Water quality assessment of the river network in Wenzhou city using PCA-BP neural network model

F Zhou¹, C C Tian¹², J L Xu¹ and J Wei¹
¹Zhejiang Design Institute of Water Conservancy and Hydroelectric Power, Hangzhou 310002, China
²E-mail: tianchuanchong@163.com

Abstract. The water quality assessment is often challenged by how to determine the index weight and simplify the evaluation model. This paper proposed an integrated model called PCA-BP neural network model to solve the problem, which combines the dimensionality reduction capability of the improved principal component analysis (PCA) method with the self-learning ability of back propagation (BP) artificial neural networks. And its application for the plain river network in the main districts of Wenzhou city indicated that the evaluation result of PCA-BP was consistent with the single-factor method and PCA method in overall trend. It was demonstrated that PCA-BP model could evaluate the water quality of the study area more reasonably and accurately, as it avoided the disadvantage of a certain factor completely covering the information of the other ones in the single-factor method and the risk of over-optimism in PCA method as well.

Keywords: Principal component analysis; Artificial neural network; Single-factor method; Water quality; River network; Wenzhou; China

1. Introduction
With the rapid development of economy, the dependence on water resources has been higher than any time before, and China is faced with the situation that the shortage of water due to poor quality is just as serious as deficient quantity. Therefore, how to assess the water quality synthetically and objectively is an important issue which needs to be solved urgently, and it is crucial to protect water resources and guide the water pollution control. A few methods have been proposed and adopted for water quality evaluation in China, but that cannot meet the requirement more or less. The traditional single-factor method is so susceptible to the extreme factors that it might diverge from the reality sometimes [1], although it is easy to operate. Fuzzy mathematics method, grey theory, and analytic hierarchy process (AHP) were used to ignore the interaction between factors when determining the weights and highly subjective [2-4]. Traditional principal component analysis (PCA) standardize dataset by standard deviation all the time making the variance of the same kinds of index approach to zero, which is capable of eliminating the differences between factors and tends to result in over-optimistic result [5]. Besides, the traditional artificial neural network method regards all monitoring indicators as input data that complicates the model and reduces the efficiency [6]. The water environment of plain river network in Wenzhou City is the focus of many scholars, and there are many research results about it, Mei, K., Zhu, Y. et al optimized water quality monitoring networks of Wen-rui tang river by using continuous longitudinal monitoring data, Wang, J., Fu, L., & Xu, H. assessed river network water quality of Wenzhou based on sluice gate control strategy[7-8]. In order to deal with how to weight the index...
suitably, prevent the approaches from being complicated and arrive at a reasonable result, this paper proposed an integrated approach combining the PCA method with BP method, whose application for water quality assessment of the main districts of Wenzhou city, Zhejiang Province, revealed that this method was more simply and effectively compared to the above ones.

2. Materials and methods

2.1. Materials

The main urban area of Wenzhou is located in the southeast of Zhejiang Province, China with jurisdiction over Lucheng District and Ouhai District. It lies in the lower reaches of the Oujiang river that belongs to the Wenruitang river plain water system. Its precipitation is significantly affected by monsoons and typhoons with an uneven distribution throughout the year, which generally concentrates from May to September accounting for about 60% -70% of the whole year [9]. In addition, there is an actually dense river net in the study area composed of dozens of rivers, mainly including Fengmen, Guoxi, Xianmen, Qinfen, Xiongmen, Wenruitang, Puzhou, and Shitan River, etc, but with an insufficient water mobility. Therefore, the study area is faced with rather serious problems of water resources management and water environment deterioration [10-11]. There are 13 water quality monitoring sites in the region totally, that are Panqiao, Xianmen, Qinfen, Huiqiao, Dongshuichang, Guangming, Wutian, Baixiang, Yutian, Guoxi, Xiaodan, Jiushan and Yangfushan, and water quality data are monitored every month. Figure 1 shows the study area with water system, as well as the water quality monitoring sites’ distribution.

![Figure 1. Water system and water quality monitoring sites’ distribution of study area.](image)

2.2. Methods

The following figure shows the main calculation process involved in research process (Figure 2).

![Figure 2. The main calculation process involved in research process.](image)
2.2.1 Improved PCA. The principal component analysis (PCA) method works like this: It commits to find out a set of linearly independent principal components that can represent the primary information of the original data, by using orthogonal transformation to reduce the dimensionality of the original datasets [12-13]. The specific steps are as follows:

1. Divide the initial data by mean value to standardize the dataset (Formula 1) and get rid of the dimensional interference while keeping the differences between the same kinds of variables [14].

\[ Z_{ij} = \frac{X_{ij} \text{Mean}(X_i)}{\text{Mean}(X_i)} \]  

(Formula 1)

Where \( Z_{ij} \) is the standardized result of the j'th value of the i'th index; \( X_{ij} \) is the initial value of the j'th data of the i'th index; \( \text{Mean}(X_i) \) is the mean of the i'th sample.

On the contrary to the most of the negative indicators of the water quality, we need to standardize the positive indicators by Formula 2, which means the larger those are, the better the water quality is.

\[ Z_{ij} = \frac{(\text{Max}(X_i) - X_{ij})}{\text{Mean}(\text{Max}(X_i) - X_{ij})} \]  

(Formula 2)

Where \( \text{Max}(X_i) \) is the maximum value of the i'th index.

2. Calculate the correlation coefficient matrix \( R \) of standardized \( \{Z_{ij}\} \).

3. Compute the eigenvalues \( \lambda_i \) and the eigenvectors of \( R \), and arrange the eigenvectors from large to small.

4. Determine the number of principal components through calculating the cumulative contribution rate of variance, and take the proportion of the first n eigenvalues' sum in total values as the cumulative variance contribution \( K \), which is more than or equal to 85% usually.

5. Calculate the composite score \( F \) of the principal components based on figuring out the \( F_i \) value of each one by Formula 3. And the larger the score, the worse the water quality.

\[ F = \sum_{1 \leq i \leq n} \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n} F_i \]  

(Formula 3)

2.2.2 BP artificial neural network. BP neural network is a multi-layer feedforward neural network, which can form a nonlinear dynamic system relying on the complex connections of a large number of neurons [15]. It is generally composed of input layer, hidden layer and output layer. While the network is running, the input data propagate from front to back between layers, and the output of each layer can only affect the next one. Besides, the neurons in the same layer neither connect directly nor have any interference. Figure 3 illustrated the Structure of BP neural network. On the contrary, the error between neurons is bound to propagate from back to front, and it is not until the error meets the setup precision that the network will stop adjusting the weights and threshold according to the error between the actual output and the expected one.

Here is how the BP neural network is set up:

1. Standardize the input dataset \( \{X_i\}(1 < i < n) \) and target dataset \( \{Y_j\}(1 < j < m) \) and eliminate the dimensional interference.

2. Determine the number of neurons in hidden-layer \( N \). As a rule of thumb, \( N = p + \sqrt{n + m} \), where \( p \) ranges from 1 to 10.

3. Establish the neural network and set the transfer function of each layer, maximum step size and the model precision.

4. Save the model for water quality evaluation when it meets the stated accuracy.
3. Results and discussion

3.1. Evaluation result by PCA

There are 11 indicators chosen to evaluate the water evaluation, including PH (X1), DO (X2), NH\textsubscript{3}-N (X3), COD\textsubscript{Mn} (X4), COD (X5), BOD\textsubscript{5} (X6), TP (X7), TN (X8), F\textsuperscript{-} (X9), VP (volatile phenol) (X10) and S\textsuperscript{2-} (X11). The specific operation is as follows:

1. The data were first processed with equalization and the corresponding correlation coefficient matrixes were calculated, shown as Table 1. The results show that most of the correlation coefficients between the indicators is above 0.3, which indicates there must exist information overlapping between indicators, so it is urgently necessary to reduce the dimensionality of the initial datasets through PCA.

![Figure 3. Structure of BP neural network.](image)

**Table 1.** Correlation coefficient matrix.

|     | X1 | X2   | X3   | X4   | X5   | X6   | X7   | X8   | X9  | X10 | X11 |
|-----|----|------|------|------|------|------|------|------|-----|-----|-----|
| X1  | 1.000 | 0.413 | -0.077 | -0.130 | -0.411 | 0.175 | -0.102 | 0.030 | 0.362 | -0.394 | 0.516 |
| X2  | 0.413 | 1.000 | -0.796 | -0.545 | -0.618 | -0.446 | -0.837 | -0.270 | -0.203 | -0.462 | 0.060 |
| X3  | -0.077 | -0.796 | 1.000 | 0.602 | 0.329 | 0.744 | 0.963 | 0.438 | 0.581 | 0.218 | 0.305 |
| X4  | -0.130 | -0.545 | 0.602 | 1.000 | 0.731 | 0.636 | 0.599 | 0.881 | 0.060 | 0.629 | -0.145 |
| X5  | -0.411 | -0.618 | 0.329 | 0.731 | 1.000 | 0.271 | 0.338 | 0.553 | -0.360 | 0.906 | -0.587 |
| X6  | 0.175 | -0.446 | 0.744 | 0.636 | 0.271 | 1.000 | 0.753 | 0.750 | 0.624 | 0.125 | 0.227 |
| X7  | -0.102 | -0.837 | 0.963 | 0.599 | 0.338 | 0.753 | 1.000 | 0.431 | 0.543 | 0.210 | 0.306 |
| X8  | 0.030 | -0.270 | 0.438 | 0.881 | 0.553 | 0.750 | 0.431 | 1.000 | 0.227 | 0.409 | -0.151 |
| X9  | 0.362 | -0.203 | 0.581 | 0.060 | -0.360 | 0.624 | 0.543 | 0.227 | 1.000 | -0.574 | 0.654 |
| X10 | -0.394 | -0.462 | 0.218 | 0.629 | 0.906 | 0.125 | 0.210 | 0.409 | -0.574 | 1.000 | -0.676 |
| X11 | 0.516 | 0.060 | 0.305 | -0.145 | -0.587 | 0.227 | 0.306 | -0.151 | 0.654 | -0.676 | 1.000 |

2. Through calculating the eigenvalues and contribution rate of the principal components, we got the corresponding explained variance, eigenvalues and cumulative contribution rate of variance of each principal component (Table 2).

**Table 2.** Eigenvalue and cumulative contribution rate of variance.

| Number | Eigenvalue | Proportion of the initial eigenvalue’s variance (%) | Cumulative contribution rate (%) |
|--------|------------|----------------------------------------------------|---------------------------------|
| 1      | 5.039      | 45.805                                             | 45.805                          |
| 2      | 3.398      | 30.890                                             | 76.695                          |
| 3      | 1.322      | 12.022                                             | 88.717                          |
| 4      | 0.553      | 5.030                                              | 93.747                          |
| 5      | 0.320      | 2.913                                              | 96.661                          |
| 6      | 0.171      | 1.555                                              | 98.215                          |
The principal component analysis of the 11 evaluation factors indicates that the top 3 principal components \((F_1, F_2 \text{ and } F_3)\) have already accumulated the 88.717\% of the original data information, of which the variance contribution rates reach 45.805\%, 30.809\% and 12.022\%, respectively. It can be seen from the load matrix of principal component shown as Table 3 that DO, NH\(_3\)-N, COD\(_{\text{Mn}}\), COD, BOD\(_5\), TP and TN have a greater influence on \(F_1\) and \(F_2\), VP and S\(_2\) on \(F_2\), and PH, DO on \(F_3\).

Table 3. Load matrix of principal components.

|     | \(X_1\) | \(F_1\) | \(F_2\) | \(F_3\) |
|-----|---------|---------|---------|---------|
| \(X_1\) | -0.227  | 0.573   | 0.572   |
| \(X_2\) | -0.814  | 0.034   | 0.521   |
| \(X_3\) | 0.846   | -0.146  | -0.271  |
| \(X_4\) | 0.883   | 0.401   | 0.326   |
| \(X_5\) | 0.721   | -0.628  | 0.063   |
| \(X_6\) | 0.776   | 0.446   | 0.277   |
| \(X_7\) | 0.852   | 0.388   | -0.300  |
| \(X_8\) | 0.749   | -0.011  | 0.600   |
| \(X_9\) | 0.271   | 0.895   | -0.026  |
| \(X_{10}\) | 0.581   | -0.748  | 0.073   |
| \(X_{11}\) | -0.087  | 0.885   | -0.082  |

According to Table 3, the corresponding scores of the first three principal components could be calculated, and then the composite score was figured our referring to Formula 4. Eventually, we succeeded in working out the result of principal component analysis of each monitoring section (Table 4).

\[
F = 0.123ZX_1 - 0.119ZX_2 + 0.238ZX_3 + 0.214ZX_4 + 0.055ZX_5 + 0.295ZX_6 + 0.234ZX_7 + 0.241ZX_8 + 0.228ZX_9 + 0.001ZX_{10} + 0.138ZX_{11} \quad \text{(Formula 4)}
\]

Table 4. Result of PCA.

| Monitoring section | \(F_1\) | \(F_2\) | \(F_3\) | \(F\) | Water quality | Evaluation single-factor method |
|--------------------|---------|---------|---------|-------|---------------|-------------------------------|
| Xiaodan            | -3.039  | -1.144  | -0.738  | -2.067| I             | IV                            |
| Yangfushan         | -1.914  | -1.308  | 0.054   | -1.436| I             | V                            |
| Xianmen            | 0.056   | 1.602   | -0.475  | 0.522 | V+            | V+                           |
| Xinqiao            | 0.556   | 1.006   | -1.896  | 0.379 | V+            | V+                           |
| Dongshuichang      | 0.672   | -1.645  | -0.404  | -0.281| IV            | V+                           |
| Baixiang           | 0.858   | 3.157   | 1.084   | 1.689 | V+            | V+                           |
| Guoxi              | -3.515  | 0.434   | 0.248   | -1.629| I             | IV                           |
| Guangming          | 1.195   | -1.415  | -1.069  | -0.021| V+            | V+                           |
| Wutian             | 2.375   | 3.092   | -0.036  | 2.297 | V+            | V+                           |
| Qinfen             | 1.802   | -2.174  | -0.358  | 0.124 | V+            | V+                           |
The water quality was classified by $F_i$ value according to China environmental quality standards for surface water (GB3838-2002), and the upper limit values of class I, II, III, IV and V are -0.617, -0.511, -0.357, -0.222, and -0.139, respectively.

As shown in Table 4, PCA promoted the water quality of section Xiaodan and Guoxi to class I and section Yanfushan and Panqiao to class I compared to what was class IV and class V in the single factor evaluation, however the water quality of other sections V+ remained the same class. Overall, PCA does synthesize all the evaluation factors, which can help the complete information to avoid being covered with a certain prominent factor entirely, but carry the risk of over-optimistic assessment that was vividly proved in section Panqiao, the water quality of which was deemed to be class I by PCA instead of class V worked out by the single factor evaluation.

The top 3 principal components obtained by reducing the dimensionality of the initial 11 indicators through PCA can reflect the 88.717% information of the original datasets, which greatly reduced the difficulty of information processing. What is worth noting in Formula 4 is that NH$_3$-N, COD$_{Mn}$, BOD$_5$, TP and TN are the main pollutants in the study area, which will be taken as the input data of the neural network model next.

3.2. Evaluation result by BP neural network

3.2.1. Generate the sample data. The sample size and differences are what especially affect the simulation precision of neural network. So as to get plenty of samples, it is suggested to take the minimum detectable concentration of each index in Chinese national GB standard method as the lower limit of class I. Meanwhile, the random interpolation was carried out between the upper and lower limits of water quality classification at all levels that were stipulated in Environmental Quality Standard of Surface Water (GB3838-2002). It was the water quality of most of the waters in the study area can't reach class V that call for an extra classification in addition to the normal five. Thus, we totally set up six water quality classification (from I to V+). Besides, the random interpolation between each two classifications generates 500 samples, with a total of 3000.

3.2.2. Pre-process the sample data. In this paper, the function "tansig" was selected as the activation function in the hidden layer and the linear function "purelin" in the output layer. Through the function "mapminmax" of MATLAB, each element of sample matrix ($P$) was normalized to [-1, 1]. The sample set that was treated as the input sample must be a matrix of $5 \times 3000$, in which each column represents a sample, with a total 3000 samples.

3.2.3. Determine the target matrix. There were six classifications of the water quality evaluation, so 6 neurons were picked up for the output layer. The $6 \times 1$ matrix $(1,0,0,0,0,0)^T$, $(0,1,0,0,0,0)^T$, $(0,0,1,0,0,0)^T$, $(0,0,0,1,0,0)^T$, $(0,0,0,0,1,0)^T$, $(0,0,0,0,0,1)^T$ presented the water quality of class I, II, III, IV, V and V+ respectively. Each input sample corresponded to an output matrix, so the target set (T) was a $6 \times 900$ matrix.

3.2.4. Establish the neural network. After inputting the sample set [P, T] to MATLAB, 70% of that was used to train the network, 15% to verify and the rest 15% to test. The number of neurons in the hidden layer was supposed to range from 4 to 15 according to the experimental equation. Thus, we started at 4 and increased the number of neurons for training one at a time until it reached 15, which finally worked out that it should be 8 on the basis of considering the relationships between the number of neurons and the mean square error as well as the training step. It was clear that the model ran a total of 98 epochs,
and the best validation performance occurred at epoch 92, as illustrated in Figure 4. Additionally, the structure of neural network adopted in the study was the three-layer network 5-8-6.

![Best Validation Performance is 2.054e-05 at epoch 92](image)

**Figure 4.** Best validation performance of neural network model when N=5.

3.2.5. **Evaluate the water quality.** After normalizing of the data of the 13 monitoring sections, the output of the neural network could be obtained by instructing " result=net (x) ", and Table 5 shows the result of water quality evaluation.

**Table 5.** Water quality evaluation result of neural network method.

| Monitoring section | Evaluation of neural network method | Evaluation of PCA method | Evaluation of single-factor method |
|--------------------|-------------------------------------|--------------------------|-----------------------------------|
| Xiaodan            | 0.0001 0.9999 0.0000 0.0000 0.0000 | II I IV                  |                                  |
| Yangfushan         | 0.0003 0.0000 0.0010 0.9986 0.0000 | IV I V                   |                                  |
| Xianmen            | 0.0000 0.0000 0.0000 0.0003 0.0008 | V+ V+ V+                 |                                  |
| Xinqiao            | 0.0000 0.0000 0.0000 0.0004 0.0009 | 0.9988 V+ V+ V+          |                                  |
| Dongshuichang      | 0.0000 0.0000 0.0000 0.0011 0.0004 | 0.9985 V+ IV V+          |                                  |
| Baixiang           | 0.0000 0.0000 0.0000 0.0000 0.0001 | 0.9999 V+ V+ V+          |                                  |
| Guoxi              | 0.0149 0.9851 0.0000 0.0000 0.0000 | 0.0000 II I IV           |                                  |
| Guangming          | 0.0000 0.0000 0.0000 0.0003 0.0004 | 0.9994 V+ V+ V+          |                                  |
| Wutian             | 0.0000 0.0000 0.0000 0.0000 0.0001 | 0.9999 V+ V+ V+          |                                  |
| Qinfen             | 0.0000 0.0000 0.0000 0.0000 0.0042 | 0.9958 V+ V+ V+          |                                  |
| Jiushan            | 0.0000 0.0000 0.0023 0.1330 0.3179 | 0.5467 V+ V+ V+          |                                  |
| Huiqiao            | 0.0000 0.0000 0.0000 0.0000 0.0001 | 0.9999 V+ V+ V+          |                                  |
| Panqiao            | 0.0000 0.0000 0.4073 0.5927 0.0000 | 0.0000 IV I V             |                                  |
From Table 5, it was indicated that the evaluation result of PCA-BP neural network method was generally in accordance with that of the single factor method and PCA method in overall trend. Further, comparing the result of PCA-BP method to the single factor method, it became evident that the water quality grades of all sections were exactly the same except for Yangfushan, Xianmen, Xinqiao, Dongshuichang, and Panqiao section of which the results were upgraded by one level and Xiaodan and Guoxi section by two levels. Similarly, comparing the result of PCA-BP method to PCA method, we can see that the water quality grades of all sections were also completely consistent except for Yangfushan and Panqiao section whose results were changed from class I to IV and Xiaodan and Guoxi section from class I to II. We can find out that the PCA-BP neural network method could adjust what was worked out by PCA method in a sense to avoid the over-optimistic evaluation.

4. Conclusion
This study has employed several methods to assess the water quality of the river network in the main districts of Wenzhou mainly reflected by 13 monitoring sections. We first adopted the improved PCA method to pick out the main pollutants that reduced the dimensionality of complex datasets. Then, integrated with the BP artificial neural network, the PCA-BP neural network model was set up, of which the evaluation result was verified by PCA method. It was demonstrated that the evaluation result of PCA-BP neural network method was more objective because it overcame the disadvantage of the single-factor method that a prominent evaluation factor is very possible to cover other factors’ information completely, and got rid of the risk of over-optimism that usually occurs in PCA as well. Therefore, it was concluded that PCA-BP neural network model that is capable of fixing a series of problems, such as determination of the reasonable index weight, the complexity of the evaluation model, and the unreasonable evaluation results, could provide scientific advice for water resources conservation and management in Wenzhou city.

Acknowledgments
This research is supported by the Science and Technology Project of Zhejiang Water Conservancy Department (RB1915).

References
[1] Pan L, Huang X and Wei X 2019 A Comparative Analysis of Three Common Water Quality Evaluation Methods China Rural Water and Hydropower 2019(06) 51-5
[2] Zhu J, Lian X and Ke A 2018 Application of Fuzzy Synthetic Evaluation Method in Nanxi River Water Quality Evaluation Zhejiang Hydrotechnics 46(06) 8-13+28
[3] Sun D 2017 Application of analytic hierarchy process in evaluation of reservoir water environment quality Water Resources & Hydropower of Northeast China 35(02) 33-4+71
[4] Pi J, Ouyang S and Zhang D 2017 Study on Water Quality Evaluation Model for Poyang Lake Based on Grey Theory Water Power 43(06) 5-8
[5] Gao J and Zhao Z 2019 Application of Principal Component Analysis in Water Quality Assessment of MaoWei Sea Guangdong Chemical Industry 46(06) 98-100+76
[6] Wang F and Tang F 2019 Application of Artificial neural networks in Water Quality Assessment of Daqinghe Water Resources & Hydropower of Northeast China 37(06) 25-6+35
[7] Wang J, Fu L and Xu H 2019 Assessment of river network water quality based on sluice gate control strategy in Wenzhou Iop. C. Ser. Earth Env. 227 062023
[8] Mei K, Zhu Y, Liao L, Dahlgren R, Shang X and Zhang M 2011 Optimizing water quality monitoring networks using continuous longitudinal monitoring data: a case study of Wen-Rui Tang River, Wenzhou, China J. Environ. Monitor. 13(10) 2755-62
[9] Tian C C, Wei J and Huang D J 2019 Study on the Scale of Water Diversion of Plain River Network in the Main Urban Area of Wenzhou City Based on the Hydrodynamic and Water Quality United-control Model Water Resour.Power 37(04) 44-7
[10] Huang H, Shang X, and Mei K 2019 River water quality assessment based on Monte Carlo
simulation: A case study of Wen-Rui Tang River. *China Environ. Sci*. 39(05) 2210-8

[11] Zhou F, Chen L and Tian C C 2018 Research of water diversion and joint control of water quantity and quality for pollution dilution in coastal plain. *Yangtze River* 49(14) 61-6

[12] Wang H, Chen Q and Zhang H 2018 Research on the Water Source Bearing Capacity in Ninxia Based on Principle Component Analyses. *China Rural Water and Hydropower* 2018(11) 30-4

[13] He Y, Tang Y and Chen L 2018 Evaluation of Water Quality Based on BP Neural Network and Study of the Environmental Protection Science Temporal and Spatial Evolution Trend of Water Quality. *Environmental Protection Science* 44(03) 114-20+126

[14] Liu Q, Li Y and Pu X 2017 Application Research of the Improved Principal Component Analysis on Reservoir Water Quality Evaluation. *Sichuan Enviroment* 36(06) 116-22

[15] Zhou Z, Liu J and Fan X 2015 Applied Research on LM-BP Neural Networks for Water Quality Assessment: Taking the Weihe River of Xinxiang City for Example. *China Rural Water and Hydropower* 2015(07) 64-7+70