Application of a PCA based water quality classification method in water quality assessment in the Tongjiyan Irrigation Area, China

Xue-feng Tao[1,2 a], Tao Huang[1,b], Xiao-feng Li[1,c], Dao-ping Peng[1,d]

[1]Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu 611756, China

[2]Chongqing Municipal Research Institute of Design, Chongqing 400020, China

[a]53843819@qq.com, [b]taohuang70@126.com, [c]495302996@qq.com, [d]pdp0330@swjtu.cn

Keywords: Water quality assessment; principal component analysis (PCA); water quality classification

Abstract: Applied principal components analysis(PCA) to assess the water quality of Tongjiyan River in 2014, based on monitoring data of 8 indicators, such as COD_{Mn}, NH_3-N, DO, etc.. As PCA could not classify water samples according to their water quality, a PCA based water quality classification method, which was similar to Nemerow approach, was proposed to overcome this problem. Classification results were compared with other methods, such like Fuzzy Evaluation, Nemerow index method and improved Nemerow index method. Result showed that PCA could present an intuitive description of river's pollution patterns in different months. Based on PCA results, we used PCA based water quality classification method to classify water samples so that we could get a deeper understanding of water pollution degree.

Introduction

River plays as one of the most important roles in socio-economic activities, such as drinking water supply, agriculture, aquaculture and industrial activities. However, due to anthropogenic activities many rivers have been contaminated by the pollutant inputs and the water quality was deteriorated.

In order to monitor and assess the river water quality, many water quality assessment methods have been applied in China, such as the Nemerow pollution index, fuzzy comprehensive evaluation, principle component analysis and so on[1]. The Nemerow pollution index is a water pollution index taking extreme values into account using a weighted environmental quantity index and frequently used in water quality assessments around the world. However, this method tends to overemphasize the influence of the maximum evaluation factor (i.e., most serious pollutant factor). Thus, the comprehensive score will be increased in situations where the index value for one evaluation factor is much higher than those of others [2]. Hence, there exists the potential problem that the assessment results may disagree with the overall water quality status. The fuzzy comprehensive evaluation is the process of evaluating an objective utilizing fuzzy set theory, which comprehensively considers the contributions of multiple related indicators according to weights and decreases the fuzziness by using membership functions. The fuzzy set comprehensive evaluation method can improve understanding of the diverse processes and complex phenomena involved in environmental studies, which is why it has been successfully used to assess pollution levels for water quality. This method can give us the assessment result but cannot compare with water samples [3]. Principal component analysis (PCA) is designed to convert the original variables into new, uncorrelated variables (axes), called the principal components. The PCA provides information on the most meaningful parameters, which describes the whole data set interpretation, provides data reduction, and summarizes the statistical correlation among water quality constituents with

minimum loss of the original information. It has been frequently employed for the purpose of evaluating water quality [4]. While the PCA method can figure out the quality of various water samples but could not tell us the pollution extents (water quality classification). However, PCA limitations include ignoring the degree of data dispersion and a weakness in processing nonlinear data. Thus, principle component analysis may not have good accuracy and reliability.

In this study, a modified PCA water quality classification method was carried out based the Nemerow theory. By setting up three classification principles, this method could classify the water sample data by normalization and solved the limitation of PCA's weakness in water quality classification. A case study of Tongjiyan River was carried out and classification results were compared between different methods including the modified PCA method, Nemerow index method and fuzzy comprehensive evaluation.

Materials and methods

Study Area and Sampling. The Tongjiyan Irrigation Area (TIA) was firstly constructed around 25 AD, which was another important large-scale hydraulic project in ancient Sichuan besides the world-famous Dujiangyan Irrigation System [5]. The TIA is situated between a latitude of 29°53'-30°24' N and a longitude of 103°42'-103°51' E in the southwest edge of Chengdu Plain, with an area of 880 km^2. The TIA has been an important tributary of the Minjiang River drainage. The main river in the Tongjiyan Irrigation Area is Tongjiyan River, which flows from the west of TIA to the Minjiang River at Qinglong. To characterize the water quality of the TIA before emptying into the Minjiang River, monthly samplings were carried out at three sampling sites along the mainstream of, from January to December, 2014. Triplicate water samples from1 to 5 m below the water surface of each sampling site were collected using a portable water sampler (LB-8000E, Qingdao Shouhang Instrument Co., Ltd, China). The measurement of water quality was conducted within 24 h after sampling.

Water Quality Classification Method.

Nemerow pollution index. The mathematical formula for the Nemerow comprehensive index calculation is as follows:

$$P = \sqrt{\frac{\left(\frac{1}{n}\sum_{i=1}^{n} P_i\right)^2 + (P_i)^2_{\max}}{2}} \tag{1}$$

where P is the Nemerow comprehensive pollution index, n is the total number of water quality parameters, P_i is the pollution index of parameter i, and $(P_i)_{\max}$ is the maximum pollution index. The following formulas are used to calculate P_i:

$$P_i = X_i / M_0;$$

And for DO,

$$P_{\mathrm{DO}} = \begin{cases} \dfrac{|C_{DOf} - X_i|}{C_{DOf} - M_0} & X_i \geq M_0 \\ 10 - 9X_i / M_0 & X_i < M_0 \end{cases}$$

X_i is the measured value of parameter i, M_0 is the desired water quality standard value (GB3838-2002) of parameter i, and C_{DOf} is the saturated dissolved oxygen concentration.

Principal component analysis. The first step of PCA is to normalize the measured values by the following formula:

$$X_i^{'} = (X_i - \overline{X}_i) / \sqrt{\sigma_i} \quad (2)$$

where, X'_i is the normalized value of parameter i, $\overline{X}_\iota$ is the mean of X_i, σ_i is the variance of parameter i.

Then do KMO test and Bartlett test of sphericity to verify data dependence before PCA. The principal component can be expressed as:

$$\begin{aligned} z_k &= a_{k1}X_1 + a_{k2}X_2 + ... + a_{kn}X_n \\ Z &= (z_1V_1 + z_2V_2 + ... + z_mV_m) / (V_1 + V_2 + ... + V_m) \end{aligned} \quad (3)$$

Where, z is the component score, a is the component loading, X is the measured value of a parameter, k is the component number, n is the total number of parameters, Z is the comprehensive score, V is the total variance of each component, and m is the total number of components.

PCA based water quality classification.

(1) Classification principles

According to the Nemerow theory, a series of water quality classification principles was set up as follows:

I For arbitrary i and j, if $M_{i,j} \geq M_{i,j+1}$, then $M'_{i,j} \geq M'_{i,j+1}$ and vice versa;

II For arbitrary i and j, if $M_{i,j} \geq X_i \geq M_{i,j+1}$,then $M'_{i,j} \geq X_i \geq M'_{i,j+1}$ and vice versa;

III For arbitrary i and j, if $M'_{i,j}$ is decided, then $\sum_{j=1}^{5} \left| M_{i,j}^{'} - \overline{X}_\iota^{'} \right|$ has minimum value.

where, $M_{i,j}$ is the Class j water quality standard value of parameter i, $M'_{i,j}$ is the normalized water quality standard value of parameter i, X_i is the measured value of parameter i, X'_i is the normalized value of parameter i, $\overline{X}_\iota^{'}$ is the mean of X'_i.

Principle I is to ensure the range consistency between the original standards and the normalized ones. Principle II is to make sure that if the measured value of parameter i X_i satisfies Class j+1 but not Class j according to original standards, the normalized value X'_I should also satisfies Class j+1 but not j according to normalized ones. Principle III is to ensure the normalized standard values are as close as possible to the means of normalized values. The classification result will destabilized due to a large deviation between the normalized standard values and the means of normalized values.

(2) Classification method

Similar like the Nemerow method, the original water quality standard values were normalized by the normalization formula (2), then a new normalized water quality standard was set up for classification. It is easily to prove that the normalization satisfies Principle I and II, as it is just a mathematical translation of original standard values. In order to satisfy Principle III and be consistent with Principle I and II, we suggest that once the Class j water quality standard value of parameter i $M_{i,j} \geq X_{i\max}$ or $M_{i,j} \leq X_{i\min}$, we use the $X_{i\max}$ or $X_{i\min}$ to substitute the water quality standard value.

As a result, we carried out the PCA based water quality classification method as follows:

$$\begin{cases} M'_{i,j} = \dfrac{M_{i,j} - \overline{X_i}}{s_i}, & if \quad X_{i,\min} \le M_{i,j} \le X_{i,\max} \\ M'_{i,j} = \dfrac{X_{i,\max} - \overline{X_i}}{s_i}, & if \quad M_{i,j} \ge X_{i,\max} \\ M'_{i,j} = \dfrac{\overline{X_i} - X_{i,\min}}{s_i}, & if \quad M_{i,j} \le X_{i,\min} \end{cases} \tag{4}$$

where $\overline{X_i}$ is the mean of X_i, s_i is the standard deviation of X_i, $X_{i,\min}$ is the minimum value of X_i, $X_{i,\max}$ is the maximum value of X_i.

Results and discussion

Data analysis. Standardization and independence test were carried out to the monthly mean concentrations of 8 indexes, including COD_{Mn}(X_1), NH_3-N(X_2), DO(X_3), Se(X_4), As(X_5), Zn(X_6), Pb(X_7) and Cu(X_8).

After standardization, the KMO test and Bartlett sphericity test were used to check the feasibility for PCA. If the KMO result is larger than 0.5 and Bartlett result is smaller than 0.05, this indicates the non-mutual independence of data and can be applied for PCA [6]. In this study, the data's KMO test is 0.674 and Bartlett sphericity test is smaller than 0.001, which means the feasibility for PCA.

By using the SPSS 20.0, we gained the eigenvalues, as shown in Fig 1.

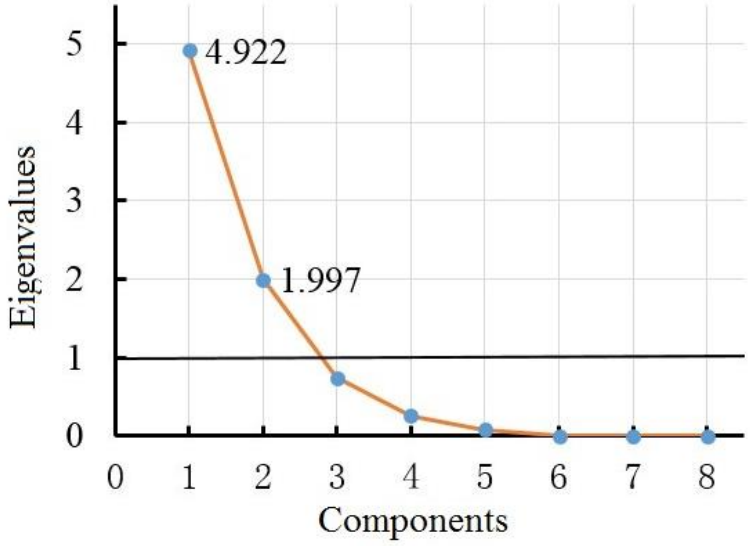

Fig. 1 Scree plot of eigenvalues

The eigenvalues of Principal Component 1 and 2 (short for PC1 and PC2) are 4.922 and 1.997, both of which are larger than 1. And the cumulative % of variance of PC1 and PC2 is 86.484%, which is larger than 85%. These indicates that PC1 and PC2 have basically included the information of raw data, which could be replaced by PC1 and PC2 [7].

Principal component loadings. The corresponding initial factor loadings of the PC1 and PC2 can be calculated by the SPSS 20.0, and by using the formula below, we can get the principal component loadings,

$$L_m = V_m / SQR(\lambda_m)$$

where, V_m and λ_m are the initial factor loading and eigenvalue of principal component m (m=1 and 2), respectively. L_m is the principal component loading of principal component m. The Principal component loadings of PC1 and PC2 are shown in Fig. 2.

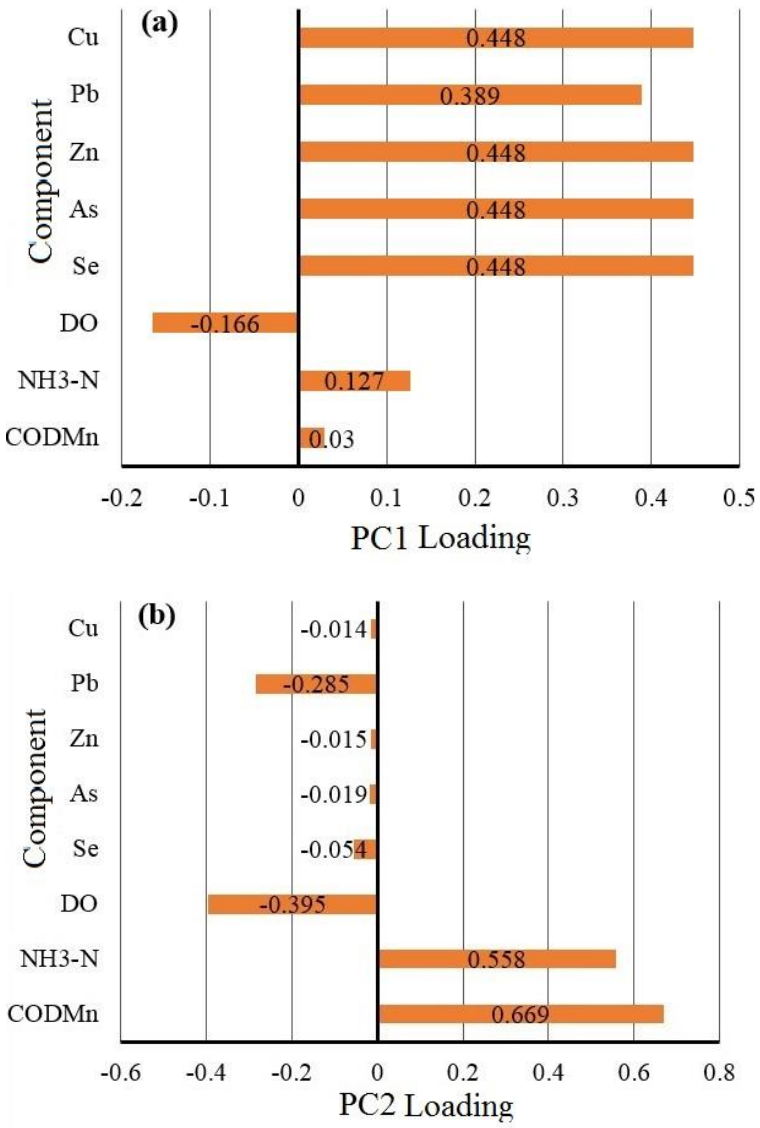

Fig. 2 PC loadings (a) PC1 loading (b) PC2 loading

As shown, the correlations among PC1 and metal indexes like Se, As, Zn, Pb and Cu, are from 0.389 to 0.448, this indicates the PC1 has mainly revealed the situation of heavy metal indexes; while, the correlations among PC2 and COD_{Mn} and NH_3-H are 0.669 and 0.558, this indicates the PC2 has mainly revealed the situation of COD_{Mn} and NH_3-H. And the correlations among the 2 principal components and DO are -0.166 and -0.395, this reveals a negative correlation, indicating the larger DO, the better water quality.

Then we can get the principal component score functions by eq. (3),

$$F_1 = 0.030X_1 + 0.127X_2 - 0.166X_3 + 0.448X_4 + 0.448X_5 + 0.448X_6 + 0.389X_7 + 0.448X_8 \quad (5)$$

$$F_2 = 0.669X_1 + 0.558X_2 - 0.395X_3 - 0.054X_4 - 0.019X_5 - 0.015X_6 - 0.285X_7 - 0.014X_8 \quad (6)$$

And the comprehensive score function is,

$$F = 0.615F_1 + 0.250F_2 \quad (7)$$

Water quality by PCA. According to eq. (5), (6) and (7), the principal scores of each month in 2014 can be calculated, where a higher score a heavier pollution, shown in Fig. 3. In Fig.3 (a), we found that except October, November and December, the PC1 scores ranged from 0.739 to 1.524 in the rest months, indicating the TIA had severe heavy metal pollution in these 9 months; while for the PC2 scores, higher scores appeared from March to June, indicating severe COD_{Mn} and NH_3-H pollution occurred in the TIA in these months; and totally, the most severe pollution conditions

appeared from March to June. In Fig.3 (b), we sorted the pollution severity by the comprehensive scores month by month and we found the TIA was most polluted in April, while November had the lightest pollution.

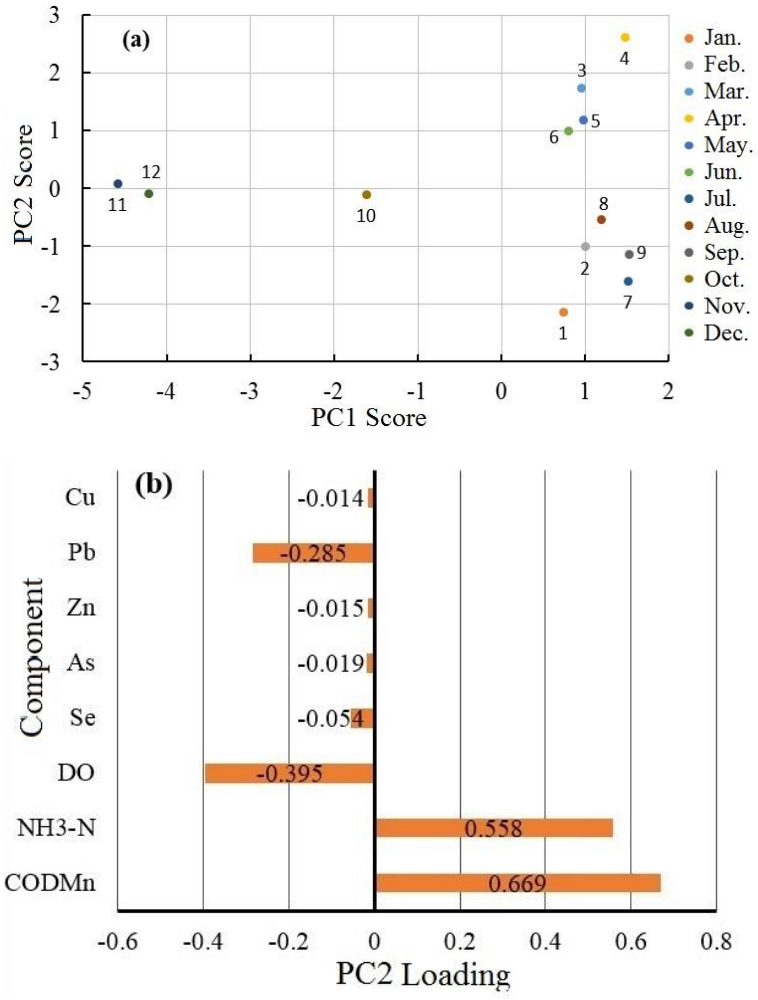

Fig.3 PC scores (a) PC1 versus PC2 scores (b) PC comprehensive score

Water quality classification. The PCA can directly describe the pollution characteristics month by month, but is disable to tell the water classifications. This study suggested a Water quality classification method based on PCA results.

According to the water quality data and eq. (4), (5), (6) and (7), the PCA classification standards were calculated, shown in Table 2.

Table 2 PCA classification standards

Category	I	II	III	IV	V
Score	-2.75	-1.11	1.31	1.93	1.99

Based on Fig.3 (a), we can get the classification plot according to Table 2 and eq. (7), shown in Fig. 4. We can see that the water quality in November was Class I and Class II for December; in April it was Class IV, which was the worst.

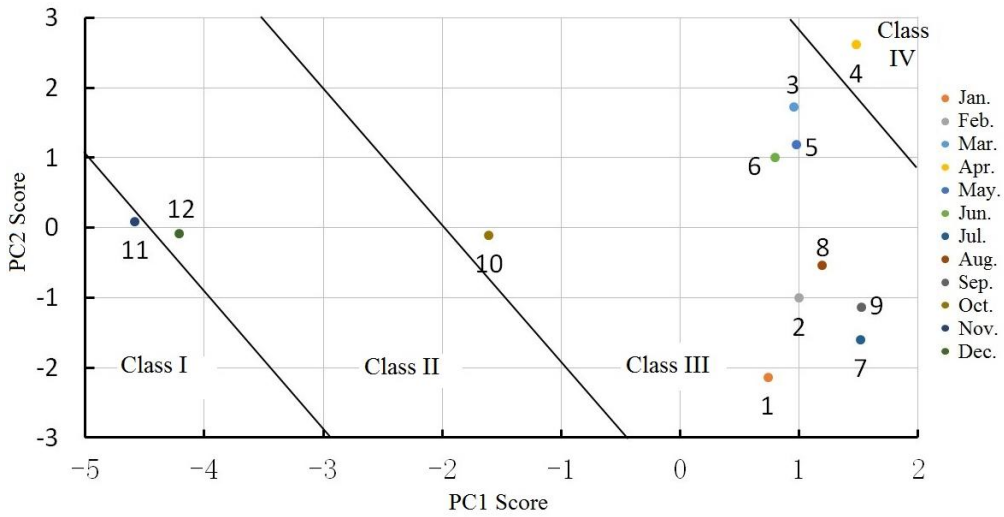

Fig.4 PCA classification result

Moreover, we compared the PCA classification results with Fuzzy Evaluation, Nemerow index method and improved Nemerow index method, shown in Fig.5. It can be found that the PCA classification had similar results with other methods, except in January and February, which was sorted as Class III by PCA method while Class I or II by other methods. We found that the PC2 scores for January and February were -2.136 and -1.006 in Fig.3 (a), while the PC1 scores were 0.739 and 1.225. By eq. (7), the PC1 has a higher weight of 0.615, larger than PC2, resulted in a higher comprehensive scores in January and February. In comparison to other methods, the PCA classification method can avoid the affecting of extreme data and is able to reflect the contribution of most indexes in the classification results.

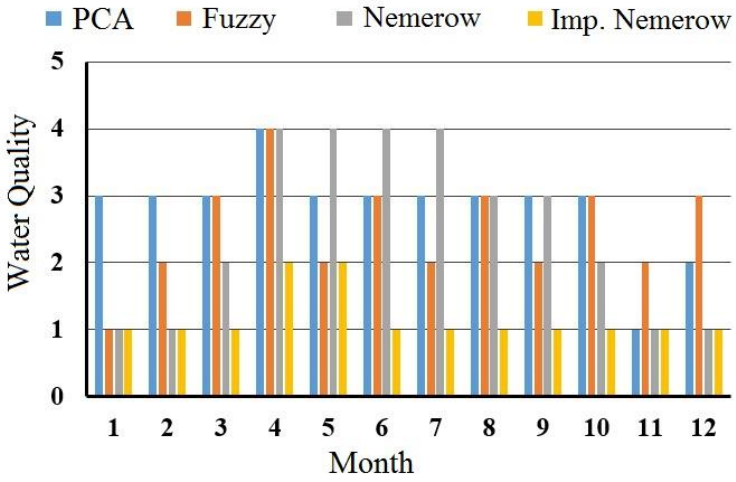

Fig.5 Classification results with 4 methods

Conclusions

This study applied the PCA based water quality classification method to assess the water quality of Tongjiyan Irrigation Area in Sichuan Province, China. The PCA classification method in this study was based on the pollution characteristics and the 86.484% data information derived from PCA. And the classification results were impacted by the contribution of different principal components. It was found that the TIA was most polluted in April, while November had the lightest pollution in 2014. The water quality in November was Class I and Class IV in April.

Acknowledgements

This work was financially supported by the Fundamental Research Funds for the Central Universities (2682016CX095).

References

[1] B. Wu, S.Y. Zang, X.D. Na: J. Saf. & Env. Vol.5 (2012), p.134-137
[2] N.L. Nemerow: *Scientific stream pollution analysis* (Scripta Book Co, USA 1974)
[3] X.G. Han, T.L. Huang, X.Z. Chen: Acta Scientiae Circumstantiae. Vol.33 (2013), p. 1513-1518
[4] R.L. Olsen, R.W. Chappell, J.C. Loftis: Water Res. Vol.46 (2012), p.3110-3122
[5] G. Liu: Sichuan Water Cons. No.1 (2015), p.50-52
[6] Q.Q. Du, K. Yan: J. Water Res. Water Eng. Vol.24 (2013), p. 212-214
[7] X.Y. Lu: Natural Science Journal of Harbin Normal University. Vol. 31 (2015), p. 156-161