ICIC Express Letters
Part B: Applications
Volume **12**, Number **1**, January **2021**

ICIC International ©2021 ISSN 2185-2766
pp. 51–57

PREDICTION OF MARINE ECONOMIC DEVELOPMENT TREND BASED ON SUPPORT VECTOR MACHINE REGRESSION MODEL

Zuoliang Lv and Along Li

School of Economics and Management
Dalian University
No. 10, Xuefu Street, Jinzhou New District, Dalian 116622, P. R. China
gtqyzx@163.net

Received April 2020; accepted July 2020

Abstract. *In this paper, the support vector machine regression model is used to predict the marine economy; the national marine data from 2007 to 2015 (except Guangxi and Hainan) were selected as the basic data for model training and testing, and the principal component analysis was used to determine the final data of the modeling; the cross validation and grid search methods are used to ensure the generalization ability of the model to data; the accuracy of the model is tested by using the scores and relative errors; the experimental results show that the model has high accuracy and meets the requirements of prediction; the marine economy of Tianjin will maintain a steady growth trend, which is based on the forecast of the output value of the marine economy of Tianjin from 2016 to 2023.*

Keywords: Marine economic forecast, Support vector machine regression, Principal component analysis, Cross validation, Grid search

1. **Introduction.** The status of marine economy in the national economy is higher and higher. With the national economic growth slowing down and the economic structure facing a stage of transformation and adjustment, the marine economy will undoubtedly become a new growth point for China's economic development in the future.

Domestic scholars have conducted a large amount of research on marine economy prediction. Ma and Yi [1] predicted the green GDP of marine economy in Jiangsu province from 2015 to 2024 by using the gray prediction method; Zhu [2] verified the feasibility of gray prediction model in the field of marine economy by using relevant data of Guangxi province's marine economy, and predicted Guangxi province's marine economy; Luo and Li [3] studied the sustainable development of Fujian's marine economy with the principal component analysis method; Li and Wang [4] established a new neural network model based on prior knowledge, analyzed and studied the indicators affecting the development of marine economy through time series and multivariate modeling, and finally applied it to the development of marine economy in Zhejiang province; Lin et al. [5] studied marine economy by using gray model, ARIMA model and multiple linear combination model; Liu et al. [6] used GM $(1,1)$ gray prediction model to predict the added value of marine economic output in Zhoushan new area from 2013 to 2015, and put forward reasonable suggestions for the development of marine economy; Qiu et al. [7] predicted the total marine economic value of Jiangsu province during the 13th five-year plan period by using the auto-regressing moving average (ARMA) model in the time series analysis method; Bai [8] used the weighted average method of various forecasts to study the forecast of various indicators affecting the development of marine economy, and finally used the gray forecast model to predict the output value of Guangdong marine economy.

DOI: 10.24507/icicelb.12.01.51

In recent years, with the rapid development of computer technology and the effective combination of statistics and computer science, the machine learning algorithms, which were put forward very early, have been applied more and more in the field of economy. Some algorithms, such as neural network, decision tree and support vector machine, are favored and have excellent performance in solving some nonlinear problems. This paper adopts support vector machine regression model based on support vector machine. Compared with traditional machine learning algorithms (such as neural network and Bayesian algorithm), it is good at solving problems with small sample size, non-linear and high dimension, and it can avoid the problems of overfitting, local optimum, poor ability of local optimum, difficulty of adjusting parameters and slow convergence.

At present, most of the prediction of marine economic development adopts regression statistical methods such as linear or exponential model. However, the relationship between marine economic development indicators is nonlinear, and it is difficult for traditional mathematical models to predict accurately.

Firstly, this paper verifies the feasibility of SVM regression model in the field of marine economic prediction. Secondly, it forecasts the output value of marine economy of Tianjin and provides countermeasures and suggestions for the development of marine economy of Tianjin, so as to realize the sustainable development of marine economy of Tianjin.

2. Indicator Selection and Data Processing.

2.1. Indicator selection.

In view of the complexity of factors affecting the development of marine economy, on the basis of learning from predecessors [9], this paper determines the preliminary evaluation index according to the actual situation of marine economic development. This paper selects 9 coastal provinces and cities in China from 2007 to 2015 (except Guangxi and Hainan), and uses the number of people involved in the sea (X1), the number of marine science and technology projects (X2), the number of marine scientific research practitioners (X3), the number of marine scientific research institutions (X4), the cargo throughput of coastal ports (X5), and the fixed asset investment of the whole society in coastal areas (X6) as the six indicators to forecast the total marine economic output (Y) of Tianjin from 2016 to 2023.

2.2. Data processing.

The data were collected from China statistical yearbook, China marine statistical yearbook and coastal statistical yearbook, and the data of the six indicators from 2016 to 2023 are calculated according to the statistical yearbook data of relevant indicators from 2007 to 2015. Because there are different orders of magnitude, the situation will affect the analysis of the data and the final modeling of the data. In order to eliminate the influence of different dimensions between indicators, we need to standardize the data to get the data needed for data analysis and data modeling. The formula for data standardization is as follows:

$$z = \frac{x_i - u}{s} \tag{1}$$

x_i is the sample data, u is the mean of the sample, and s is the standard deviation of the sample.

3. Model Construction and Model Evaluation.

3.1. Support vector machine regression model.

In order to achieve linear fitting, support vector machine (SVM) regression algorithm is mapped to high-dimensional feature space by mapping all the data points. We are mainly mapping to the high-dimensional nonlinear characteristic space by mapping the nonlinear data points, and the linear fitting of the original space is realized in the high dimensional characteristic space. That is:

$$f(x) = \omega \cdot \Phi(x) + b \tag{2}$$

To fit the data (x_i, y_i), $i = 1, 2, 3, \ldots, n$. The mapping function $\Phi(x_i)$ is a predetermined higher-dimensional function. Use the dot product kernel of the mapping function $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$. The kernel function used in this paper is Gaussian kernel function. The calculation formula is as follows:

$$K(x_i, x_j) = e\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{3}$$

where $f(x)$ is the output and determines ω and b. Get the following formula:

$$\begin{gathered} R(\omega, \xi_i, \xi_i^*) = \frac{1}{2}\omega \cdot \omega + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{s.t.} \begin{cases} y_i - f(x_i) \le \varepsilon + \xi_i \\ f(x_i) - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* > 0 \end{cases} \end{gathered} \tag{4}$$

where ξ_i, ξ_i^* are a relaxation factor. When there is an error in the partition, ξ_i, ξ_i^* are all greater than zero, the error is not less than zero. ε is the sensitivity function, indicating that training accuracy can be achieved to form a pipe with a width of 2ε in the two-dimensional feature space. Constant $C > 0$ represents the penalty degree for the sample exceeding the error. The maximum function can be obtained by solving it as follows:

$$\begin{gathered} L_{dual} = -\frac{1}{2}\sum_{i=1,j=1}^{n}(\partial_i - \partial_i^*)\left(\partial_j - \partial_j^*\right) \cdot K(x_i, x_j) + \sum_{i=1}^{n}(\partial_i - \partial_i^*)\,y_i - \sum_{i=1}^{n}(\partial_i + \partial_i^*)\,\varepsilon \\ \text{s.t.} \begin{cases} \displaystyle\sum_{i=0}^{n}(\partial_i - \partial_i^*) = 0 \\ 0 \le \partial_i, \partial_i^* \le C \end{cases} \end{gathered} \tag{5}$$

The parameters are calculated to get the parameter b, the parameter ω does not need to be explicit.

$$\begin{aligned} b = \frac{1}{n}\Bigg\{ &\sum_{0<\partial_i<C}\left[y_i - \sum_{i,j=1}^{n}\left(\partial_j - \partial_j^*\right)K(x_j, x_i)\right] \\ &+ \sum_{0<\partial_i^*<C}\left[y_i - \sum_{i,j=1}^{n}\left(\partial_j - \partial_j^*\right)K(x_j, x_i)\right]\Bigg\} \end{aligned} \tag{6}$$

Therefore, the linear fitting function obtained from the sample points is

$$f(x) = \omega \cdot x + b = \sum_{i=1}^{n}(\partial_i - \partial_i^*)\,K(x_j, x_i) + b \tag{7}$$

3.2. **Model test.** After the establishment of the model, it is necessary to test the reliability of the model selected in the paper. The paper uses the relative error test method and explained_variance_score test method.

Relative error test:

$$\delta = \Delta/L \times 100\% \tag{8}$$

Δ is the absolute error between the real result and the predicted result, and L is the true result.

The explained_variance_score test:

$$\text{explained\_variance}\,(y, \tilde{y}) = 1 - \frac{Var\,\{y - \tilde{y}\}}{Var\{y\}} \tag{9}$$

The closer the explained_variance_score is to 1, the better the model is;

The closer the explained_variance_score is to 0, the worse the model is.

3.3. Marine economic output prediction process based on support vector machine regression model.

i. Formula (1) is used to standardize the original data;

ii. Using principal component analysis algorithm for the completed data to determine the final index of support vector machine regression model;

iii. The final data is segmented to prepare the data for the establishment of the model;

iv. Using the training data to construct the model;

v. In the training data, the method of cross-validation is used to improve the generalization ability of the data to the model, and the method of grid search is used to determine the basic range of parameters required by the model;

vi. In the parameter range that satisfies the model, the grid search method is used to test the test data and determine the final maximum parameter combination range;

vii. Using the optimal parameter combination and test data, the regression algorithm of support vector machine was used to predict the predicted value, and using the relative error test and explained_variance_score test to verify the feasibility of support vector machine regression model in the field of marine economic prediction;

viii. Using the final determined optimal combination parameters to predict the output value of Tianjin marine economy from 2016 to 2023.

4. Empirical Analysis.

4.1. Statistical analysis of data.

In order to measure the degree of correlation between indicators, this paper uses Pearson correlation coefficient to measure the degree of correlation between indicators. The specific Pearson correlation coefficient matrix is shown in the following table.

Table 1. Correlation coefficient matrix between indicators

	X1	**X2**	**X3**	**X4**	**X5**	**X6**	**Y**
X1	1	0.5323	0.3381	0.8338	0.6334	0.2206	0.7354
X2	0.5323	1	0.7804	0.5192	0.2278	0.4969	0.7816
X3	0.3381	0.7804	1	0.5994	0.3328	0.2997	0.7299
X4	0.8338	0.5192	0.5994	1	0.6458	0.1872	0.7138
X5	0.6334	0.2278	0.3328	0.6458	1	0.4112	0.6625
X6	0.2206	0.4969	0.2997	0.1872	0.4112	1	0.4920
Y	0.7354	0.7816	0.7299	0.7138	0.6625	0.4920	1

According to Pearson correlation coefficient matrix in Table 1, the degree of correlation between each indicator is different and all of them are greater than zero. The results showed that the correlation linear relationship between the indicators was different, but they are positively correlated; the correlation degree between indicators is less than 0.3, it indicates that the degree of correlation between indexes is not linearly dependent; the degree of correlation between indicators is also greater than 0.8, and it indicates that the degree of correlation between indexes is strongly linear.

The correlation degree of each index to the output value of marine economy: the number of marine science and technology projects has the highest correlation with the output value of marine economy, which indicates that marine science and technology plays an increasingly important role in the development of marine economy; the number of sea-related employment is the second most correlated with the output value of marine economy, which represents a growing industry in the ocean.

4.2. **Tianjin marine economic forecast under support vector machine regression model.** This paper uses scikit-learn toolkit to complete the construction and debugging of the model. Firstly, six indexes were analyzed by principal component analysis; when the value of principal component is 0.95, the cumulative contribution of the four indicators has reached the requirement of 95%, and then the four indicators are used to construct the model. Firstly, 10% of the data samples are randomly selected as test samples (it is finally used to evaluate the regression effect of support vector machine regression model); secondly, in the remaining 90% of the training samples, the method of grid search was adopted to combine different C and $Gamma$ (C and $Gamma$ are independent of each other); in order to improve the generalization ability of the model, using 10 fold cross-validation method to determine the combination range of parameters C and $Gamma$ satisfying the model, the model is evaluated using the explained_variance_score value. The experimental results showed that when $C \geq 10000$ and $Gamma \geq 0.08$, the regression effect of support vector machine regression model satisfies the regression effect. Considering that the value range of C affects the final regression effect, a reasonable C value is very important.

From a risk perspective, in order to make the target function smaller the value of C weighs the empirical risk (fitting ability of samples) and structural risk (prediction ability of test samples); as C gets bigger, the smaller the regularization term, which indicates that the greater the structural risk is, the smaller the empirical risk is, and the greater the possibility of over-fitting is; on the contrary, the smaller C is, the lower the complexity of the model is and the lower the structural risk is. In order to weigh the two risks, this paper determines the final value of C is $10000 \leq C \leq 15000$.

After $Gamma$ was selected as a particular kernel function, we need a parameter to optimize. The value of $Gamma$ indirectly affects the distribution of the original data mapped to the new feature space. When the value of $Gamma$ is larger, the original data is more dispersed in the new feature space, the fewer support vectors. So the magnitude of the $Gamma$ is related to the number of support vectors, the number of support vectors affects the computational speed of the model.

In this paper, the final $Gamma$ is determined to be $0.08 \leq Gamma \leq 0.12$. The relationship between parameters σ and $Gamma$ in Gaussian kernel function is as follows:

$$Gamma = \frac{1}{2 * \sigma^2} \tag{10}$$

Finally, using 10% test samples, the final model parameter combination is determined by using grid search method within the range of model parameters satisfying regression effect. So that is the final $C = 14307$, $Gamma = 0.1$. The explained_variance_score values corresponding to the different parameter combinations are shown in Figure 1.

Under the model of optimal parameter combination, the comparison between predicted results and real results of 10% randomly selected test sample points is shown in Table 2.

By comparing the predicted results of experimental data with the actual results, the marine economic prediction based on the regression model of support vector machine meets the prediction requirements; although the relative error is too large, from the point of view of the explained_variance_score, the explained_variance_score score is close to 1, which meets the precision requirements of model regression. Therefore, it is allowed that the relative error of some prediction results is large. Using the model parameters completed by commissioning to predict the marine economy of Tianjin from 2016 to 2023, the results obtained are shown in Table 3.

The model parameters are used to predict the marine economy of Tianjin from 2016 to 2023, and the results are shown in Figure 2.

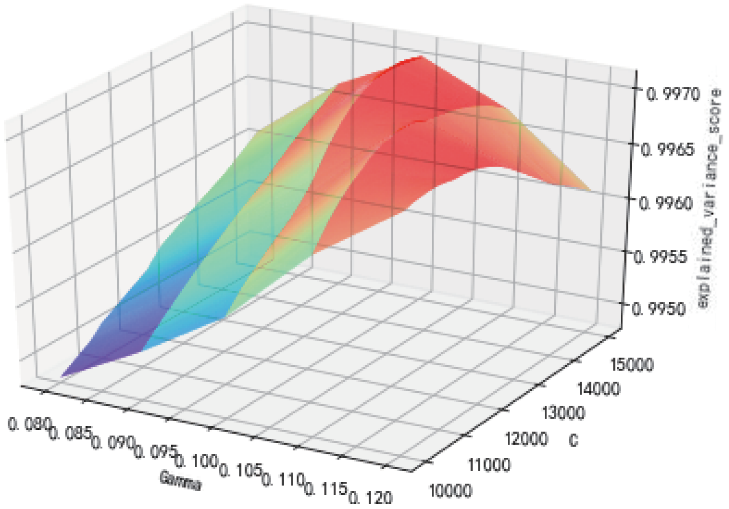

FIGURE 1. The value of the explained_variance_score corresponding to different *Gamma*, *C* combinations

TABLE 2. Comparison of prediction results with real results

Forecast result	1611.1	3522	4215.3	5492.9	5975.9
Real result	1622	3345.5	4554.1	5437.7	6016.6
Relative error	0.7%	−5.3%	7.4%	−1%	0.7%
Forecast result	6983	9402	11051.9	12895.7	
Real result	7074.5	9191.1	11283.6	13229.8	
Relative error	1.3%	−2.3%	2.1%	2.5%	
Explained_variance_score	0.9971				

TABLE 3. Forecast of marine economic output value in Tianjin from 2016 to 2023

Years	2016	2017	2018	2019	2020	2021	2022	2023
Forecast result	5414.1	5750.8	5997	6162.7	6348.5	6516.7	6657.8	6920.8

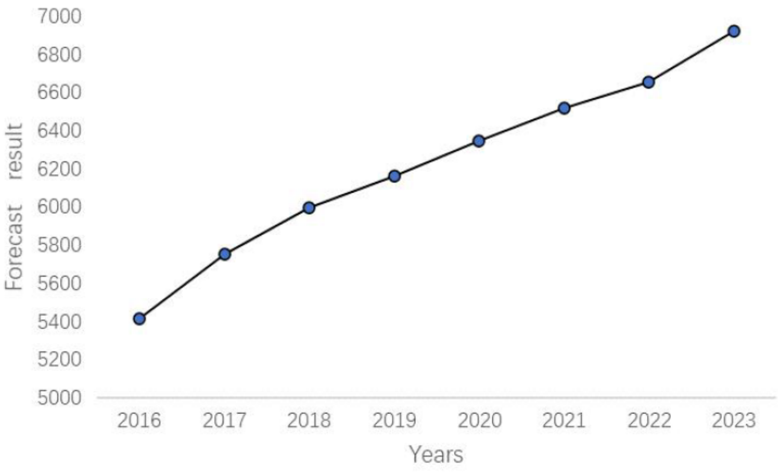

FIGURE 2. Trends of marine economic output forecast in Tianjin from 2016 to 2023

5. Conclusions and Countermeasures.

5.1. Conclusion.

Due to the complex and multivariate characteristics of marine economy, there are nonlinear, nonequilibrium and evolutionary characteristics at the same time. If scientific decisions are to be made about marine economic development in a complex marine economic environment, it is necessary to analyze various indicators that affect the development of marine economy to predict the future development of marine economy. By checking the predicted results and real results of the test data, the predicted results are not much different from the real results, meeting the requirements of prediction, which shows that the regression model of support vector machine is feasible in marine economic prediction.

The regression model of support vector machine is used to predict the output value of Tianjin marine economy from 2016 to 2023. The forecast results show that the output value of marine economy in Tianjin maintains a steady growth trend from 2016 to 2023.

5.2. Countermeasures.

i. Tianjin develops marine industries with regional advantages according to the advantages and development status of marine industries according to the actual basic situation of marine nature to develop the marine industry with regional advantages; at the same time, we should also pay attention to the coordinated development of land and sea, and formulate policies for marine economic development from the economic development of the whole coastal area to ensure the rapid and stable development of marine economy.

ii. Tianjin should attach importance to the influence of marine science and technology on the development of marine economy, and increase the input of marine science research, and make the development of marine economy more healthy.

iii. Tianjin should attach importance to the modernization construction of the port, so that the modernization construction of the port matches with the rapid development of marine economy.

REFERENCES

[1] C. Ma and L. Yi, Macroeconomic dynamic prediction based on Bayesian network model, *Statistics and Decision Making*, vol.14, pp.64-67, 2018.

[2] N. Zhu, Prediction of Guangxi's marine economic added value based on grey model, *Practice and Understanding of Mathematics*, vol.46, no.1, pp.102-109, 2016.

[3] P. Luo and Y. Li, Empirical analysis on sustainable development of marine economy in Fujian Province based on principal component analysis, *Forum on Industry and Science and Technology*, vol.17, no.5, pp.22-24, 2018.

[4] Z.-B. Li and L.-L. Wang, Analysis on the relationship between the amount of port logistics to coastal marine economic growth and gorecasting based on priori neural network: Case of Zhejiang Province, *Value Engineering*, vol.35, pp.144-147, 2012.

[5] P. Lin, X. Ni and Y. Liu, Combined prediction of marine economic development based on cross validation method, *Systems Science and Mathematics*, vol.38, no.7, pp.823-829, 2018.

[6] C. Liu, M. Hu and M. Wang, Prediction and suggestions on marine economic development trend of Zhoushan Islands New Area, *Water Transport Management*, vol.36, no.1, pp.17-19, 2014.

[7] Y. Qiu, X. Song and F. Luo, Prediction of marine GDP of Jiangsu Province during the 13th Five Year Plan period, *Marine Development and Management*, vol.35, pp.9-12, 2018.

[8] F. Bai, Bai Fuchen's dilemma and countermeasures in constructing marine ecological compensation mechanism in Guangdong, *New Economy*, no.1, 2013.

[9] Z. Han, X. Wang and F. Peng, Dynamic analysis and prediction of TFP in China's marine economy, *Geography and Geographic Information Science*, vol.35, no.1, pp.95-101, 2019.