ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 76
7th International Conference on Education, Management, Information and Mechanical Engineering (EMIM 2017)

Time Series Analysis in the Prediction of Water Quality

Qi An[1, a] and Min Zhao[2, b, *]

[1]Management School, Shanghai University, Shanghai, 200444, China

[2]Sydney Institute of Language and Commerce, Shanghai University, Shanghai, 201800, China

[a]an7i1994@163.com, [b]zhaomin1226@163.com

Keywords: Time series analysis; ARIMA model; Water quality prediction

Abstract. The time series analysis method and the actual situation of the real-time monitoring system of the coastal waters are used to select the real-time monitoring data of the dissolved oxygen (DO) in the water quality monitoring data from 24th to 28th March 2016 as a research sample, the ARIMA model was fitted in the Eviews. The model was used to predict the data on March 29th and 30th and compared with the actual measured data. The results show that the relative error between the predicted value and the real value is 4-12% and the average error is about 4.79%. It is proved that the time series analysis is good enough in the water quality prediction. By comparing the results of static prediction and dynamic prediction, it is discussed that the limitation and future research of time series analysis method in water quality prediction problem.

Introduction

Water resource is one of the most important natural resources in the world, which is the basic condition for human survival and development. The sustainable utilization of water resources is an important guarantee for the sustainable development of society and economy. Therefore, water quality monitoring and forecasting is one of the main subjects in the current social research: the data obtained by real-time monitoring to water quality, can reflect the impact of influencing factors on water environment system [1]. And then using the current science and technology to predict the evolution trend of the quality of the monitoring waters can timely take appropriate measures such as water quality management, disaster warning [2].

In this paper, considering the actual situation of the real-time monitoring system of coastal water quality of Shanghai Ocean University, dissolved oxygen (DO) as an important parameter in water quality monitoring data is selected as the research sample, and ARIMA (3,1 (1,2,4)) model was established by using time series analysis method. The experimental results show that the prediction method is better, and it can be applied to the water quality monitoring and early warning.

Time series analysis is an important part of mathematical statistics. Its basic idea is "let the data speak itself". As long as the current value of the observed variable is seen as its own past values and external random interference factor function, based on the historical data of the observed variables, the model with dynamic regularity of the time series is established, and the future development trend is forecasted [5].And usually the water quality parameters of the monitoring data will have a sound record, so the time series analysis of the required data acquisition is very convenient, and data processing is not very complicated. And the random interference is controlled, so the short-term forecast is very accurate [4].

To overcome the deficiencies of deterministic factorization, G.E.P. Box and G. M. Jenkins [6] proposed the Autoregressive Integrated Moving Average (ARIMA) model. The non-stationary time series is transformed into a stationary differential sequence by using the difference method, and then the ARMA model is fitted by using the stationary sequence analysis method.

Prediction Model

Data Processing

Dissolved oxygen (DO) refers to the molecular oxygen dissolved in water, which is one of the main living conditions of aquatic organisms, and can reflect the degree of water pollution, especially the

degree of organic pollution, so it is an important indicator to measure water self-purification capacity, but also one of the important indicators to evaluate the degree of water pollution. Therefore, this study is based on the real-time monitoring system of coastal water quality of Shanghai Ocean University, and selects the dissolved oxygen (DO) from 24th to 28th March 2016 as a research sample.

By plotting the timing diagram of sequence, it is found that the original image has a certain seasonal trend and an intercepted graph, so it is initially determined as a nonstationary sequence. The ADF test shows that the unit root statistic is greater than the critical value, with the probability P> 0.05, so accept the original hypothesis, that is, the original sequence is non-stationary sequence.

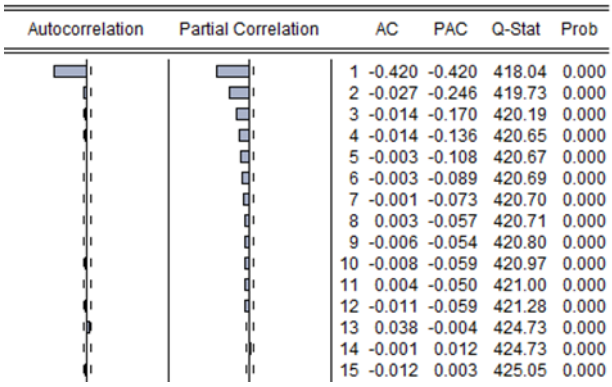
Using the autocorrelation diagram to further assist in the test determines whether the sequence has analytical value. It is found that the Q statistic is very significant and the concomitant probability is less than 0.05. The rejection sequence is the original hypothesis of the white noise sequence, indicating that there is an autocorrelation in the original sequence.

The essence of ARIMA is to combine differential operation with stationary time series analysis method [8]. When there is some deterministic factor in the above sequence, the trend is eliminated first, and then the appropriate order is selected to deal with the adjusted sequence and transform it into a suitable time series for appropriate analysis.

Model Designing

(1) Stationary Processing. First of all, the original sequence DO is subjected to logarithmic first order difference processing. It is found that the sequence is always fluctuating around a fixed value, and there is no obvious trend or periodicity. Observing the autocorrelation graph of the dlndo sequence found that the autocorrelation function and the partial correlation function gradually decay and tend to zero (figure1), the Q statistic is very large and the concomitant probability P is 0. In the unit root test, the concomitant probability is 0, and the null hypothesis is rejected. Therefore, it can be concluded that the dlndo sequence after the original sequence is processed by logarithmic difference is stationary.

(2) Model Identification. Based on the previous analysis, the sequence is taken as the first order differential smooth non-white noise sequence, so consider the establishment of d = 1 ARIMA model. Observe the autocorrelation diagram of the dlndo sequence (figure1), q value optional 1, 2, 3, 4. dlndo partial correlation function was slow decay trend, 1-12 orders are more than 2 times the standard deviation, so the p value can choose 1-12.



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.420	-0.420	418.04	0.000
		2	-0.027	-0.246	419.73	0.000
		3	-0.014	-0.170	420.19	0.000
		4	-0.014	-0.136	420.65	0.000
		5	-0.003	-0.108	420.67	0.000
		6	-0.003	-0.089	420.69	0.000
		7	-0.001	-0.073	420.70	0.000
		8	0.003	-0.057	420.71	0.000
		9	-0.006	-0.054	420.80	0.000
		10	-0.008	-0.059	420.97	0.000
		11	0.004	-0.050	421.00	0.000
		12	-0.011	-0.059	421.28	0.000
		13	0.038	-0.004	424.73	0.000
		14	-0.001	0.012	424.73	0.000
		15	-0.012	0.003	425.05	0.000

Figure 1 ACF and PACF of dlndo series

(3) Model Fitting and Parameter Estimation. LS is used to estimate the parameters. In the case of d = 1, the least squares estimation (LS) is used to consider the combination of different values of p and q, And finally two optional models that passed the parameter significance test: sparse coefficient model ARIMA (3,1 (1,2,4) when p = 1, 2, 3; q = 1, 2, 4) and sparse coefficient model ARIMA (3,1, (1,3,4) when p = 1, 2, 3; q = 1,3, 4).

(4) Model Establishment and Optimization. For the comparison and selection of models often

can not simply look at a certain indicator, but to consider all aspects of the situation, to make a comprehensive judgment. Comprehensive analysis of various types of indicators on the conditions we have a comparative analysis of the model (see Table 1). The Sparse coefficient model ARIMA (3,1, (1,2,4)) is the optimal model. And then observe the residual sequence of the autocorrelation diagram, found that all the associated probability greater than 0.05, that is, the residual sequence of pure random sequence, indicating that the original observation sequence implied relevant information has been sufficient extract.

Table 1 comparison of evaluated indexes in two models

	R-squared	Log likelihood	AIC	DW
A: ARIMA(3,1,(1,2,4))	0.293478	5209.336	-4.395721	1.999444
B: ARIMA(3,1,(1,3,4))	0.293461	5209.308	-4.395697	1.997374
Optimal	A	A	A	A

Prediction and Analysis

According to the above analysis, the sparse coefficient model ARIMA (3,1,(1,2,4)) is established. First of all, the static prediction of do original sequence (1-2371) was used to test the model. The prediction part is obtained, and the solid line part is the predicted sequence diagram, and the dotted line is the 95% upper and lower confidence interval. According to the results, we can see that the prediction error of the model is very small, and the fitting effect is very good. The model structure of ARIMA (3,1, (1,2,4)) is as follows.

$$(1-B)(\ln do)_t = 0.000027 + 1.451280(\ln do)_{t-1} - 0.346661(\ln do)_{t-2} - 0.161144(\ln do)_{t-3} + \varepsilon t - 2.090768\varepsilon t-1 + 1.210169\varepsilon t-2 - 0.102110\varepsilon t-4 \quad (1)$$

(1) Dynamic Prediction. Dynamic prediction of DO concentration data from March 29, 30 shows that the relative error of forecasting and real value is 4% -12% and the average relative error is about 4.79%, which indicates that the forecast is more accurate.

(2) Static Prediction. In order to further verify the accuracy and practicability of the model, the model was used as the prediction sample for the 29-day dissolved oxygen concentration sequence. Static prediction is said that each forecast once, with the real value instead of the forecast value, added to the estimated interval, and then forward forecast, that is static prediction is the actual value of the original sequence to achieve the forecast. It can be seen from the figure 2 that the simulation curve can reflect the actual trend of dissolved oxygen concentration. The average relative error is 17% and the minimum absolute error is 0.003%, MAPE = 0.604%, good prediction accuracy.

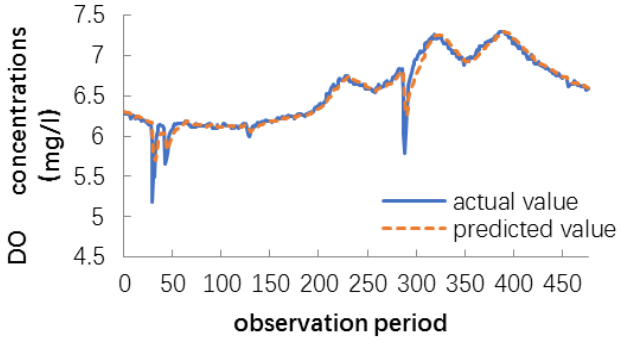


Figure 2 the contrast of predicted and truth value for DO on 29th

(3) Prediction Result Analysis. The results of the analysis can be found to be very accurate for the short-term prediction of the model, but with the increase of the forecast period, the prediction error will gradually increase. This is mainly due to the following reasons.

Firstly, because the ARIMA model only considers the law of the development of the time series itself, it does not take it into account that the water quality parameters will be subject to many

uncontrollable internal and external factors, and these unexpected factors in the model can only be ε_t random interference item to indicate that there is no way to show in the forecast.

Secondly, the time series analysis requires data to be as continuous as possible, the more the sample, the more obvious the law, and the more accurate the fitting model. The sample in this paper is only 5 * 480 data of five days, can not be a good show of water quality parameters of the monthly, quarterly changes in the law. This increased difficulty to the work of this article to a certain extent.

Thirdly, comparing the results of static prediction with dynamic prediction, it is considered that if the samples are supplemented in future research, and if the newly received monitoring data can be added to the original sequence in time, it is possible to update the model in real time and reduce the error of long-term prediction to improve the prediction accuracy of the model.

Conclusion

The study of water quality monitoring and forecasting is very large and complicate system engineering. The complexity of influencing factors and the richness of the research needs are far from what we imagine. The future research trend can be mainly aimed at the following aspects.

Firstly, we carry out the combinatorial model combined with time series model. At present, the single water quality prediction method is an effective way to improve the accuracy and reliability of the prediction. Although the application has been widely used, but each has its own shortcomings, and the method of combining and coupling with different methods is effective, such as time series analysis and artificial neural network prediction method to enhance the ability to deal with nonlinear problems, and by virtue of BP neural network strong adaptive ability to optimize the model to improve the long-term prediction accuracy.

Secondly, on the basis of time series analysis, combined with new technologies, such as time series analysis data calculation, modelling and forecasting based on the strong spatial analysis capabilities of GIS to complement each other, to improve water quality monitoring, predictive simulation capabilities, a single data table into a vivid image of the graphical way to help people water environment control and decision.

References

[1] G. H. Sun, Y. Shen, Y.M. Xu, et al, Time series analysis and forecast model for water quality of Yellow River based on Box-Jenkins method Journal [J]. Agro-Environment Science, 30(2011)1888-1895(in Chinese)

[2] K.S. Parmar, R. Bhardwaj, Statistical, time series, and fractal analysis of full stretch of river Yamuna (India) for water quality management [J]. Environmental Science and Pollution Research, 22(2015)397-414.

[3] X. Du, G. Wu, D. Xu, Prediction Methods Analysis of the Water Quality Based on the ARMA Model [J]. Chinese Agricultural Science Bulletin, 29(2013)221-224. (in Chinese)

[4] Z. B. Shi, Z.H. Zou, Applied study of ARIMA model based on wavelet analysis on water quality prediction [J]. Chinese Journal of Environmental Engineering, 8(2014)4550-4554. (in Chinese)

[5] P. Cortez, M. Rocha, Neves J. Evolving Time Series Forecasting ARMA Models [J]. Journal of Heuristics, 10(2004)415-429.

[6] D. Ö. Faruk, A hybrid neural network and ARIMA model for water quality time series prediction [J]. Engineering Applications of Artificial Intelligence, 23(2010)586-594.

[7] G.E.P. Box, G.M. Jenkins, Time Series Analysis: Forecasting and Control [M]. San Francisco: HoldenDay. 1970

[8] B. Krishna, Y.R.S. Rao, P.C. Nayak. Time Series Modeling of River Flow Using Wavelet Neural Networks [J]. Journal of Water Resource & Protection, 3(2011)50-59.