Analysis on influential factors of water quality in key sections of Yangtze River basin in China

Dong Wang¹, Youdong Wen², a

Jiangxi Agricultural University, Nanchang Business School Economics Department, Jiangxi Nanchang 330013, China
School of Statistics, Jiangxi University of Finance and economics Jiangxi Nanchang 330013, China

ajxufewyd@163.com

Abstract. The Yangtze River Valley is the largest of China's seven major water systems, the longest water system, and the length of 6300 kilometers, and ranked third in the world. Its water quality is not only closely related to the residents' drinking water safety, but also to the ecological environment of China has an important impact. Based on the panel data and stochastic forest modeling of the key sectional water quality monitoring data of the Yangtze River basin since 2018, NH₃-N (ammonia nitrogen content) has the most important effect on water quality, and its value directly determines the quality of the water quality.

1. Introduction

The water network system composed of all rivers and lakes in the river basin is called drainage China is a wide geographical area, the climate and topography between different regions, the main flow of rivers in the Pacific Ocean and the Indian Ocean, a small amount of flow is the Arctic Ocean. China's seven major water systems: The Pearl River, the Yangtze River, the Yellow River, the Huai River, Liao River, Hai River and Songhua River. The Yangtze River is the longest, full length of 6300 kilometers, ranked third in the world. The main tributaries are Yalong, Minjiang, Jialing River, Wujiang, Yuanjiang, Hanjiang and Ganjiang, and so on, their average flow is above 1000 cubic meters/sec (all over the Yellow River water), of which, the river basin area to the Jialingjiang as the largest, to 160,000 square kilometers; the length of the Hanjiang, the longest, 1577 km; To 87.7 billion cubic meters .Most of the Yangtze River valley in the subtropical monsoon climate zone, warm and humid, more than 1100 mm average annual precipitation, more than average annual water intake of nearly 1 trillion cubic meters, accounting for China's River runoff volume of about 36%, equal to 20 Yellow River.

The quality of water is not only closely related to the residents' drinking water safety, but also to the ecological environment of China has an important impact. R X Chen et al. [1] According to the Yangtze River basin water quality monitoring data, the representative main factor PH, DO, permanganate index, NH3-N, using the comprehensive Water Quality identification index model of the Yangtze River basin for quantitative comprehensive evaluation, the evaluation results show that The main pollution factors affecting the water quality in the Yangtze River basin are NH3-N and permanganate index, and the quantitative comprehensive evaluation shows that the overall water
quality of the Yangtze River basin is type water, but it has the tendency of deteriorating to type water. P Xiong et al. [2] In order to study the water quality situation and development trend of the Changjiang river system, the projection Pursuit (PPC) model of water quality comprehensive evaluation was established according to the data of 8 water quality monitoring indexes (515 samples) and groundwater quality evaluation standard of 103 national control sections in 2006-2010. The results show that the overall water quality is better in the state-controlled section of the Changjiang River system, and the water quality of category I and II is 60% and 22.7% respectively. The upstream water quality is better than the downstream water quality; the water pollution of tributaries and rivers in the more developed areas of chemical industry is serious, and in 2006-2009 the water quality was improved year in 2010, but the water quality was the worst. Z Y mu et al. [3] mainly studied the evaluation and prediction of the Yangtze River water quality, a multilevel fuzzy comprehensive evaluation method based on analytic hierarchy process is established to quantitatively evaluate the water quality of each region. Through the one-dimensional model of river in the calculation of pollutant capacity in water area, the section calculation of the Yangtze River is carried out to estimate the self-purification ability of water body, According to the difference between the predicted theoretical value and the actual value, the quantity increase of pollutant in the corresponding REACH is calculated, and the pollution source and location of the main pollutant in the Yangtze River are analysed. J J Y [4] based on the analysis of the basic advantages of environmental protection and economic development in the Yangtze River basin, the paper puts forward the necessity and countermeasures of sustainable development in order to provide reference for the construction and ecological environment protection of the Yangtze River Economic Belt. Z H Wang et al. [5] It is believed that some areas in the Yangtze River basin, especially in the vast western mountainous area, still have corresponding difficulties, the rural water protection and water quality detection work is weak, water supply project construction supervision system is not in place, follow-up operation and maintenance mechanism is not perfect, resulting in rural drinking water security

In summary, although there are more scholars on the Yangtze River Basin water quality and ecological environmental protection research, but the use of comparative cutting-edge statistical methods for analysis of the study is still less. Based on the data of continuous monitoring week published by China Bureau of Statistics, this paper analyses the factors affecting the water quality in the Yangtze River basin with stochastic forest algorithm in machine learning method, hoping to give effective suggestions.

2. Monitoring data and method selection

2.1. Monitoring data

Research data from the environmental quality of the Ministry of Ecological Environmental Protection published by the National main watershed water quality automatic monitoring weekly, the data obtained by the week of 2018 2nd, 5th, 8th, 9th, 11th, 12th, such as 6 weeks of data. The selected watershed is the Yangtze River basin, comprising a total of 21 points and a total of 126 monitoring data. The main indexes of evaluation were PH (cutting) value, DO value (dissolved oxygen content), CODMn (chemical oxygen demand) value and NH3-N (ammonia nitrogen content), the classification label was mainly water quality grade. The level of water quality is based on the environmental function and protection objectives of surface water, divided into five categories according to the function level: I-V. According to the five types of water functions of surface water, the standard values of surface water environmental quality standard are divided into five categories, and the corresponding standard values of different functional categories are carried out respectively.
Table 1. 2018 the 2nd week of the Yangtze River Basin water quality monitoring of key broken sections

| river | time  | level | PH   | DO(mg/L) | CODMn(mg/L) | NH3-N(mg/L) |
|-------|-------|-------|------|----------|-------------|-------------|
| 1     | 201812 | 2     | 7.89 | 9.16     | 1.9         | 0.24        |
| 2     | 201812 | 4     | 6.98 | 6.23     | 3.6         | 1.05        |
| 3     | 201812 | 1     | 7.71 | 9.92     | 1.2         | 0.05        |
| 4     | 201812 | 2     | 8.36 | 8.9      | 2.6         | 0.13        |
| 5     | 201812 | 2     | 7.74 | 9.88     | 1.1         | 0.22        |
| 6     | 201812 | 2     | 7.76 | 9.93     | 2.2         | 0.11        |
| 7     | 201812 | 2     | 7.92 | 8.5      | 2.2         | 0.17        |
| 8     | 201812 | 2     | 7.81 | 9.16     | 2.7         | 0.18        |
| 9     | 201812 | 2     | 7.83 | 9.31     | 2.1         | 0.2         |
| 10    | 201812 | 2     | 8.04 | 8.81     | 2.4         | 0.18        |
| 11    | 201812 | 2     | 7.56 | 6.92     | 1.6         | 0.35        |
| 12    | 201812 | 2     | 7.25 | 11.9     | 3.3         | 0.21        |
| 13    | 201812 | 1     | 8.11 | 8.58     | 1.8         | 0.15        |
| 14    | 201812 | 1     | 8.07 | 11.5     | 1.6         | 0.05        |
| 15    | 201812 | 2     | 7.36 | 10.7     | 3.8         | 0.19        |
| 16    | 201812 | 2     | 7.99 | 11.7     | 2.5         | 0.16        |
| 17    | 201812 | 1     | 8.42 | 11.8     | 1.8         | 0.03        |
| 18    | 201812 | 2     | 7.24 | 7.51     | 2.3         | 0.19        |
| 19    | 201812 | 2     | 8.4  | 10.7     | 2.6         | 0.42        |
| 20    | 201812 | 2     | 7.79 | 8.37     | 3.9         | 0.43        |
| 21    | 201812 | 2     | 7.59 | 10.2     | 2.5         | 0.37        |

Through the analysis of 126 data, we know that the monitoring results of the first level to level four have 19, 94, 11 and 2, the second level most, applicable to the centralized drinking water surface water source of a protected area, the habitat of rare aquatic organisms, fish and shrimp spawning grounds, juvenile larvae, such as bait.

2.2. Method Selection

Methods the panel data regression and stochastic forest algorithm were used in econometrics.

Panel data [6], which is also called "parallel data", refers to taking multiple sections on a time series and selecting sample data for sample observations on these sections. Or it is an m*n data matrix, which records a data metric of M objects on N time nodes. Panel data analysis method is a new statistical method developed in recent decades, the panel data can overcome the time series analysis by multiple collinearity, can provide more information, more changes, less collinearity, more freedom and higher estimation efficiency, The unit root test and cointegration analysis of panel data is one of the most advanced fields.

Random forest [7] refers to a classifier that trains and predicts samples using multiple trees, which were first proposed by Leo Breiman and Adele Cutler and registered as trademarks. The stochastic forest algorithm combines the classification tree into random forest, that is, the use of variables (columns) and the use of data (rows) are randomized, many classification trees are generated, and the results of the classification trees are summarized. The prediction accuracy of stochastic forest is improved without significant increase of computational volume. Stochastic forest is insensitive to multivariate collinearity, and the result is robust to missing data and unbalanced data, which can predict the function of up to thousands of explanatory variables well. One important advantage of random forests is that there is no need to cross-check it or use an independent test set to obtain an unbiased estimate of the error. It can be evaluated internally, that is, an unbiased estimate of the error can be established during the generation process. When building each tree, we used different bootstrap
samples (random and put-back) for the training set. So for each tree (assuming for the K tree), about 1/3 of the training instances are not involved in the generation of the K tree, they are called OOB samples of the K tree.

3. Empirical analysis

3.1. Panel return

Using R language statistic software to carry on the regression analysis of panel data, first need to carry on the Hausman test, the test result's p value is 0.366, cannot reject the model to be the random effect the original hypothesis, and therefore should establish the stochastic effect model.

\[
\begin{array}{cccc}
\text{Estimate} & \text{Std. Error} & \text{t-value} & \text{Pr(>|t|)} \\
(Intercept) & 1.0982569 & 0.2542483 & 4.3196 & 3.222e-05 \\
\text{pH} & 0.0011546 & 0.0412601 & 0.0260 & 0.9777 \\
\text{DO} & -0.0179983 & 0.0216894 & -0.8298 & 0.4083 \\
\text{CODMn} & 0.2663825 & 0.0437839 & 6.0840 & 1.416e-08 \\
\text{NH3-N} & 1.8649203 & 0.1964238 & 9.4944 & 2.579e-16 \\
\end{array}
\]

**Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1**

Total Sum of Squares: 16.822  
Residual Sum of Squares: 7.6695  
R-Squared: 0.54409  
Adj. R-Squared: 0.52902  
F-statistic: 36.1008 on 4 and 121 DF, p-value: < 2.22e-16

**Figure 1.** Panel data regression results Chart (ph means PH, do means DO, codmn means CODMn, nh means NH3-N)

From figure1, the factors that have a significant effect on water quality are codmn (COD) and NH3-N (ammonia-nitrogen content).

3.2. Stochastic forest algorithm modeling

Using the Random forest program of R language to analyze the data, we first need to do 10 cross-validation, examine the classification effect of the model, get the average misjudgment rate of 0, and show that the modeling effect is good.

From figure2, the two most important variables affecting water quality are PH (cotting) and NH3-N (ammonia-nitrogen content).

Combined with the modelling results of Figure1 and Figure2, it is known that the most important factor affecting water quality is the NH3-N (ammonia nitrogen content), which has a significant effect on the dependent variable or the classified label.
Figure 2. Importance extraction Diagram of variable of stochastic forest algorithm

4. Conclusion
Based on the panel data and stochastic forest modeling of the water quality monitoring data of the Yangtze River basin since 2018, NH3-N (ammonia nitrogen content) has the most important effect on water quality, and its value directly determines the quality of the water quality. Ammonia nitrogen has a harmful effect on aquatic organisms mainly free ammonia, its toxicity is dozens of times larger than ammonium salt, and increases with the increase of alkaline. Ammonia nitrogen toxicity and water temperature have a close relationship between the PH value and water temperature, generally, the higher the toxicity, the risk of the fish is similar to nitrite. Ammonia-nitrogen has acute and chronic damage to aquatic organisms. The harm of chronic ammonia-nitrogen poisoning is: lower food intake, slower growth, tissue damage, lower oxygen delivery between tissues. Fish are more sensitive to ammonia nitrogen in water and can cause fish death when the content of ammonia is high. The hazards of acute ammonia-nitrogen poisoning are: aquatic life performance is high, loss of balance in the water, convulsions, severe even death. Therefore, the relevant State departments should pay attention to the environmental assessment of sewage enterprises, reduce their ammonia nitrogen emissions, the greatest degree of protection of China's ecological environment.
Acknowledgements
This article obtains Science and technology project of Jiangxi Provincial Education Commission, Study on the Combining efficiency of agricultural science and technology with finance in Jiangxi, Grant No. GJJ151514, Humanities and Social Sciences project of Universities in Jiangxi Province, Study on Jiangxi’s Energy Efficiency and Its Spatio-temporal Variation Based on DEA-ESDA, Grant No. JC162015 funding.

References
[1] Chen Yun Sheep, Chianguo. Evaluation and prediction of water quality in Yangtze River basin [J].
[2] Environmental Science and Technology, 2008, 31 (12): 169 - 173.
[3] Bear, Lou Wengao. Comprehensive evaluation model of Yangtze River basin water quality based on projection pursuit classification and its application model [J]. Journal of water Resources and water Engineering, 2014, 25 (06): 156 - 162.
[4] Miu Ziyang, Zhu Shennan, Hu Jianhui. Evaluation and prediction model of Yangtze River water quality [J]. Technology and innovation, 2017 (16): 75 - 77.
[5] Yang Jingjing. Ecological environment Protection and sustainable development of Yangtze River basin economy [J]. Green technology, 2017 (23): 21 - 22.
[6] Zhenhua, Liqingyun, Chen. Current situation, problems and countermeasures of rural drinking water safety in the Yangtze River Basin [J]. Research on Water Conservancy development, 2017, 17 (12): 13 - 16.
[7] Okui R, Yanagi T. Panel Data Analysis with Heterogeneous Dynamics [J]. Social Science Electronic Publishing, 2018.
[8] Lee S, Moon N. Location recognition system using random forest [J]. Journal of Ambient Intelligence & Humanized Computing, 2018 (11): 1 - 6.