Research on Container Throughput Forecast Based on ARIMA-BP Neural Network

Yifei Zhang\textsuperscript{*}, Yuhui Fu\textsuperscript{1} and Genghua Li\textsuperscript{1}

\textsuperscript{1}Navigation College, Dalian Maritime University, Dalian, LiaoNing, 116026, China
\textsuperscript{*}Corresponding author’s e-mail: yifeizhanger@163.com

Abstract. In order to improve the accuracy of the container throughput, the paper proposed a prediction method based on ARIMA-BP neural network for the container throughput, and compared with the combined prediction method based on ARIMA-BP neural network, from the perspective of simple weighting and residual optimization. It is applied to the container throughput prediction of the Qingdao port statistics for a total of 24 quarters from 2014-2019. The results show that the prediction accuracy of the combination prediction method based on residual optimization was the highest. Compared with other prediction models, the evaluation indexes RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and MAPE (Mean Absolute Percentage Error) were 15.95, 13.31 and 2.52\% respectively and the prediction accuracy based on the BP neural network was lowest. The prediction method proposed in this paper for container throughput can provide guidance for the related personnel.

1. Introduction

Container throughput is an important reflection of the value of a port. At present, the prediction methods were commonly used by scholars on container throughput mainly include gray theory [1], BP neural network [2-3], Grey-Markov [4], ARIMA [5] and other methods. Song et al constructed a prediction method of support vector machine model and verified its feasibility [6]; FENG et al. decomposed the characteristic components of monthly container throughput and proved the operability of the indirect prediction method [7]; LE et al. used genetic programming to predict the container throughput of Ningbo Port and opened up a new way to solve the nonlinear prediction [8]; among them, some scholars used combined forecasting methods to predict and analyze container throughput. WANG et al. [9] used the combination of the three-dimensional exponential smoothing method and the Markov model to predict container throughput. Their prediction accuracy is compared with a single model. The prediction accuracy is significantly improved; JIA et al. used factor analysis and curve fitting methods to construct the model, and finally used ARIMA to indirectly predict the container throughput. Its fit and prediction accuracy are both high; but each type of model may not be performed well in all cases [10]. In order to overcome the limitations, some scholars have proposed a combined prediction method, which aims to take advantage of the unique advantages of each different type of model. The common practice is to decompose the time series into two forms, linear and non-linear, and then apply the appropriate types to them respectively. For example, the literature [11-14] have overcome the limitations of a single prediction method to some extent good effect.

Therefore, this paper collects and summarizes 24 quarters of container throughput data collected by Qingdao Maritime Safety Administration from 2014 to 2019, and uses it as a sample to apply it to the two different combined prediction models based on ARIMA and BP neural network. To explore, select the container throughput data of the first 20 quarters as training, and build a container throughput
prediction model based on two combined prediction models. The container throughput data of the next 4 quarters analyzes the prediction results and explores the container throughput changes with the seasons, with a view to providing relevant personnel with reference for port development decisions.

2. Theoretical model

2.1. ARIMA model

ARIMA (p, d, q) is called the Autoregressive Integrated Moving Average Model. Among them, AR stands for autoregression, I stands for difference, MA stands for moving average; p is for autoregression, d is the number of differences made when the time series becomes stationary, and q is the number of moving averages, so it is also called Mixed model. The essence is to make the most primitive non-stationary sequence show differential characteristics through differential operation, and then organically combine with ARMA model to establish ARIMA (p, d, q) model.

Using the ARIMA model, the first step is to perform a stationary test on the time series. If the original time series is non-stationary, the time series needs to be differentially processed until it is a stable time series, and then it is modeled and predicted. The form of ARIMA model is as follows:

\[
\begin{align*}
(\Phi(B)\nabla^dX_t) &= \theta(B)a_t \\
E(a_t) &= 0, \text{Var}(a_t) = \sigma_a^2, \text{Var}(a_t) = 0, s \neq t \\
\text{Var}(s) &= 0, \forall s < t 
\end{align*}
\]

\[
\Phi(B) = 1 - \varphi_1B - \varphi_2B^2 - \cdots - \varphi_pB^p
\]

is the autoregressive correlation coefficient polynomial of the time series ARIMA model; \( \varphi_i (i = 1, 2, 3, \cdots, p) \) is the autocorrelation coefficient; p is the order of autoregressive terms; \( \nabla^d = (1 - B)^d \) is the high-order difference; d is the difference order; \( X_t \) is the time series; \( \theta(B) = 1 - \theta_1B - \theta_2B^2 - \cdots - \theta_qB^q \) is the moving average coefficient polynomial of the time series ARIMA model; \( \theta_i (i = 1, 2, 3, \cdots, q) \) is the moving average coefficient; \( q \) is the moving average term order; \( a_t \) is the residual term; B represents the lag operator; \( \text{Var}(a_t) \) is the variance of the residual sequence.

2.2. BP neural network

BP neural network is a classic artificial neural network, specifically a multi-layer neural network trained according to the error back propagation algorithm, and its learning ideas are: forward propagation of signals and backward feedback of errors. The input layer, hidden layer and output layer constitute the structure of the BP neural network. The neurons of these three layers are connected in sequence, but the neurons of the same layer are not related. Increasing the number of hidden layers can improve the accuracy of the model, but inevitably increases the computational burden. The core of the BP neural network is the reverse propagation of errors. When the information passes from the input layer through the hidden layer and finally reaches the output layer, the error between the output value and the expected value is compared and calculated. If the requirements are met, then the learning will end, otherwise the error will be layer by layer Backward propagation to the input layer, while adjusting the values of various parameters, iterate the above process continuously until convergence.

\[
\begin{align*}
Z_j &= f_1(\sum_{i=1}^{n} \omega_{ij}X_i) \\
Y_k &= f_2(\sum_{i=1}^{n} \omega_{jk}Z_j)
\end{align*}
\]

\( Y_k \) represents the output, \( X_i \) represents the input, \( \omega_{ij} (i = 1, 2, \cdots, N_1; j = 1, 2, \cdots, N_2) \) is the weight between the input layer and the hidden layer, \( \omega_{jk} (k = 1, 2, \cdots, N_3) \) are the weights between the hidden layer and the output layer; \( f_1, f_2 \) are both BP neural network activation functions.
3. Construction of combined forecasting model

The analysis of previous studies shows that the time series ARIMA model is highly suitable for extracting the linear part of the sequence, and the BP neural network is extremely sensitive to the nonlinear factors in the data. From this point of view, the advantages of the two models are combined to improve the prediction accuracy. Therefore, it is proposed to construct a model of container throughput combination forecasting from two perspectives of simple weighted forecasting combination and residual optimization forecasting combination and apply it to the research of container throughput forecasting.

3.1. Combination prediction method based on simple weighting

Two single models were used to fit and analyze the container throughput data, and the weights of the ARIMA model prediction result and the BP neural network prediction result were obtained by a simple weighting method to obtain the container throughput prediction value. The essence of simple weighting is to arrange the variance of the error of the prediction results of each single model. The larger the variance, the lower the weight, and the higher the weight. The basic process of simple weighting is: if the container throughput is m sets of data, \( X_t(t = 1,2,3,\cdots,m) \) is the container throughput sequence collected and sorted; \( X_{1t}, X_{2t}(t = 1,2,3,\cdots,m) \) are the predicted container throughput values based on ARIMA and BP neural networks; \( Y_t(t = 1,2,3,\cdots,m) \) are the final predicted container throughput values based on the simple weighted combined forecast method; \( E_{1t}, E_{2t} \) are the container throughput prediction error values based on ARIMA and BP neural networks, so the final forecast value of container throughput based on the simple weighted combination forecasting method of ARIMA-BP neural network is:

\[
Y_t = k_1X_{1t} + k_2X_{2t}(k_1 + k_2 = 1, k_1 \geq 0, k_2 \geq 0)
\]

According to the simple weighting method, the respective weights \( k_1 \) and \( k_2 \) of the two single prediction models are calculated. The method flow is as follows:

a. Firstly, the stationarity test is performed on the container throughput data series. If it is a non-stationary series, the number of differences \( d \) needs to be determined. Then, the AIC (Akaike Information Criterion) criterion is used to order the model, and the parameters of the ARIMA model are established to obtain the container based on the ARIMA prediction model. Throughput prediction results.

b. Determine the number of structural layers, transfer functions, and nodes of the BP neural network, select the three factors that affect container throughput as GDP, cargo throughput, and total foreign trade export as input, and compare actual container throughput data as output, rolling training Based on the maximum training times and the maximum error, adjust the relevant parameters and thresholds of the BP neural network to establish the BP neural network and obtain the container throughput prediction results based on the BP neural network.

c. The simple weighting method is used to analyze the prediction errors of the two single prediction methods, and the optimal weight values \( k_1 \) and \( k_2 \) are determined, and the final predicted value of the container throughput \( Y_t = k_1X_{1t} + k_2X_{2t} \) is obtained to realize the container throughput prediction based on the simple weighting method Value research.

3.2. Combination prediction method based on residual optimization

Using ARIMA to extract the linear part of the container throughput data sequence, the collected container throughput data sequence is \( X_t \), assuming that the container throughput prediction value is \( X'_{1t} \), the container throughput prediction error is \( e_t \), and \( e_t = X_t - X'_{1t} \). The forecast error \( e_t \) includes the impact of a series of nonlinear factors such as GDP, cargo throughput and total foreign trade export on container throughput. The nonlinear part of the container throughput data sequence is extracted using BP neural network to extract the prediction result is \( X'_{2t} \), then the final prediction result of the combined model with the prediction error optimized by the BP neural network residual is \( Y_t = X'_{2t} \).
The method flow is as follows:

a. First, the stationarity test is performed on the container throughput data series. If it is a non-stationary series, the number of differences d needs to be determined. Then, the AIC (Akaike Information Criterion) criterion is used to order the model, and the parameters of the ARIMA model are established to obtain the container based on the ARIMA prediction model. Throughput data prediction results.

b. The predicted value obtained by ARIMA is compared with the true value \( X_t \) of the container throughput obtained by collecting and sorting, and the predicted error \( e_t \) is calculated.

c. Using BP neural network for optimization training, the weight parameters and thresholds of BP neural network are also adjusted according to the maximum training times and maximum errors, and the final predicted value of container throughput \( Y_t = X_{2t} \) after BP neural network residual optimization is obtained to achieve Research on prediction value of container throughput data based on residual optimization.

3.3. Forecast results evaluation method

In order to verify the rationality of the proposed method and compare its effectiveness, it is proposed to use three indicators such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) to verify the prediction results of each model. That is, the smaller the index, the higher the accuracy and the best effect.

4. A case

4.1. Data Sources

In order to verify the effectiveness of the proposed model, the container throughput data for Qingdao Port from 2014 to 2019 and three factors affecting container throughput were collected and compiled, which are data such as cargo throughput, GDP, and total foreign trade export. As shown in Table 1.

| Quarter | Cargo throughput (10^4) | GDP (Billion) | Total export value (Billion) | Container throughput (10^4) |
|---------|------------------------|--------------|-----------------------------|-----------------------------|
| 14-1    | 12293                  | 1678.40      | 1163.30                     | 415.36                      |
| 14-2    | 11891                  | 2254.40      | 1317.90                     | 420.99                      |
| 14-3    | 11058                  | 2202.90      | 1177.40                     | 415.30                      |
| 14-4    | 11067                  | 2556.40      | 1247.20                     | 421.21                      |
| 15-1    | 12297                  | 1801.00      | 1003.90                     | 422.54                      |
| 15-2    | 12313                  | 2386.70      | 1098.10                     | 436.66                      |
| 15-3    | 11879                  | 2403.10      | 1040.70                     | 438.23                      |
| 15-4    | 11804                  | 2709.27      | 1217.10                     | 446.27                      |
| 16-1    | 12469                  | 1937.53      | 921.40                      | 443.04                      |
| 16-2    | 12606                  | 2582.77      | 1059.70                     | 450.37                      |
| 16-3    | 12546                  | 2630.65      | 1116.40                     | 455.46                      |
| 16-4    | 12463                  | 2860.34      | 1232.60                     | 452.32                      |
| 17-1    | 12686                  | 2460.61      | 1190.00                     | 451.24                      |
| 17-2    | 12794                  | 2614.47      | 1324.90                     | 458.24                      |
| 17-3    | 12709                  | 2908.74      | 1251.90                     | 463.91                      |
| 17-4    | 12609                  | 3053.46      | 1257.40                     | 456.62                      |
| 18-1    | 20196                  | 2845.57      | 1185.40                     | 476.91                      |
| 18-2    | 13256                  | 3139.83      | 1206.20                     | 483.57                      |
| 18-3    | 13730                  | 3100.70      | 1428.40                     | 496.47                      |
| 18-4    | 14117                  | 2915.42      | 1496.10                     | 493.14                      |
| 19-1    | 13813                  | 3190.10      | 1363.70                     | 493.00                      |
4.2. Data analysis and preprocessing

Container throughput abnormal data processing: choose in accordance with the law to collect information preprocessing, the principle is if a quarter of the container throughput data of minus the time series data of the mean of the absolute value of difference is more than 3 times of its standard deviation, so the data is handled as abnormal data, data is replaced with the two quarter of the average.

4.3. Container throughput forecasting

Figure 1 shows a time series diagram of container throughput after preprocessing. It can be seen from Figure 1 that container throughput is an upward trend as a whole, and it is unstable.

Multiple experiments have shown that ARIMA (1,2,1) parameters are the best, and use this to predict container throughput data to obtain container throughput residual data and prediction data; for the construction of BP neural network, first use the trial The algorithm determines that the optimal number of hidden layer nodes is 9, check the changes in accuracy to determine the number of hidden layers, establish a 3-9-1 BP neural network structure, take GDP, cargo throughput, and total foreign trade export as input, the actual container throughput data is used as output, that is, the final predicted value of container throughput; the determination of weights in the optimal weight model ranks the error variance of the results of each single prediction model from large to small according to the basis shown in 3.1, and finally determines the proportion of ARIMA to be $2/3$, the proportion of BP is $1/3$; the residual optimization model calculates the prediction error of ARIMA according to the steps described in 3.2 and uses BP neural network to enjoy the optimization, and also determines the optimal number of hidden layer nodes as 8 according to the trial algorithm , Establish the BP neural network structure of 3-8-1, and obtain the predicted value of container throughput based on residual optimization.

To verify the prediction effect of the proposed combined forecasting method, successively use BP neural network, ARIMA, simple weighted combined forecasting method and residual-based optimized combined forecasting method to forecast the container throughput of Qingdao Port in 2014-2019, and the forecast results as shown in table 2.
Table 2. The predicted results of each model are compared with the real values

| Quarter | 19-1 | 19-2 | 19-3 | 19-4 |
|---------|------|------|------|------|
| Actual value | 493.00 | 537.00 | 539.00 | 532.00 |
| ARIMA | 502.90 | 507.98 | 515.11 | 521.75 |
| BP neural network | 490.40 | 509.26 | 509.40 | 472.28 |
| ARIMA-BP optimal weights | 498.68 | 508.41 | 513.21 | 505.26 |
| ARIMA-BP residual optimization | 503.60 | 511.67 | 522.79 | 530.89 |

4.4. Model evaluation

Table 3 shows the statistical results of the four models under the three indicators of RMSE, MAE and MAPE. As shown in the table, compared with the four prediction models, the prediction error is small, and the optimal prediction method is the residual prediction method. The RMSE, MAE, and MAPE are 15.95, 13.31, and 2.52%, which is the same as ARIMA. Compared with the prediction models, RMSE, MAE and MAPE decreased by 4.15, 4.95 and 0.92% respectively; compared with the BP neural network single prediction model, RMSE, MAE and MAPE decreased by 20.17, 4.95 and 3.09% respectively. In addition, based on the simple weighted combination prediction model and the BP neural network single prediction model, RMSE, MAE and MAPE decreased by 12.51, 8.26 and 1.54%, respectively, but compared with the ARIMA prediction model, RMSE, MAE and MAPE were increased by 3.51 and 3.44, respectively. It is basically the same as 0.63%. Finally, compared with the two combined forecasting models, the residual optimized joint forecasting model has a simpler weighted combined forecasting method whose RMSE, MAE and MAPE are reduced by 7.66, 8.39 and 1.55%, respectively. Through statistical data comparison, the prediction results and evaluation indicators show that the two combined forecasting methods proposed in Qingdao Port Container Throughput Forecast, the combination forecasting method based on residual optimization is the best, and the combination forecasting method based on simple weighting is the best. For the BP neural network prediction method, the ARIMA prediction model is superior to the BP neural network prediction model.

On the whole, the container throughput prediction results based on the combined prediction model have high accuracy. Among them, the combined prediction method based on residual optimization predicts the container throughput with the highest accuracy and the smallest error, and the root mean square error (MAPE) is controlled within 3%.

Table 3. Analysis of prediction results of different prediction models

| Model                          | MAPE  | MAE   | RMSE  |
|--------------------------------|-------|-------|-------|
| ARIMA                          | 3.44% | 18.26 | 20.10 |
| BP neural network              | 5.61% | 29.96 | 36.12 |
| ARIMA-BP optimal weights       | 4.07% | 21.70 | 23.61 |
| ARIMA-BP residual optimization | 2.52% | 13.31 | 15.95 |

5. Conclusion

This article combines ARIMA and BP neural network in different ways to predict and analyze the container throughput of Qingdao Port. Compared with the two combined forecasting models, the results show that the combined forecasting model based on residual optimization has a better forecasting effect on the container throughput of Qingdao Port. Secondly, for the result analysis, it can also be shown that the container throughput time series does have both linear parts and nonlinear factors, because the prediction results and the three evaluation indicators can show that the combined forecasting method based on residual optimization is better than the other three models in the prediction of container throughput, and it can also verify that the different combinations of ARIMA model and BP neural network model can predict the rationality of container throughput.
References

[1] Chen C Y, Dai R, Yang T T. (2016) Study on Maritime Traffic Accidents Prediction Based on Improved Grey Model[J]. Ship & Ocean Engineering, 45: 67-71.

[2] Chen J W, Lan P Z. (2019) Port Throughput Prediction Based on Improved BP Neural Network[J]. Journal of Jimei University(Natural Science), 38: 352-357.

[3] Fan Y Y, Yu S Q. (2015) Port container throughput forecast based on NARX neural network[J]. Journal of Shanghai Maritime University, 36: 1-5.

[4] Du B S, Zhu P F, Liang C M, et al. (2019) Container Throughput Prediction of Shenzhen Port Based on Grey-Markov Model[J]. Journal of Zhejiang Ocean University(Natural Science), 38: 180-186.

[5] Wu C. (2019) Forecast of Port Container Throughput Based on Time Series Model[J]. Pearl River Water Transport, (5): 73-74.

[6] Song C L, Jin L J, Guan F, et al. (2019) Major cargos throughput prediction in port of Dalian based on SVM model[J]. Journal of Dalian Fisheries University, 34: 752-756.

[7] Feng H X, GRIFOLL M, AGUSTI M, et al. (2019) Container Throughput Forecasting Model for Shanghai Port Based on Data Decomposition Method[J]. Navigation of China, 42: 132-138.

[8] Le M L, Fang Y. (2003) Application of Genetic Programming to Prediction of Containers Handling Capacity[J]. JOURNAL OF SHANGHAI JIAOTONG UNIVERSITY, 37: 1246-1250.

[9] Wang Z Z, Chang D F, Zhu Z L. (2019) Analysis and Prediction of Port Quarterly Container Throughput with ES-Markov Model[J]. Navigation of China, 42: 125-130.

[10] Jia F Y, Han X L. (2019) Prediction of container throughput based on factor analysis and curve fitting[J]. Journal of Shanghai Maritime University, 40: 18-22.

[11] Peter G, Zhang. (2003) Time series forecasting using a hybrid ARIMA and neural network model[J]. Neurocomputing, 50: 159-175.

[12] Khashei M, Bijari M. (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting[J]. Applied Soft Computing Journal, 11: 2664-2675.

[13] Babu C N, Reddy B E. (2014) A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data[J]. Applied Soft Computing, 23: 27-38.

[14] Büyükşahin, Ümit Çavuş, Ertekin, Şeyda. (2018) Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition[J]. Neurocomputing, 361: 151-163.