ARIMA-based Freight Forecasting and Network Optimization in E-commerce Logistics

Tianyou Wu *, Jinkai Deng #, Chengxi Chu #, Jinghao Luo #

School of Transportation and Logistics, Southwest Jiaotong University, Chengdu China, 611756

* Corresponding Author Email: wutianyou@my.swjtu.edu.cn

#These authors contributed equally.

Abstract. Considering the proliferation of e-commerce platforms exemplified by Taobao and JD, the paradigm of online shopping has evolved into an integral facet of contemporary societal existence. To enhance network transport performance, it is essential to predict freight volume, assess existing capacity, and establish new sites to alleviate network pressure. This study focuses on a logistics transportation network, utilizing daily freight turnover data from January 1, 2021, to December 31, 2022. The ARIMA time series method is employed to forecast freight volumes at different sites within the e-commerce logistics network. Taking three logistics site route combinations, namely DC14→DC10, DC20→DC35, and DC25→DC62, as examples, two differencing operations are applied to meet the stationarity requirement of the ARIMA model. Autocorrelation and partial autocorrelation coefficients are used for preliminary model order determination. By minimizing the Akaike Information Criterion (AIC) value, an ARIMA (5) model is established. Single and multiple-step predictions are conducted, resulting in forecast curves for future freight volumes. Subsequently, an evaluation of the transportation network is performed, considering the importance of nodes and routes. The entropy weight method is applied to determine the weights of evaluation indicators. Importance indices for nodes and routes are calculated, leading to a ranking. For the assessment and selection of new site capabilities, Fisher's discriminant function is employed to classify different sites. The valuation level for site selection is determined, providing a scientific basis for the systematic choice of new sites.

Keywords: E-commerce logistics, ARIMA time series forecasting, Transportation network evaluation, Logistics node selection.

1. Introduction

As the "meridian" of the national economy, the modern logistics system is an important support for extending the industrial chain, upgrading the value chain and building the supply chain. Promotional activities and holidays cause e-commerce users to place a surge in the amount of orders, bringing enormous overload pressure on the logistics network, the need to predict the flow of goods in advance, arranging transportation and sorting plans to prevent system paralysis, but also the need to optimize the structure in order to reduce costs and improve the efficiency of logistics [1].Traditional prediction methods are based on ARIMA model and BP neural network model, and furthermore there are Same-stage and cross-stage cascade failure model [2-3], underload cascade failure model that integrally considers fresh food transportation efficiency and freshness loss, network node importance identification algorithm based on the entropy of neighboring information considering the influence of neighboring nodes' information on node load [4-7].

2. ARIMA time series forecasting of freight volume

2.1. Smoothing process through double differencing operations

This study employs the ARIMA time series method to forecast future freight volumes between different sites within the e-commerce logistics network. Analysis is conducted using daily freight turnover data from a specific logistics network spanning from January 1, 2021, to December 31, 2022.

Three logistics site route combinations, namely DC14→DC10, DC20→DC35, and DC25→DC62, are selected for showcasing the forecasting results [8].

The data reveals significant fluctuation in the freight turnover of the DC10→DC14 route during the period from January 1, 2021, to December 31, 2022. Given the requirement of the ARIMA model for stationarity, where the mean and variance of the sequence need to remain stable, double differencing operations are performed.

Initially, the first differencing operation is applied to eliminate the trend in the time series data is as follows.

$$y_t' = y_t - y_{t-1} \tag{1}$$

Subsequently, the second differencing operation is conducted to satisfy the stationarity requirement of the ARIMA model as follows.

$$y_t'' = y_t' - y_{t-1}' \tag{2}$$

After performing two differencing operations, sequence plots for the route were generated for comparison. The red and blue lines represent the original and processed data, respectively. It is evident from the plots that the two differencing operations successfully eliminated the trend and periodicity in the data, rendering the data sequence more stationary, as shown in Figure 1.

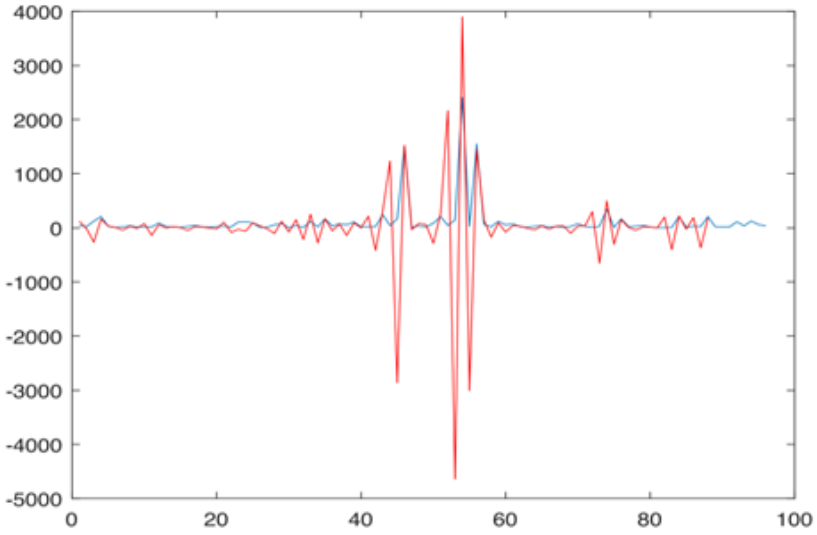

Figure 1. Comparison between the sequence plot of the doubly differenced series and the original data.

2.2. Determination Ordering

Subsequently, autocorrelation coefficients and partial autocorrelation coefficients were calculated to make preliminary determinations regarding the order of the ARIMA model.

The calculation method for autocorrelation coefficients is as follows.

$$\rho = \frac{Cov(y_t, y_{t-k})}{\sqrt{Var(y_t)Var(y_{t-k})}} \tag{3}$$

Where $Cov(y_t, y_{t-k})$ represents the covariance of time series data at moments t and t-k. $Var(y_t)$ And $Var(y_{t-k})$ respectively denote the variances of time series data at moment's t and t-k.

The calculation method for partial autocorrelation coefficient is as follows.

$$\Phi_{kk} = \frac{Cov(y_t, y_{t-k} - \sum_{i=1}^{k-1} \Phi_{k-1,i} \cdot y_{t-i})}{Var(y_t \mid y_{t-1}, y_{t-2} ..., y_{t-k+1})} \tag{4}$$

Φ_{kk} Is the partial autocorrelation coefficient with a lag order of k. $Cov(y_t, y_{t-i})$ represents the covariance of time series data at moment's t and t-i $Var(y_t \mid y_{t-1}, y_{t-2} ..., y_{t-k+1})$ Represents the variance of time series data at moment t.

The computation results of autocorrelation coefficients and partial autocorrelation coefficients are depicted in Figure 2(a) and Figure 2(b), respectively.

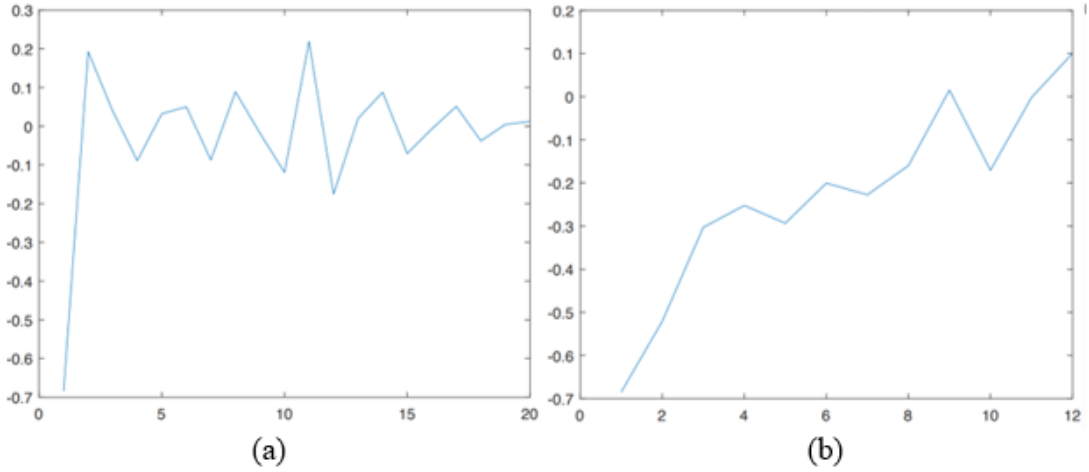

Figure 2. (a) autocorrelation coefficients (b) partial autocorrelation coefficients.

Based on the truncation of the partial autocorrelation function, the preliminary determination of the model order is 5. Parameter estimation using the least squares method is required, calculating the model's residual variance and AIC values within the range of 1 to 10 orders. The model order is then determined by minimizing the Akaike Information Criterion (AIC).

Rooted in information theory, AIC assesses model performance by considering both the goodness of fit and model complexity. Its calculation formula is as follows:

$$AIC = -2\ln(L) + 2k \tag{5}$$

Where L the maximum likelihood is function of the model, and k is the number of free parameters in the model.

In this study, the model with the minimum AIC value among different orders is chosen as the final forecasting model. The calculation formula is as follows:

$$p = \arg\min_{p} \{AIC(p)\} \tag{6}$$

Where p represents the order corresponding to the minimum AIC value, and p represents the order.

Throughout the computation process, the AIC values obtained are as follows: 175.7353, 164.6330, 159.3559, 143.6745, 141.7495, 143.5362, 143.8233, 145.4632, 148.3037, 150.5327. 141.7495(The order corresponding to the minimum AIC value), is selected, establishing the model order as 5. Therefore, based on the principles, the model order is confirmed as five, affirming that the developed forecasting model is an ARIMA (5) model.

Initially, single-step predictions are conducted for the subsequent six data points. Subsequently, the predicted values undergo two reverse differencing operations to obtain an approximated forecasting curve, as illustrated in Figure 3. Following this, multi-step predictions are carried out, utilizing the last 5 data points as input, and leveraging the fully trained model to predict the subsequent 6 values. Ultimately, the predicted freight volume is obtained, as depicted in Figure 4.

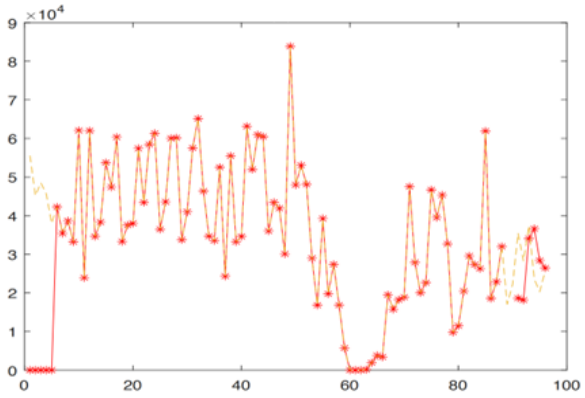

Figure 3. Approximated forecasting curve.

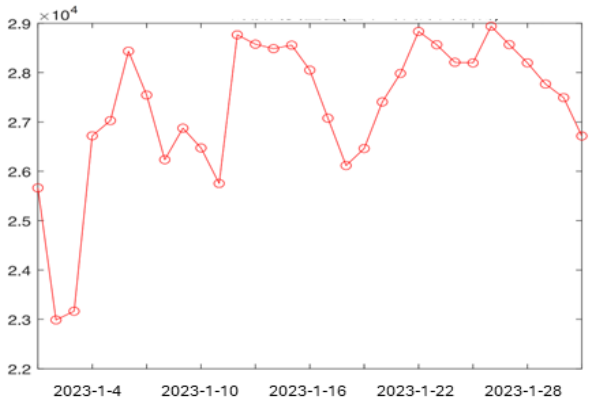

Figure 4. Predicted freight volume.

Similarly, the same method was applied to analyze and compute predictions for the DC20→DC35 and DC25→DC62 routes, obtaining partial results as shown in Table.1.

Table 1. Freight volume forecasting results.

Date	DC14→DC10	DC20→DC35	DC25→DC62
2023/1/1	25664.00	38.00	6320.00
2023/1/2	22987.50	48.50	9932.00
2023/1/3	23164.67	74.00	8185.67
2023/1/4	26720.00	64.75	9586.00
2023/1/5	27027.80	74.00	10552.20
……	……	……	……
2023/1/29	27769.17	59.55	8992.38
2023/1/30	27493.00	58.33	8813.97
2023/1/31	26715.94	56.87	8882.81

3. Transportation Network Evaluation

The transportation network comprises nodes and routes. Evaluating and optimizing it essentially involves assessing the importance of sites and routes [9]. Data filtering is needed to be performed and processed, determine evaluation indicators, and construct a comprehensive evaluation system, as shown in Table 2. Where In-degree refers to the number of edges pointing towards a particular node, and Out-degree refers to the number of edges originating from that node and pointing towards other nodes.

Table 2. Evaluation indicators for logistics nodes and transportation routes.

Serial number	Evaluation indicators for logistics nodes	Serial number	Evaluation indicators for transportation routes
1	In-degree and Out-degree (N_1)	1	Maximum transport Capacity (T_1)
2	Average load (N_2)	2	Average load (T_2)
3	Maximum storage capacity (N_3)		

Before conducting the importance evaluation, a preliminary statistical analysis was performed on the relevant data of historical logistics nodes and transportation routes, as shown in Tables 3 and 4.

Table 3. Statistical data for historical transportation nodes.

Logistics nodes Nodes	Sum degree	Average load	Storage capacity
DC10	39	0.03495	136780.4
DC27	15	0.66667	61070.8
DC18	6	0.3939	48764.0
……	……	……	……

Table 4. Statistical data for historical transportation routes.

Transportation route	Maximum transport capacity	Average load
DC11-DC10	9190	0.487
DC17-DC12	2032	0.121
DC20-DC14	30502	0.257
……	……	……

To make an objective assessment of the evaluation indicators, the specific algorithm of the entropy weight method is as follows: calculate the information entropy for each indicator, determine the information utility value, and normalize to obtain each indicator.

$$\mathrm{e}_j = \mathrm{k} \cdot \sum_{i=1}^{n} y_{ij} \cdot \ln y_{ij} \tag{7}$$

To ensure that e_j can lie within the interval [0, 1], the typical value for k is usually taken as $\frac{1}{\ln n}$.

The calculation of the information utility value (information entropy redundancy) is as follows: The greater the information utility, the larger the information quantity. The method is as follows:

$$d_j = 1 - e_j \tag{8}$$

When calculating the weight coefficients of the indicators, the information utility values are normalized. The weight for each indicator can be obtained through the following formula:

$$w_j = \frac{d_j}{\sum_{i=1}^{m} d_i} \tag{9}$$

Using the entropy weight method, the weights for the indicators of node importance and transportation route importance are calculated, as presented in Table.5 and Table.6.

Table 5. Weights of transportation node importance indicators.

Indicator	N1	N2	N3
W_{1i}	0.51	0.23	0.26

Table 6. Weights of transportation route importance indicators.

Indicator	T1	T2
W_{2i}	0.62	0.38

Subsequently, the quantities N_i and T_i are quantitatively processed, and the node and route importance indices S_1 and S_2 are calculated using the following formula, as presented in Table.7 and Table.8, respectively.

$$S_1 = \sum_{i=1}^{n} W_{1i} \cdot N_i \tag{10}$$

$$S_2 = \sum_{i=1}^{n} W_{2i} \cdot T_i \tag{11}$$

Table 7. Nodes importance.

Logistics nodes	Importance indices	rankings
DC10	35552.8	1
DC6	15886.0	2
DC19	12681.8	3
……	……	……

Table 8. Route importance.

Transportation routes	Importance indices	rankings
DC11-DC10	5698.0	2
DC17-DC12	1259.9	3
DC20-DC14	18911.3	1
……	……	……

4. Assessment of the capacity for new site establishment

Due to the absence of historical data during the establishment of new sites, this study relies on indicators such as freight volume, the number of routes, and the freight volume of connected upstream and downstream sites. It classifies all sites into different hierarchical stages, ranging from 1 to 9, with stages 1-2 considered as extremely important and the remaining levels following suit. When predicting and determining site levels, factors such as freight volume, the number of routes, and variance are primarily considered. Using SPSS software for data analysis, the weight of each indicator was ascertained, and the discriminant coefficient of the Fisher discriminant function was obtained. For the establishment of a new site, inputting the new predicted indicator data allows the corresponding score level to be obtained, facilitating site selection [10].

The calculation method for the sample mean values of the ten-stage site valuation levels in the d-dimensional feature space is as follows:

$$M_i = \frac{1}{n}\sum y_k \in Y_i y_k \quad , i = 1, 2, ...9 \tag{12}$$

The calculation method for the mean values of each class after mapping through the transformation ω into a one-dimensional feature space is as follows:

$$m_i = \frac{1}{n_i}\sum y_k \in Y_i y_k \quad , i = 1, 2, ...9 \tag{13}$$

After mapping, the definition of "within-class scatter" for each class is as follows:

$$S_i^2 = \sum y_k \in Y_i (y_k - m_i)^2 \quad , i = 1, 2, ...9 \tag{14}$$

Clearly, after mapping, a larger freight volume among the ten class means, and smaller within-class scatter for each class were desired. Therefore, the calculation method for the Fisher discriminant function is as follows:

$$J_F(\omega) = \frac{(m_1 - m_2)^2}{S_1^2 + S_2^2 + S_3^2} \tag{15}$$

The solution ω that maximizes JF is the optimal solution vector, which represents Fisher's linear discriminant.

$$\omega = S_\omega^{-1}(M_1 - M_2 - M_3) \tag{16}$$

Where S_ω is the total within-class scatter matrix?

The determination coefficients and constants for the ten discriminant functions are solved and the results are shown in Table.9.

Table 9. The determination coefficients and constants.

	1	2	3	4	5	6	7	8	9
Freight volume	3.55 E-11	-4.98 E-12	-1.33 E-11	-1.46 E-11	-2.49 E-11	-4.93 E-13	-7.31 E-12	-5.35 E-11	-1.51 E-12
Number of routes	-8.51 E-11	2.26 E-11	7.09 E-11	5.10 E-11	8.96 E-11	1.92 E-11	3.36 E-11	1.04 E-10	8.69 E-11
Freight difference	0.00	0.00	0.00	0.00	-0.03	-5.37 E-5	0.00	0.00	-0.04
Route difference	6.48 E-11	-9.39 E-11	-1.69 E-10	-1.69 E-10	-3.06 E-11	-8.58 E-11	-1.03 E-10	-1.54 E-10	-3.54 E-11
Variance fluctuation	0.019	0.028	0.016	0.016	0.02	0.01	0.02	0.02	0.06
Total volume	-2.66 E-11	4.10 E-11	3.59 E-11	4.49 E-11	7.52 E-11	1.01 E-11	2.20 E-11	1.03 E-10	3.45 E-11
Constant	-3.33	-4.33	-3.47	-3.03	-4.41	-2.53	-2.78	-4.75	-11.2

Finally, the classification formula is expressed as the following equation.

$$F_1 = 3.55E^{-11}x_1 - 8.51E^{-11}x_2 + 6.48E^{-11}x_4 + 0.019x_5 - 2.66E^{-11}x_6 - 3.33 \quad (17)$$

$$F_2 = -4.98E^{-12}x_1 + 2.26E^{-11}x_2 - 9.39E^{-11}x_4 + 0.028x_5 + 4.10E^{-11}x_6 - 14.33 \quad (18)$$

$$F_9 = -1.51E^{-12}x_1 + 8.69E^{-11}x_2 - 0.04x_3 - 3.54E^{-11}x_4 + 0.06x_5 + 3.45E^{-11}x_6 - 11.2 \quad (19)$$

Five transportation nodes were selected for Fisher discriminant analysis and presented the results in Table 10.

Table 10. The results predicted by the Fisher discriminant function.

Logistics nodes	Valuation level
DC34	8.00
DC67	6.00
DC18	2.00
DC47	7.00
DC42	1.00

For establishing a new site, one needs to input the new data for the six indicators into the ten discriminant functions separately, obtaining ten function values. If F_1 is the largest, the site should be classified into category 1. Similarly, if F_2 is the largest, the site should be classified into category 2. If F_9 is the largest, the site should be classified into category 9. Subsequently, site selection for new construction can be conducted based on the assigned category.

5. Conclusions

In this study, by applying the ARIMA time series analysis method, a specific e-commerce logistics network was used to forecast the shipment volume between different sites in the coming period. The process includes the preprocessing of data, modeling, parameter estimation, and final prediction. By comparing the prediction results with the actual data, the effectiveness and feasibility of the ARIMA model in the prediction of freight transportation volume in e-commerce logistics networks were

demonstrated. In addition, the study also involves the evaluation of the transportation network, determining the weights of the evaluation indexes through the entropy weighting method and calculating the importance indexes of the nodes and routes, which provides a scientific basis for the optimization of the logistics network and the establishment of the new sites. Through the Fisher discriminant analysis method, it further provides an effective tool for the selection of new sites.

Overall, this study not only provides a scientific method for freight volume prediction of e-commerce logistics network, but also puts forward valuable suggestions for the optimization and development of the network, which lays the foundation for future research and practical application.

References

[1] Ma Youhong. Countermeasures for the development of modern logistics [J]. Cooperative Economy and Technology, 2024, (03): 87-89.DOI: 10.13665/j.cnki.hzjjykj.2024.03.046.

[2] Shen Suhao. Research on cascading failure problem based on real logistics network [D]. Xi'an University of Electronic Science and Technology, 2021.DOI: 10.27389/d.cnki.gxadu.2021.003631.

[3] Sheng Hu, Zhang Yuxue. Research on network traffic modeling and prediction based on ARIMA [J]. Communication Technology, 2019, 52(4): 903-907.

[4] Wang Jialong. Research on the stability of supply chain network of group-type enterprises base d on cross-stage cascade failure [D]. Zhejiang Gongshang University, 2020.DOI: 10.27462/d.cnki. ghzhc.2020.000542.

[5] Li Shumin, Wang Xu. Impact of underloaded cascade failure on network destructive resistance - an example of fresh produce supply chain network [J]. Science, Technology and Engineering, 2022, 22(18):7746-7756.

[6] Zhao Zhigang, Zhou Gengui, Du Hui. Study on cascading destruction resistance of complex weighted supply chain networks [J]. Small Microcomputer Systems, 2019, 40(12):2591-2596.

[7] Gang Hu, Xiang Xu, Hao Gao, et al. Network node importance identification algorithm based on neighbor information entropy [J]. Systems Engineering Theory and Practice, 2020, 40(03):714-725.

[8] Hu Jieqiong,LI Zhenping. Forecasting and analyzing the whole society freight transportation volume based on time series [J]. Logistics Technology, 2014, 33(09):128-130.

[9] Zheng Maolin,Xia Xiaohong, Wang Xiaorong. Research on transportation network security performance system and comprehensive evaluation method [J]. Computer Knowledge and Technology, 2023, 19(01): 95-97.DOI: 10.14004/j.cnki.ckt.2023.0006.

[10] Huang Ying-Yi,Jin Chun,Rong Li-Li. A cascading failure model for logistics networks considering the integrated importance of nodes [J]. Operations Research and Management, 2014, 23(06):108-115.