Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

ISSN: 0972-6268	Vol. 14	No. 4	pp. 855-858	2015

Original Research Paper

Evaluation of Water Quality Using Principal Component Analysis

Shuquan An*, Xiufan Xie* and Ying Ma†**
*Wenzhou Vocational College of Science and Technology, Wenzhou, Zhejiang, 325006, P. R. China
**School of Water Resources, North China University of Water Resources and Electric Power, Zhengzhou, 450045, China
†Corresponding author: Ying Ma

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com
Received: 12-08-2014
Accepted: 16-10-2014

Key Words:
Water quality
Principal component analysis
Comprehensive index

ABSTRACT

Principal component analysis is a way to reduce original dimension, to make multiple variables into a few comprehensive index. According to the characteristics of water quality evaluation model, principal component analysis method is developed to evaluate surface water quality using SPSS software at representative sections. By the combination of variables index, adjusting the combinatorial coefficient to make the new variables representative independent. The process is introduced in the paper in detail. The results indicate that the principal component model is suitable for water quality evaluation. By analysis, it is important to pay attention to bring into effective measures for pollution control.

INTRODUCTION

Water quality evaluation is to distinguish the grade of water quality according to the water quality standard. This is a typical problem of pattern recognition. A traditional water quality evaluation method usually adopts an accurate mathematic model to describe. But there are many factors influencing water quality, the relationship between evaluation indicators and standards is nonlinear (Peng 2011). The relationship among every grade in evaluation standard is fuzzy and grey. There is much difficulty if adopting a certain and uncertain method (Wan 2009, 2010). While the principal component analysis will make several index standards into a few comprehensive index, and simplify the structure of statistical analysis. With no loss of the original information, the influence index of water quality will be combined into a new group index to reflect the comprehensive index, in order to achieve dimension reduction, simplified data and improve the reliability analysis of the results (Chen 2008).

The author uses the principal component analysis, with the help of SPSS, to evaluate the water quality and find out the main influencing index of water quality, to prevent and control the reasonable exploitation and utilization of water resources and provide decision-making guidance.

PRINCIPLE OF PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a way to reduce the original dimension, to make multiple variables into a few comprehensive indices. By the combination of variables index, adjusting the combinatorial coefficient to make the new variables representative independent.

For n samples, the matrix of p observation data ($x_1, x_2, \cdots x_p$) is as follows:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \left(x_1, x_2, \cdots x_p\right)$$

$$\text{Where } x_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}, \quad j = 1,2,\cdots p \qquad \ldots(1)$$

Principal component analysis is to make p comprehensive observation variables as new variables (variables), i.e.

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p \\ \cdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p \end{cases} \qquad \ldots(2)$$

This equation can be simplified as :

$$F_j = \alpha_{j1}x_1 + \alpha_{j2}x_2 + \cdots + \alpha_{jp}x_p \quad j = 1,2,\cdots, p$$

The model should satisfy the following conditions:

1. F_i, F_j is discrete ($i \neq j, i, j = 1,2,\cdots, p$)

2. Variance of F_1 variance is larger than F_2 and greater than F_3.

3. $a_{k1}{}^2 + a_{k2}{}^2 + \cdots + a_{kp}{}^2 = 1 \qquad k = 1,2,\cdots p$.

So, F_1 is the first principal component, F_2 the second principal component. By analogy, there are p principal components. a_{ij} is coefficient of principal component. The model can be expressed as a matrix:

$$F = AX ,$$

Where $F = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_p \end{pmatrix}$ $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix}$$

A is called principal components coefficient matrix.

The process is as follows:

1. In order to eliminate the influence caused by different orders of magnitude and dimension, the original data should be standardized.

$$X_{ij} = \frac{Y_{ij} - EY_j}{\sqrt{DY_j}} \quad ...(3)$$

Where,

$$EY_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij} , DY_j = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - EY_j)^2$$

（i＝1,2，……,n;j＝1,2,……,p）

2. After normalization treatment of the original data, obtain the standard data tables $Y(x_{ij}^{'})$ can be obtained. Correlation coefficient matrix $R = (r_{ij})_{p*p}$ can be calculated.
3. Eigen values and eigen vectors: According to the correlated coefficient matrix R, eigen value λ_i(i=1,2,……,m) can be calculated by characteristic equations |λI-R|=0. According to the order of size, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$: then calculating e_i(i=1,2, …, m).
4. Principal components contribution and the cumulative contribution.

   Principal components contribution: $e_i = \lambda_i / \sum_{k=1}^{m} \lambda_k$ (i=1,2, ……, m).

   the cumulative contribution: $E_k = \sum_{i=1}^{p} \lambda_i / \sum_{k=1}^{m} \lambda_k$

   calculation of principle component $z_j = \sum_{j=1}^{p}\sum_{i=1}^{n} u_{ij} x_{ij}^{'}$
5. Comprehensive analysis: If the former *p* principal components can provide 85%~95% information, the eigen values can be taken λ_1, λ_2,…, λ_p corresponding *p* (p≤m) components, variables are cut from *m* to *p* , producing reduced-order effect.

APPLICATION

Selection of monitoring indicator: The situation of water environment of river is a complex system composed of a plurality of water quality indexes, and each index of water quality are closely linked. In the process of analysis, due to collinearity problem of water quality index, correct conclusions cannot be derived. The main aims of principal component is to make multiple indexes into a few independent comprehensive indexes through a linear transformation, to simplify and improve the reliability of data analysis. In this paper, 10 main quality indexes are selected as principal component.

Standardization: Combined with the local environmental quality, pH value, conductivity, suspended solids, chloride, sulphate, dissolved oxygen, ammonia nitrogen, nitrite nitrogen, nitrate nitrogen and permanganate index are selected as monitoring and analysis project. Due to the concentration difference between the 10 indicators, the data should be standardized which can be seen in Table 1.

Eigen values and feature vectors and cumulative contribution rate: The original monitoring data were normalized to get the correlation coefficient matrix, SPSS software was used to calculate the correlation coefficient matrix and eigen value of 10 indicators, to determine the main factor evaluation number (Table 1). According to the feature values of cumulative contribution of variance, the number of selected principal components (Table 2) was determined.

According to Table 2, characteristics of the first, second, third, the fourth or fifth principal component values are 2.865, 2.065, 1.513, 1.313, 1.024 respectively, which are larger than 1. The variance contribution rate are 28.653%, 20.647%, 15.128%, 13.126%, and 10.239%. The cumulative variance rate of 87.793%, indicates that they include most of the information of 10 indicators. Among them, the first principal component is the most important as most information is contained and has the biggest influence on water quality.

$$F1 = 0.049x_1 + 0.264x_2 - 0.068x_3 + 0.245x_4 + 0.212x_5 - 0.09x_6 + 0.086x_7 + 0.260x_8 + 0.082x_9 + 0.279x_{10}$$

$$F2 = -0.104x_1 + 0.074x_2 + 0.226x_3 - 0.250x_4 - 0.319x_5 + 0.064x_6 + 0.418x_7 + 0.172x_8 - 0.041x_9 + 0.209x_{10}$$

Table 1: Standardized data of water quality indicators in Kuangmenkou station water quality.

		Correlations									
		pH	Electrical conductivity	Suspended solids	Cl^-	SO_4^{2-}	DO	Ammonia nitrogen	Nitrite nitrogen	Nitrate nitrogen	Permanganate index
pH	Pearson Correlation	1	.167	.226	-.041	.257	-.192	-.197	.054	.195	-.042
	Sig. (2-tailed)		.459	.312	.857	.249	.392	.379	.810	.384	.853
	N	22	22	22	22	22	22	22	22	22	22
Electrical conductivity	Pearson Correlation	.167	1	.160	.506*	.413	-.170	.416	.402	.104	.433*
	Sig. (2-tailed)	.459		.478	.016	.056	.450	.054	.064	.644	.044
	N	22	22	22	22	22	22	22	22	22	22
Suspended solids	Pearson Correlation	.226	.160	1	-.362	-.187	-.121	.442*	-.229	.044	-.161
	Sig. (2-tailed)	.312	.478		.098	.404	.591	.039	.305	.845	.475
	N	22	22	22	22	22	22	22	22	22	22
Cl^-	Pearson Correlation	-.041	.506*	-.362	1	.797**	-.068	-.126	.175	.121	.289
	Sig. (2-tailed)	.857	.016	.098		.000	.763	.577	.436	.592	.192
	N	22	22	22	22	22	22	22	22	22	22
SO_4^{2-}	Pearson Correlation	.257	.413	-.187	.797**	1	-.133	-.305	.100	.233	.123
	Sig. (2-tailed)	.249	.056	.404	.000		.556	.168	.658	.297	.586
	N	22	22	22	22	22	22	22	22	22	22
DO	Pearson Correlation	-.192	-.170	-.121	-.068	-.133	1	.277	-.236	.166	-.188
	Sig. (2-tailed)	.392	.450	.591	.763	.556		.212	.290	.459	.402
	N	22	22	22	22	22	22	22	22	22	22
Ammonia nitrogen	Pearson Correlation	-.197	.416	.442*	-.126	-.305	.277	1	.284	.034	.478*
	Sig. (2-tailed)	.379	.054	.039	.577	.168	.212		.201	.881	.024
	N	22	22	22	22	22	22	22	22	22	22
Nitrite nitrogen	Pearson Correlation	.054	.402	-.229	.175	.100	-.236	.284	1	.041	.885**
	Sig. (2-tailed)	.810	.064	.305	.436	.658	.290	.201		.858	.000
	N	22	22	22	22	22	22	22	22	22	22
Nitrate nitrogen	Pearson Correlation	.195	.104	.044	.121	.233	.166	.034	.041	1	.152
	Sig. (2-tailed)	.384	.644	.845	.592	.297	.459	.881	.858		.501
	N	22	22	22	22	22	22	22	22	22	22
Permanganate index	Pearson Correlation	-.042	.433*	-.161	.289	.123	-.188	.478*	.885**	.152	1
	Sig. (2-tailed)	.853	.044	.475	.192	.586	.402	.024	.000	.501	
	N	22	22	22	22	22	22	22	22	22	22

*. Correlation is significant at the 0.05 level (2-tailed).**. Correlation is significant at the 0.01 level (2-tailed).

$$F3 = 0.471x_1 + 0.208x_2 + 0.503x_3 - 0.078x_4 + 0.129x_5 - 0.161 + 0.051x_7 - 0.167x_8 + 0.197x_9 - 0.157x_{10}$$

$$F4 = -0.169x_1 + 0.098x_2 + 0.020x_3 + 0.198x_4 + 0.142x_5 + 0.617x_6 + 0.253x_7 - 0.245x_8 + 0.385x_9 - 0.105x_{10}$$

$$F5 = 0.362x_1 - 0.343x_2 - 0.237x_3 - 0.299x_4 - 0.133x_5 + 0.119x_6 - 0.181x_7 + 0.280x_8 + 0.626x_9 + 0.215x_{10}$$

$$F = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \cdots + \lambda_5} F_1 + \frac{\lambda_2}{\lambda_1 + \lambda_2 + \cdots + \lambda_5} F_2 + \cdots + \frac{\lambda_5}{\lambda_1 + \lambda_2 + \cdots + \lambda_5} F_5$$

$$= 0.326F_1 + 0.235F_2 + 0.172F_3 + 0.150F_4 + 0.117F_5$$

Thus, each principal component scores can be calculated, according to the comprehensive evaluation function, section water pollution comprehensive score can be calculated. Greater the score, indicates more the serious pollution.

Seen from the principal component scores, the water quality in the year 1999 is the worst, comprehensive score reached 1.199 and the overall water quality fluctuates with time.

In the first principal component, electrical conductivity, chloride, sulphate, nitrate nitrogen and permanganate index play the leading role in the water quality evaluation; in the second principal components, the absolute maximum suspended matter and nitrogen; and in the third principal components, suspended solids and pH value of absolute value. Seen from Table2, the first , the second and the third principal component contribution rate is 28.653%,20.647% and 15.128% respectively. The water is polluted by organic and inorganic pollutant.

CONCLUSIONS

There are many methods for comprehensive assessment of water environmental quality, which have their advantages and disadvantages.

Table 2: Eigenvalues, contribution rates and accumulated contribution rates of principal comonents.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.865	28.653	28.653	2.865	28.653	28.653
2	2.065	20.647	49.300	2.065	20.647	49.300
3	1.513	15.128	64.427	1.513	15.128	64.427
4	1.313	13.126	77.554	1.313	13.126	77.554
5	1.024	10.239	87.793	1.024	10.239	87.793
6	.615	6.153	93.946			
7	.310	3.097	97.042			
8	.175	1.753	98.796			
9	.089	.889	99.685			
10	.032	.315	100.000			

Extraction Method: Principal Component Analysis.

Table 3: Loading values of principal components.

	Component				
	1	2	3	4	5
Permanganate index	.799	.431	-.238	-.137	.221
Electrical conductivity	.758	.153	.314	.128	-.352
Nitrite nitrogen	.744	.355	-.253	-.322	.287
Cl^-	.701	-.515	-.118	.260	-.306
Ammonia nitrogen	.246	.862	.077	.332	-.186
SO_4^{2-}	.607	-.658	.195	.186	-.136
Suspended solids	-.195	.467	.761	.026	-.243
pH	.141	-.214	.713	-.222	.371
DO	-.257	.132	-.244	.810	.122
Nitrate nitrogen	.236	-.084	.297	.506	.641

1. Principal component-based water quality model has the advantage of simple learning, good compatibility, high ability of classification and so on. The evaluation results indicate that the water quality is poor.
2. Principal component-based water quality model is established. This model can evaluate water quality in Kuangmenkou station well, which avoids the cumbersome theory and parameter identification. The results indicate that the method is reliable.

Table 4: Component score coefficient matrix.

	Component				
	1	2	3	4	5
pH	.049	-.104	.471	-.169	.362
Electrical conductivity	.264	.074	.208	.098	-.343
Suspended solids	-.068	.226	.503	.020	-.237
Cl^-	.245	-.250	-.078	.198	-.299
SO_4^{2-}	.212	-.319	.129	.142	-.133
DO	-.090	.064	-.161	.617	.119
Ammonia nitrogen	.086	.418	.051	.253	-.181
Nitrite nitrogen	.260	.172	-.167	-.245	.280
Nitrate nitrogen	.082	-.041	.197	.385	.626
Permanganate index	.279	.209	-.157	-.105	.215

Table 5: Evaluation outcomes of water qualities.

1st principal component	2nd principal component	3rd principal component	4th principal component	5th principal component	synthesis score
-0.93977	0.54174	-0.90823	-1.99042	-1.30574	-0.78553
0.9487	-1.53036	-0.45024	-0.05675	-0.96745	-0.24925
-0.16095	0.3587	2.52797	0.01217	0.72854	0.554185
-0.48865	0.34248	0.04805	-0.2031	-0.86626	-0.20199
-0.35195	0.8418	-0.31641	0.59262	-1.98979	-0.11478
0.14368	2.15735	0.24095	0.69636	-1.03843	0.578832
-0.14763	1.77671	0.15323	1.6759	0.45967	0.700257
-0.26991	-0.06644	-0.4913	0.41291	1.69846	0.071421
3.33131	1.58632	-1.13138	-1.44222	1.27961	1.198763
-0.52292	-0.20394	-2.09469	0.77943	-0.35213	-0.50404
-1.11209	-0.15896	-1.38883	0.54253	1.37638	-0.39796
-0.76501	0.12439	-0.36853	-0.16699	0.6097	-0.23774
-0.34488	0.63152	1.66776	-0.62783	0.17166	0.249509
-0.52394	-0.57835	0.59027	-0.37738	1.62934	-0.07172
-0.66221	-0.22712	0.71106	-0.52599	0.40663	-0.1782
-0.93033	-0.08168	-0.40503	0.04381	0.24663	-0.35726
-0.32031	0.0161	0.48559	-0.22228	-0.49691	-0.10823
0.55307	-0.9627	0.22401	-1.03667	-0.52932	-0.22406
0.92465	-0.9616	0.98935	-0.45509	-1.09156	0.050698
1.39054	-1.12201	0.25907	2.59572	-0.42739	0.572701
0.66949	-1.30798	0.17066	0.66851	0.23088	0.067085
-0.42087	-1.17598	-0.51331	-0.91521	0.22747	-0.61266

3. The results evaluated by principal component can provide references for environmental departments.

REFERENCES

Chen, Shuqin, Xu, Qiujin, Yan, Changzhou and Chu, Zhaosheng 2008. Application of principal components analysis method to weight determination in fuzzy recognition model for water quality assess. Journal of Kunming University of Science and Technology, 33(2): 77-80.

Peng, Qian and Zheng, Yanxia 2011. A study of the water quality assessment of Minxinhe of Shijiazhuang city based on the major element analysis method. Journal of Huaihua University, 30(8): 92-96.

Wan, Jinbao, Zeng, Haiyan and Zhu, Banghui 2009. Application of principal component analysis in evaluation of water quality of Lean river. China Water & Wasterwater, 25(16): 104-108.

Wan, Jinbao, He, Huayan, Zeng, Haiyan and Li, Yuanyuan 2010. Application of principal component analysis in evaluating water quality of Poyang lake. Journal of Nanchang University, 32(2): 113-116.