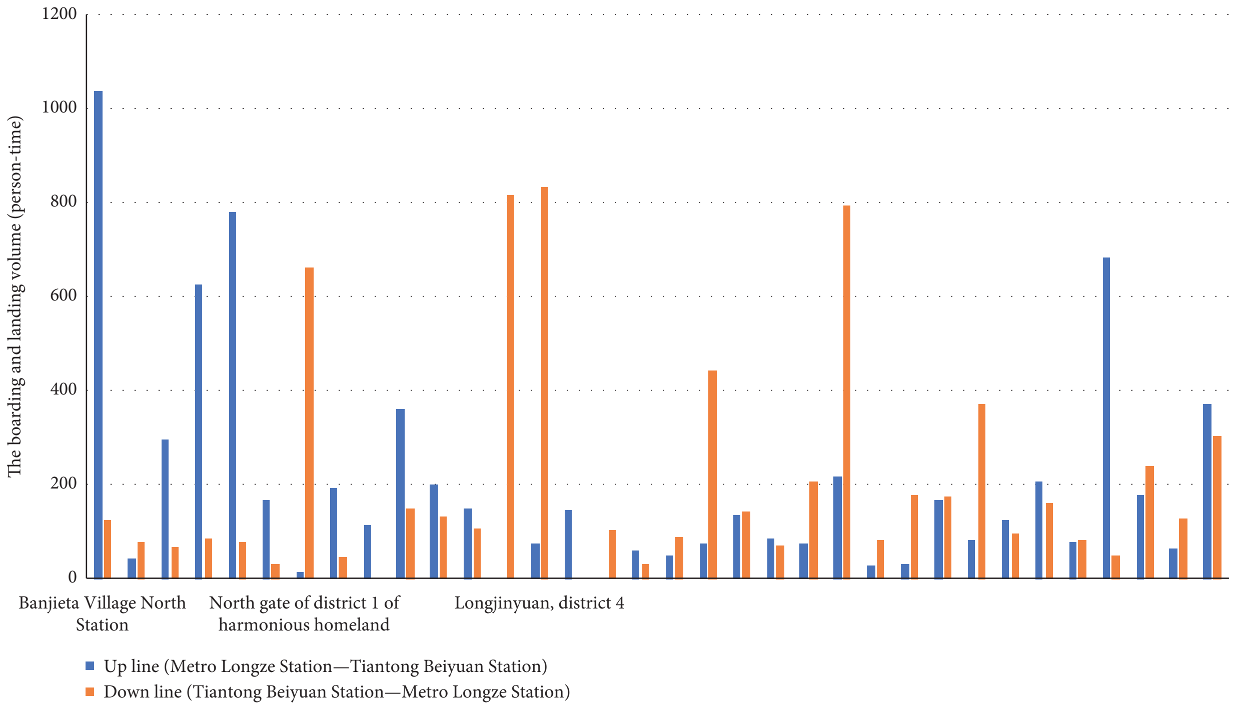

Figure 2: The boarding and landing volume.

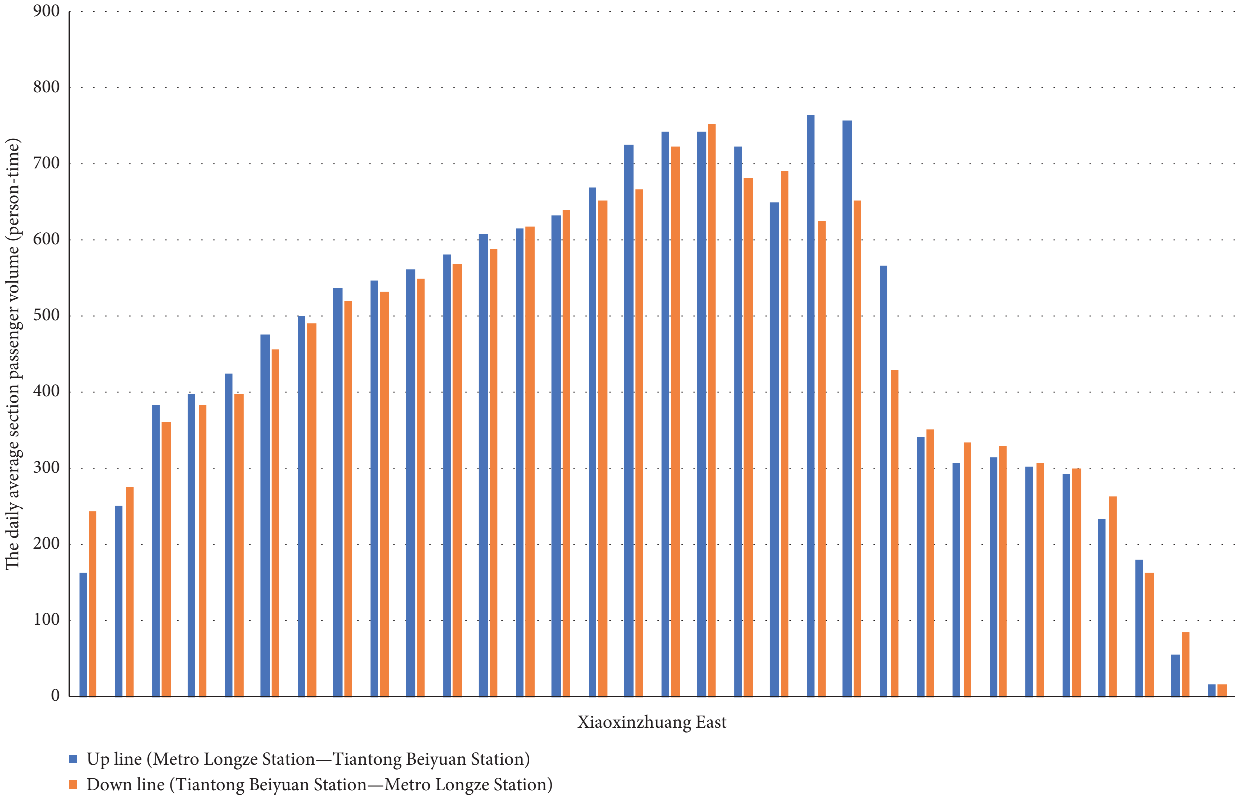

Figure 3: Daily average section passenger volume of the number 428 bus.

3.4. Build the Regression Integration Model. Integrated learning is an idea rather than a specific algorithm in machine learning. The core of this method is to combine multiple models called weak learners into a more accurate model. The integrated model uses different sampled data to train these weak learners continuously, adjusts the weak

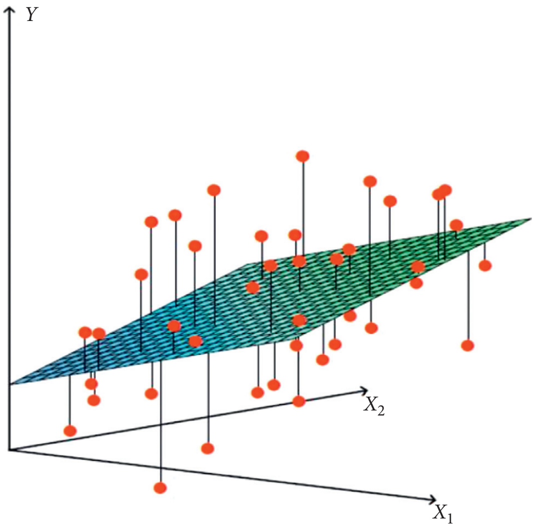

Figure 4: Schematic diagram of linear regression.

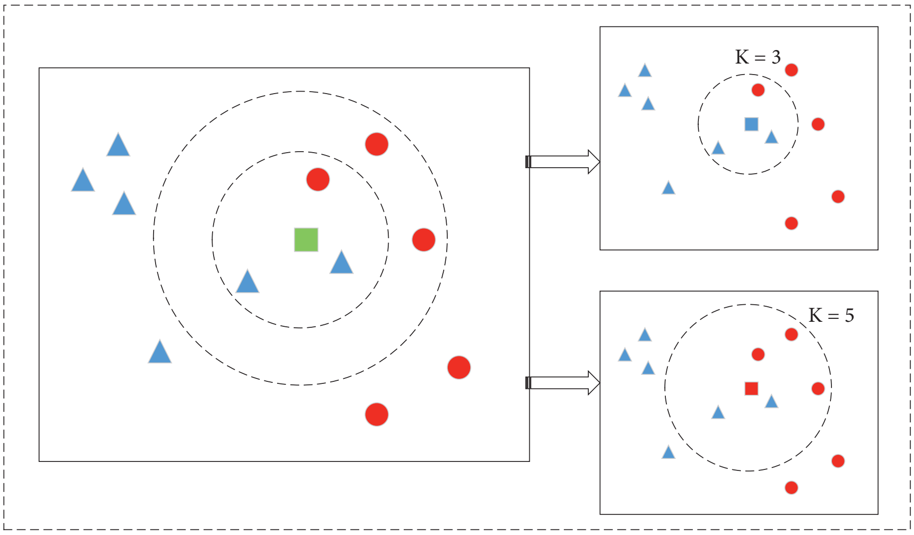

Figure 5: Schematic diagram of the KNN algorithm.

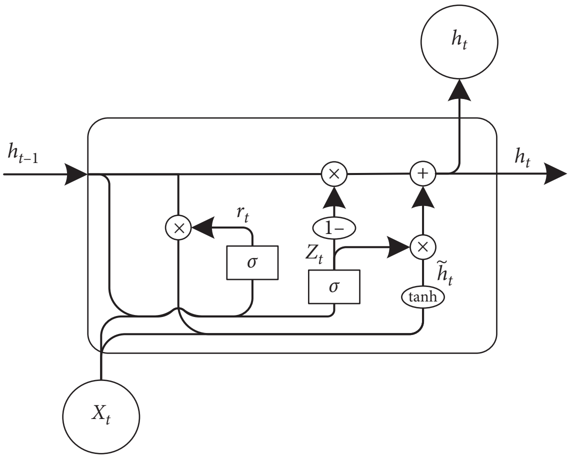

Figure 6: Internal structure of GRU.

learners through errors, and effectively combines the predicted results of the weak learners to a certain extent.

3.4.1. The Advantages of Integrated Learning. If the individual model is compared to a decision-maker, the integrated learning approach is equivalent to multiple decision-makers working together to make a decision. The advantages of ensemble learning are as follows: (1) overall, ensemble learning has a high accuracy rate; (2) the introduction of randomness makes it not easy to overfit, has good antinoise ability, is not sensitive to outliers of abnormal points, and can handle high-dimensional data without making feature selection; and (3) it can process both discrete data and continuous data. In addition, the data set does not need normalization, so the overall training speed is considerable.

3.4.2. The Regression Integration Model Based on GBDT. In this part, we combine the prediction results of the four-seed model with the regression model. This paper selected the regression model is Gradient Boosting Decision Tree (GBDT); this algorithm is based on the integration of learning. The passenger flow results predicted by each submodel are input into the GBDT model as an independent variable and the real value of passenger flow as a dependent variable for a new round of learning. Some nonlinear relations between the predicted results of the submodels and the real values can be learned through regression models, and the advantages of different submodels can be brought into play to make up for the disadvantages of different models. The model is shown in Figure 7.

3.5. Set Evaluation Index. In order to more comprehensively compare the different prediction results caused by the selection of different parameters in the same model, this paper selects Root Mean Square Error (RMSE) as the objective function of the optimization model and selects Mean Absolute Error (MAE) as the index of the evaluation model. Its definitions are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\mathrm{true}} - y_{\mathrm{pred}}\right)^{2}}, \quad (1)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_{\mathrm{true}} - y_{\mathrm{pred}}\right|, \quad (2)$$

y_{true} represents the actual value, y_{pred} represents the predicted value, and N represents the predicted sample number. Both indicators reflect the size of the error between the predicted value and the actual value, but the former is more able to amplify the error, while the latter reflects the true error. The smaller the values of RMSE and MAE, the closer the predicted value to the actual value and the higher the prediction accuracy of the model.

4. Results

The passenger flow of swiping card and the boarding and landing volume reflects the passenger flow of a certain line or a certain station, and the passenger flow of section reflects the passenger flow between two adjacent stations on the line. The three indexes correspond to the essential basic data for optimizing the design of the route network and deploying vehicles in the public transport system, as well as the important basis for planning the bus dispatching frequency and considering whether to set interregional buses. Therefore, this paper selects three basic indicators of passenger flow—section passenger flow, card-swiping passenger flow, and boarding and landing volume for short-term prediction, providing the basis for rational planning of bus network, allocation of bus station facilities, and preparation of its operation plan.

4.1. The Prediction of Section Passenger Volume. This part selects the passenger volume data of the section with a grain size of 15 minutes from January 1, 2020, to May 31, 2020, in the upward direction of Xiaoxinzhuang East Station of number 428 bus in the “Huitian Area”. Excluding the data not in the bus operation time, there are a total of 76 pieces of data in a day, with a total of 11,552 pieces of data. The time step was selected as a comparison of 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, and 6 hours. In other words, time step values were 1, 2, 4, 8, 12, and 24. In addition, the data ratio of the training set, verification set, and test set is 7 : 1 : 2, with 8086, 1156, and 2310 pieces of data, respectively.

It can be seen from the comparison of MAE and RMSE precision in Tables 1 and 2 that the regression integration prediction effect is the best in all different time steps. The prediction effect of different time steps is shown in Figures 8–13.

4.2. The Prediction of Card-Swiping Passenger Volume. This part selects the card-swiping passenger volume data with a grain size of 15 minutes from January 1, 2020, to May 31, 2020, in the upward direction of number 428 bus in the “Huitian Area.” The time step was also selected to compare 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, and 6 hours. The data ratio of the training set, verification set, and test set was 6 : 2 : 2, with 8812, 2938, and 2938 pieces of data, respectively.

It can be seen from the comparison of MAE and RMSE precision in Tables 3 and 4 that the regression integration prediction effect is the best in all different time steps. The prediction effect of different time steps is shown in Figures 14–19.

4.3. The Prediction of Boarding and Landing Volume. In terms of the boarding and landing volume, 81,400 pieces of data have been collected from the North Station of Banjieta Village from the 15-minute ascending direction of bus number 428 from January 1, 2020, to May 31, 2020. In the same way, the data was converted into a supervised sequence according to the set time step; the time step was 15 minutes, 30 minutes, 1 hour, 2 hours, 3 hours, and 6 hours, respectively, and the data ratio of the training set, verification set, and test set was 6 : 2 : 2, with 48,840, 16280, and pieces of 16280 data, respectively.

It can be seen from the comparison of MAE and RMSE precision in Tables 5 and 6 that the regression integration prediction effect is the best in all different time steps. The prediction effect of different time steps is shown in Figures 20–25.

5. Discussion

According to the “no free lunch” theorem in machine learning theory, there is no algorithm that can solve all problems perfectly. Many factors such as the size and structure of the data set will affect the final result. For specific data sets and actual needs, we should consider how to choose a suitable algorithm. This paper proposes a method for

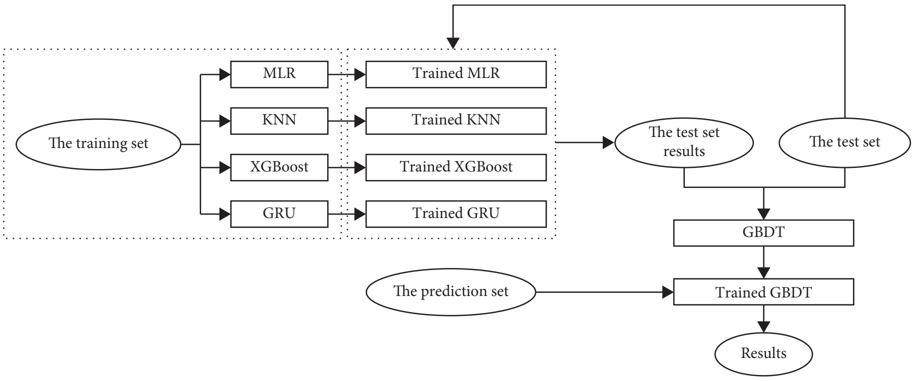

Figure 7: Ensemble flowchart based on the regression model.

Table 1: Multimodel MAE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	5.00	4.71	4.71	4.65	4.62	4.68
KNN	5.27	5.21	4.97	4.95	4.96	5.22
XGBoost	5.11	5.30	5.10	4.83	4.78	4.68
GRU	4.97	4.71	4.70	4.78	4.62	4.67
REG	4.94	4.61	4.70	4.62	4.57	4.67

Table 2: Multimodel RMSE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	7.40	6.69	6.72	6.66	6.65	6.81
KNN	7.68	7.52	7.14	6.99	7.05	7.45
XGBoost	7.51	7.77	7.33	6.91	6.84	6.72
GRU	7.33	6.89	6.75	6.75	6.54	6.65
REG	7.30	6.64	6.71	6.59	6.50	6.64

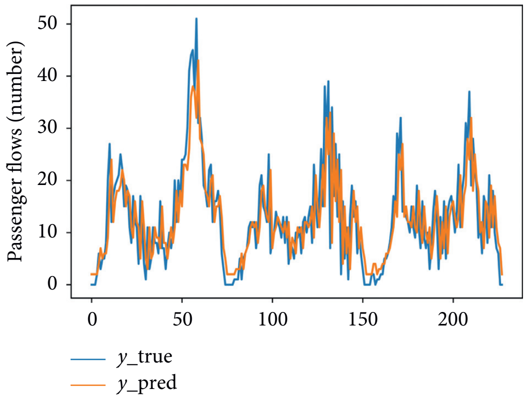

Figure 8: Regression integration prediction results when the time step is 15 min.

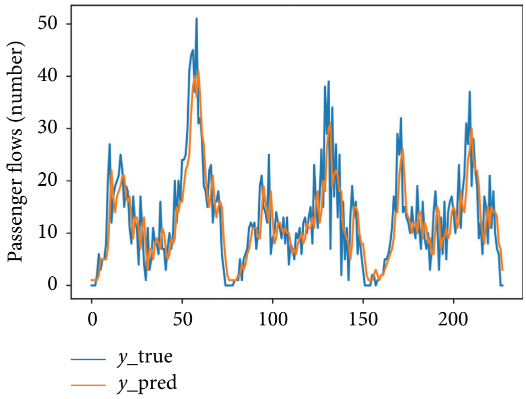

Figure 9: Regression integration prediction results when the time step is 30 min.

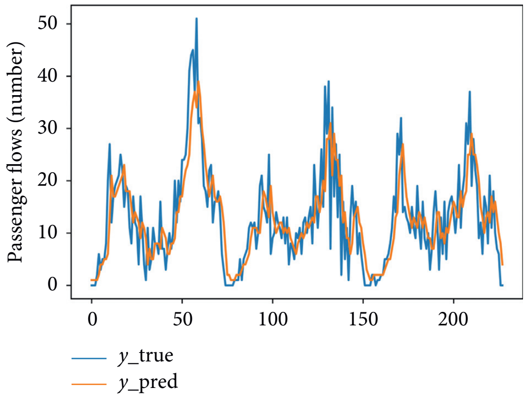

Figure 10: Regression integration prediction results when the time step is 2 h.

selecting the optimal model in regression prediction. The focus of this method is not the final specific model, but the selection process of the optimal model. Therefore, it is not limited to being used on a given data set. This is exactly the innovation of this article. The integrated model selection part is the focus of this article. Realize regression prediction for boarding and landing volume, cross section passenger flow, and card-swiping passenger flow. The same regression integration algorithm can be used to predict all MAE and RMSE. The submodels can also be divided into several categories, and different algorithms can be selected for

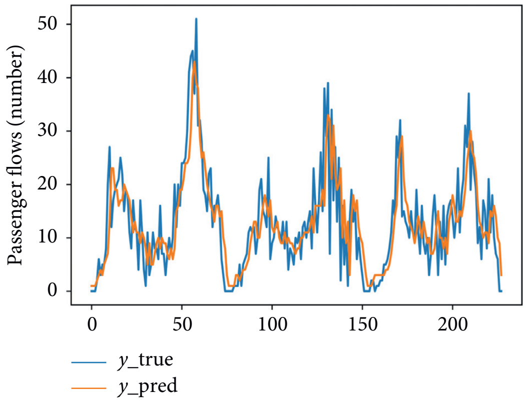

Figure 11: Regression integration prediction results when the time step is 2 h.

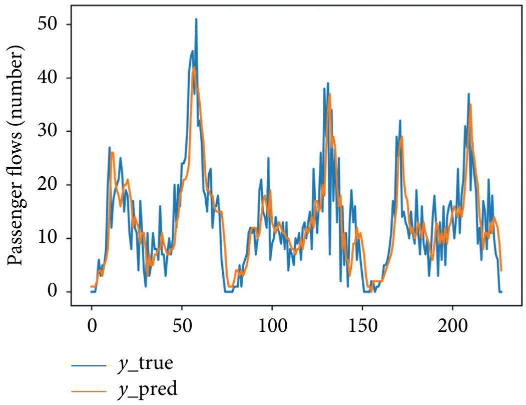

Figure 12: Regression integration prediction results when the time step is 3 h.

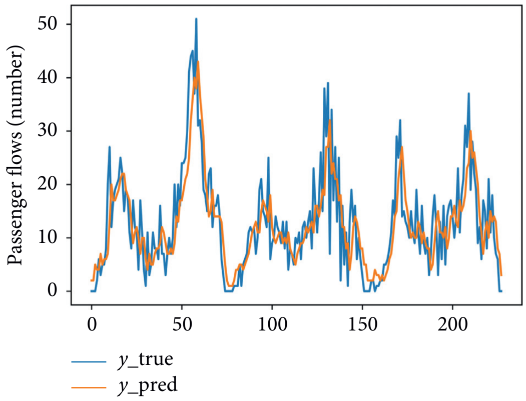

Figure 13: Regression integration prediction results when the time step is 6 h.

Table 3: Multimodel MAE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	17.87	17.67	16.85	16.93	17.53	20.74
KNN	19.84	18.84	18.93	21.37	18.24	18.71
XGBoost	19.91	20.47	18.72	18.75	17.45	17.51
GRU	18.05	18.04	16.81	19.39	18.34	18.64
REG	17.71	17.56	16.80	16.09	16.31	17.03

Table 4: Multimodel RMSE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	17.87	17.67	16.85	16.93	17.44	19.42
KNN	19.84	18.84	18.93	21.37	17.82	18.09
XGBoost	19.91	20.47	18.72	18.75	18.43	18.47
GRU	18.05	18.04	16.81	19.39	17.25	17.53
REG	17.71	17.56	16.80	16.09	16.90	17.22

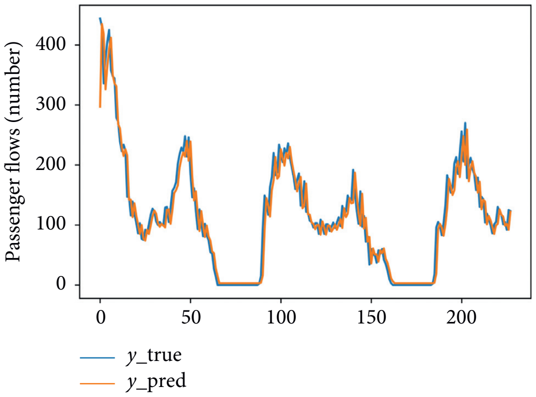

Figure 14: Regression integration prediction results when the time step is 15 min.

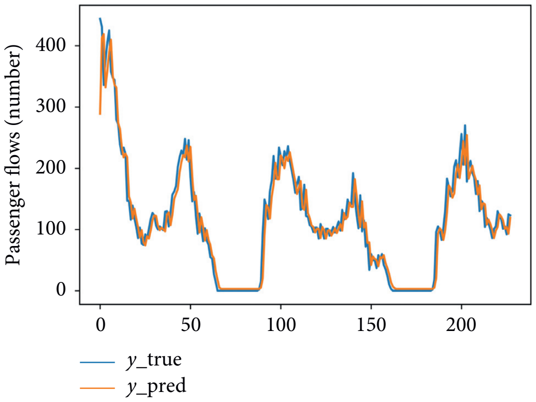

Figure 15: Regression integration prediction results when the time step is 30 min.

different types of indicator data sets. If different indicators are classified and predicted, there are problems of how to classify and which algorithm to choose. The method proposed in this paper is to use multiple algorithms to predict each index separately, select the optimal integrated model, and propose a comparative model to verify whether the selected optimal integrated model performs best. In the empirical study, four machine learning algorithms of KNN, LR, XGBoost, and GRU were used to predict boarding and landing volume, cross

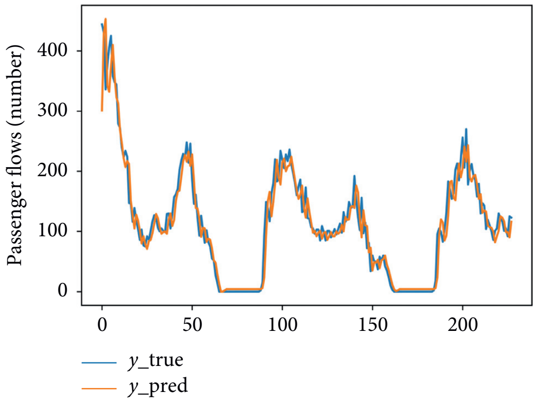

FIGURE 16: Regression integration prediction results when the time step is 1 h.

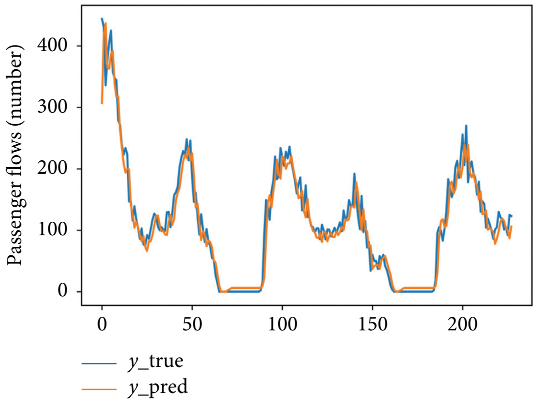

FIGURE 17: Regression integration prediction results when the time step is 2 h.

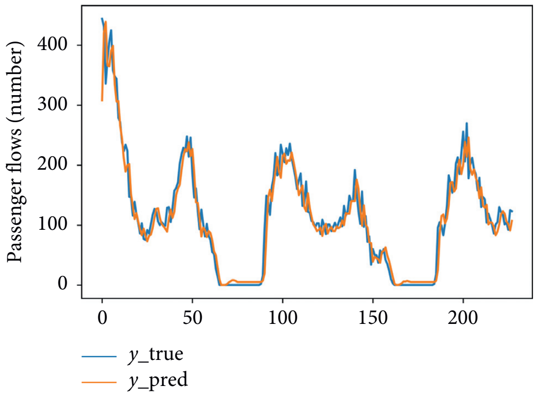

FIGURE 18: Regression integration prediction results when the time step is 3 h.

section passenger flow, and card-swiping passenger flow, respectively, finally comparing the prediction results of linear regression integration algorithms.

By comparison, in the cross section passenger flow prediction, the prediction results of LR and GRU at each step in the four submodels have lower MAE and RMSE values; the

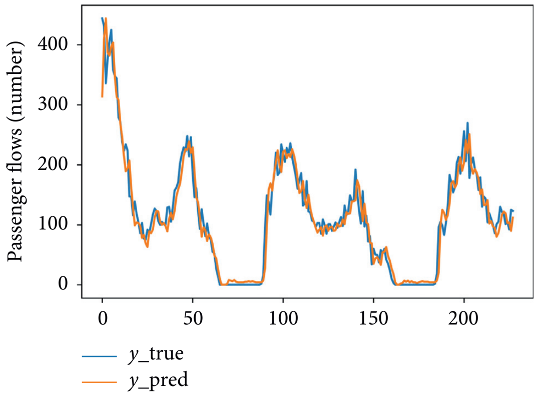

FIGURE 19: Regression integration prediction results when the time step is 6 h.

TABLE 5: Multimodel MAE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	5.79	5.19	5.04	5.05	5.05	5.04
KNN	5.95	5.65	5.44	5.05	4.96	5.00
XGBoost	5.71	5.37	5.41	5.06	4.91	4.86
GRU	5.70	5.04	4.94	4.73	4.82	4.72
REG	5.32	4.95	4.81	4.70	4.68	4.67

TABLE 6: Multimodel RMSE comparison table for different step sizes.

Models	15 min	30 min	1 h	2 h	3 h	6 h
LR	9.51	8.56	8.46	8.36	8.33	8.31
KNN	9.86	9.70	9.12	8.53	8.28	8.32
XGBoost	9.41	9.47	9.18	8.37	7.91	7.81
GRU	9.43	8.56	8.40	8.10	7.80	7.80
REG	9.37	8.54	8.37	7.95	7.70	7.78

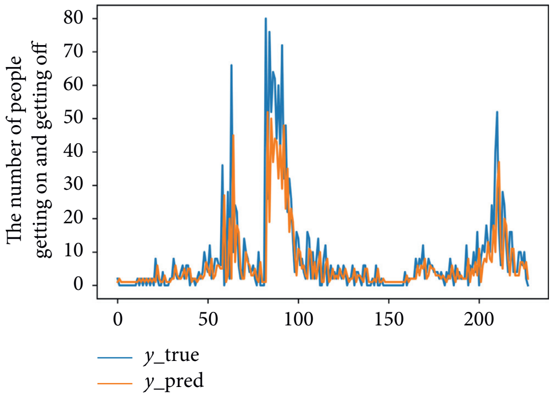

FIGURE 20: Regression integration prediction results when the time step is 15 min.

prediction results are more accurate; each submodel when the step length is 8 and 12 has relatively low MAE and RMSE values; and the results are more accurate. In comparison with

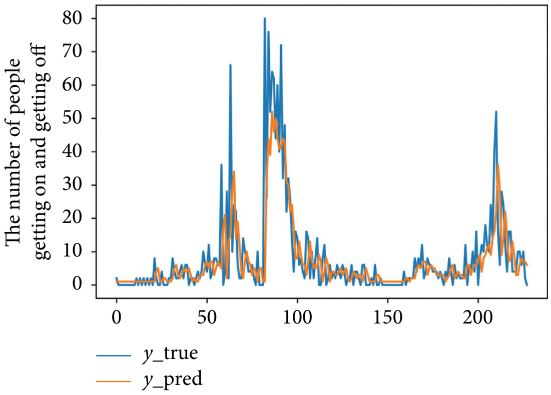

Figure 21: Regression integration prediction results when the time step is 30 min.

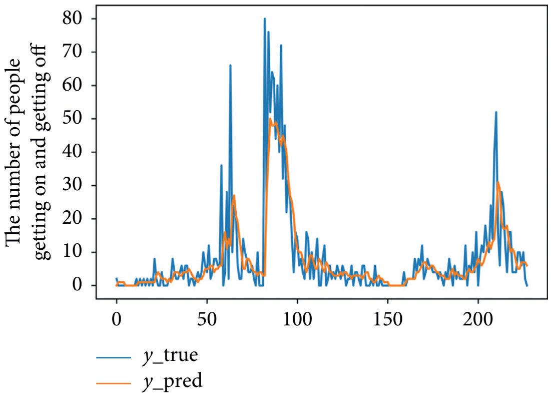

Figure 22: Regression integration prediction results when the time step is 1 h.

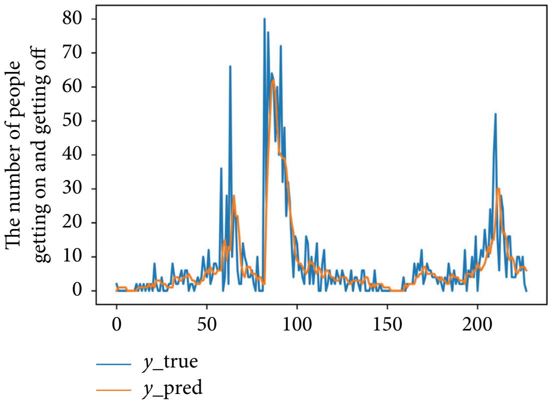

Figure 23: Regression integration prediction results when the time step is 2 h.

the prediction results of the regression integrated model, the integrated model has the lowest MAE and RMSE values at each step, indicating the result of using the regression integrated model to predict the most accurate. When the step size of the regression ensemble model is 12, the MAE value is 4.57 and the RMSE is 6.50, which are both the lowest values at each step size, indicating that the regression ensemble model has good prediction accuracy when the step size is 12.

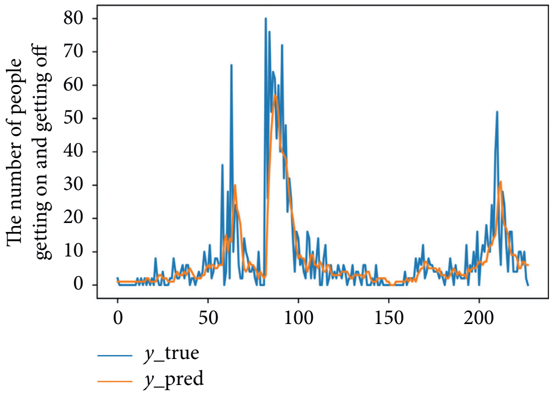

Figure 24: Regression integration prediction results when the time step is 3 h.

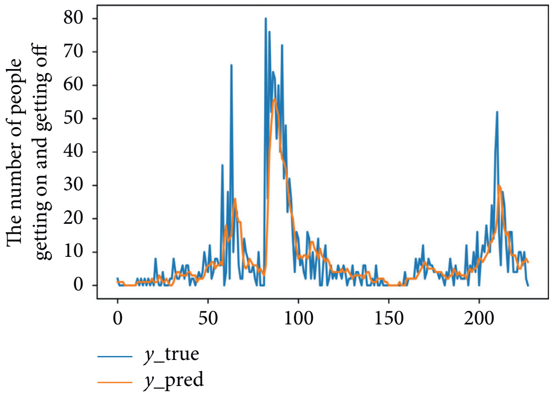

Figure 25: Regression integration prediction results when the time step is 6 h.

In the prediction of passenger flow by swiping cards, the prediction results of LR in the four submodels at each step have lower MAE and RMSE values, and the prediction results are more accurate. When the step is 4, each submodel has relatively low MAE and the RMSE value is more accurate. In comparison with the prediction results of the regression ensemble model, the ensemble model has the lowest MAE and RMSE values at each step, indicating that the prediction results of the regression ensemble model are the most accurate. When the step size of the regression ensemble model is 8, the MAE value is 16.09, and the RMSE is 16.09, both of which are the lowest values at each step size. It shows that the regression ensemble model has good prediction accuracy when the step size is 8.

In the prediction of landing volume, the prediction results of GRU under each step size in the four submodels have lower MAE and RMSE values, and the prediction results are more accurate. When the step size is 12, each submodel has relatively low MAE and RMSE. The results are more accurate. In comparison with the prediction results of the regression integrated model, the integrated model has the lowest MAE and RMSE values at each step, indicating that the prediction results using the regression integrated model are the most accurate. When the step size of the regression ensemble model is 8, the MAE value is 4.68 and the RMSE is 7.70, which are the lowest values at each step size. It shows that the regression ensemble model has good prediction accuracy when the step size is 8.

6. Conclusions

The core of urban bus network operation management is to effectively allocate and use system resources according to changes in the bus network passenger flow, adjust operation strategies in time, and ensure that the bus network safely completes transportation service tasks. Short-term passenger flow prediction and analysis is the basis of operation management. It can provide a basis for emergency management and response and is also an important decision-making index for public transportation service level and system operation status evaluation. Short-term passenger flow prediction is an important decision data for urban public transportation operation and management, and its prediction accuracy will directly affect urban public transportation decision-making, adjusting the scientificity and accuracy of the operation plan.

This paper analyzes the operational monitoring data of 428, a typical line in the Huitian area, from the perspective of the urban public transport network in the Huitian area, including traffic capacity, as well as the boarding and landing volume and cross-sectional passenger flow of each station. At the same time, based on objective bus operation data, the lr, KNN, Xgboost, and GRU four-seed models and the regression integration model based on the four-seed model were used to predict three different passenger flow indicators. From the prediction results, it can be seen that the regression integration is compared with the other four submodels and the model has a higher degree of fit. For passenger flow prediction, the result of this integrated model has a high degree of credibility.

The reliability of the prediction results reflects the availability and effectiveness of the prediction methods and models to a certain extent and also ensures the availability of the final short-term passenger flow prediction results. According to the reliable prediction results, once the passenger flow prediction value is greater than the preset threshold, decision-makers can activate emergency management plans. Secondly, operational planning can be dynamically adjusted based on passenger flow fluctuations. Managers can effectively control short-term passenger flow changes, adjust network operation strategies in a timely manner, rationalize the use of public transportation resources, and reduce operating costs. At the same time, the result of the short-term passenger flow forecast is used as a positive feedback of the line network monitoring, which can assist the manager in obtaining more effective information from the daily bus line network monitoring, so as to improve the control and management of the bus line network.

Since the research in this paper focuses on the construction and verification of the basic model, there are still certain shortcomings and limitations. Based on these shortcomings and limitations, the following prospects and suggestions can be provided for future related work:

(1) The impact of traffic policies on individual travel characteristics is a long-term impact. At the same time, traffic data can accurately record the long-term travel activities of each individual; therefore, urban big data such as traffic card data is very suitable for analyzing the impact of changes in urban traffic policies, the influence of individual travel characteristics. In the later period, we can use the data over a long period of time to analyze the impact of urban traffic policy changes on individual travel characteristics from a longitudinal perspective.

(2) The addition of more source data and the improvement of richer individual attribute information: the addition of mobile phone data and other data including complete travel chain data can significantly improve the identification of passenger activity locations. This will improve the analysis of the generation mechanism of rail transit passenger flow and enrich individual attribute information more accurately. In addition, the data including the complete travel chain is also of great help to the research on the route selection of individual passengers in the rail transit network.

(3) Joint analysis of multicity data to improve the universality and robustness of the model: this paper takes Beijing as an example to check and verify the parameters of each model. From the results, it can be seen that the model framework has ideal prediction accuracy. However, the applicability of the model parameters to other cities and the robustness of the model's prediction accuracy to other cities cannot be estimated. Therefore, in order to improve the universality and robustness of the model and make it more suitable for engineering practice, later studies can use data from multiple cities to conduct spatial and horizontal joint analysis and verify the model parameters.

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 work was partly supported by the National Nature Science Foundation of China–Young Scientists Fund (Grant no. 71103014), the Beijing Municipal Philosophy Office (Grant no. 14JGC095), the Science and Technology Project of the Beijing Traffic Commission (Grant no. B17M00080), and the Science and Technology Project of Beijing Transportation Industry (Grant no. 201905-ZHJC2).

References

[1] Y. Wei and M.-C. Chen, "Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks," *Transportation Research Part C: Emerging Technologies*, vol. 21, no. 1, pp. 148–162, 2012.

[2] B. M. Williams, P. K. Durvasula, and D. E. Brown, "Urban freeway traffic flow prediction: application of seasonal autoregressive integrated moving average and exponential smoothing models," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1644, no. 1, pp. 132–141, 1998.

[3] S. Lee and D. B. Fambro, "Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1678, no. 1, pp. 179–188, 1999.

[4] C. Cai, E. Yao, and M. Wang, "Passenger flow prediction of inbound and outbound stations of urban rail transit based on product ARIMA model," *Journal of Beijing Jiaotong University*, vol. 38, no. 2, pp. 135–140, 2014.

[5] M. Milenković, L. Švadlenka, and V. Melichar, "SARIMA modelling approach for railway passenger flow forecasting," *Transport*, vol. 33, no. 5, pp. 1–8, 2015.

[6] Y. Wang, B. Han, and Q. Zhang, "SARIMA model-based passenger flow prediction of Beijing subway station," *Transportation System Engineering and Information*, vol. 15, no. 6, pp. 205–211, 2015.

[7] D. Q. Wu, M. Dong, H. Y. Li, and F. Li, "Vehicle routing problem with time windows using multi-objective co-evolutionary approach," *International Journal of Simulation Modelling*, vol. 15, no. 4, pp. 742–753, 2016.

[8] J. M. Munoz-Guijosa, E. Riesco, and M. Olmedo, "Neural network and training strategy design for train drivers' vibration dose simulation," *International Journal of Simulation Modelling*, vol. 16, no. 1, pp. 72–83, 2017.

[9] J. Guo, W. Huang, and B. M. Williams, "Adaptive kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 50–64, 2014.

[10] H. Deng, X. Zhu, and Q. Zhang, "Short-term bus passenger flow prediction based on multi-core least-squares support vector machine," *Journal of Transportation Engineering and Information*, vol. 10, no. 2, pp. 84–88, 2012.

[11] D. Y. Zhang and H. N. Yang, "Passenger flow analysis in subway using a kind of neural network," *Applied Mechanics and Materials*, vol. 715, pp. 2284–2287, 2015.

[12] S. Wang, R. Zhou, and L. Zhao, "Forecasting Beijing transportation hub areas's pedestrian flow using modular neural network," *Discrete Dynamics in Nature and Society*, vol. 2015, 6, pp. 1–6, 2015.

[13] B. Tadic, M. Zivkovic, and G. Simunovic, "The influence of vacuum level on the friction force acting on the pneumatic cylinder sealing ring," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 4, pp. 970–976, 2019.

[14] W. F. Yu, G. S. Hou, and P. C. Xia, "Supply chain joint inventory management and cost optimization based on ant colony algorithm and fuzzy model," *Tehnicki Vjesnik-Technical Gazette*, vol. 26, no. 6, pp. 1729–1737, 2019.

[15] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[16] Y. Bai, Z. Sun, B. Zeng, J. Deng, and C. Li, "A multi-pattern deep fusion model for short-term bus passenger flow forecasting," *Applied Soft Computing*, vol. 58, pp. 669–680, 2017.

[17] G. D. Thomas, "Machine learning research: four current directions," *Ai Magazine*, vol. 18, no. 4, pp. 97–136, 1997.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[19] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[22] A. Ledezma, R. Aler, A. Sanchis, and D. Borrajo, "GA-stacking: evolutionary stacked generalization," *Intelligent Data Analysis*, vol. 14, no. 1, pp. 89–119, 2010.

[23] Y. Son and G.-G. Jin, "A nonlinear PD controller design and its application to MOV actuators," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 5–12, 2019.

[24] G. I. Y. Mustafa, H. Wang, and Y. Tian, "Model-free adaptive fuzzy logic control for a half-car active suspension system," *Studies in Informatics and Control*, vol. 28, no. 1, pp. 13–24, 2019.