Journal of Environmental Sciences Vol. 16, No. 6, pp. 994—995, 2004

ISSN 1001—0742
CN11—2629/X

Article ID: 1001-0742(2004)06-0994-02 **CLC number**: X703.1; X832 **Document code**: A

Projection pursuit cluster model and its application in water quality assessment

WANG Shun-jiu[1,*], YANG Zhi-feng[1,2], DING Jing[3]

(1. School of Environment, Beijing Normal University, Beijing 100875, China. E-mail: wsjbnu@163.com.cn; 2. The State Key Laboratory of Water Environment Simulation, Beijing Normal University, Beijing 100875, China; 3. School of Hydropower Engineering, Sichuan University, Chengdu 610065, China)

Abstract: One of the difficulties frequently encountered in water quality assessment is that there are many factors and they cannot be assessed according to one factor, all the effect factors associated with water quality must be used. In order to overcome this issues the projection pursuit principle is introduced into water quality assessment, and projection pursuit cluster(PPC) model is developed in this study. The PPC model makes the transition from high dimension to one-dimension. In other words, based on the PPC model, multifactor problem can be converted to one factor problem. The application of PPC model can be divided into four parts: (1) to estimate projection index function $Q(\vec{a})$; (2) to find the right projection direction $\hat{a}$; (3) to calculate projection characteristic value of the i th sample z_i, and (4) to draw comprehensive analysis on the basis of z_i. On the other hand, the empirical formula of cutoff radius R is developed, which is benefit for the model to be used in practice. Finally, a case study of water quality assessment is proposed in this paper. The results showed that the PPC model is reasonable, and it is more objective and less subjective in water quality assessment. It is a new method for multivariate problem comprehensive analysis.

Keywords: projection pursuit; cluster; model; water quality assessment

Introduction

Water quality assessment is usually related to multiple-factor comprehensive analysis. Presently, it is difficult to resolve complex high dimensional problem directly. If there is an effective way to reduce the dimensionality, multidimensional space problems can be resolved on visual space, such as three-dimensional space, two-dimensional space even one-dimensional space.

Friedman (Friedman, 1974) developed a projection pursuit principle. It is able to find a right projection direction that can retain the main feature of data according to a projection index. On the basis of the right projection direction, high dimensional problem can be converted to low dimensional problem such as one-dimension. The objective, high dimensional data characteristics can be analyzed on one-dimensional or two-dimensional space, and many ordinary methods used on low dimensional space can be used to analyze high dimensional problems can be achieved. Based on the consideration mentioned above, the projection pursuit cluster(PPC) model is developed. During the development, the PPC model is brought into steady improvement, which has been used in many fields(Jin, 2001; Zhang, 2001a; 2001b; 2002; Wang, 2002a). The PPC model and its application will be introduced in detail in the following sections.

1 PPC model

Projection pursuit is a statistical method that can be used to analyze complex multivariate problems. This paper describes a linear projection. High dimensional data is projected onto one-dimensional space, and the feature of high dimensional data was studied through the projected characteristics of the one-dimensional space (Friedman, 1974).

If x_{ij}^0 ($i = 1, \cdots, n; j = 1, \cdots, m$; n is the total number of samples; m is the total number of effect factors of sample) is the jth factor original value of the ith sample, the steps of developing PPC model can be illustrated as follows.

(1) Data standardization. In order to eliminate the effect from dimension and ranges, the original value will be transformed before it is used by PPC model. Here, the transformation formula is $x_{ij} = \frac{x_{ij}^0}{x_{j\max}^0}$, where $x_{j\max}^0$ is the original maximum of the jth factor.

(2) Linear projection. A linear projection is described in this paper. If $\vec{a}$ is a m-dimensional unit vector, and z_i is the projection characteristic value of the ith sample, the linear projection of the ith sample can be expressed by Formula (1),

$$z_i = \sum_{j=1}^{m} a_j x_{ij}. \quad (1)$$

(3) Objective projection index function. Water quality assessment and cluster analysis are the same in essence. Based on the principle of cluster analysis, the projection index $Q(\vec{a})$ can be written as Formula (2),

$$Q(\vec{a}) = s(\vec{a}) \cdot d(\vec{a}). \quad (2)$$

Here, $\vec{a}$ is the projection axis, $s(\vec{a})$ measures the spread of data, and $d(\vec{a})$ describes the "local density" of the points after projection onto $\vec{a}$. The bigger the value of $Q(\vec{a})$ is,

Foundation item: The National Hi-Tech Research and Development Programa(973) of China(No. G19990436-05); the National Natural Science Foundation of China(No. 50279023); China Postdoctoral Science Foundation(No. 20040350277); * Corresponding author

the more remarkable the cluster will be.

The value of $s(\vec{a})$ is calculated by the variance of z_i, as Formula (3),

$$s(\vec{a}) = \left[\sum_{i=1}^{n}(z_i - \bar{z})^2/n\right]^{1/2}. \quad (3)$$

Here, $\bar{z}$ is the average value of z_i in axis $\vec{a}$.

On the other hand, $d(\vec{a})$ is calculated through the distance r_{ik} between two points' projection characteristic value. Let $r_{ik} = |z_i - z_k|\ (k = 1, \cdots, n)$, and $d(\vec{a})$ can be expressed by Formula (4),

$$d(\vec{a}) = \sum_{i=1}^{n}\sum_{k=1}^{n}(R - r_{ik})f(R - r_{ik}). \quad (4)$$

Here, $f(R - r_{ik})$ is a jump function, $f(R - r_{ik}) = 1$ when $R > r_{ik}$, or else $f(R - r_{ik}) = 0$. R is the parameter in PPC model, which is called cutoff radius according to Friedman's model (Friedman, 1974). There are indications that the scope of cutoff radius is $\max(r_{ik}) + \frac{m}{2} \leqslant R \leqslant 2m$, and let $R = m$ in practice(Wang, 2002b).

(4) Optimizing projection direction. The estimation of the optimum projection direction $\vec{a}$ is the key problem of PPC model. According to the analysis above, we know that $\vec{a}$ will be the right projection direction when Formula (2) reaches the maximum value. The issue to find $\vec{a}$ can be expressed by the following optimum problem,

$$\begin{cases}\max Q(\vec{a}) \\ \|\vec{a}\| = 1.\end{cases} \quad (5)$$

The problem described by Formula (5) can be resolved by many optimum methods; however, projection index function may be differentiable and continuous when projection direction is optimized with traditional optimum method. Because it is difficult to develop a suitable projection index function for some problems in practice, it may affect the development of projection pursuit technique. In order to overcome this issue, genetic algorithm is used to estimate optimum projection direction $\vec{a}$ in this paper (Zhang, 2001a).

(5) Comprehensive analysis. Because z_i is able to reflect the comprehensive feature of samples, we can analyze data feature and draw a right conclusions according to the discrepancy of z_i.

2 Application

A case study of water quality assessment will be given in the following section. Data on water quality standard and assessment sample (only one) are shown in Table 1 (Li, 1995).

First, establishing PPC model according to the data of water quality standard, where $n = 5$ (not 6), $m = 5$. We can estimate the parameter of PPC model and get the cluster standard about z_i (Table 1). Cluster standard based on z_i are: type Ⅰ ($z_i \leqslant 0.0861$), type Ⅱ ($0.0861 < z_i \leqslant 0.1235$), type Ⅲ ($0.1235 < z_i \leqslant 0.3012$), type Ⅳ ($0.3012 < z_i \leqslant 0.7679$), and type Ⅴ ($0.7679 < z_i \leqslant 2.235$).

Second, the projection characteristic value of observed sample could be computed using above model, $z = 0.2331$ (Table 1).

Table 1 Value of water quality standard and observed sample (mg/L)

Index	Water quality standard					Observed data
	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	
COD	1	2	4	8	20	2.1
NH_3-N	0.2	0.4	0.5	1	5	0.38
Volatile phenol	0.002	0.005	0.01	0.1	0.5	0.003
Cyanide	0.002	0.02	0.05	0.2	0.5	0
Total degree of limewater	30	40	80	160	300	106
z_i	0.0861	0.1235	0.3012	0.7679	2.235	0.2331

Finally, according to the comparison between z and z_i, we can draw a conclusion that the water quality type of observed sample is belong to type Ⅲ. It is consistent with that obtained by Li(Li, 1995).

3 Conclusions

The study in this paper led to four major conclusions: (1) an operational cluster model based on projection pursuit principle is purposed. The results showed that the PPC model is reasonable and functional. The cluster results according to the characteristics of data would be more objective and less subjective; (2) the empirical formula of estimating cutoff radius is developed, which is benefit for PPC model to be used in practice; (3) the genetic algorithm, which can overcome the shortcoming of traditional methods, is used to find the right projection direction in this paper. It can extend the application and dissemination of the PPC model; (4) there are some questions that need to have further investigation, e.g., the principle of projection index function section and the theory of cutoff radius estimation.

References:

Friedman J H, Tukey J W, 1974. A projection pursuit algorithm for exploratory data analysis[J]. IEEE Transactions on Computers, C-23: 881—890.

Jin J L, Wei Y M, Fu Q *et al.*, 2001. Projection pursuit model for comprehensive evaluation of agricultural productive capacity [J]. System Sciences and Comprehensive Studies in Agriculture, 17(4): 241—243.

Li Z Y, 1995. Study on comprehensive assessment of water quality using B-P network[J]. Environmental Engineering, 13(2): 51—53.

Wang S J, Zhang X L, Hou Y *et al.*, 2002a. Projection pursuit model for evaluating of flood events[J]. Hydrology, 22(4): 1—4.

Wang S J, Zhang X L, Ding J *et al.*, 2002b. Projection pursuit cluster model and its application[J]. Journal of Yangtze River Scientific Research Institute, 19(6): 53—55, 61.

Zhang X L, Ding J, Wang S J, 2001a. Projection pursuit method for assessing analogy basins[J]. Advances in Water Science, 12(3): 356—360.

Zhang X L, Ding J, Wang S J, 2001b. The use of projection trace in classification of Karstic water quality [J]. Geotechnical Investigation and Surveying, (5): 26—28.

Zhang X L, Wang S J, Ding J, 2002. Application of projection pursuit in environmental impact assessment of project management [J]. Systems Engineering - Theory & Practice, 22(6): 131—134.

(Received for review November 25, 2003. Accepted December 29, 2003)