# Study on pollution of different tributaries of Bai Ta Pu river based on principal component analysis

Shanshan Ma [1,2], Shicong Geng [2], Guofeng Wang [1], Shuo Yang [1], and Yu Gao[1]

[1]Shenyang Institute of Engineering, Shenyang 110136, China
[2]Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China

**Abstract:** On the basis of understanding the indexes of different tributaries of Bai Ta Pu river and according to the contribution rate of each principal component, the comprehensive score of different tributaries was calculated through principal component analysis. Each tributary was ranked according to the score, and the evaluation results were analyzed. It is pointed out that the tributary with the highest comprehensive index of Bai Ta Pu river is in Shi Jia Zhai bridge and the tributary with the lowest comprehensive index is in De Sheng Tun bridge.

## 1 Principal component analysis

### 1.1 Principle of principal component analysis

Puckett et al. evaluated the water quality of some rivers in Virginia for the first time with principal component analysis in 1992, which was the first application of principal component analysis in water quality evaluation [1-4]. Principal component analysis is a method of dimensionality reduction by recombining the original indexes with certain correlation into a group of new unrelated comprehensive indexes.

### 1.2 The mathematical model of principal component analysis

$$F_1 = a_{11}ZX_1 + a_{21}ZX_2 \ldots\ldots + a_{p1}ZX_p$$
$$F_2 = a_{12}ZX_1 + a_{22}ZX_2 \ldots\ldots + a_{p2}ZX_p$$
$$F_m = a_{1m}ZX_1 + a_{2m}ZX_2 + \ldots\ldots + a_{pm}ZX_p$$

$a_{1i}$, $a_{2i}$, and $a_{pi}$ are the eigenvectors of X's Covariance matrix Σ corresponding to his eigenvalue, $ZX_1$, $ZX_2$, and $ZX_p$ are the normalized value of the original variables. In practice, there are always different dimensions of indexes, so the influence of dimensions should be eliminated before calculation and the original data should be standardized (data standardization in this paper refers to Z standardization)。

$A=(a_{ij})_{p\times m}$, $R_{ai}=\lambda_i \times a_i$, R is the correlation coefficient matrix, $\lambda_i$ and $a_i$ are eigenvalue and unit eigenvector, respectively ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$).

### 1.3 The basic steps of principal component analysis

The main steps are as follows [5-7]. We selected indicators and data according to research questions firstly, and Standardized indicator data and converted to dimensionless data using statistic package for social science software secondly. Then we determined correlation between indicators thirdly, determined the number of principal components fourthly, expressed principal component Fi fifthly, name the principal component Fi sixthly. Finally we evaluated principal component and comprehensive principal component [6-8].

## 2 Results and discussion

### 2.1 indicator selection

11 indicators of Bai Ta Pu river were selected in this paper. X1, X2, X3, X4, X5, X6, X7, X8, X9, X10 and X11 were water temperature, pH, chemical oxygen demand, ammonia nitrogen, nitrate nitrogen, nitrite nitrogen, total nitrogen, total phosphorus, total suspended solids, dissolved oxygen and acute toxicity.

### 2.2 Specific operation steps

Factor process from SPSS statistical analysis software was used to conduct principal component analysis of Bai Ta Pu river's indicators. The specific operation steps were as follows. Firstly, analysis- dimension reduction-factor analysis-popup factor analysis dialog box. Secondly, we selected $X_1$~ $X_{11}$ into the variable box. Thirdly, description- correlation matrix- factor analysis- clicked "continue"- returned to factor analysis. Fourthly, we clicked "OK".

The factor analysis process was invoked by SPSS and the raw data was automatically standardized. Therefore, the variables after the calculation result were all the variables had been standardized. SPSS didn't give standardized data directly. If standardized data was needed, descriptive statistical analysis procedure was called for calculation. We could do this through the "analysis- descriptive statistics- description" dialog box. After popping descriptive statistical analysis dialog box, selected X1~X11 into the variable box, saved the normalized value as a variable, clicked "OK". Standardized data was automatically filled in the data window and named that started with Z.

The principle of extracting the number of principal components was the first m principal components whose eigenvalue was bigger than 1. To some extent, the eigenvalue can be regarded as an indicator of the influence strength of principal components. If the eigenvalue was smaller than 1, it indicated that the explanatory power of the principal component was not as strong as the average explanatory power of directly introducing an original variable. Table 1 was the extraction and analysis table of variance decomposition principal components. It can be seen from Table 2 that 4 principal components were extracted (i.e. m = 4).

Table 2 was the loading matrix of initial factor. CODcr, NH3-N, T-N, and T-P contained higher load on the first principal component. It indicated that the first principal component basically reflected the information of these indicators. T, pH, SS, and DO contained higher load on the second principal component. It indicated that the second principal component basically reflected the information of these four indicators. $NO_3$-N and $NO_2$-N contained higher load on the third principal component.

Table 1 The extraction and analysis table of variance decomposition principal components

| | **Initial Eigenvalues** | | | **Quadratic Sum Extraction** | | |
|---|---|---|---|---|---|---|
| | **Total** | **Var** | **Accum-ulation** | **Total** | **Var** | **Accum-ulation** |
| 1 | 3.609 | 32.807 | 32.807 | 3.609 | 32.807 | 32.807 |
| 2 | 2.430 | 22.095 | 54.901 | 2.430 | 22.095 | 54.901 |
| 3 | 1.693 | 15.390 | 70.291 | 1.693 | 15.390 | 70.291 |
| 4 | 1.028 | 9.344 | 79.635 | 1.028 | 9.344 | 79.635 |
| 5 | 0.963 | 8.758 | 88.394 | | | |
| 6 | 0.475 | 4.316 | 92.709 | | | |
| 7 | 0.383 | 3.478 | 96.188 | | | |
| 8 | 0.241 | 2.187 | 98.375 | | | |
| 9 | 0.112 | 1.017 | 99.391 | | | |
| 10 | 0.052 | 0.472 | 99.863 | | | |
| 11 | 0.015 | 0.137 | 100.000 | | | |

Table 2 The loading matrix of initial factor

| | Components | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| 1 | 0.26 | 0.71 | -0.40 | -0.31 |
| 2 | 0.31 | 0.85 | -0.13 | 0.26 |
| 3 | 0.39 | -0.28 | 0.24 | -0.17 |
| 4 | 0.87 | -0.34 | -0.20 | -0.14 |
| 5 | -0.09 | 0.06 | 0.81 | -0.02 |
| 6 | 0.44 | 0.01 | 0.67 | 0.43 |
| 7 | 0.89 | -0.32 | -0.14 | -0.11 |
| 8 | 0.91 | -0.18 | 0.11 | -0.22 |
| 9 | 0.42 | 0.54 | 0.44 | -0.39 |
| 10 | 0.55 | 0.65 | -0.07 | 0.38 |
| 11 | 0.47 | -0.40 | -0.29 | 0.52 |

It indicated that the third principal component basically reflected the information of these two indicators. Acute toxicity inhibition contained higher load on the fourth principal component. It indicated that the fourth principal component basically reflected the information of this indicator. The extracted four principal components can basically reflect the information of all indicators. So these 4 new variables were decided to replace the 11 original ones. The expression of these four new variables cannot be directly obtained from the output window, because each load quantity of the initial factor load matrix represented the correlation coefficient between the principal component and the corresponding variable.

Divided the data by the eigenvalues corresponding to the principal components and took the square root to get the coefficients corresponding to each index in the four principal components (namely the eigenvectors). The four columns of data in the initial factor load matrix were variables B1, B2, B3, and B4. A1 = B1/SQR (characteristic value of 3.609)(Note: the second principal component SQR was followed by the second eigenvalue 2.430, while the third and fourth principal components were 1.693 and 1.028, respectively). We got the eigenvector A1. And similarly we got the eigenvectors A2, A3, and A4. Multiplied the obtained eigenvector and the normalized data, and then the principal component expression can be obtained.

$$F_1=0.14ZX_1 + 0.16ZX_2 + 0.21ZX_3 + 0.46ZX_4 - 0.05ZX_5 + 0.23ZX_6 + 0.47ZX_7 + 0.48ZX_8 + 0.22ZX_9 + 0.29ZX_{10} + 0.24ZX_{11}$$

$$F_2=0.46ZX_1 + 0.55ZX_2 - 0.18ZX_3 - 0.22ZX_4 + 0.04ZX_5 + 0.01ZX_6 - 0.20ZX_7 - 0.11ZX_8 + 0.35ZX_9 + 0.41ZX_{10} - 0.26ZX_{11}$$

$$F_3= - 0.31ZX_1 - 0.10ZX_2 + 0.19ZX_3 - 0.16ZX_4 + 0.62ZX_5 + 0.52ZX_6 - 0.10ZX_7 + 0.09ZX_8 + 0.34ZX_9 - 0.05ZX_{10} - 0.22ZX_{11}$$

$$F_4= - 0.31ZX_1 + 0.26ZX_2 - 0.17ZX_3 - 0.14ZX_4 - 0.02ZX_5 + 0.43ZX_6 - 0.10ZX_7 - 0.21ZX_8 - 0.39ZX_9 + 0.38ZX_{10} + 0.52ZX_{11}$$

Table 3 Comprehensive principal component value and ranking

| Fields | F1 | F2 | F3 | F4 | F |
|---|---|---|---|---|---|
| Shi Jia Zhai bridge | 2 | 2 | 6 | 8 | 1 |
| South canal | 5 | 3 | 1 | 13 | 2 |
| Li Xiang Xin Cheng bridge | 4 | 10 | 2 | 1 | 3 |
| Bang Shi Tai bridge | 1 | 13 | 12 | 12 | 4 |
| Gao Ba Zhai bridge | 7 | 1 | 13 | 2 | 5 |
| North canal | 3 | 8 | 9 | 4 | 6 |
| Ying Cheng Zi bridge | 6 | 5 | 5 | 3 | 7 |
| Science and engineering bridge | 10 | 6 | 3 | 11 | 8 |
| Wang Shi Lan bridge | 8 | 12 | 4 | 6 | 9 |
| Source (Reservoir) | 11 | 4 | 10 | 10 | 10 |
| Grass pond ditch bridge | 13 | 7 | 7 | 5 | 11 |
| Estuary | 9 | 11 | 8 | 9 | 12 |
| De Sheng Tun bridge | 12 | 9 | 11 | 7 | 13 |

Note: F1 is the first principal component, F2 is the second principal component, F3 is the third principal component, F4 is the fourth principal component, F is the comprehensive principal component.

The proportion of the eigenvalue corresponding to each principal component to the sum of the total eigenvalues of the extracted principal component was taken as the weight to calculate the principal component synthesis model.

$$F= (\lambda_1\times F_1+\lambda_2\times F_2+\lambda_3\times F_3+\lambda_4\times F_4)/(\lambda_1+\lambda_2+\lambda_3+\lambda_4)$$

$\lambda_1=3.609$, $\lambda_2=2.430$, $\lambda_3=1.693$, $\lambda_4=1.028$. $\lambda_1+\lambda_2+\lambda_3+\lambda_4=8.760$

The principal component synthesis model can be obtained.

$$F= (3.609\times F_1 + 2.430\times F_2 + 1.693\times F_3 + 1.028\times F_4)/8.760$$

According to the principal component synthesis model, the comprehensive principal component values can be calculated and sorted according to the comprehensive principal component values. The comprehensive evaluation and comparison can be made for each tributary. The results were shown in Table 3.

This study was to comprehensively analyze various indicators of Bai Ta Pu river, such as ammonia nitrogen, nitrate nitrogen, nitrite nitrogen, total phosphorus, chemical oxygen demand, total suspended solids, acute toxicity and so on[9-12]. Principal component analysis was used to determine the dimensionality reduction of each index of the river to a few indexes, and to comprehensively evaluate the advantages and disadvantages of sampling points of different tributaries, so as to provide a basis for understanding the characteristics of water quality distribution of Bai Ta Pu river[13-15].

It can be seen that the tributary with the highest comprehensive index was Shi Jia Zhai bridge, followed by the south canal, and the sampling point with the lowest comprehensive index is De Sheng Tun bridge. It indicated that the pollution sources near Shi Jia Zhai bridge had a serious impact on the sub-watershed water system. Bai Ta Pu river entered the middle reaches from Shi Jia Zhai village. The place was the transition zone from low hills to Hun River flood plain. The terrain was relatively flat, with an average elevation of about 50 meters. The water pollution composite index of Shi Jia Zhai section rose dramatically, which was divided into rural section and urban section, and the water pollution in urban areas was more serious than in rural areas. Sources of pollution were living and non-point sources and sewage disposal from sewage treatment plants[16]. Hu Jiguo also found that water quality of the Bai Ta Pu river was bad mainly due to tributaries pollution. The tributary of Shi Jia Zhai was seriously polluted, which should be reorganized in order to reduce the pollution flowing into the main stream [16].

## References

1. H. Zhang, C. Hao, M. Hao, et al. The application of principal component analysis in river quality evaluation. *Journal of Hebei Institute of Architectural Engineering*. **33**, 3 (2015)
2. Z. Ji, L.Fang, J. Zhang, et al. Operation of principal component analysis in SPSS software and application in river quality evaluation. *Environmental Protection and Technology*. **18**,4 (2012)
3. W. Yin, X. Xin, J. Liang, et al. Evaluation of water quality in tributaries of Dan Jiang Kou reservoir based on principal component analysis. *Water Resources and Power*. **33**,1 (2015)
4. T. Wang, S. Xu, C. Han. Application of improved principal component analysis method in quality evaluation of Nan Fei river. *Water Resources and Power*. **30**,10(2012)
5. C. Qian, W. Mu, K.Wang, et al. Fuzzy comprehensive evaluation of groundwater quality based on principal component analysis. *Water Resources and Power*. **34**,11( 2016)
6. S. Li. Study on evaluation method of surface water quality: a case study of Liu Xi river. *South China University of Technology.*(2013)
7. P. Zhao, J. Liu, S. Hu. Application of principal component analysis based on SPSS software in water quality evaluation. *Science and Technology Entrepreneurship Monthly*. **10**(2016)
8. J. Liu. Application of principal component analysis in water quality evaluation of Sha River. *Hai He Water Resources*. **4(**2019)
9. H. Fang, S. Sun, Y. Zhu. Application and analysis of principal component analysis in water quality evaluation. *Environmental Science and Management.***34**, 12(2009)
10. H. Li, H. Xing. Fuzzy comprehensive evaluation of Liu Yang River quality based on principal component analysis. *Hunan Agricultural Science*. **8**(2019)
11. X. Zhang. Research status and trend of water quality evaluation methods. *Business Management and Technolog.***5** (2017)
12. J. Xing, Y. Zhang, Y. Chen. Evaluation of water quality in national controlled section of Yellow River basin based on principal component analysis. *Water Saving Irrigation*. **10**(2013)
13. G. Cai, J. Zhang, T. Liu. Water quality evaluation of a reservoir in south China based on principal component analysis. *Environmental Science and Technology*. **41,**(S2)2018
14. X. Wang. Principal component analysis method for comprehensive evaluation of water quality. *Mathematical Statistics and Management*. **4** (2001)
15. H. Bai. Application of principal component analysis in SPSS - a case study of herb community in the forest of Wen Yu river bank. *Scientific and Technological Information Development and Economy*.**19**(9)2009
16. J. Hu. Study on environmental quality analysis and assessment of Bai Ta Pu river. Northeastern University, 2015