Prediction of container throughput of Dalian Port Based on factor analysis and ARIMA model

Weikai Yao1,*

1School of transportation, Shanghai Maritime University
*Corresponding author: ywk402179@163.com

Abstract. With the development of global economy, the competition of container throughput in ports in the world is becoming more and more fierce. Therefore, it is necessary to forecast the container throughput and determine the policies and measures earlier. In this paper, we use factor analysis and regression model to analyze the container throughput of Dalian port from 1985 to 2019 and get the main factor. We use ARIMA prediction model to predict the main factor sub coefficient of 2020-2025 and calculate the container throughput of Dalian port from 2020 to 2025. The results show that the throughput growth rate will slow down in the next six years, which is in line with the expectation of the 14th five year plan.

1. Introduction
With the development of global economy, the competition among ports in the world is gradually towards the competition of comprehensive strength with container throughput as the core. The accurate prediction of container throughput is an important basis for the reasonable planning of port terminals. Therefore, if we can accurately predict the trend of container increase and decrease in the future and formulate effective strategies and policies, we can occupy a favorable position in the world container market.

2. Journals reviewed
The linear regression method[1] is difficult to simulate nonlinear data or polynomial regression with correlation between data features, and can not well represent highly complex data; The exponential smoothing method[2] gives a smaller proportion in the long term and a larger proportion in the short term; Artificial neural network method[3] can't tell you how it produces the results, let alone why it produces such results; The grey prediction method[4-5] is based on the index rate, which can not consider the randomness of the system. Therefore, this paper proposes a container throughput prediction model based on factor analysis and curve fitting. Firstly, the main influencing factors of container throughput are determined, and then the multicollinearity among factors is eliminated through factor analysis. Then, the comprehensive economic development value of each year is obtained, and then the curve fitting between the comprehensive economic development value and container throughput is carried out, Finally, the ARIMA model is used to predict the comprehensive economic development value in the future, and the value is substituted into the above model to make accurate prediction.

3. Theory and method
Step1. Analyze the problems, collect data, and select the indicators that affect the container throughput.
Step 2. Judge whether the index matrix is suitable for factor analysis. Generally, correlation coefficient matrix calculation, KMO test and Bartlett sphericity test are used to judge. If it is suitable, go to step 3, otherwise go to step 1.

Step 3. Determine the main factor: Let \( x_1, x_2, ..., x_n \) be \( n \) influencing factors. When the cumulative variance contribution rate of the current \( k \) (\( k < n \)) influencing factors is not less than 85\%, the first \( k \) influencing factors can be taken as the main factors. The formula of cumulative variance contribution rate is:

\[
m = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}
\] (1)

In the formula: \( \lambda_i \) is the proportion of the variance of the main factor to the total variance.

Step 4. If the \( k \) main factors cannot be determined or the actual meaning is not obvious, the factors need to be rotated to obtain more obvious actual meaning. \( N \) original variables are expressed as a linear combination of \( k \) principal factors. The first \( k \) principal factors to be found are \( E_1, E_2, ..., E_K \). According to the relationship between the principal factors and the original variables, the mathematical model of factor analysis is established:

\[
X = A \cdot E + \varepsilon
\] (2)

In the formula: \( a_{ij} \) is the linear correlation coefficient between the \( i \)-th variable and the \( j \)-th principal factor. The model is expressed in matrix form as:

\[
X = AE + \varepsilon
\] (3)

In the formula: \( E \) is the main factor vector, \( E_1, E_2, ..., E_K \) can be understood as \( K \) mutually perpendicular coordinate axes in high-dimensional space; \( A \) is the factor load matrix; \( \varepsilon \) is a special factor vector.

Step 5. Determine the comprehensive economic development value (also known as the score of comprehensive factors). Taking the variance contribution rate of main factor as the weight, the main factor score function is obtained by the linear combination of each main factor, and then the comprehensive economic development value of each main factor is determined. The weight of variance contribution rate of main factor is (the higher the value is, the more important the main factor is):

\[
S_j = \sum_{i=1}^{n} a_{ij}^2
\] (4)

Step 6. then uses curve fitting method to fit the comprehensive economic development value and container throughput, and finds out the best prediction model.

Step 7. The autoregressive integrated moving average model is used to predict the comprehensive economic development value in the future, and then it is substituted into the best prediction model to predict the future container throughput.

4. Case analysis

4.1 Data collection and index selection

This paper selects Dalian port as the research object. There are many factors that affect the container throughput \( y \) of Dalian port. In view of the availability of data, six indicators are selected as the influencing factors of container throughput, including GDP \( X_1 \), total cargo transportation \( X_2 \), total import and export \( X_3 \), industrial added value above designated scale \( X_4 \), coastal port cargo throughput \( X_5 \), and total berths \( X_6 \) of Dalian port. This paper uses the data of container throughput and its influencing factors of Dalian port from 1985 to 2019 for empirical research.

Through the correlation analysis of the influencing factors of Dalian port container throughput, we can see that there is a high linear correlation between the indicators. At the same time, Bartlett sphericity test statistical observation value is 329.981, KMO value is 0.827, all pass the test, because the established index matrix is suitable for factor analysis.
4.2 Factor analysis
Using SPSS software for factor analysis, we can draw the following conclusions: (1) when extracting the main factors, the common degree of X1, X2, X3, X4, X5 and X6 are 0.975, 0.906, 0.484, 0.851, 0.967 and 0.819 respectively, the factor loading values are 0.987, 0.952, 0.695, 0.923, 0.983 and 0.905 respectively, and the characteristic root of the extracted main factors is 5.002, And it can explain 83.36% of the total variance of the six indicators, so the effect of extracting the main factor is better.

4.3 Curve fitting
In the above factor analysis, the comprehensive factor scores of different years reflect the comprehensive economic development value of different years, so we can use the comprehensive factor scores to predict the future container throughput of Dalian port. By drawing the scatter diagram with the comprehensive economic development value x as the abscissa and the container throughput of Dalian Port y as the ordinate (see Figure 1), it can be preliminarily judged that there are linear model, quadratic curve model and cubic curve model between them. The curve fitting function of SPSS software is used to evaluate the fitting effect of the three models on the comprehensive economic development value and container throughput (see Table 1). R2 is the decision coefficient, DF1 is the degree of freedom of the model, SIG is the significance, and B1, B2 and B3 are the first, second and third coefficients of the curve respectively.

![Fig 1.Scatter plot of comprehensive economic development value and container throughput](image)

| Model       | R^2   | df1 | sig | constant | b1   | b2   | b3   |
|-------------|-------|-----|-----|----------|------|------|------|
| Linear model| 0.966 | 1   | 0.00| 337      | 373  |      |      |
| Conic       | 0.972 | 2   | 0.00| 293      | 346  | 45.49|      |
| Cubic curve | 0.978 | 3   | 0.00| 264      | 416  | 112  | -61.86|

According to the coefficient R^2 of each model, the cubic curve model has the best fitting degree. Therefore, the estimation equation of comprehensive economic development and container throughput can be obtained as follows:

\[ Y = 264 + 416X + 112X^2 - 61.86X^3 \]  

4.4 Forecast the comprehensive economic development value (ARIMA).
The time series diagram of comprehensive economic development is drawn, as shown in Figure 2. From 1985 to 2019, the time series of comprehensive economic development shows a fluctuating rise, that is, the series is in a non-stationary state. Due to the rising trend of sample data, differential processing is needed to eliminate the possible heteroscedasticity of sample data. After the first-order difference, the sequence tends to be stable, and the ACF and PACF are shown in Figure 3 and Figure 4: The ACF and PACF of the first-order difference sequence show tailing attenuation characteristics when the second-order difference reaches the peak value, so the first-order difference time series is a stationary white noise series, so ARIMA (3,1,0) model can be constructed. The R2 of the model is 0.988, and the
standardized BIC is -3.772, so it can be considered that the fitting effect of the model is relatively good. The comprehensive economic development values of container throughput of Dalian port from 2021 to 2025 predicted by the model are 2.149, 2.1, 2.145, 2.342 and 2.425 respectively.

Fig 2. Scatter plot of year and economic development value

Fig 3. ACF

Fig 4. PACF

4.5 Fitting effect test of cubic curve model
The 1985-2019 comprehensive economic development value calculated by ARIMA (3,1,0) model is substituted into the cubic curve model to fit the container throughput of Dalian port, and compared with the actual value. The results are shown in Figure 5. It can be seen from Figure 5 that the fitting effect is very good, and the model can be used to predict the container throughput of Dalian port from 2021 to 2025.
4.6 Forecast container throughput

By substituting the predicted value of comprehensive economic development from 2020 to 2025 into the cubic curve model, the predicted value of container throughput of Dalian port is 1039.2, 1061.3, 1058.6, 1061.1, 1057.9 and 10493 million TEU respectively, and the month on month growth rates from 2020 to 2025 are 18.6%, 2.1%, -0.3%, 0.24%, -0.3% and -0.8% respectively. It can be seen that the average annual growth rate is 3.3%, which is in line with the "normal" that the container throughput growth of China's ports has entered the medium and low speed growth during the "fourteenth five year plan".

5. Conclusion

Based on factor analysis and curve fitting, the container throughput of Dalian port is predicted. The results show that: first, the model has good fitting effect and high prediction accuracy, which can be applied to the prediction of container throughput. Second, while the container throughput of Dalian port is growing steadily, its growth rate will obviously slow down. Therefore, Dalian port should actively realize the transformation and upgrading of port specialization and intensive operation, increase port service value and enhance core competitiveness.

References

[1] Chien-Chang Chou, Ching-Wu Chu, Gin-Shuh Liang. (Mathematical and Computer Modelling, 2008)
[2] Slawek Smyl, International Journal of Forecasting. (2020)
[3] Nicolas Couellan, Neural Networks. (2021)
[4] Huimin Zhu, Applied Mathematical Modelling. (2021)
[5] Zhu Huimin, Applied Mathematical Modelling. (2021)