Prediction of the Ammonia Nitrogen Content with Improved Grey Model by Markov Chain

Yiheng Fang *

Department of Maths, Ohio State University, Ohio, US

* Corresponding Author Email: fang.1016@osu.edu

Abstract. Water pollution prediction plays a crucial role in environmental protection and sustainable development. This study proposes an innovative approach to enhance the accuracy of water pollution prediction by combining the grey prediction model (GM) with Markov chain analysis. This research focuses on predicting the concentration of ammonia nitrogen (NH3-N) in Dongting Lake, a significant water body. Grey prediction models (GM) are utilized to forecast NH3-N content, addressing the challenge posed by incomplete or insufficient data. However, due to the dynamic nature of water quality indicators, GM models may have limitations in terms of accuracy. To overcome this issue, this study introduces the concept of Markov chains, incorporating historical state transitions into prediction models to achieve more precise forecasts. The research demonstrates a novel method for water pollution prediction that integrates GM models with Markov chain analysis, resulting in improved accuracy when predicting NH3-N concentrations. A comparison with traditional GM predictions highlights the effectiveness of this approach. The model's performance was evaluated using actual data from the China Automated Water Quality Monitoring Report. Combining grey prediction models with Markov chains outperforms traditional methods when it comes to predicting water pollution levels. The result contributes to advancing the field of water pollution forecasting by enhancing forecasting accuracy and providing informed decision support for environmental protection and management purposes.

Keywords: Water pollution prediction; grey prediction model; Markov chain analysis; ammonia nitrogen; Dongting Lake.

1. Introduction

Water pollution prediction aims to study the trends of the changes in water quality and also to predict future water quality. It has important research significance and value in environmental protection and sustainable development. It acts as a reference for governments and public health sectors to design policies and allocate resources to protect the environment. Water pollution prediction research history has evolved from the early rule of thumb to a stage based on mathematical models and modern data-driven methods. Early water pollution prediction mainly relied on field monitoring and empirical methods, but with the development of statistics, mathematical modeling, and computer technology, researchers have gradually proposed more accurate and complex prediction methods, covering time series analysis, material balance model, mathematical modeling, machine learning, etc. [1-5].

The grey prediction model is a forecasting method suitable for incomplete or insufficient data. In water pollution forecasting, multiple factors often affect data and may be missing or incomplete [4]. The grey prediction model can accurately predict future water quality change by analyzing the trend and law in the data series. Its key advantages include being suitable for small samples or short sequences, considering system trends, and being simple and easy to understand. Therefore, when solving the problem of water pollution prediction, the grey prediction model can make up for the shortcomings of insufficient data, considering the system trend and simple modeling, and providing a powerful tool for improving the accuracy and reliability of prediction [5].

Now, the grey prediction model is widely used in a lot of areas of science [6-9]. Using the grey prediction model to predict the different indexes for water qualities in specific water bodies is widely researched [8]. However, because the indexes for water quality, especially the concentration of pollutant in water is widely distributed and frequently changing, the grey prediction model cannot

usually predict future trends precisely. Therefore, the Markov chain can be introduced to facilitate a grey prediction model to better predict the concentration of pollutants in water bodies. In a grey prediction model, the residual term is a representation of the model error. With the Markov Chain being utilized to classify the residuals from the GM (1,1) model, the trends of the change in water quality can be predicted more accurately [10, 11]. In this article, the grey prediction model is used to predict the content of Ammonia nitrogen (a pollutant) in Dongting Lake, and the Markov chain is utilized to improve the precision.

2. Data and Method

The data collected in this research comes from weekly reports on automatic water quality monitoring in China. The three positions of the river detected are Yiyang South mouth, Yueyang antlers, and Yueyang Tower. The concentration of NH3-N in the last report each year is recorded in Table 1. The average value is the average concentration of three positions, representing the estimation of the average concentration of NH3-N in Dongting Lake. The following model is based on the average values of the concentration of NH3-N from 2008 to 2017, and the value in 2018 is used in estimation. The basic idea of the GM (1,1) model is to transform the original data series into a gray differential equation model by establishing a gray differential equation, and then use the model to make predictions. The main feature of the model is to decompose the original data series into an accumulated generation sequence and an anti-cumulated generation series. By analyzing the two series, the state equation and output equation of the system are obtained, to realize the prediction.

Table 1. The concentration of NH_3-N in each position of Dongting Lake mg/L

Area\Year	2012	2013	2014	2015	2016	2017	2018
Yiyang South mouth	0.13	0.15	0.18	0.19	0.21	0.18	0.22
Yueyang antlers	0.25	0.19	0.21	0.28	0.22	0.22	0.39
Yueyang Tower	0.29	0.24	0.25	0.29	0.19	0.35	0.1
Average	0.22	0.19	0.21	0.25	0.21	0.25	0.24

3. Results and Discussion

3.1. GM(1, 1)

GM (1,1) Model can be used when there is a series $X^{(0)} = \{x^{(0)}(1),\ x^{(0)}(2),\ x^{(0)}(3), \dots,\ x^{(0)}(N)\}$. In this case, the $X^{(0)}$ series would be:

$$X^{(0)} = \{0.22,\ 0.19, 0.21, 0.25, 0.21, 0.25\} \tag{1}$$

Then, the accumulated generation sequence has the formula:

$$X^{(1)}(k) = \sum_{m=1}^{k} x^{(0)}(m)(k = 1,2, \dots, n) \tag{2}$$

In this case, the accumulated generation sequence would be:

$$X^{(1)}(k) = \{0.22,\ 0.41, 0.62, 0.87, 1.08, 1.33\} \tag{3}$$

$$X^{(1)} = \{x^{(1)}(1), x^{(1)}(2), x^{(1)}(3), \dots, x^{(1)}(N)\} \tag{4}$$

$$X^{(1)}(k) = \sum_{m=1}^{k} x^{(0)}(m)(k = 1,2, \dots, n) \tag{5}$$

Adjacent mean generation Sequence can be formulagted:

$$Z^{(1)} = \{Z^{(1)}(2), Z^{(1)}(3), \dots, x^{(1)}(N)\} \tag{6}$$

$$Z^{(1)} = \frac{1}{2}\left(X^{(1)}(k-1) + X^{(1)}(k)\right) \tag{7}$$

Then, the matrices B and Y can be constructed:

$$\mathbf{B} = \begin{bmatrix} -\frac{1}{2}\left(\boldsymbol{X^{(1)}(1) + X^{(1)}(2)}\right) & \mathbf{1} \\ -\frac{1}{2}\left(\boldsymbol{X^{(1)}(2) + X^{(1)}(3)}\right) & \mathbf{1} \\ \cdots & \cdots \\ -\frac{1}{2}\left(\boldsymbol{X^{(1)}(n-1) + X^{(1)}(n)}\right) & \mathbf{1} \end{bmatrix} \tag{8}$$

$$\mathbf{Y} = \begin{bmatrix} X^{(0)}(2) \\ X^{(0)}(3) \\ \cdots \cdots \\ X^{(0)}(n) \end{bmatrix} \tag{9}$$

Therefore, the differential equation can be formulated:

$$\frac{\mathrm{dX}^{(1)}}{\mathrm{dt}} + \mathrm{a}X^{(1)} = \mu \tag{10}$$

$$\hat{X}^{(1)}(k+1) = \left[X^{(0)}(1) - \frac{\mu}{a}\right]e^{-ak} + \frac{\mu}{a} \tag{11}$$

$$a = (B^T B)^{-1} B^T Y \tag{12}$$

The formula is solved by and prediction results and statistics are given in Table 2 and Table 3.

$$\mu = 0.190,\ a = -0.0408, \hat{X}^{(1)}(k+1) = [4.877]e^{0.0408k} - 4.657 \tag{13}$$

From the table, it is clear that the prediction with grey prediction is not accurate enough. That is why Markov Chain is introduced to improve the precision of the model.

Table 2. Prediction results.

years	true value	predict value	residuals	relative residuals (%)
2012	0.22	0.22	0	0
2013	0.19	0.20	0.01	-6.7
2014	0.14	0.21	-0.05	0.50
2015	0.25	0.22	0.05	12
2016	0.21	0.23	-0.01	-9.1
2017	0.25	0.24	0.01	4.5
2018	0.24	0.25	-0.01	-3.6

Table 3. Prediction statistics.

Grade	Average relative error	Ratio of MSE	Residual probability
I	<0.01	<0.35	>0.95
II	<0.05	<0.50	<0.80
III	<0.10	<0.65	<0.70
IV	<0.20	>0.65	<0.60

Table 4. State descriptions.

	State 1	State 2	State 3
State 1	0	2	0
State 2	2	0	1
State 3	1	0	0

3.2. Markov Chain and Fuzzy Classification

A Markov Chain is a stochastic process that describes random behavior in a state space through a series of state transitions. Markov chains have the characteristic of "memoryless", that is, given the current state, the future state depends only on the current state and is not affected by the past state. This property makes Markov chains ideal for modeling situations in which the future state is influenced by the existing state and is constantly evolving. Fuzzy classification is a kind of classification method based on fuzzy logic, which is used to deal with information uncertainty and fuzziness. Unlike traditional binary classification (yes/no), fuzzy classification allows objects to belong to different categories to varying degrees, rather than just being divided into two discrete categories. In fuzzy classification, each object is assigned to one or more categories, and its membership to each category is given. This degree of membership represents the degree of ambiguity between an object and each class, that is, the extent to which an object has the characteristics of a certain class. To improve the precision of the GM (1,1) model, the relative residuals are modeled as the states in the Markov matrix. State one is when the relative residual is negative and very large: relative residual < -0.05. State 2 is when the relative residual is relatively small: -0.05 < relative residual < 0.05. State 3 is when the relative residual is relatively small: 0.05 < relative residual. The summary is given in Table. 4. Then, the statistics of state step transfer frequency:

- Markov matrix:

$$P = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{bmatrix} \tag{14}$$

- One-step markov matrix is constructed:

$$P_{(1)} = \begin{bmatrix} 0 & 1 & 0 \\ 0.67 & 0 & 0.33 \\ 1 & 0 & 0 \end{bmatrix} \tag{15}$$

- Two-step markov matrix is constructed:

$$P_{(2)} = \begin{bmatrix} 0.67 & 1 & 0.33 \\ 0.33 & 0.67 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{16}$$

- Three-step markov matrix is constructed:

$$P_{(3)} = \begin{bmatrix} 0.33 & 0.67 & 0 \\ 0.44 & 0.33 & 0.22 \\ 0.67 & 0 & 0.33 \end{bmatrix} \tag{17}$$

Therefore, using this Markov chain to amend for gry prediction model, the results are given in Table. 5.

Table 4. Prediction results.

years	true value	predicted value by GM (1, 1)	predicted value by GM (1,1) improved by Markov chain
2012	0.22	0.22	0.22
2013	0.19	0.20	0.21
2014	0.14	0.21	0.20
2015	0.25	0.22	0.21
2016	0.21	0.23	0.23
2017	0.25	0.24	0.23
2018	0.24	0.25	0.25

4. Limitations and Prospects

Although the gray-Markov chain model has shown promising results in improving the prediction accuracy of ammonia nitrogen (NH3-N) content in Dongting Lake, there are still some limitations and future research directions worthy of attention. The accuracy of any predictive model depends heavily on the quality and quantity of available data. Inaccurate or incomplete data may affect the reliability of forecasts. Future research could explore strategies to improve data collection and validation processes. Applying the gray-Markov chain model to different water bodies and pollutants may yield different results due to different dynamics and influencing factors. When extending the model to other scenarios, careful validation and adjustment is required.

For further study, omproving forecast accuracy may involve taking into account external factors such as weather conditions, industrial activity, and land use changes. Incorporating these variables into the model allows for a more complete understanding of pollution dynamics. Integrating policy and management factors: Combining predictive models with environmental policies and management strategies can help develop targeted interventions to mitigate water pollution and improve overall water quality.

5. Conclusion

In this paper, the prediction of ammonia nitrogen content in Dongting Lake is studied based on the gray-Markov chain model. By combining the grey prediction model with the Markov chain, this study shows a new method to improve the accuracy of water pollution prediction. By comparing the observed and predicted values, the conclusions are drawn. First, the grey prediction model shows certain effects in the prediction of ammonia nitrogen content in Dongting Lake. However, in the face of a complex water pollution situation, there may be some deviations in the forecast results, which is not enough to meet the needs of practical applications. Therefore, Markov chains are introduced to improve the precision. Secondly, by applying the Markov chain to the residuals of the GM(1,1) model, one successfully increases the accuracy of the prediction of ammonia nitrogen content. Markov chains can use historical state change information to model residual terms and incorporate more accurate trends into forecasts, thus making up for the shortcomings of gray prediction models. In the experiment, this study compares the prediction results of the classical grey prediction model with those of the Markov chain. The results show that the Markov chain model has higher accuracy and reliability in predicting the ammonia nitrogen content in Dongting Lake.

However, there are still some limitations in this study, such as the integrity of the data and the consideration of external factors. Future studies can further explore how to integrate more influencing factors to improve the accuracy of the prediction model. Additionally, the method can be applied to the prediction of other water bodies and different pollutants to verify its applicability and universality. In summary, the gray-Markov chain model provides an effective method for water pollution prediction, which has high prediction accuracy and practical application potential. In future research and practice, it is confident that through further improvement and optimization, this method will be applied to a wider range of environmental protection fields and contribute to water quality management and sustainable development.

References

[1] Wang Y, Li M, Li H. A novel hybrid prediction model for water quality based on the improved grey wolf optimizer and long short-term memory neural network. Environmental Science and Pollution Research, 2021, 28(15): 19263-19276.

[2] Zhang L, Zhou Y, Wang S, Wang H. Short-term water quality prediction using a hybrid model based on grey wolf optimizer and least squares support vector machine. Science of the Total Environment, 2021, 770: 145260.

[3] Hu X, Zhang X, Shi P. Water quality prediction based on a hybrid model of singular spectrum analysis and least squares support vector machine. Water, 2019, 11(7): 1452.

[4] Liu W, Shang H, Xue L, Li Y. Prediction of water quality parameters using long short-term memory network and deep belief network. Water, 2019, 11(12): 2551.

[5] Li S, Lu W, Li P. A hybrid model for water quality prediction combining a variational mode decomposition with optimized support vector regression. Environmental Science and Pollution Research, 2018, 25(15): 14675-14691.

[6] Chen Y, Lu C, Li J, Wu D. Combining grey model GM (1, 1) and back propagation neural network for predicting the solar energy storage. Energy, 2020, 196: 117058.

[7] Zhang G, Sun W, Chen W. Prediction of soil organic matter content in arid regions based on the Grey-ARIMA model. CATENA, 2021, 196: 104898.

[8] Zheng C, Yang W. Estimation of earthquake casualties using a grey forecasting model: The case of the Wenchuan Earthquake. International Journal of Disaster Risk Reduction, 2020, 46: 101494.

[9] Li Z, Zhou J, Lai K K. Prediction of tourism demand using a new grey model considering the characteristics of Big Data. Tourism Management, 2018, 64: 278-289.

[10] Jia Z Q, Zhou Z F, Zhang H J, Li B, Zhang Y X. Forecast of coal consumption in Gansu Province based on Grey-Markov chain model. Energy, 2020, 199: 117444.

[11] Hu Y, Peng J, Lee P. Forecasting tourism demand by incorporating neural networks into Grey–Markov models, Journal of the Operational Research Society, 2019, 70: 12-20.