Railway Traffic Volume Prediction Method Based on Hadoop Big Data Platform

Pei Su(✉)

Wuhan Railway Vocational College of Technology, Wuhan 430000, China
udsfg68@163.com

Abstract. In order to improve the accuracy and efficiency of railway traffic volume prediction, a railway traffic volume prediction method based on Hadoop big data platform is proposed. Firstly, the traffic big data preprocessing mainly includes three parts: redundant data processing, numerical abnormal data processing and missing data processing. Then the spatial cross-correlation characteristics of traffic flow are calculated. Finally, a combined prediction model based on multi features and multifractals is established to realize the railway traffic volume prediction based on Hadoop big data platform. The experimental results show that the prediction method in this study has high prediction accuracy, reduces the prediction time, and meets the needs of method design.

Keywords: Hadoop big data platform · Railway transportation · Transportation volume forecast · Redundant data · Threshold method · Relevance

1 Introduction

Railway transportation is a mode of transportation that uses railway trains to transport passengers and goods. It plays an important role in the process of social material production. It is characterized by large transportation volume, high speed, low cost and generally not limited by climate conditions. It plays an important role in China's economic development and residents' lives. Railway transportation often has problems such as congestion and equipment overload operation. Accurate prediction of railway passenger volume can not only quickly arrange train dispatching and prevent congestion and equipment overload operation, it can also reduce transportation in idle time and achieve the goal of energy conservation and emission reduction.

In order to solve the problems of congestion and equipment overload operation, many scholars have carried out research on the prediction method of railway traffic volume. Among them, reference [1] proposed a medium and long-term high-speed railway network passenger flow OD and channel traffic volume prediction method, which is based on Logit the model builds a passenger flow distribution model, and uses an iterative weighting method to solve the problem to realize the forecast of transportation volume. With the increasing amount of data involved in intelligent transportation, traditional methods cannot achieve better application effects.

S.-H. Wang and Y.-D. Zhang (Eds.): ICMTEL 2022, LNICST 446, pp. 470–485, 2022.
https://doi.org/10.1007/978-3-031-18123-8_36

The emergence of Hadoop can well analyze and process these data. Hadoop is a distributed architecture, which is studied and developed by the Apache foundation. Users do not need to thoroughly understand the implementation process at the very bottom of the system, so they can write corresponding applications in common programming languages. Use clusters for fast computing and storage. An important part of Hadoop is the system file distributed Hadoop (HDFS). One of the advantages of HDFS is its high fault tolerance and very low hardware requirements. It provides high data rate for application data and is suitable for applications with large data sets. HDFS has wide requirements for POSIX. In the file system, the data reading operation is carried out through streaming. In urban traffic, a large number of traffic information data is generated every day. The emergence of Hadoop HDFS can make good use of these traffic information to reasonably induce urban traffic. Therefore, this paper proposes a railway traffic volume prediction method based on the Hadoop big data platform to achieve high prediction accuracy of railway traffic volume, so as to alleviate the pressure of urban traffic congestion, provide convenience for people's daily life and work and travel.

2 Traffic Big Data Preprocessing

The traditional traffic data preprocessing method is to check whether there is redundancy and missing in the data by manual comparison or using conventional data processing tools, and then find out whether there is numerical abnormal data according to certain criteria. If problems are found, delete, modify and fill them manually according to relevant standards, so as to complete the preprocessing of traffic data. This method is completely feasible when dealing with conventional data, but it may not be fully applicable when dealing with traffic big data.

According to the definition of traffic big data, its scale is large and the data storage files may be scattered. Using traditional data restoration methods to detect and repair quality problems of traffic big data has a large workload and low efficiency, which can not meet the timeliness requirements of traffic management. At the same time, human operation errors may occur, resulting in problem data or threatening the security of data. In addition, due to the data protection policy of HDFS distributed file system, the data cannot be modified directly. Therefore, when processing traffic big data, we should rely on big data technology and traditional data preprocessing methods to detect and repair data.

According to the actual situation of the traffic big data file used in the research, this paper uses Hadoop big data technology to preprocess the traffic big data on the principle of maintaining the original characteristics of the data as much as possible. The processing sequence is carried out according to three steps: data redundancy processing, numerical abnormal data processing and missing data processing.

2.1 Redundant Data Processing

In traffic data, the form of redundant data is that each data item in one data element is exactly the same as the data items in other data elements. The redundant data detection

method for conventional traffic data with a small amount of data is to compare each row of data elements with the data elements of all other rows, and the time complexity is O(n2). When the amount of data is small, the processing method can work normally. However, when the amount of data is very large, this processing method will have serious problems in terms of memory overhead and time overhead. Therefore, the redundant processing of big data still needs to rely on big data technology [2]. Hadoop has its own character statistics function, which can be used to detect and repair redundant data.

1) Use data collection equipment number, collection date and time serial number as keywords for character statistics.
2) The data element with the count of 1 time is normally output once, and the redundant data element with the count of more than 1 time is only output for the first time, and a new data file is generated [3].
3) Replace the old data file with the newly generated data file to achieve the purpose of eliminating redundant data [4].

The specific flow of redundant data processing of traffic big data used in the experiment is shown in Fig. 1.

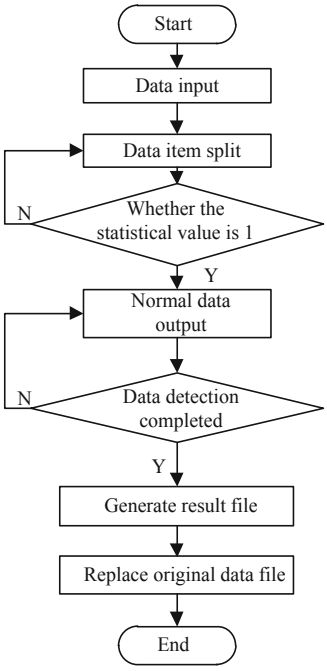

Fig. 1. Preprocessing method for redundant data of traffic big data

This method reduces the comparison times and improves the detection efficiency of redundant traffic data by classifying data items. At the same time, by generating

new data files to replace the original files, it not only conforms to the data protection strategy of HDFS system, but also avoids the problems of low efficiency and human operation errors that may occur in traditional redundant data processing methods. The corresponding pseudocode for redundant data processing is expressed as:

```
df.duplicated() #
df.drop_duplicates(inplace=True) #
```

2.2 Processing of Numerical Abnormal Data

Traffic data is an objective reflection of road traffic operation. In reality, the traffic flow of a specific section cannot exceed the saturation flow rate of its section. Similarly, the average travel speed of the traffic flow will not be much greater than the road speed limit. The data processing method of threshold method is relatively simple, that is, set a reasonable value interval [5] for the corresponding data items, and determine the data items exceeding the reasonable value interval as numerical abnormal data items. However, the threshold method can only detect the data items with significant abnormal values, and the recognition degree of abnormal data contained in the normal value range is not high. However, when the amount of data is large enough, using the threshold method to process data has three advantages. Firstly, it can eliminate the numerical abnormal data items with relatively large interference to the follow-up research, and meet the requirements of data reliability as a whole; Second, the original data characteristics can be retained as much as possible [6] to reduce the generation of artificial noise data; The third is that the algorithm has low complexity, improves the efficiency of data processing, and meets the requirements of practical application timeliness. Therefore, this paper uses the threshold method to determine the numerical abnormal traffic data.

The threshold method discriminates the abnormal vehicle speed data and defines the value range of the average vehicle speed v:

$$0 \leq v \leq c \times v_l \tag{1}$$

In formula (1), v_l represents the legal speed limit of the road and c is the average speed correction factor.

The threshold method discriminates the abnormal traffic flow data and defines the value range of traffic flow q:

$$0 \leq q \leq \varepsilon \times C \times t/60_v \tag{2}$$

Because the HDFS file system has a data protection strategy and cannot directly modify the value of abnormal data, a new method must be adopted for the processing of traffic big data. Firstly, threshold method and traffic flow mechanism method are used to test whether the value of traffic flow parameters is within a reasonable range. For the data elements determined as abnormal values, record the detector number, acquisition date and time serial number as keywords, generate abnormal data identification files, locate the abnormal traffic data, and deal with them together with the missing data in the

next work. This method not only conforms to the data protection strategy of the system, but also improves the efficiency of traffic big data preprocessing. The corresponding pseudocode for numerical abnormal data processing is expressed as:

```
import pandas as pd
import numpy as np
test_dict = {'id':[1,2,3,4,5,6],'name':['Alice','Bob','Cindy','Eric','Helen','Grace
'],'math':[90,'\\N',99,78,97,93],'english':[89,94,80,94,94,90]}

df = pd.DataFrame(test_dict)
df
df.loc[df['math']=='\\N']
df.loc[df['math']=='\\N','math'] = df.drop(1).math.mean()
df
```

2.3 Missing Data Processing

Judging whether the data is missing is the premise of missing data processing. Since the data used in the experiment is the data collected by the microwave detector at a fixed time interval of 2 min, it can be seen that under normal working conditions, the standard data collection volume of each detector per day should be 720 pieces. Therefore, the missing data detection method for traffic big data is to detect whether there is a null value in each row of data elements with the equipment number of microwave detector and data acquisition date as keywords after data redundancy processing. If none is empty, the statistical value is added by 1. The statistical value is based on 720. If it is lower than this value, it indicates that the node is missing data. There have been many studies on how to fix the problem of missing traffic information data. The common processing methods are historical average method, moving average method and exponential smoothing method. The moving average method uses adjacent data items to fill in the missing data, and the historical average method uses the periodic law of traffic flow parameters to repair the missing data with the data of the previous day or week. The exponential smoothing method uses the trend of time series to repair missing data. In the relevant research on missing traffic data repair methods, experiments have proved that in the case of a small amount of missing data, the exponential smoothing method is not much different from the moving average method and the historical average method [7]. Considering comprehensively, this paper adopts the moving average method and historical average method to deal with the missing data of experimental traffic big data.

The formula of moving average method is as follows:

$$y(t) = \frac{[y(t+n) + y(t+n-1) + \ldots + y(t-n)]}{n} \tag{3}$$

The missing data processing method in Hadoop is also a way to generate new files after data repair to replace old files, meet the system data protection strategy, reduce workload and improve efficiency. The corresponding pseudocode for missing data processing is expressed as:

```
df.isnull().sum()
#axis=0 means delete this line, =1 means to delete this column
df.dropna(axis=0,inplace=True)
df.fillna(0, inplace=True) #
df.fillna(df.mean(),inplace=True) #
df.fillna(value={'edu_deg_cd': train_tag['edu_deg_cd'].mode()[0], #
                                'deg_cd':train_tag['deg_cd'].mode()[0],
                                    'atdd_type':
train_tag['atdd_type'].mode()[0]},inplace = True)
```

3 Calculation of Spatial Cross-Correlation Characteristics of Traffic Flow

Traffic flow data is a typical time series and spatial geographic data, which has strong correlation in time and spatial dimensions. Therefore, urban road traffic flow has strong temporal and spatial distribution characteristics, showing not only temporal variation characteristics, but also spatial variation characteristics. The spatial variation characteristics of traffic flow mainly include the transverse variation characteristics of the same section and the longitudinal variation characteristics of different sections. Lateral variation characteristics are also simply called spatial correlation, which refers to the variation characteristics of traffic flow in the same section and different lanes; Longitudinal variation characteristics, also known as spatial time lag characteristics, refer to the variation characteristics of traffic flow between upstream and downstream detection sections of the same lane or the same section. In this paper, spatial statistical analysis and correlation analysis will be used to study the spatial correlation and spatial time delay of traffic flow.

Spatial correlation [8] is an important property of spatial geographic data. Its concept is similar to the autocorrelation of time series. It describes the correlation characteristics of a spatial location and its adjacent spatial location in the value of research attributes. Spatial correlation analysis of traffic flow data refers to the spatial correlation characteristics of traffic flow attribute values (such as flow, density or speed). This paper also uses Pearson correlation coefficient to analyze the spatial correlation of traffic flow. If the adjacent spatial location is close in the value of the research attribute, the value of Pearson coefficient is large, indicating that the correlation between the two spatial locations is strong; Otherwise, the spatial location has weak or no relevance. Suppose $X\{x_i|i = 1, 2, \cdots n\}$ and $Y\{y_i|i = 1, 2, \cdots n\}$ represent two traffic flow sequences at different spatial locations respectively, $\bar{x}$ and $\bar{y}$ represent x and y mean values respectively, in which the spatial correlation number of d order delay is an extension of Pearson correlation coefficient [9], which is expressed as:

$$\rho_{xy} = \frac{\sum_{n}^{i=1} (x_i - \bar{x})(y_{i-d} - \bar{y})}{\sqrt{\sum_{n}^{i=1} (x_i - \bar{x})^2} \sqrt{\sum_{n}^{i=1} (y_{i-d} - \bar{y})^2}} \tag{4}$$

Based on the above process analysis, the spatial correlation analysis of traffic volume provides a basic basis for traffic flow prediction.

4 Realization of Railway Traffic Volume Prediction

The modeling process is shown in Fig. 2:

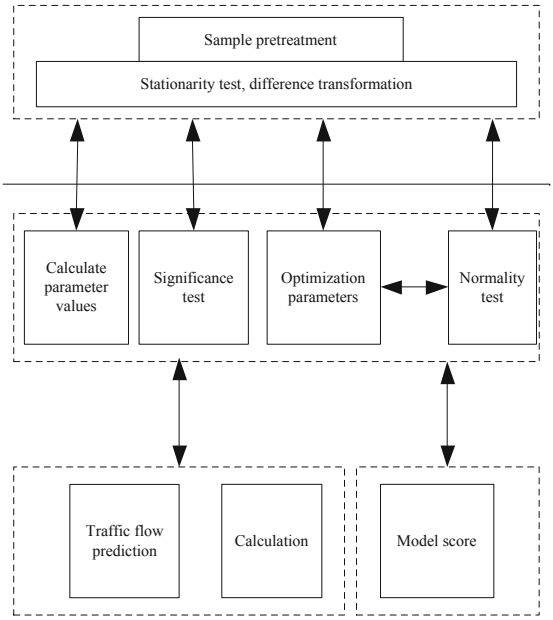

Fig. 2. Modeling process

The conditional variance of traffic flow time series exists, that is, the possibility of Heteroscedasticity in the actual traffic flow series. The GARCH model conducts modeling research on residual variance, solves the modeling problem of residual variance, and can simulate and predict the nonlinear volatility of traffic flow sequence. The GARCH model was proposed by Bolseslev in 1986. It is an extension of the autoregressive conditional heteroscedasticity model ARCH. It solves the problem of the high order of the ARCH model and the difficulty of estimating the parameters due to the long-term autocorrelation of the residual sequence conditional heteroscedasticity in practical applications. The problem is a very important model for processing time series data [10]. Express the calculation formula as:

$$\begin{cases} \sigma_t^2 = \alpha_0 + \sum\limits_{u_t=\sigma_t e_t}^{j=1} \beta_j \sigma_{t-j}^2 + \sum\limits_{v}^{i=1} \alpha_i \varepsilon_{t-i}^2 \end{cases} \tag{5}$$

In formula (5), α represents the autoregressive order, β_j represents the residual sequence, σ is the lagging sample variance coefficient, and e is the lagging conditional variance coefficient.

The periodicity, trend, linear correlation and random volatility of traffic flow and the complexity of traffic system jointly determine that accurate prediction can not be solved by a single model or method. A single prediction model has its advantages and disadvantages, scope of application and application conditions under specific circumstances, it is impossible to deeply mine the characteristics of traffic flow and make more accurate prediction. So far, no prediction method can show better performance than all other methods. Combining each advantageous single model according to a certain method, comprehensively maximizing the use of the useful information of a single model and comprehensively understanding the predicted object will help to improve the prediction effect and improve the prediction accuracy. Moreover, to predict traffic flow more accurately, it not only needs methodological innovation, such as cloud computing, big data, deep learning, the internet of things and combined forecasting [11]; It also needs to have a deeper understanding of the internal characteristics of traffic flow, such as spatio-temporal analysis [12], multifractal analysis, statistical analysis and so on. Therefore, this paper proposes a combined traffic flow forecasting method that integrates the time series forecasting models ARIMA and SARIMA and the time series fluctuation forecasting model GARCH, using spectrum analysis, time series and statistical fluctuation analysis methods to fully explore the temporal and spatial characteristics of traffic flow. According to the characteristics of self-similarity, long-term memory, and self-similarity of the traffic flow itself, it is proposed to decompose the traffic flow time series into periodic items, trend items and random fluctuation items, and combine and forecast the characteristics of different items thought of. This method is different from the traditional method, which simply and subjectively assumes that the time series of traffic flow meets a certain mathematical model or regular distribution, but uses the periodic and random fluctuation characteristics of traffic flow itself, uses different models to predict respectively, and finally recombines. Because after the non-stationary time series are transformed into stationary time series by difference, the model is constructed by returning the dependent variable only to the present value and lag value of its lag term and random error term, which is very suitable for the prediction of non-stationary single variable time series. Therefore, ARIMA and SARIMA are combined with GARCH.

A combined forecasting model based on multi feature and multifractal is established. The model considers the periodicity, trend, linear and nonlinear characteristics of traffic flow caused by many factors, and analyzes the combined forecasting method of modeling for each sub item. The schematic diagram of the combined model is shown in Fig. 3.

Firstly, the method decomposes the traffic flow sequence into four sub items: the periodic sub item is represented by $P(t)$, the trend sub item is represented by $T(t)$, the linear sub item is represented by $L(t)$ and the nonlinear sub item is represented by $N(t)$. Then a traffic flow sequence can be represented by the cumulative sum of $X(t)$ sub items, as shown in the following formula:

$$X(t) = P(t) + T(t) + L(t) + N(t) + \varepsilon_t \tag{6}$$

In formula (6), $P(t)$ and $T(t)$ are the determined components of traffic flow, ε_t represents the error term, and $L(t) + N(t) + \varepsilon_t$ is the uncertain component of traffic flow [13], representing the random fluctuation of traffic flow.

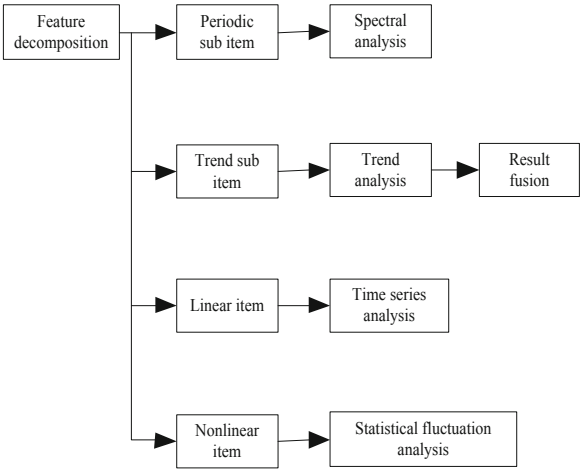

Fig. 3. Schematic diagram of combined model

Due to many random factors, a large amount of information with many influencing factors and complex periodic rules is divided into several modules, and a multi module weighted neural network prediction model based on the characteristics of time series is established. Different from the traditional RBF neural network, RBF neural network mixes the data information with different time series characteristics as an independent processing unit, which is input respectively for multi module comprehensive analysis and training learning. The feature layer output of each module is weighted to obtain the final prediction result. It avoids the nonlinear optimization problems such as blindness and local optimization in the design of single kernel function, improves the prediction accuracy and solution performance.

The principle of model improvement is as follows:

(1) Given historical data and preprocessed according to the predicted demand;
(2) Make statistical analysis on the change trend of historical data and specific cycle law, and consider the impact of various influencing factors on passenger flow;
(3) The input layer receives data samples with certain time series characteristics and establishes the corresponding prediction model single kernel function;
(4) Nonlinear analysis is carried out on the relevant feature data, and the weights between the input layer and the feature layer are partially connected, rather than all connected, that is, each module is trained and learned separately;
(5) The integration layer integrates the data and information with certain temporal characteristics. The number of integration layers is the arithmetic average of the number of neurons in the feature layer and the output layer, which is completely connected to the feature layer and the output layer to realize the exchange of information;
(6) The neurons in the integration layer are completely connected with the output layer to produce the final output of the network. The RBF network takes the predicted value of the modular model as the input value, and the actual output value is the training value of the RBF network until the predetermined error is reached. Finally, the multi modular network structure is determined.

The detailed prediction process of the model is as follows:

Step 1: Define error function:

$$E = \frac{1}{2}\sum_{k}\left(y_k - y_k'\right)^2 \tag{7}$$

In formula (7), y_k represents the expected output of the k sample point; y_k' represents the actual output of the k sample point.

Step 2: Given preset error.

Step 3: Select the radiation basis kernel function. On this basis, Gaussian function is selected as the kernel function:

$$\phi_i(x) = \exp\left[-(x - c_i)^2/2\sigma_i^2\right], i = 1, 2, \ldots, h \tag{8}$$

In formula (8), x represents the input sample, c_i represents the center of the i unit in the feature layer, and σ_i^2 represents the width of the $_i$ unit in the feature layer.

Step 4: Model training. The improved model is trained to obtain the weight from the integration layer to the output layer. Specific algorithms:

$$y = \sum_{m}^{l=1} w_l\phi_1(x), y/w^T\varphi \tag{9}$$

In formula (9), y represents the expected output of sampling, φ represents the number of neurons in the composite layer, and w^T represents the weight of the T neuron output layer in the composite layer.

Through the training of the model, the trained model is used for prediction, and the prediction results are obtained. The model comprehensively considers the long-term invariance and short-term time-varying characteristics of traffic flow, and can show and mine the evolution law and fluctuation characteristics of traffic flow to a great extent.

5 Experimental Comparison

In order to verify the effectiveness of the proposed railway traffic volume prediction method, experiments are carried out. The data used in this experiment is China Statistical Yearbook network, which mainly includes railway passenger volume, total population, number of domestic tourists and other data. The experiment is carried out based on the support of these data. Due to the large amount of data, in order to improve the experimental speed and accuracy, the experimental environment is established, as shown in Table 1.

The experimental process is shown in Fig. 4.

Table. 1 Experimental hardware configuration

Database server	Client computer	Network environment	Isolation device
CPU: Pentium core4 I7	CPU: Pentium core2 I7	100M/10M network card	Reverse physical isolation device
Memory: DDR3–1667	Memory: DDR3–1667		
Hard disk: 2TG	Hard disk: 1TG		

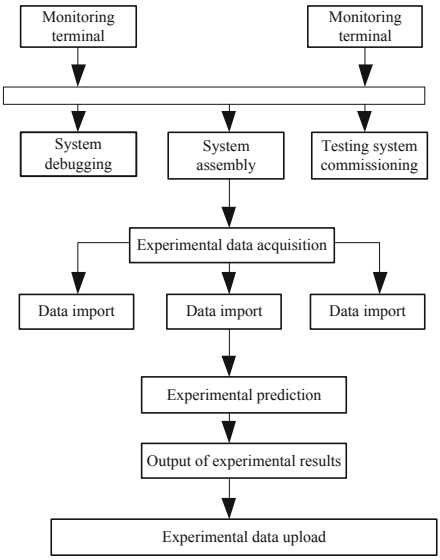

Fig. 4. Experimental process

Based on the above unified experimental environment, the railway traffic volume prediction method based on Hadoop big data platform is taken as the research method, and the medium and long-term high-speed railway network passenger OD and channel traffic volume prediction method proposed in reference [1] is taken as the traditional method. According to the above process experiment, the research method is compared with the traditional method. The specific experimental contents are as follows.

5.1 Comparison of Vehicle Speed Prediction Results

Draw a broken line diagram between the predicted speed and the actual data, and the comparison of speed prediction results is shown in Fig. 5.

Based on Fig. 5, it can be seen that the road speed at the location of the detector shows an upward, downward and upward trend during the morning peak hours, indicating that there is traffic congestion, and then it gradually returns to smooth. The predicted speed is generally consistent with the actual situation, and the changes of traffic operation during morning peak hours can be predicted.

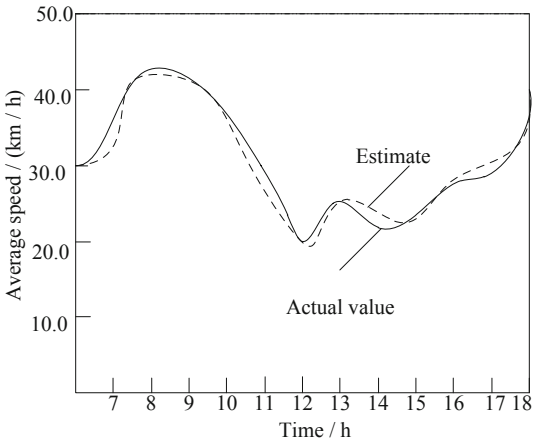

Fig. 5. Comparison of vehicle speed prediction results

5.2 Comparison of Prediction Deviation of Maximum Continuous Traffic Operation

The maximum continuous prediction deviation between the prediction results and the actual traffic operation conditions is shown in Table 2.

Table 2. Comparison of prediction deviation of maximum continuous traffic operation

Time	Predicted speed (km/h)	Actual speed (km/h)
9:04	47.70	57.4
9:06	60.55	53.1
9:08	49.68	66.2
9:10	54.69	60
9:12	49.51	64.5
…	…	…
9:30	52.91	63.1
9:32	49.75	63.9
9:34	55.91	64.3
9:36	55.57	57.5
9:38	53.04	64.4
9:40	50.93	59.1
9:42	45.30	59.6

Based on Table 2, it can be seen that there are 12 continuous time points of prediction deviation between the prediction results of traffic operation conditions and the actual situation, and the fluctuation trend of actual speed and predicted speed shows a sawtooth deviation.

The comparison results of the maximum continuous traffic operation prediction deviation between the studied method and the traditional method are shown in Fig. 6

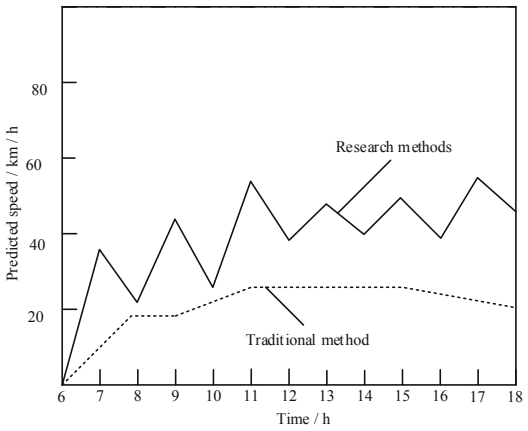

Fig. 6. Comparison results of prediction deviation of maximum continuous traffic operation

According to Fig. 6, the prediction speed of the research method is consistent with the actual speed, while the prediction speed of the traditional method is quite different from the actual speed. Therefore, the prediction deviation of the maximum continuous traffic operation condition of the research method is small.

5.3 Comparison of Prediction Errors of Traffic Volume

The comparison results of the prediction errors between the studied method and the traditional method are shown in Fig. 7.

Based on Fig. 7, it can be seen that the prediction error of the research method is relatively small, which can be less than 5, and has high prediction accuracy, while the prediction error of the traditional method is high, which has no good application effect than the proposed prediction method. Because the research method processes the data in advance and calculates the spatial cross-correlation characteristics of traffic flow, the prediction error is reduced.

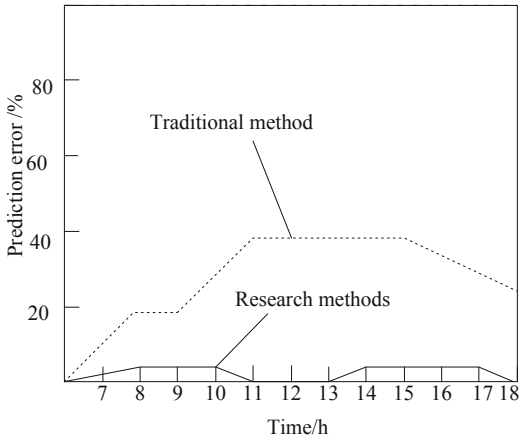

Fig. 7. Comparison of prediction errors of traffic volume

5.4 Comparison of Prediction Time

The prediction time of the research method and the traditional method is compared, and the comparison results are shown in Fig. 8.

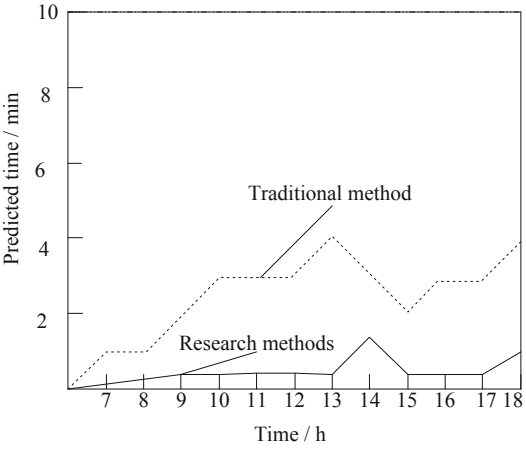

Fig. 8. Comparison of prediction time

Based on Fig. 8, the prediction time of the research method is less, and the prediction can ensure a shorter prediction time in each time period. The prediction stability of the traditional method is low, the time spent in some time is less, and the time spent in some time prediction is more, which is worse than the prediction effect of the proposed traffic volume prediction method. The reason for the poor effect of traditional methods may be that there is no preprocessing of redundant data, which is greatly disturbed by the data, thus reducing the application effect.

5.5 Comparison of Prediction Accuracy

The prediction accuracy of the research method and the traditional method are compared, and the comparison results are shown in Table 3.

Table 3. Comparison of prediction accuracy between the research method and the traditional method

Time	Predicted speed (%)	Actual speed (%)
9: 04	98.6	82.6
9: 06	97.8	81.9
9: 08	96.5	83.5
9: 10	96.8	82.3
9: 12	97.1	80.9
…	…	…
9: 30	98.2	81.4
9: 32	98.5	82.6
9: 34	97.6	82.4
9: 36	96.8	80.8
9: 38	96.4	81.6
9: 40	97.6	83.6
9: 42	98.5	84.7

According to Table 3, the prediction accuracy of the research method is higher, and the prediction can guarantee a higher prediction accuracy in each time period, while the prediction accuracy of the traditional method is lower. It can be seen that the research method can effectively improve the prediction accuracy.

6 Conclusion

This paper proposes a prediction method for railway traffic volume based on Hadoop big data platform. By preprocessing traffic big data and calculating the spatial cross-correlation characteristics of traffic flow, a combined prediction model based on multi-feature and multi-fractal is established to realize railway traffic volume prediction. It is verified by experiments that the research method not only improves the accuracy of forecasting, but also improves the efficiency of traffic volume forecasting, which has certain practical application significance.

References

1. Long, W.: Prediction method of high-speed rail passenger OD flow and traffic volume in medium and long-term high-speed railway network plan. J. Beijing Jiaotong Univ. **44**(4), 76–85 (2020)

2. Jiang, Y.: Simulation of multi-dimensional discrete data efficient clustering method under big data analysis. Comput. Simul. **36**(02), 205–208 (2019)
3. Iqbal, B., Iqbal, W., Khan, N., Mahmood, A., Erradi, A.: Canny edge detection and hough transform for high resolution video streams using hadoop and spark. Cluster Comput. **23**(1), 397–408 (2019). https://doi.org/10.1007/s10586-019-02929-x
4. Teng, L., Li, H., Yin, S., Sun, Y.: A modified advanced encryption standard for data security. Int. J. Network Secur. **22**(1), 112–117 (2020)
5. Chawla, S., Shahu, J.T., Gupta, R.K.: Design methodology for reinforced railway tracks based on threshold stress approach. Geosynth. Int. **26**(2), 111–120 (2019)
6. Sun, G., He, S., Fu, H., Xie, J., Zheng, L.: Study on shaking table test method for seismic responses of bridge-tunnel lapped structure in weak surrounding rocks. Tiedao Xuebao/J. China Railway Soc. **41**(1), 117–125 (2019)
7. Zhang, J.: Research on adaptive recommendation algorithm for big data mining based on hadoop platform. Int. J. Internet Protoc. Technol. **12**(4), 213–220 (2019)
8. Li, R., Huang, Y., Wang, J.: Long-term traffic volume prediction based on type-2 fuzzy sets with confidence interval method. Int. J. Fuzzy Syst. **21**(7), 2120–2131 (2019)
9. Gao, K., Han, F.R., Wen, M.F., Du, R.H., Li, S., Zhou, F.: Coordinated control method of intersection traffic light in one-way road based on v2x. J. Central South Univ. **26**(9), 2516–2527 (2019)
10. Yuan, W., Wang, J.: High mobility sparse channel estimation method based-on DCS-KF. Tiedao Xuebao/J. China Railway Soc. **41**(1), 74–79 (2019)
11. Liu, S.B., W, & Liu, G.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
12. Liu, S., Liu, G., Zhou, H.: A robust parallel object tracking method for illumination variations. Mobile Networks Appl. **24**(1), 5–17 (2018). https://doi.org/10.1007/s11036-018-1134-8
13. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools Appl. **76**(4), 5787–5802 (2014). https://doi.org/10.1007/s11042-014-2408-1