ISSN: 2664-9640

Prediction of Shanghai Port throughput Based on Typical Factors

Wenwen Li

College of Transport & Communications, Shanghai Maritime University, Shanghai 201306, China.

Abstract

As a node of cargo transportation, the port is a junction of land transportation and sea transportation, a base of industrial activities, a center of comprehensive logistics, and a growth point of urban development, which plays a promoting role in social and economic development. Throughput is a basic index to measure the port, which plays a vital role in the planning and reasonable layout of the port. In this context, this paper analyzes the nine factors affecting the container throughput of Shanghai port, USES the clustering method to obtain the three representable factors, and USES multiple linear regression, combined with the gray system theory and exponential smoothing to predict the container throughput of Shanghai port in the next year. The test results show that the predicted results are more accurate than the original data, and the predicted results show that the container throughput of Shanghai port still has a growing trend in 2019.

Keywords

Cluster analysis; multiple linear regression; Grey System theory; port throughput forecast.

1. Introduction

In recent years, the maritime transportation industry of our country has developed rapidly, and there are more and more factors affecting the development of port throughput. To promote the formulation of port-related policies, it is necessary to study a port throughput forecasting method that can fully consider the influence of many factors and appropriately simplify the forecasting model. Therefore, it is urgent to study appropriate analysis methods of port throughput influencing factors to establish a scientific and effective port throughput forecasting model. Nowadays, the commonly used port throughput forecasting methods mainly include the time series method, exponential smoothing method, regression analysis method, gray forecasting method, etc. Zhu [1] used the time series-causality combination method to predict the throughput of Guangzhou Port, and the relative error between the predicted value and the actual value was controlled within 10%; Huang et al. [2] took the cargo throughput of a port in the past 15 years as Based on the original data, a three-time exponential smoothing prediction model is established, a computer program is written in VB language, and the best smoothing coefficient is calculated by running it. Enter the prediction year to obtain the corresponding throughput prediction result. Wang et al. [3] applied the grey system forecasting theory to predict the cargo throughput and container throughput of Wuhan Port from 2004 to 2008 with the GM(1,1) model and the residual correction GM(1,1) model. Huang et al. [4] used multiple linear regression analysis to predict the passenger throughput of a civil transportation airport and obtained high-precision prediction results, which provided a practical reference for decision-makers in the construction of the second airport in the city where the airport is located. Based on statistical analysis, Lu Chunyu [5] used the moving average method, the quadratic exponential method, and the elastic coefficient method to predict the non-linear change throughput of Yunnan Shuifu Port in 2020, 2025, and 2030. At the same time, it is concluded that the elastic coefficient method is more accurate and can predict port throughput well.

There are many influencing factors in port throughput. To take into account many influencing factors and control the number of independent variables in the forecasting model to achieve effective prediction, the systematic clustering method is used to determine the typical factors affecting the container throughput of Shanghai Port, and then multiple linear regression analysis is applied. Method, with typical factors as independent variables, establish a forecast model of typical factors of port throughput. In this way, the influence of many factors can be considered, and the types of influencing factors can be adjusted according to the actual situation and forecasting demand, to control the number of independent variables of the forecasting model, and make the forecasting model difficult to moderate and controllable. Since the change of port throughput is not a time-oriented event, it is not feasible to use the obtained multiple linear regression equation to directly predict the throughput of the next year. Based on the multiple linear regression analysis methods, the gray forecasting method and exponential smoothing method should be used. Predict the future data of the main influencing factors, and then bring the predicted main factor data back to the multiple linear regression equation to calculate the container throughput of Shanghai Port in 2019. Through the analysis of the container throughput of Shanghai Port, the validity of the model is verified, and it provides a reference for the forecast and planning of the port's future throughput.

2. Typical factors forecasting model of container throughput at Shanghai Port

2.1. Determine the influencing factors

The forecast of port throughput is affected by many factors, some of which have similar characteristics, while some are completely unrelated. In this article, our team uses clustering to determine typical factors and classifies the influencing factors one by one according to the degree of similarity. The greater the degree of similarity, the priority will be aggregated until all the influencing factors have been aggregated, forming a relationship between the influencing factors. Hierarchical clustering diagram of the output relationship. According to the hierarchical cluster diagram, independent and representative influencing factors can be obtained as typical factors. The system clustering process is as follows:

Data preparation. Assume that m factors are influencing port throughput, $\mathrm{X} = (x_1, x_2, \dots, x_m)^T$. Observed values of each influencing factor for n years, $\mathrm{X}_i = (x_{i1}, x_{i2}, \dots, x_{in}), i = 1,2, \dots, m, x_{ij}$ Indicates the observed value of the i influencing factor in the j year. To avoid the dependence of the value of each influencing factor on the unit of measurement, Z-Score standardization of the observed value is required.

$$z_{ij} = \frac{x_{ij} - \bar{x}}{s_i}, \quad i = 1,2, \dots, m; j = 1,2, \dots, n \tag{1}$$

Where, z_{ij} is the Z-score standardized value of x_{ij} ; $\overline{x_\iota}$ is the mean value of influencing factors x_i; s_i is the standard deviation of the influencing factor x_i.

Determine the similarity between influencing factors. The measure of similarity between various influencing factors is Euclidean distance.

$$d_{ik} = \sqrt{\sum_{j=1}^{n} \left(z_{ij} - z_{kj}\right)^2}, \; i, k = 1,2, \dots, m \tag{2}$$

Determination of similarity between classes. Let G_p and G_q be the two types of influencing factors that have been clustered and merged, then define the similarity between the clusters:

$${D_{pq}}^2 = \frac{1}{h_p h_q} \sum_{i=1}^{h_p} \sum_{k=1}^{h_q} {d_{ik}}^2, x_i \in G_p, x_k \in G_q, 1 \le h_p, h_q \le m, \tag{3}$$

Where: ${D_{pq}}^2$ is the average squared distance between the two factors in the two categories of $G_p$and G_q; $h_p$and $h_q$are the number of influencing factors in the category $G_p$and category G_q, respectively.

Clustering of influencing factors. Clustering is performed using the intra-group join method in the systematic clustering method. Calculate the distance between various influencing factors by formula (2), merge the two influencing factors with the smallest distance into a new category, and each other influencing factors into each category; calculate the distance between classes according to formula (3), And merge the two categories with the smallest distance to form a new category until all the factors are merged into one big category; finally, a hierarchical clustering graph representing the affinity and distancing relationship of the influencing factors is generated according to the hierarchical clustering process.

Determination of typical factors. Determine the classification number of influencing factors according to actual needs and combining with the systematic clustering diagram, and obtain independent and representative influencing factor categories. In each influencing factor category, an influencing factor with the smallest total similarity distance L value is selected as a typical factor to replace the remaining influencing factors.

$$L_{xi} = \frac{\sum {d_{ik}}^2}{h-1} x_k \text{ and } x_i \text{are similar influencing factors} \tag{4}$$

Where: L_{xi} is the total similarity distance of the influencing factors x_i; h is the number of influencing factors in this category, $1 \le h \le m$.

2.2. Throughput prediction mode

The multiple linear regression analysis methods can take into account the influence of multiple factors on port throughput. Combining with the systematic clustering method can effectively predict the port throughput and improve the prediction accuracy of port throughput. The typical factors for setting port throughput are$x_1, x_2, \dots, x_l$ and the application of multiple linear regression analysis methods to establish a typical factor prediction model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_l x_l + \varepsilon \tag{5}$$

In the formula: y is the port throughput; $\beta_l$is the undetermined parameter; ε is the random interference term, which obeys the normal distribution.

2.3. Factor value prediction method based on GM(1,1) model

The GM(1,1) model is the most commonly used predictive model based on gray system theory. Solving the gray parameters is the key and difficult point in establishing the GM(1,1) model. This paper chooses Excel to solve the gray parameters.

For a sequence of $X = \left(x(1), x(2), \dots, x(\mathrm{m})\right)$, $X^{(0)} = \left(x(1), x(2), \dots, x(m)\right)$ is the original sequence, and its cumulative generating sequence for X(0)is $X^{(1)} = x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(m))$, $X^{(1)}$ The immediate mean value generation sequence is $\mathrm{Z}^{(1)} = \left(z^{(1)}(2), z^{(1)}(3), \dots, z^{(1)}(m)\right)$, where $\mathrm{Z}^{(1)}(\mathrm{K}) = 0.5x^{(1)}(k) + 0.5x^{(1)}(k-1), k = 2,3, \dots, m$, then there is an equation$x(k) + az^{(1)}(k) = b$.

If $\hat{a} = [a, b]^T$ is the parameter column, and the constant term vector Y and the accumulation matrix B are respectively.

$$Y = \begin{bmatrix} x(2) \\ \vdots \\ x(m) \end{bmatrix} \qquad B = \begin{bmatrix} -z^{(1)}(2) & 1 \\ \vdots & \vdots \\ -z^{(1)}(m) & 1 \end{bmatrix} \tag{6}$$

Then the least square estimation parameter sequence of gray differential equation $x(k) + az^{(1)}(k) = b$ satisfies

$$\hat{a} = [a, b]^T = （BB^T)B^T Y \tag{7}$$

The predicted value should meet

$$\hat{x}^{(1)}(k+1) = (x(1) - \frac{b}{a})e^{-ak} + \frac{b}{a} \quad (8)$$

$$\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - x^{(1)}(k) \quad (9)$$

The steps to establish GM are: construct the constant vector Y and the accumulation matrix B according to formula (6); solve the gray parameter $\hat{a}$ according to formula (7); put the value of the gray parameter $\hat{a}$ into the formula (8) to obtain GM (1,1) Model; According to formula (9), it is reduced and reduced to the fitted value or the predicted value.

However, not all data are applicable to the GM(1,1) model. After the GM(1,1) model is used to predict, the prediction accuracy needs to be tested. Suppose the variances of the original sequence X(0) and the residual sequence ak are S_α and S_β respectively, then the posterior difference ratio is $C = \frac{S_\beta}{S_\alpha}$, and then judge whether the predicted value is reliable according to the gray model prediction accuracy (Table 1).

Table 1 Grey model Forecast Accuracy table

Forecast Accuracy	Level 1 (Good)	Level 2 (qualified)	Level 3 (barely qualified)	Level 4 (unqualified)
C	C<0.35	0.35≤C<0.50	0.50≤C<0.65	C≥0.65

3. Data collection and analysis

3.1. Data collection

The port of Shanghai connects the north and south coasts of China and the oceans of the world in the front and then runs through the Yangtze River Basin, the inland rivers of Jiangsu, Zhejiang, and Anhui, and the Taihu Lake. It is committed to building an international shipping center and is an important port in our country. Considering that the hinterland economic development level, economic structure, economic development vitality, hinterland traffic conditions, economic foreign trade level, and collection and distribution capabilities will all have an impact on the throughput of Shanghai Port, the following selections have a greater impact on the throughput of Shanghai Port. The big influence factors are analyzed in Table 2.

Table 2 Important influencing factors of Shanghai Port container throughput

influencing factor	Quantitative representation (expressed in Xi)
The overall level of hinterland economic development	Hinterland GDP value (x1)
Hinterland economic structure	GDP of secondary and tertiary industries (X2 X3)
The vitality of the hinterland economy	Retail Sales of Consumer Goods (X4)
Traffic in hinterland	Value added by transportation, warehousing, and postal services (X5)
Hinterland economic level of foreign trade	Total Foreign Trade Imports and Exports (X6)
Hinterland collection and distribution capacity	Collecting and distributing freight volume (road x7, railway x8, aviation x9)

By consulting the relevant data from 2007 to 2018, the relevant data of the above influencing factors are collected, as shown in Table 3**.**

Table 3 Relevant data of each influencing factor

Year	x1	x2	x3	x4	x5	x6	x7	x8	x9
2007	12001.16	5675.49	6223.83	3847.79	724.58	2829.73	35634.00	1122.44	290.14
2008	13698.15	6235.92	7350.43	4537.14	769.64	3221.38	37430.00	985.00	305.00
2009	14900.93	5939.96	8847.15	5172.88	642.13	2777.31	37745.00	941.32	298.25
2010	16872.42	7139.96	9618.31	6036.86	746.41	3688.69	40890.00	958.54	372.31
2011	19195.69	7959.69	11111.06	6777.11	913.60	4374.36	42685.00	887.88	356.22
2012	20101.33	7912.77	12060.76	7387.32	895.31	4367.58	42911.00	825.29	337.96
2013	21602.12	8027.77	13445.07	8019.05	935.06	4413.98	43809.00	694.09	334.98
2014	23560.94	8164.79	15271.89	8718.65	1044.46	4666.22	42848.00	548.96	361.39
2015	24964.99	7940.69	16914.52	10055.76	1130.88	4517.33	40627.00	471.28	370.88
2016	27466.15	7994.34	19362.34	10946.57	1160.27	4338.05	39055.00	460.51	386.92
2017	30133.86	9251.40	20783.47	11830.27	1344.24	4761.23	39743.00	471.89	423.18
2018	32679.87	9732.54	22842.96	12668.69	1533.36	5139.47	39595.00	468.38	417.57

Data source: Shanghai National Economic and Social Development Statistical Bulletin, Shanghai Bureau of Statistics

3.2. Data standardization and factor clustering

To eliminate the influence of the dimensional difference of the original data, the data is standardized according to formula (1), and then

$$Z=\begin{bmatrix} -1.44 & -1.61 & \dots & -1.48 \\ -1.18 & -1.16 & \cdots & -1.14 \\ \vdots & \vdots & & \vdots \\ 1.71 & 1.68 & \dots & 1.45 \end{bmatrix}$$

According to formulas (2) and (3), through the calculation of similarity between influencing factors and influencing factors, between classes and the merging of classes, a systematic clustering diagram of influencing factors is obtained, as shown in Fig 2.

Combining the actual situation of the container throughput of Shanghai Port and the forecast demand, the vertical straight line is shifted left and right in Figure 2 and the three types of influencing factors can be determined when the straight line is stopped at the abscissa 5 and 15. At this time, each horizontal line that has an intersection with the straight line can be determined. The influencing factor corresponding to the line is one category, that is, the various influencing factors included in the left end of the horizontal line are members of this category. See Table 4 for specific classification results.

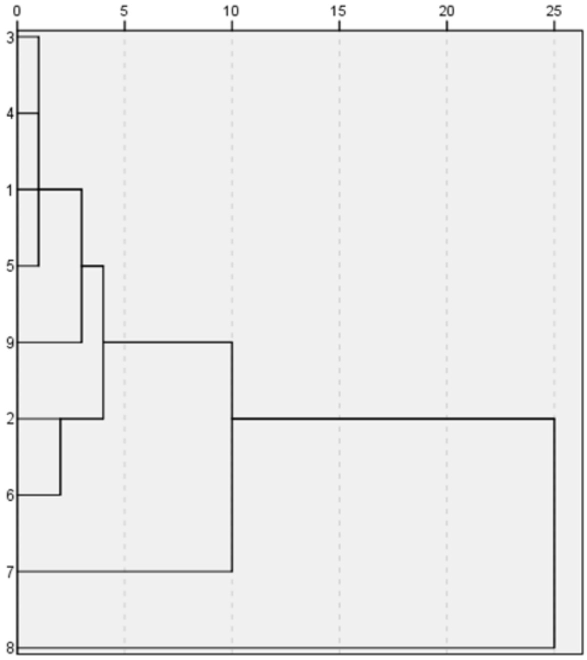

Fig. 2 Clustering diagram of influencing factors

Table 4 Classification results of influencing factors

Category Classification	1	2	3
influencing factor	x1, x2, x3, x4, x5, x6, x9	x7	x8

3.3. Identify typical influencing factors

Calculate the L value in the first type of influencing factors according to formula (4), L_{x_1}=1.094, L_{x_2}=1.656, L_{x_3}=1.395,L_{x_4}=1.261 , L_{x_5}=1.716, L_{x_6}=2.602, L_{x_9}=2.413. The comparison shows that the L value of the influencing factor x_1 is the smallest, so the typical factor in the first category of influencing factors is determined to be the GDP value of the hinterland. Because the characteristics of the various factors in this category are quite similar, and their contribution rates to the forecasted objects are the same, the GDP value of the hinterland selected by calculation can fully represent the impact of the remaining factors in this category on the container throughput of Shanghai Port.

In summary, it can be determined that the typical factors affecting the container throughput of Shanghai Port are the GDP value of the hinterland, the freight volume of road collection and distribution, and the volume of railway collection and distribution. These three typical factors obtained after systematic cluster analysis are independent and representative of each other, and can comprehensively represent the impact of other factors on the container throughput of Shanghai Port.

4. Result

4.1. Factor value calculation

The statistics of Shanghai Port container throughput, hinterland GDP value, road collection, and distribution freight volume, and railway collection and distribution freight volume from 2007 to 2018 are shown in Table 5.

Table 5 Statistics of 4 aspects of Shanghai Port from 2007 to 2018

Year	Container throughput (10,000 tons)	Hinterland GDP (100 million yuan)	Road collection and distribution freight volume (10,000 tons)	Freight volume of railway collection and distribution (10,000 tons)
2007	2615.2	12001.16	35634.00	1122.44
2008	2800.6	13698.15	37430.00	985.00
2009	2500.2	14900.93	37745.00	941.32
2010	2906.9	16872.42	40890.00	958.54
2011	3173.9	19195.69	42685.00	887.88
2012	3252.9	20101.33	42911.00	825.29
2013	3361.68	21602.12	43809.00	694.09
2014	3528.5	23560.94	42848.00	548.96
2015	3653.7	24964.99	40627.00	471.28
2016	3713.31	27466.15	39055.00	460.51
2017	4023.31	30133.86	39743.00	471.89
2018	4201.02	32679.87	39595.00	468.38

According to formula (5), the three typical factors of hinterland GDP value, road collection and distribution freight volume, and railway collection and distribution freight volume are used as independent variables, and the multiple linear regression analysis methods are used to establish the forecast of the typical factors of Shanghai port container throughput. model:

$$y = 0.0810x_1 + 0.0109x_7 + 0.0437x_8 + 1105.662$$

The corresponding data from 2007 to 2018 had been taken as a test and then brought in the multiple linear regression equation. The image obtained is shown in Figure 3. It can be found that the degree of the fitting is better.

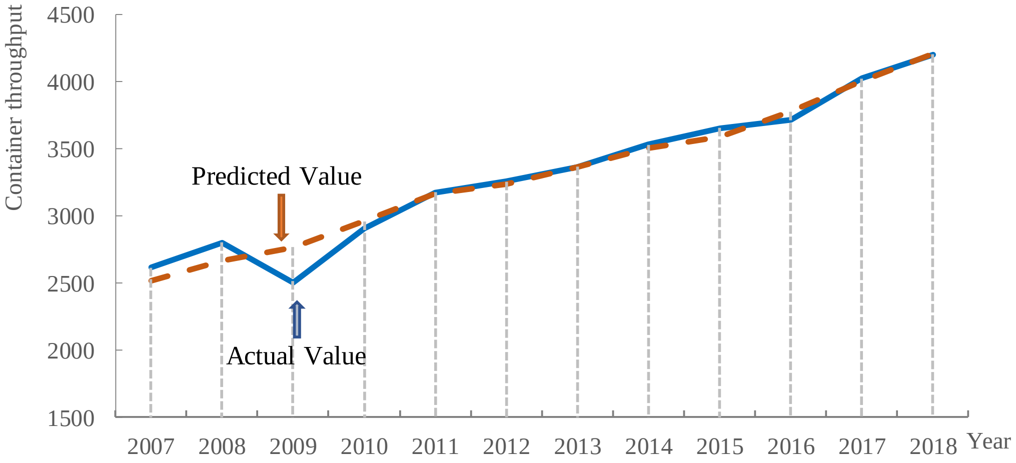

Fig. 3 Linear regression prediction image

Three typical factors, i.e., GDP value of hinterland during 2007-2018, freight volume of road transportation, and railway transportation were used as original data to predict the data of the three influencing factors in 2019 by using the grey system theory.

Taking the GDP value of Hinterland as an example, the original data can be obtained x1(0)={12001.16, 13698.15, 14900.93, 16872.42, 19195.69, 20101.33, 21602.12, ‘’x1(1)=

{12001.16, 25699.31, 40600.24, 57472.66, 76668.35, 96769.68, 118371.80, 141932.74, 166897.73, 194363.88, 224497.74, 251711.61} and generating series next to the mean Z(1)= {18850.24, 33149.78, 49036.45, 67070.51, 86719.02, 107570.74, 130152.27, 154415.24, 180630.81, 209430.81, 240837.68}.

Construct matrices Y and B, namely

$$Y=\begin{bmatrix}13698.15\\14900.93\\16872.42\\19195.69\\20101.33\\21602.12\\23560.94\\24964.99\\27466.15\\30133.86\\32679.87\end{bmatrix} \qquad B=\begin{bmatrix}-18850.24 & 1\\-33149.78 & 1\\-49036.45 & 1\\-67070.51 & 1\\-86719.02 & 1\\-107570.74 & 1\\-130152.27 & 1\\-154415.24 & 1\\-180630.81 & 1\\-209430.81 & 1\\-240837.68 & 1\end{bmatrix}$$

It can be obtained by the least square method $[a, b] = [-0.083, 12622.566]$. The prediction model is $x^{(1)}(k+1) = (x(1) + 152079.108)e^{0.083k} - 152079.108$, k=1, 2,..., n

4.2. Accuracy check

Substitute the original data from 2007 to 2018 into the above formula to obtain the simulated value of the corresponding year. The actual value and the simulated value are compared in Table 6, where: the residual is the difference between the actual value and the simulated value, which can reflect the prediction accuracy to a certain extent.

Table 6 Comparison of actual value and simulated value of Shanghai Port container throughput from 2007 to 2018

Observed value	Predicted value	Residuum	Std. residual	Actual value
2008	14191.05266	492.9026596	1.140720794	13698.15
2009	15380.88608	-479.9560761	-1.110758616	14900.93
2010	16702.78152	169.6384789	0.392593013	16872.42
2011	18203.35575	992.3342455	2.296551432	19195.69
2012	19838.26525	263.0647462	0.608808698	20101.33
2013	21573.29164	28.82836106	0.066717252	21602.12
2014	23452.25131	108.688689	0.251537388	23560.94
2015	25471.11951	-506.1295081	-1.171331587	24964.99
2016	27652.45969	-186.3096923	-0.431175073	27466.15
2017	30048.84505	85.01494542	0.196749427	30133.86
2018	32662.14153	17.72846992	0.04102886	32679.87

It has been calculated that $C_1 = \frac{S_\alpha}{S_\beta} = 0.004944$. According to table 1, $C_1 < 0.35$ has better accuracy. The forecast results show that the GDP of the hinterland in 2019 will be 3,543,7298 billion yuan. The predicted value of X7 road freight volume and X8 railway freight volume in 2019 can be calculated in the same way. But in the calculation, $C_7 \geq 0.65$, Accuracy is poor. Therefore, the exponential smoothing method has been adopted to forecast the cargo volume of highway collection and distribution freight volume.

4.3. Throughput calculation

According to calculations, the predicted values of x1 hinterland GDP value, x7 highway collection and distribution freight volume and x8 railway collection and distribution freight volume in 2019 are respectively 3,543,729,768 billion, 3,960,472,77 million, and 3,949,853,135 million. Bringing the obtained data back to the typical factor forecasting model of the container throughput of the seaport,

$$y = 0.0810x_1 + 0.0109x_7 + 0.0437x_8 + 1105.662$$

It is available that the forecast value of Shanghai Port container throughput in 2019 is 43,810,220 TEU.

5. Conclusion

From the prediction results of the model, the container throughput of Shanghai Port still has an upward trend in 2019, which can increase to 44,234,958 TEU.

The container throughput of Shanghai Port is affected by multiple indicators, among which the GDP value of the hinterland, the freight volume of road collection and distribution, the volume of railway collection, and distribution freight have the most prominent influence. It can be proved that: (1) The GDP of Shanghai is closely related to Shanghai Port container throughput; (2)Roads and railways have a greater impact on port throughput among various transportation methods. The reason is that Shanghai Port is the largest port in our country. As an international shipping port in our country, a large number of goods in the hinterland of China are imported and exported through Shanghai.

Therefore, it can be inferred that: (1) Improving the level of collection and distribution under various transportation modes in Shanghai will further increase the development space of Shanghai Port container throughput; (2) During the data collection process, it is found that the data of Shanghai sea-rail combined transport is not Completely (not only because of the lack of statistics but also because of the low level of Shanghai sea-rail combined transportation), it can be considered that promoting the connection of various transportation modes in the Shanghai area can also greatly increase the throughput of Shanghai Port.

References

[1] Wu Chen.Prediction of port container throughput based on time series model[J].Pearl River Water Transport,2019(05):73-74.

[2] Huang Rongfu, Qi Huale, Cai Jun. Research on the application of the cubic exponential smoothing method in port throughput forecasting[J]. Water Transport Engineering, 2007(6): 13-14.

[3] Wang Zaiming, Wang Hongbo. Application of Grey System Theory in Port Throughput Forecasting[J]. Journal of Wuhan University of Technology (Transportation Science and Engineering Edition), 2005, 29(3):456-459.

[4] Huang Bangju, Lin Junsong, Zheng Xiaoyu, et al. Passenger throughput prediction of civil transportation airport based on multiple linear regression analysis[J]. Mathematics in Practice and Knowledge, 2013, 43(4):172-178.

[5] Lu Chunyu. Research on Throughput Forecast of Jinsha River Water Rich Port[J]. People's Pearl River, 2018, 39(4): 17-20.