Water pollution index evaluation of lake based on principal component analysis

Dalin Liu, Faming Huang, Wenjing Kang, Yaoxian Du, Zhongshan Cao *
School of Civil Engineering and Architecture, Nanchang University, Nanchang, China

*Corresponding author e-mail: cuczs@sina.com

Abstract. Poyang Lake is the largest freshwater lake in China, with a wide area and abundant species resources. It is a serious issue to protect and monitor the water quality of Poyang lake. This paper proposes to use principal component analysis (PCA) to evaluate the water pollution index of Poyang Lake. The input variables of PCA are the weekly monitoring water pollution factors including dissolved oxygen (DO), chemical oxygen demand (CODMn) and NH4+-N. The water quality monitoring station is in Hukou County of China from 2004 to 2014. Finally, a series of new water pollution indexes are generated by PCA to reflect the change characteristics of lake water pollution. The results can provide support for the comprehensive evaluation of lake water quality. Meanwhile, the results also discuss the variation in water pollution, which is practical and innovative.

1. Introduction
Lake is the most important source of fresh water on Earth, which is closely related to the economy and ecology[1]. The emission of various pollutants and uneven rainfall distribution have sharpened the contradiction between freshwater supplement and requirement[2, 3]. According to the results of the monitoring and evaluation of 118 important lakes in 2016 years, the proportion of lake water quality in the I-III class is 23.7%, in the IV-V class is 58.5%, and in inferior V class is 17.8%. It can be seen that lake environmental protection and pollution control become a very important content of China. Thus, it is imperative to strengthen lake water pollution assessment, water pollution process excavation and management [4, 5].

Mathematics model is applied to describe the lake water pollution processes. These models can be utilized for lake water pollution assessment, prediction, analysis of influencing factors, sewage control and so on[6, 7]. These models can be classified as a single water quality index, coupled water quality index and aquatic ecological model according to the influence factors. According to the time characteristics of water pollution, these models can be classified as stable state and unstable state models which do not change with time. According to the dimension of water quality evaluation, the models can be classified as zero-dimensional, one-dimensional, two-dimensional and three-dimensional models.

Generally speaking, since the mid-1920s, scholars around the world have begun to study the mathematical model of pollutant diffusion and transport in rivers. After the exploration in decades, the research of the water quality evaluation model has made remarkable progress, and a large number of reliable and comprehensive water quality models have appeared successively, especially the one-dimensional model developed considerably. Traditionally, the acknowledged water pollution
assessment models mainly include the Nemerow index, comprehensive index method, multiple regression method, fuzzy mathematics theory, gray evaluation method, analytic hierarchy process, entropy method, and other models [8-13]. In recent years, with the development of mathematical model and machine-learning, some new water quality evaluation models have appeared, including BP neural network, fuzzy neural network, harmony degree equation, Markov model, matter-element extension, random forests algorithm, Bayesian theory, cloud model and extreme learning machine [14-20].

As an important tool in pollution control and decision analysis, the lake water pollution evaluation model can provide a scientific basis for comprehensive improvement and management. In this study, a water pollution index is calculated by the principal component analysis. Making full use of a new index to study protection and utilization of natural resources.

2. Study area and data sources preparation

2.1. Introduction of Poyang Lake Basin

Poyang Lake Basin is mostly located in the Jiangxi Province. It covers an area of 162,200 km², accounting for 96.9% of the total drainage area. The Poyang Lake Basin is positioned on the south bank of the middle and lower reaches of the Yangtze River. It spans 24°29' to 30°05' north latitude and 113°34' to 118°29' east longitude. The Poyang Lake Basin is surrounded by mountains on three sides, and the area consisting of five sub-basins and the Poyang Lake area is 160,000 km². Most of the basin area has a subtropical monsoon climate, with abundant water resources. The Poyang Lake Basin has abundant rainfall and annual precipitation. The average annual runoff is 156.5 billion m³. However, there are problems when rainfall and interannual variation is large, the distribution is uneven during the year, which increases the difficulty of water resource utilization. The overall water quality in the basin is good. In 2008, the length of rivers that reached or exceeded Class III water quality in the basin accounted for 87.9% of the total evaluation length, of which Class I water bodies accounted for 6.5%, Class II waters accounted for 64.2%, and Class III waters accounted for 17.2%. However, with the development of the social economy in Jiangxi Province, the water quality of the Poyang Lake Basin has deteriorated. Therefore, It is of great significance for water resources management and pollution control in Poyang Lake Basin [21-22].

![Figure 1. Poyang Lake Basin](image-url)
2.2. Basic data
The water quality data of the Zhouyang Lake in the lake outlet of Poyang Lake from 2004 to 2014 is obtained from China Environmental Monitoring Station. The study selected three factors of dissolved oxygen (DO), permanganate index (CODMn), and ammonia nitrogen (NH4+-N) to analyze the time variation of lake water quality. The DO content is vital to the growth of aquatic vegetation. To some extent, it can also represent the overall status of water quality. Relevant studies have shown that NH4+-N concentration is a key factor to determine the regression of aquatic vegetation. Additionally, CODMn can directly reflect the extent of organic pollution caused by lake outflow.

![Figure 2. DO, CODMn and NH4+-N standard data in 2004~2014.](image)

3. Water Pollution of Evaluation
3.1 Principal Component Analysis
Principal Component Analysis (PCA) is mainly used to study how to illustrate the structural characteristics of variance-covariance by a few components. The application of PCA in water quality assessment has two main aspects. For one thing, a comprehensive evaluation system is established to evaluate and rank the relative pollution degree of each monitoring section. For another, it is to judge the role of each individual factor in the evaluation system, which determines the main ingredients and ignores the secondary factors. The detailed steps are as follows.

1. Standardized collection of p-dimensional random vectors from raw index data: 
   \[ x = (x_1, x_2, \ldots, x_p)^T \] set with a total of \( n \) samples: 
   \[ x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T, \quad i = 1, 2, \ldots, n \] \( n > p \),
   Sample array is constructed and the normalized transformation of sample array elements is carried out as follows:

   \[ z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, p \]  
   (1)
\[ x_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \]

Among them, \( \bar{x}_j \), then get the normalized matrix Z.

2. Calculate the correlation coefficient matrix for standardized matrix Z.

\[ R = \left[ r_{ij} \right]_{p \times p} = \frac{z^T z}{n-1}. \tag{2} \]

Among them, \( r_{ij} = \frac{\sum z_{ij} \cdot z_{kj}}{n-1}, i, j = 1, 2, \ldots, p \).

3. Solve the characteristic equation of Sample Correlation Matrix R, \( |R - \lambda I_p| = 0 \), and get p characteristic roots and determine principal components. According to \( \sum_{j=1}^{p} \lambda_j \geq 0.85 \), determine \( m \) to make the utilization rate of information more than 85\%. For each \( \lambda_j, j = 1, 2, \ldots, m \), unit eigenvector \( b_j^o \) is obtained by solving system of equations \( Rb = \lambda_j b \).

4. Convert standardized index variables into main components.

\[ U_j = z^T b_j^o, \quad j = 1, 2, \ldots, m \tag{3} \]

\( U_1, U_2 \) are respectively the first and second principal components, \( U_p \) is principal component P.

5. Synthetically evaluate of \( m \) principal components: The final evaluation value is obtained by weighted summation of \( m \) principal components. The weight is the variance contribution rate of each principal component.

4. Evaluation of Water Pollution Index of Poyang Lake

1. Raw data normalization. The water quality data of DO, CODMn and NH4+-N in weekly units in the section of Poyang Lake outlet from 2004 to 2014 are standardized, as shown in Fig.3.

\[ \text{Figure 3. DO, CODMn and NH4+-N standard data in 2004~2014.} \]
(2) Determine the principal components. As can be seen from Table 1, when the number of principal components is 2, the contribution rate of cumulative variance of a cross-section is 80.38% and more than 80%. It shows that these two principal components reflect 80.38% information provided by original variables, so the number of principal components of a cross-section is 2.

| component | Initial eigenvalue | Extracting sum of squares |
|-----------|------------------|--------------------------|
|           | Total variance%  | Cumulative%              | Total variance%  | Cumulative%              |
| 1         | 1.38             | 46.05                    | 1.38             | 46.05                    |
| 2         | 1.03             | 34.33                    | 1.03             | 34.33                    |
| 3         | 0.59             | 19.62                    | 0.59             | 19.62                    |

(3) Determine the comprehensive evaluation formula. According to the cross-sectional load in Table 2, the first principal component is DO and the second principal component is CODMn. It is indicated that the main pollutants in the lake section water bodies are dissolved oxygen and chemical oxygen demand.

The eigenvectors can be obtained by dividing the data of two columns of the factor load matrix by the square root of their corresponding eigenvalues. The principal component expression can be obtained by multiplying the obtained eigenvectors with the normalized data.

Table 2. Initial Factor Load Matrix

| factor     | component | 1    | 2    |
|------------|-----------|------|------|
| DO         |           | 0.84 | 0.09 |
| CODMn      |           | 0.20 | 0.95 |
| NH4+-N     |           | 0.79 | -0.34|

Table 3. Principal component expression

\[ F_1 = 0.84Z_1 - 0.18Z_2 - 0.80Z_3 \]
\[ F_2 = 0.10Z_1 + 0.96Z_2 - 0.33Z_3 \]

Comprehensive evaluation formula: \[ F = 0.57F_1 + 0.43F_2 \]

(4) Obtain the evaluation index of water quality. According to the comprehensive evaluation formula, the principal component scores of the water quality at the outlet of Poyang Lake from 2004 to 2014 can be obtained as follows: Fig. 4.
5. Discussion

It is necessary to use the water pollution evaluation model to assess the quality of water bodies. The evaluation model can provide effective technical support and promote the modernization of management technology.

(1) Control water pollution from the source and strictly control sewage discharge. According to the above data analysis, it is recommended to adjust the water pollution control of Poyang Lake from the end treatment to the source control, which can fundamentally cut off the water quality decline of Poyang Lake. Poyang Lake brings together 97% of the runoff in Jiangxi Province. Only by controlling the source of pollution can we control the river with the source and protect the lake with the river. Focus on high-pollution industries such as paper mills, mining plants. In addition, the regulation of domestic sewage discharge should be strengthened. To ensure that the effluent water quality meets national emission standards.

(2) Control the opposite source pollution from the point of view of the large environment. The ecological forest project in Poyang Lake basin should be built to improve forest cover, so as to prevent the surface source pollution caused by soil erosion. Simultaneously, the amount of fertilizer and pesticide is strictly controlled. According to the local characteristics of the land and the needs of crops for nutrients, reasonable arrangement of fertilization, fertilization and fertilization time. In the scope of protected areas, rural sewage and waste pollution control should be carried out according to local conditions, and sewage treatment facilities should be built in small towns.

(3) The dissolved oxygen in the section of the south branch is extremely low. It can artificially aerate oxygen to the river, increase the oxygen content of the water body, promote the decomposition of pollutants by microorganisms. Construct information security water quality monitoring. The rapid development of technologies such as GIS and RS has provided support for the early warning and forecast of sudden water pollution accidents. The Poyang Lake District and the “Five Rivers” should improve the water quality information control system as soon as possible to achieve rapid and accurate positioning of sudden water pollution accidents.

6. Conclusion

This research studies the time-varying characteristics of water quality indices (DO, NH4 +, N and CODMn) in Poyang Lake. The results show that the water quality pollution indexes are calculated objectively and accurately by principal component analysis, which reflects the water quality of the outlet of Poyang Lake. The research results are of great practical and innovative significance for
comprehensive evaluation of lake water quality and predictive early warning analysis of water quality pollution.

Acknowledgments
This work was financially supported by National Natural Science Foundation of China (NO.41807285).

References
[1] WANG Mingcui, LIU Xueqin, et al. Evaluate method and classification standard on lake eutrophication[J]. Environmental Monitoring In China, 2002. 18(5): 47-49.
[2] Faming Huang, Kunlong Yin, Guirong Zhang, et al. Landslide displacement prediction using discrete wavelet transform and extreme learning machine based on chaos theory[J]. Environmental Earth Sciences, 2016, 20(75): 1-18.
[3] LUO Ming-ke, TIAN Xueda, YU Hui, et al. Vertical pollution characteristics of sediments from lakes in middle reaches of Yangtze[J]. Journal of Agro-Environment Science, 2016. 5(35): 955-962.
[4] LI Wen-hao, , ZHANG Meng, MEN Jishuai, et al. Community Structure of Macrozoobenthos in Xiannv Lake Basin and Assessment of its Water[J]. Resources and Environment in the Yangtze Basin, 2016(08): 1218-1227.
[5] Long-wei, WU Guo zheng, LI Changyou. Applying PAC-SOM Combined Model to Assess Lake Water Quality [J]. Environmental Science & Technology, 2009(3): 173-177.
[6] WAN Jin Bao, LI Yuanyuan. Progress in the Research of Lake Water Quality Model[J]. Resources and Environment in the Yangtze Basin, 2007(06): 805-809.
[7] ZHANG Yufeng, TAN Yufeng, Wenhui Qi, et al. Study on Lake Water Quality Assessment Based on Simple Modeling and Factor Analysis[J]. Journal of Ecology and Rural Environment, 2015(03): 432-439.
[8] Yang Kun, Luo Yi Xu Yufei. Dynamic monitoring and simulation of Cyanobacteria Bloom based on wireless sensor network and GIS [J]. Transactions of the Chinese Society of Agricultural Engineering, 2016(24): 197-205.
[9] Faming Huang, Jinsong Huang*, Shuihua Jiang, Chuangbing Zhou. Landslide displacement prediction based on multivariable chaotic model and extreme learning machine[J]. Engineering Geology, 2017:173–186.
[10] Faming Huang, Peng Wu*, Y.Y. Ziggah. GPS monitoring landslide deformation signal processing using time-series model[J]. International Journal of Single Processing, Image Processing and Pattern Recognition. 2016,9(03):321-332.
[11] Huang F, Yao C, Liu W*, Li Y, Liu X. Landslide susceptibility assessment in the nantian area of china: A comparison of frequency ratio model and support vector machine. Geomatics, Natural Hazards and Risk, 2018, 9(1): 919-938.
[12] SHAN Liping, TIANWei WEI Qiang. Application of Fuzzy Comprehensive Evaluation Method to Water Quality Evaluation of Chaganhu Lake[J]. Wetland Science, 2015(01): 124-128.
[13] MA Chun, MENG Weiqing JU Meiting. Application of grey relevancy model based on ameliorated entropy for comprehensive evaluation of wetland environment quality[J]. Journal of Safety and Environment, 2008(06): 80-83.
[14] Kaixiang, Wang. Taihu Lake Eutrophication Index Analysis and Prediction Based on Multiple Regression and Neural Network[J]. Journal of Nanjing Normal University(Engineering and Technology Edition), 2017(02): 70-74.
[15] Faming Huang, Kunlong Yin, Jinsong Huang, et al. Landslide susceptibility mapping based on self-organizing-map network and extreme learning machine[J]. Engineering Geology, 2017, 223: 11-22.
[16] Faming Huang, Jinsong Huang, Shui-Hua Jiang, et al. Prediction of groundwater levels using evidence of chaos and support vector machine[J]. Journal of Hydroinformatics, 2017, 19(4): 586-606.
[17] Faming Huang, Lixia Chen, Kunlong Yin, et al. Object-oriented change detection and damage assessment using high-resolution remote sensing images, Tangjiao Landslide, Three Gorges Reservoir, China[J]. Environmental Earth Sciences, 2018, 77(5): 183.

[18] Faming Huang, Xiaoyan Luo, Weiping Liu. Stability Analysis of Hydrodynamic Pressure Landslides with Different Permeability Coefficients Affected by Reservoir Water Level Fluctuations and Rainstorms[J]. Water, 2017, 9(7): 450.

[19] Yuanyao Li, Jinsong Huang, Shui-Hua Jiang, et al. A web-based GPS system for displacement monitoring and failure mechanism analysis of reservoir landslide[J]. Scientific reports, 2017, 7(1): 17171.

[20] YU Hui, SUN Baosheng, LI Ya-nan. Water quality prediction of Haihe River using grey-fuzzy-markov chain model[J]. China Environmental Science, 2014(03): 810-816.

[21] Faming Huang, Kunlong Yin*, Tao He, et al. Analysis of Influence Factors and Displacement Prediction of Reservoir Landslide-A Case Study of Three Gorges Reservoir, China[J]. Technical Gazette, 2016, 23(02):617-626.

[22] Weiping Liu, Yanxiao Luo*, Faming Huang, Mingfu Fu. Uncertainty of Soil–Water Characteristic Curve and its Effects on Slope Seepage and Stability Analysis under the Condition of Rainfall using Markov Chain Monte Carlo Method[J]. Water-SUI. 2017, 9(10), 758.