

Research on Cargo Forecasting Based on ARIMA Time Series Forecasting and LSTM Models

Yihan Li*, Zhetong Li

College of Petroleum Engineering, Liaoning Petrochemical University, Fushun, China, 113001

*Corresponding author: 15842235885@163.com

Abstract. With the prosperity of the domestic e-commerce market, the volume of express delivery business climbs dramatically, which poses an unprecedented challenge to the cargo volume prediction of the warehouse in the transit centre. This paper introduces traditional logistics sorting centre cargo volume prediction models, such as grey prediction model, Bayesian deep network generalized linear model, ARMA model, etc., as well as current common prediction methods including qualitative prediction, linear regression prediction, time series prediction and neural network prediction. Then, the improved ARIMA and LSTM models are proposed to predict the cargo volume of logistics sorting centres based on the improved ARIMA and LSTM models, and the construction and improvement methods of these two models are described in detail, including the SARIMAX model and the improved LSTM model based on the attention mechanism. Finally, this paper describes the construction method of the two combined models, and selects the most suitable prediction model through comparative analysis, and predicts the daily and hourly cargo volume in the next 30 days and carries out experiments and analyses.

Keywords: Logistics Cargo Volume, Combined Prediction, Improved ARIMA And LSTM.

1. Introduction

In recent years, with the prosperity of the domestic e-commerce market, the volume of express delivery business has climbed sharply, showing significant growth momentum. According to a research report[1] by CRRP Industry Institute, it is expected that by 2024, the volume of express delivery business will exceed the mark of 142.5 billion pieces. This huge growth in business volume poses an unprecedented challenge to the volume forecasting of warehouses in transit centres, and the forecasting capability needs to be more accurate and efficient. Moreover, the forecasting of cargo volume in transit centre warehouses is also affected by a variety of factors, including seasonal demand changes, various promotional activities and holidays. Especially during the double eleven, double twelve and other big promotions, the express business volume will have a significant growth[2], which requires the transfer centre to be able to carry out accurate cargo volume prediction in advance, so as to prepare sufficient warehouse capacity to cope with this growth trend.

The core of the logistics sorting centre volume prediction model is a model and method to accurately predict the future volume of the warehouse in the transshipment centre based on historical and current data. Qi Xingmin, Xu Hai et al[3] used grey prediction model to predict the material flow in the park; Yang Tiantian[4] based on Bayesian deep network generalised linear model for regional logistics volume forecasting research; Li Guo[5], Li Zhenzhen[6] and others chose ARMA model with multiple regression to predict the logistics volume demand; Zhong Qi[7] through the principal component analysis of the BP neural network for prediction, all of the above methods have shown good characteristics in some aspects.

Currently, the four main methods of cargo volume forecasting used in logistics transfer centres are qualitative forecasting, linear regression forecasting, time series forecasting and neural network forecasting. Qualitative forecasting is highly subjective and highly dependent on the experience and intuition of the forecaster, so the results may vary greatly due to the forecaster's subjective factors, and the prediction accuracy is limited; linear regression model assumes that there is a linear relationship between the independent variable and the dependent variable, but in practice, this relationship may be non-linear, resulting in inaccurate prediction results. Simple time series

forecasting may be difficult to accurately capture the dynamics of complex time series that contain trends and seasonal fluctuations. Moreover, different time series forecasting methods are applicable to different data characteristics, and selecting the appropriate model and adjusting the parameters may be a complex and time-consuming process; the training and inference process of neural networks usually requires a large amount of computational resources, which may lead to high forecasting costs. If there is insufficient data, it may lead to overfitting phenomenon[8-10].

The paper is based on predicting the cargo volume of logistics sorting centre based on improved ARIMA and LSTM models to further predict the cargo volume accurately so as to prepare sufficient warehouse capacity.

2. Forecasting logistics sorting centre cargo volume prediction based on improved ARIMA and LSTM models

The daily and hourly shipments of 57 sorting centres for the next 30 days were predicted by building a shipment forecasting model. Firstly, the time series forecasting model was chosen, where the ARIMA model as well as the LSTM model were selected. However, while analysing whether the models can solve the problem of this paper, it was found that both the models have certain deficiencies and hence they need to be improved separately. We hope to combine the advantages of the two improved models in a certain way to predict the daily and hourly shipments for the next 30 days. See Figure 1.

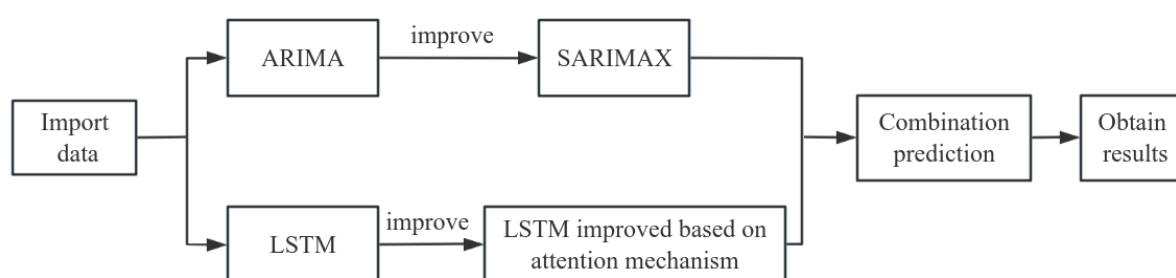


Figure 1. Flowchart of model building

2.1. Predictive Logistics Sorting Centre Volume Model

2.1.1 Model Construction

2.1.1.1 Improved ARIMA-based model - SARIMAX model

The data used in this paper to predict the cargo volume data of the day contains such special activities as double eleven, which will have a certain impact on the establishment of the ARIMA model, so based on the ARIMA model, this question proposes the SARIMAX model as an improvement.

SARIMAX model is an advanced model for time series analysis, which is called "Seasonal Autoregressive Integrated Sliding Average Model with Exogenous Variables". It is used to simulate and forecast time series data with seasonal fluctuations and external influences.

Specifically, the SARIMAX model is an extension of the ARIMA model, and an important feature of SARIMAX is its ability to include exogenous variables, which are factors that affect the time series but are not part of the time series itself. For example, price, promotional activities and economic conditions can all be used as exogenous variables when forecasting sales.

Therefore one of the forecasting models in this question is the SARIMAX model, which is developed for each of the sorting centres in Annex I with the date dimension. Based on the same treatment, the other dimensions are predicted according to the same model using the traversal concept. They are tested for smoothness and white noise, and if they fail, they are modified. The calculation process is shown in Figure 2.

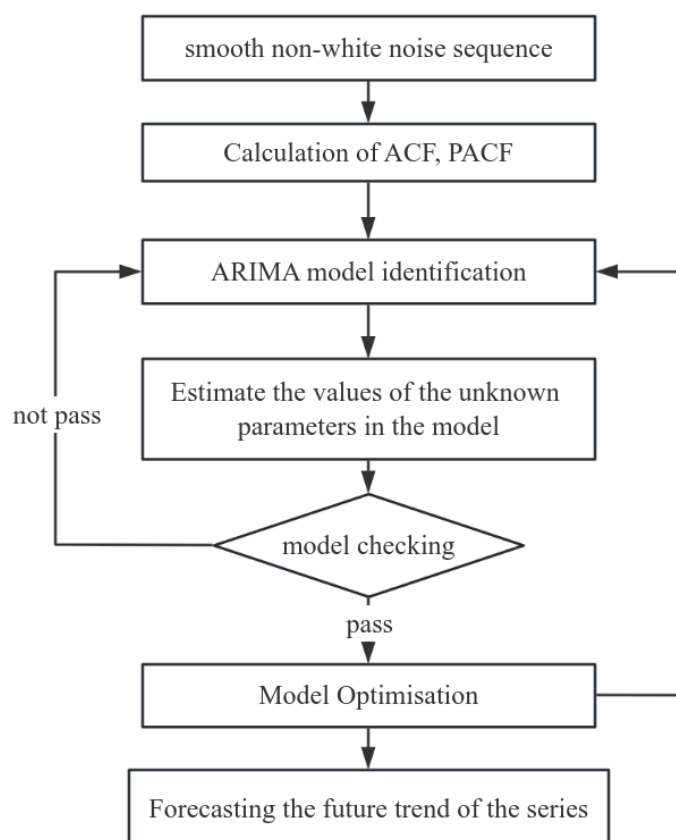


Figure 2 SARIMAX modelling flowchart

2.1.1.2 Improved LSTM based on Attention Mechanisms

Attention mechanism is a mechanism used to process information in deep learning. Its main function is to improve the performance and accuracy of the model by paying attention to different parts of the input data and assigning different weights. The core idea of the attention mechanism is to dynamically adjust the degree of attention to different parts of the input data according to their importance or relevance.

In terms of the data on which this paper is based (Data from website: <https://www.saikr.com/vse/mathorcup>), when observing the basic data, it can be seen that the volume of goods before and after 11 November is significantly higher, and similarly the volume of goods from 20:00 p.m. to 4:00 p.m. the next day is significantly higher, and in the subsequent prediction, we can give higher weight to the data of this period of time, and sequentially increase the sensitivity of the model to improve the performance of the model.

Therefore, the improved LSTM based on the attention mechanism solves the problem that traditional LSTM models may lose information when dealing with long sequences because they can only retain a limited amount of historical information. By introducing the attention mechanism, the model can selectively focus on different parts of the sequence more flexibly, thus reducing the possibility of information loss. The specific calculation flow is shown in Figure 3

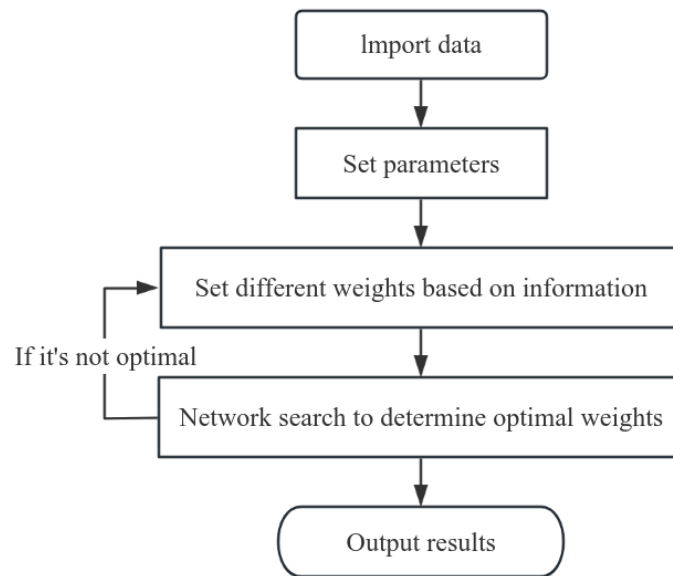


Figure 3 Improved LSTM model based on attention mechanism

The LSTM neural network structure as well as the hyperparameter settings are shown in Table 1.

Table 1 LSTM parameter settings

parametric	numerical value
numFeatures	1
numResponses	1
numHiddenUnits	100
MaxEpochs	200
Gradient Threshold	1
InitialLearnRate	0.01
LearnRateDropPeriod	100
LearnRateDropFactor	0.2

2.1.2 Establishment of combined model

The above SARIMAX model and the improved LSTM model based on the attention mechanism are analysed from the traditional statistical model and the modern deep learning respectively, next there are two model combination mechanisms, this question will be based on the fitting effect to select one for the subsequent days and hours of prediction. The combined model approach based on the inverse of the variance weights is used to calculate the hourly forecasted shipments; the segmented forecast based on different models is used to calculate the daily forecasted shipments.

Approach 1: Covariance-based inverse weighting

Combined modelling is an important approach to improve the accuracy of a model. A single model may have limitations in expressive power to solve complex problems well. Combined models, on the other hand, improve the overall performance by combining the strengths of multiple models in an integrated manner. Since the model can be refined to solve for each moment in time and give a better solution, the variance based inverse assignment of weights is chosen to solve for the amount of goods per day for the next 30 days. Here this question uses the data before 23rd November for training, predicts the data for each hour within 24th November to 30th November and compares it with the actual data to calculate the RMSE and reacts to the weights by the size of the RMSE.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2} \quad (1)$$

The calculated RMSE results of the two models for the hourly cargo volume prediction in the last week are shown in Table 2.

Table 2 Prediction results of different models

model	RMSE
SARIMAX	2298
Improved LSTM based on Attention Mechanism	2123

Therefore the weights for each model can be calculated based on the size of the two RMSEs as follows

$$\omega_1 = \frac{\frac{1}{2298}}{\frac{1}{2298} + \frac{1}{2123}} = 0.48 \quad (2)$$

$$\omega_2 = \frac{\frac{1}{2123}}{\frac{1}{2298} + \frac{1}{2123}} = 0.52 \quad (3)$$

Thus based on the above equation it can be concluded that the weights assigned to different models in the combined model are shown in Table 3.

Table 3 Weights assigned to the combined model

model	weights
SARIMAX	0.48
Improved LSTM based on Attention Mechanism	0.52

Approach 2: Segmented prediction based on different models

Since the SARIMAX model is more obvious in dealing with the occurrence of certain activities that lead to data anomalies, and at the same time based on the attention mechanism to improve the LSTM has a very obvious advantage in the prediction of long-term data, so in order to determine the optimal model, this question is to split the dataset, due to the impact of the activity is more calculated in terms of the unit of days, in the Therefore, this combination of models will be carried out to predict the volume of goods by day.

Because it is predicting the numerical results of the volume of goods per day for the next 30, the data set of the first three months is used for training in the previous training model, predicting the data of November, the data set of November is split into 4 weeks, predicting the results of each week separately to predict the results of the (RMSE) small model as the final prediction model, for four weeks of prediction results are shown in Table 4.

Table 4: Prediction results and model selection

cyclicity	Volume fluctuates.	SARIMAX	Based on the attention	Selection of models
			mechanism Improved LSTM	
11.01- 11.08	small	1097	981	Improved LSTM
11.09- 11.16	large	1864	2952	SARIMAX
11.17- 11.24	small	1674	1483	Improved LSTM
11.25- 11.31	large	1518	2476	SARIMAX

As can be seen from Table 4, when the volume of goods fluctuates greatly, SARIMAX has a better selection ability, and when the fluctuation is small, based on the attention mechanism to improve the

LSTM, at this time it is clear that in the week of 11.09-11.16 due to the double eleven activities brought about by the huge logistics effect; and in the 25th of November to the 31st of November due to the presence of Black Friday (the last Friday in November) and Cyber Monday (the last Monday of November) also brings some logistics, so for the time with a large impact of activities, SARIMAX prediction is used; and for the time with fewer activities, LSTM model is used based on the improved attention mechanism.

Similarly corresponding to the month of December, here we also need to divide the time period by week, each week manually define the size of the fluctuations in the volume of goods, the specific definition of the results shown in Table 5.

Table 5 Volume fluctuations by week in December

cyclicity	Events	Volume fluctuation size evaluation	Models to be used
12.01- 12.07	No	Small	Improved LSTM
12.08- 12.14	Double Twelve	Large	SARIMAX
12.15- 12.21	None	Small	Improved LSTM
12.22- 12.30	Christmas Day	Large	SARIMAX

2.1.3 Experiment and Analysis of Combined Prediction Models

In this paper, the model prediction will be established separately according to "days" and "hours", and the first introduction is to predict the cargo volume of 57 sorting centres in the next 30 days by day.

Firstly, the daily volume data for the next 30 days is predicted by day through mode 2 (based on different model segment prediction), and only two of the sorting centres are shown here. In this section SC40 is used as an example of data prediction.

The predicted data obtained when predicting the data from 12.01 to 12.07 is shown in Table 6.

Table 6 SC40 Partial Prediction Results

sorting centre	times	Projected results
SC40	2023/12/1	3861
SC40	2023/12/2	3735
SC40	2023/12/3	3717
SC40	2023/12/4	3996
SC40	2023/12/5	4351
SC40	2023/12/6	4625
SC40	2023/12/7	4899

In predicting the data from 12.08 to 12.14, the training is based on the daily shipment data of the previous three months, where $d=1$ is taken as the base first, and d is modified if the subsequent model fits poorly. The ACF and PACF plots obtained are shown in figure 4.

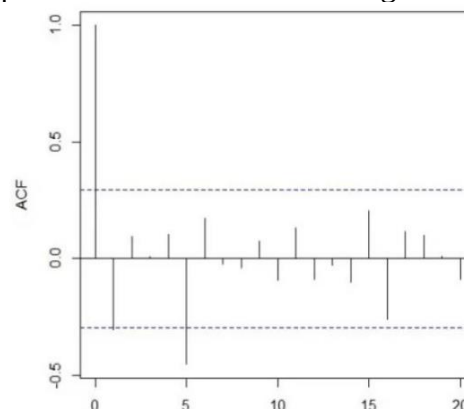


Figure 4 ACF and PACF diagrams

By looking at the autocorrelation plots, we find that the autocorrelation value of lag 1 order basically does not exceed the boundary, while the autocorrelation value of lag 5 order exceeds the boundary, but does not exceed the significant boundary on other lag orders. This suggests that when choosing model parameters, we can consider setting the maximum order of the autocorrelation and partial autocorrelation terms to 1 and 5, corresponding to the AR(1) and MA(5) models, respectively. Therefore, the ARIMA time series model obtained under the dimension of the whole time period of SC40 is ARIMA(1,1,5).

The final result can be obtained from the data results of 57 sorting centres for the next 30 days of cargo allocation, due to the relatively large amount of data, the detailed results are shown in Table 7 of the results to show only part of the content.

Table 7 Forecast Results for the Next 30 Days for 57 Sorting Centres (Partial)

sorting centre	times	times	sorting centre	times	times
SC1	2023/12/1	61065	SC10	2023/12/1	39418
SC1	2023/12/2	69184	SC10	2023/12/2	40277
SC1	2023/12/3	77959	SC10	2023/12/3	41776
SC1	2023/12/4	76330	SC10	2023/12/4	52046
SC1	2023/12/5	67705	SC10	2023/12/5	54175
SC1	2023/12/6	59761	SC10	2023/12/6	49968
SC1	2023/12/7	57496	SC10	2023/12/7	41975
SC1	2023/12/8	58398	SC10	2023/12/8	36211
SC1	2023/12/9	63115	SC10	2023/12/9	39937
SC1	2023/12/10	71809	SC10	2023/12/10	39826
SC1	2023/12/11	78539	SC10	2023/12/11	42445
SC1	2023/12/12	74251	SC10	2023/12/12	52287
SC1	2023/12/13	65162	SC10	2023/12/13	54006
SC1	2023/12/14	58525	SC10	2023/12/14	49545
SC1	2023/12/15	57530	SC10	2023/12/15	41185

As shown in Table 7, by comparing the data with similar periods, it is found that the volume of goods in the sorting centre increases significantly when it is in an active period and the predicted results are completely in line with this phenomenon, which demonstrates the accuracy of the approach 2 (based on the segmented prediction of different models).

After solving the daily volume data for the next 30 days, the hourly volume data for the next 30 days is solved by mode 1 (based on the inverse of the variance weight), and the partial prediction results are shown in figure 5, taking SC40 as an example.

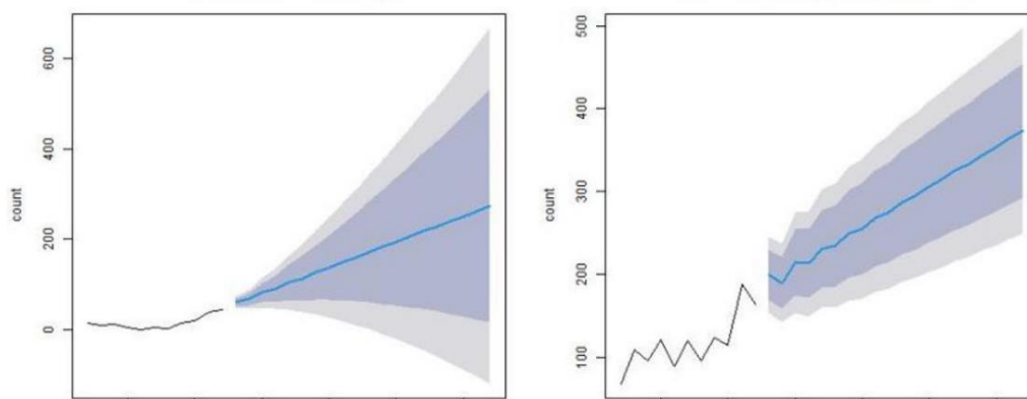


figure 5 Forecast results from 1 December to 7 December

The final forecast data SC1 two days before the hourly volume forecast results are shown in Table 8.

Table 8 Hourly cargo forecast results for the first two days of SC1

sorting centre	times	times	sorting centre	times	times	sorting centre	times
SC1	2023/12/1	0	3117	SC1	2023/12/2	0	2979
SC1	2023/12/1	1	3192	SC1	2023/12/2	1	3017
SC1	2023/12/1	2	2973	SC1	2023/12/2	2	2884
SC1	2023/12/1	3	2738	SC1	2023/12/2	3	2580
SC1	2023/12/1	4	2945	SC1	2023/12/2	4	2834
SC1	2023/12/1	5	2458	SC1	2023/12/2	5	2547
SC1	2023/12/1	6	697	SC1	2023/12/2	6	1295
SC1	2023/12/1	7	851	SC1	2023/12/2	7	1116
SC1	2023/12/1	8	1537	SC1	2023/12/2	8	1556
SC1	2023/12/1	9	1451	SC1	2023/12/2	9	1553
SC1	2023/12/1	10	1464	SC1	2023/12/2	10	1485
SC1	2023/12/1	11	1500	SC1	2023/12/2	11	1666
SC1	2023/12/1	12	2449	SC1	2023/12/2	12	2417
SC1	2023/12/1	13	2670	SC1	2023/12/2	13	2495
SC1	2023/12/1	14	954	SC1	2023/12/2	14	1339
SC1	2023/12/1	15	367	SC1	2023/12/2	15	594
SC1	2023/12/1	16	616	SC1	2023/12/2	16	687
SC1	2023/12/1	17	747	SC1	2023/12/2	17	777
SC1	2023/12/1	18	763	SC1	2023/12/2	18	743
SC1	2023/12/1	19	1268	SC1	2023/12/2	19	1036
SC1	2023/12/1	20	2187	SC1	2023/12/2	20	1795
SC1	2023/12/1	21	2625	SC1	2023/12/2	21	2352
SC1	2023/12/1	22	2535	SC1	2023/12/2	22	2300
SC1	2023/12/1	23	2612	SC1	2023/12/2	23	2318

It can be seen through Table 8: after the processing of mode 1 (based on the inverse of the variance of the weight assigned), most from the comparison of the predicted data with the actual data, has a good prediction performance, the error is relatively small, can meet the demand, can quickly predict the future data, easy to operate.

3. Conclusions

Based on the research and analysis of the express business volume growth and cargo volume forecasting model, this paper can draw the following conclusions:

Firstly, with the booming development of the domestic e-commerce market, the express business volume shows a trend of rapid growth. It is expected that the express business volume will reach a huge 142.5 billion pieces by 2024, which poses an unprecedented challenge to the cargo volume prediction of the warehouse in the transit centre. Secondly, for the cargo volume prediction of transshipment centre warehouses, this paper adopts the improved ARIMA and LSTM based models. By analysing historical and current data, we found that there are some limitations of a single model, so the ARIMA and LSTM models are improved to enhance the prediction accuracy and efficiency. Finally, through experiments and analyses, we verified the effectiveness of the two improved models in cargo volume forecasting. Both the variance-based inverse weighting and the combined model approach based on the segmented prediction of different models show good prediction performance and can quickly and accurately predict the future cargo volume situation.

In summary, the improved ARIMA and LSTM models proposed in this study have important applications in express business growth and cargo volume prediction, which provide a more accurate and efficient cargo volume prediction method for the transit centre, and help to cope with the challenges of the rapid growth of express business.

References

- [1] Zhongyan Puhua Industry Institute. Analysis Report on the Current Situation and Future Development Trends of Express Delivery Industry, 2024-2029 [R]. Beijing: Zhongyan Puhua Industry Research Institute, 2023.
- [2] Fu actually. Research on the response to the phenomenon of "double eleven" warehouse explosion in county logistics sorting centre [J]. Small and medium-sized enterprise management and science and technology, 2019(27): 15-16.
- [3] QI Xingmin, XU Hai, DUAN Chen, FAN Rui. Forecasting logistics freight volume in Xiangyang city based on grey correlation analysis and genetic neural network [J]. Logistics Technology. 2021 Sep 3; 40(7): 67.
- [4] Yang T-T. Research on regional logistics forecasting based on Bayesian deep network generalised linear model [D]. Tianjin University of Commerce, 2022.
- [5] JIA Chunmiao, FU Zhongning, MA Yaling, LI Jianguo, CHEN Hulin. Forecasting railway freight volume in Ganning District based on ARMA model and multiple regression [J]. Comprehensive Transport, 2022, 44(09): 147- 154.
- [6] Li Zhenzhen. Forecasting express business volume in Henan Province based on ARIMA simple seasonal model [J]. Value Engineering, 2019, 38(17): 271-273.
- [7] Zhong Qi. BP neural network grain yield prediction model based on principal component analysis [J]. Food Safety Guide, 2022(20): 172-174.
- [8] Xu Wensheng. Optimisation of logistics sorting centre business model based on Anylogic [D]. Anhui University of Technology, 2023.
- [9] KU Yong, YANG Zezhao. Random forest-based intermodal container freight volume forecasting [J]. Journal of Wuhan University of Technology, 2023, 45(01): 35-44.
- [10] Lin Mengman. Research on Optimisation Problems and Methods of Cross-border E-commerce Overseas Warehouse Mode Logistics Network [D]. Zhejiang: Zhejiang University of Technology, 2020.