Water Quality Evaluation Model Based on Principal Component Analysis and Information Entropy: Application in Jinshui River

MA Jianqin\textsuperscript{1*}, GUO Jingjing\textsuperscript{1} and LIU Xiaojie\textsuperscript{2}

1 North China University of Water Conservancy and Hydroelectric Power, Zhengzhou 450011, China; 2 Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China

Abstract: Water quality evaluation is important because it could provide guidance when determining water utility. But many interacting impact factors are involved in water quality evaluation systems, making water quality evaluation difficult. Principal component analysis (PCA) is widely used in water quality evaluation because it can eliminate the correlation among factors. However, PCA ignores the degree of data dispersion, which is considered by information entropy (IE). To solve this problem, a model combined PCA and IE methods to obtain the weights of indicators is proposed in this paper, and the proposed model was applied to assess the reused water quality of Jinshui River in Zhengzhou City in 2009. The evaluation results were compared with those using PCA and IE methods for the same data. The results proved that the method is feasible and practical, and it can provide a theoretical basis and decision reference for the utility of unconventional water.

Key words: impact factors; water quality evaluation; principal component analysis (PCA); information entropy (IE); weight; unconventional water

1 Introduction

The evaluation of water quality is very important, especially in most cities in China. An evaluation can provide guidance in determining the utility of different kinds of water. However, there are so many interacting impact factors involved that make water quality evaluation difficult.

At present, there are many mathematical models for assessing water quality, such as the principal component analysis (PCA), the information entropy (IE) and gray clustering method. PCA can eliminate the correlation among factors, and change a large set of indicators into a smaller one, and PCA is widely used in water quality evaluation. Kunwar et al. (2004) applied PCA to the research of spatial and temporal variations in water quality of the Gomti River in India. Besides, Wu et al. (2009) studied changes characterization of water quality in Daya Bay with PCA. Indicators correlation is considered in the PCA method, however the degree of data dispersion which is considered by IE is ignored (Li et al. 2006). IE method is applied to set pair analysis model in water quality evaluation (Meng et al. 2009). Zhang Yiding achieved indicators’ weights by IE for a study of comprehensive assessment of groundwater environment in Zhengzhou City. Those study shows that using of IE to obtain indicators’ weights is feasible, but these traditional methods failed to solve the complicated nonlinear relationship between the interacting impact factors and the water class, especially for unconventional water.

To obtain better water quality assessments and to merge the advantages of PCA and IE, a water quality evaluation model combining these two methods was established to assess the landscape water quality in Zhengzhou City. The results provided in this paper indicate both feasibility and correctness.

Jinshui River is the first landscape river to use
reclaimed water as its source in Zhengzhou City. A total of 5 × 10⁴ t treated reclaimed water is transported from Wulongkou sewage treatment plant to Jinshui River daily. An assessment of the water quality of landscape water in Jinshui River, and analysis of the existing problems can be of great importance for the increased use of unconventional water resources and landscape water safety.

2 Methodology

PCA is a multivariate statistical analysis method which could transform multidimensional factors into the same system for quantitative study. Certain correlations exist among multi-indicators, PCA attempts to transform a large set of inter-correlated indicators into a smaller set of composite indicators, and simplifies the structure of the statistical analysis system. The IE method determines weight based on the data disorder degree. Based on the information theory, entropy reflects the information disorder degree and can be used to measure the information of the factor. The more information carried by an indicator which provides a larger effect to decision, the entropy will be much smaller, which means that the system has a larger degree of disorder (Meng et al. 2009). A model combining PCA and IE for Water Quality Evaluation is proposed in the paper. The model’s calculation steps are as follows:

(1) Obtain the original data matrix of m samples with n indicators, C_i^j represents the value of the j indicator of i sample.

For favorable factors,

\[ Z_{ij} = (C_{ij} - \frac{1}{n} \sum_{i=1}^{n} C_{ij}) / S_j \]  

(1)

For negative factors,

\[ Z_{ij} = (\frac{1}{n} \sum_{i=1}^{n} C_{ij} - C_{ij}) / S_j \]  

(2)

where \( S_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (C_{ij} - \frac{1}{n} \sum_{i=1}^{n} C_{ij})^2} / n \).

(2) The indicators’ correlation coefficient matrix R_y is calculated as shown in Eq. (3):

\[ R_y = \frac{1}{n} \sum_{k=1}^{m} Z_{ki} Z_{kj} \]  

(3)

Then get the proper value \( \lambda_i \) of \( R_y \) and its eigenvector, arrange \( \lambda_i \) from large to small, then obtain contribution rate \( T_k \) and cumulative contribution rate \( D_k \) of factor k as Eq. (4) and (5):

\[ T_k = \lambda_k \sum_{j=1}^{n} \lambda_k \]  

(4)

\[ D_k = \sum_{j=1}^{k} T_k \]  

(5)

Select the corresponding principal components (PCs) of the \( \lambda_i \) when the \( D_k \geq 90\% \), then make \( D_p \) of p PCs equal to 1, calculate the \( T'_1, T'_2, \cdots, T'_p \) corresponding to \( T_1, T_2, \cdots, T_p \).

(3) Calculate the dispersion degree of p PCs:

Firstly, obtain the standard matrix of p PCs, the j factor of PCs is calculated as Eq. (6):

\[ M_{ij} = l_{ij} Z_{ij} + l_{ij} Z_{ij} + \cdots + l_{ij} Z_{ij} \]  

Then make data translated so that \( M_{ij} \geq 1 \), and obtain the entropy of the j factor as Eq. (7):

\[ e_j = -k \sum_{i=1}^{n} q_i^j \ln q_i^j \]  

(7)

where \( k = 1 / \ln(p) \), \( q_i^j = M_{ij} / \sum_{i=1}^{n} M_{ij} \).

The entropy weight of the j factor is:

\[ g_j = \frac{1 - e_j}{\sum_{i=1}^{p} (1 - e_i)} \]  

(8)

(4) The contribution rates represent the contribution of p PCs to the total samples, and entropy weights represent the information carried by p PCs. Then their weights are decided by combining contribution rates and entropy weights in the model. The final weight of the PC j is:

\[ w_j = aT'_j + (1-a)g_j \]  

(9)

where a is the proportional coefficient, \( 0 \leq a \leq 1 \).

(5) Calculate the score matrix \( Y_i \) of original indicators.

For negative indicator, \( Y_{ij} = C_{ij} / S_j \)  

(10)

For favorable indicator, \( Y_{ij} = S_j / C_{ij} \)  

(11)

where: \( S_j \) is the standard value of j. The standard values are decided based on the water quality standard for recycling water reused as landscape water (GB/T 18921-2002).

(6) Obtain the score matrix of p PCs. The score of PC j of sample i is computed as:

\[ O_{ij} = l_{ij} Y_{ij} + l_{ij} Y_{ij} + \cdots + l_{ij} Y_{ij} \]  

(12)

The comprehensive evaluation index of the sample is:

\[ I_i = \sum_{j=1}^{p} w_j O_{ij} \]  

(13)
3 Results

3.1 Evaluation process

According to the quality standard of recycling water reused as scenic water (GB/T 18921-2002), 14 types of indicators, such as suspended solids (SS), chemical oxygen demand (COD), 5-days biochemical oxygen demand (BOD), fecal coliform, total phosphorus (TP), total nitrogen (TN) and etc. were selected as evaluation indicators in this study. There were two monitoring sites following the Jinshui River channel, one was in mid-stream and the other was downstream. Samples were collected once a month from July to November in 2009. Part of data of site 1 are shown in Table 1.

Table 1 Water quality monitoring data of Jinshui River of site 1.

<table>
<thead>
<tr>
<th>Factor</th>
<th>July</th>
<th>August</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS (mg L⁻¹)</td>
<td>29</td>
<td>16</td>
</tr>
<tr>
<td>COD (mg L⁻¹)</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>BOD (mg L⁻¹)</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>DO (mg L⁻¹)</td>
<td>6.63</td>
<td>5.39</td>
</tr>
</tbody>
</table>

Using the methodology proposed in the paper, the proper value, contribution rates and cumulative contribution rates of PCs are obtained (Table 2).

Table 2 Proper value, contribution rates and cumulative contribution rates of PCs.

<table>
<thead>
<tr>
<th>PC</th>
<th>Proper value</th>
<th>Contribution Rate (%)</th>
<th>Cumulative contribution rate (%)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>4.245</td>
<td>30.320</td>
<td>30.320</td>
<td>0.328</td>
</tr>
<tr>
<td>PC2</td>
<td>2.923</td>
<td>20.875</td>
<td>51.196</td>
<td>0.226</td>
</tr>
<tr>
<td>PC3</td>
<td>2.044</td>
<td>14.597</td>
<td>65.793</td>
<td>0.158</td>
</tr>
<tr>
<td>PC4</td>
<td>1.618</td>
<td>11.554</td>
<td>77.347</td>
<td>0.125</td>
</tr>
<tr>
<td>PC5</td>
<td>1.111</td>
<td>7.936</td>
<td>85.283</td>
<td>0.086</td>
</tr>
<tr>
<td>PC6</td>
<td>1.013</td>
<td>7.234</td>
<td>92.517</td>
<td>0.077</td>
</tr>
</tbody>
</table>

The entropy weights of these six PCs are 0.151, 0.166, 0.168, 0.170, 0.173 and 0.172 by calculating their disorder degrees with IE method. According to Eq. (8), the final weights of the six PCs are 0.240, 0.196, 0.163, 0.147, 0.129 and 0.125. Calculating the water quality index of each sample, one can determine the water quality rank that each sample belongs to, based on a water quality grading standard (Zhang 2002) (Table 3). The final water quality rank results are shown in Table 4.

3.2 Results analysis

The same samples were evaluated using PCA and IE, respectively, the results are shown in Table 4 and Figure 1. From the results, it can be seen that the evaluation results of IE are most serious, those of PAC are middle, and the results for a combination of PAC and IE are the lightest.

According to the evaluation results of IE, the water quality of sample 8 is worse than sample 9, while the results following the other two methods show the opposite. It was found that only total nitrogen, fecal coliform, and chlorine, in sample 8 are higher than in sample 9, and the differences are very small, but the conditions for the other 11 indicators are much better. Overall, the water quality of sample 8 should not be significantly worse than sample 9. It would appear then that the evaluation results of IE are not consistent with the
monitoring data.

The evaluation results using the combination of PCA and IE are consistent with those using PCA sample 9 is lightly polluted evaluated with PCA, while it is still clean evaluated using the combination of PCA and IE. There are only 4 indicators that have slightly excessive levels in 14 indicators of that sample: SS 22 mg L\(^{-1}\), TP 1.02 mg L\(^{-1}\), TN 17.2 mg L\(^{-1}\), and fecal coliform 30 000 L\(^{-1}\), and the standards for the four indicators are 20 mg L\(^{-1}\), 1.00 mg L\(^{-1}\), 15.0 mg L\(^{-1}\), and 10 000 L\(^{-1}\), respectively. The conditions for the other 10 indicators is much better than the standard. It can be seen that sample 9 is still clean, so obviously the results evaluated by PCA are not as good as the combined method. Our results indicate that evaluations using the method of combining PCA and IE is preferable to the other two methods.

4 Conclusions

At present, most cities in China are faced with water crises, using unconventional water as a landscape water resource is an efficient means of easing part of the water problem. Thus, it is important to evaluate the quality of landscape water, especially whenever many interacting impact factors are involved.

A water quality evaluation model combining the PCA and IE methods is proposed in this paper. The methodology presented is easy to perform: based on the monitoring water quality data. Firstly PCA is used to calculate the contribution rate of each factor, then the IE method is used to obtain the factor’s entropy weight. Secondly contribution rate and entropy weight are combined to obtain each factor’s weight. Finally, a comprehensive evaluation index is obtained by combination. The model and its method have been applied in Zhengzhou landscape Jinshui River.

Compared to the respective individual results for PCA and IE, our method combining PCA and IE is more reasonable and feasible for landscape water. The method can be used to evaluate systems that involve many interacting impact factors.

References


