Water Resour Manage (2008) 22:1421–1429
DOI 10.1007/s11269-007-9234-4

Application of Projection Pursuit Dynamic Cluster Model in Regional Partition of Water Resources in China

Shun-Jiu Wang · Chang-Jian Ni

Received: 24 November 2006 / Accepted: 12 November 2007 /
Published online: 12 December 2007

Abstract The present study develops a projection pursuit dynamic cluster (PPDC) model by combining dynamic cluster with projection pursuit to solve the problem of regional partition of water resources in China. The procedures of the PPDC model are described as follows. Firstly, a multi-factor cluster problem can be converted into a single-factor (projected characteristic value) cluster problem according to linear projection. Secondly, a new projection index on the basis of dynamic cluster rule is set up in the PPDC model, which successfully avoids the problem of parameter calibration and makes objective cluster results. Thirdly, genetic algorithm (GA) is applied to optimize projection direction of the PPDC model. Finally, the developed PPDC model is used in a case study of regional partition of water resources in China to evaluate its application. The cluster results of the PPDC model agree well with the actual regional partition of water resources in China, indicating that the PPDC model is a powerful tool in multi-factor cluster analyses and could be a new method for regional partition of water resources.

Keywords Projection pursuit · Dynamic cluster · Projection index · Genetic algorithm · Regional partition of water resources

1 Introduction

It is well known that multi-factor cluster analysis has been widely applied in many fields such as hydrology, ecology, etc. However, it is still a challenge for hydrologist, ecologist or other

S.-J. Wang (✉)
Institute of Plateau Meteorology, China Meteorological Administration, Chengdu 610071, People's Republic of China
e-mail: wsjbnu@163.com

C.-J. Ni
Department of Atmospheric Sciences, Chengdu University of Information Technology, Chengdu 610041, People's Republic of China

scientist to directly solve the multi-factor cluster problem. As a consequence, it is necessary to propose an efficient method that is able to transfer multi-factor problem into a single one.

To analyze multi-factor cluster, we need to solve a high-dimensional problem. Friedman and Tukey (1974) developed the projection pursuit (PP) principle, a powerful tool to deal with high-dimensional problem. The principle provided a way to find out a projection direction retaining the main characteristics of data based on a projection index. According to determined projection direction, high-dimensional question can be converted into low-dimensional question (e.g., three-dimension, two-dimension or one-dimension).

Based on projection pursuit principle, some new mathematical methods for exploratory analysis of high-dimensional data have been developed, such as projection pursuit regression (PPR) (Friedman and Stuetzle 1981), projection pursuit cluster (PPC) (Hall 1989), projection pursuit density estimation (PPDE) (Friedman et al. 1984), projection pursuit learning network (PPLN) (Hwang et al. 1994), and projection pursuit wavelet learning network (PPWLN) (Lin et al. 2003), etc.

PPC model is a powerful tool in multi-factor clustering or evaluating, and has been successfully applied in many engineering fields (Zhang et al. 2000; Wang et al. 2002, 2006). However, PPC model does have disadvantage in practice as follows: (1) Being the only parameter in PPC, the cutoff radius is hard to estimate, even though it has a significant effect on the results. Nowadays, we still set the cutoff radius based on experience as well as trial and error. For example, it is always assumed to be 10% of the square root of the data variance along the projection axis as suggested by Friedman and Tukey (1974); Wang et al. (2002) set an experienced equation to calculate the cutoff radius. So far, there is no theory or common formula to calibrate the cutoff radius. (2) The cluster results cannot directly be obtained from the output of PPC model. It only provides the projected characteristic value remaining the major characteristics of data according to the projection index. In other words, other approaches such as the method of scatter points should be used to re-analyze the projected characteristic value series in order to obtain the desired cluster results.

To solve the above mentioned issues of PPC model, a projection pursuit dynamic cluster (PPDC) model is proposed, which combines dynamic cluster rule with projection pursuit principle. In the PPDC model, a new projection index is set up according to dynamic clustering principle, which is the difference between the PPDC model and PPC model.

It is one of the key issues of the PPDC model to estimate the right projection direction. Genetic algorithm (GA) is used in this study to optimize the projection direction of the PPDC model.

Regional partition of water resources is a typical multi-factor cluster problem. As a case study, the PPDC model is applied to analyze the regional partition of water resources in China and to investigate the feasibility of this method.

2 Projection Pursuit Dynamic Cluster Model

Projection pursuit, a statistical method, was developed to analyze high-dimensional data based on projection (Friedman and Tukey 1974). This paper describes an issue of linear projection. High-dimensional data are projected onto one-dimensional space and their characteristics are analyzed through the one-dimensional space (Friedman and Tukey 1974).

If x_{ij}^0 ($i = 1, \cdots, n; j = 1, \cdots, m$. n is the number of samples; m is the number of cluster factors of the samples) is the initial value of the jth factor of the ith sample, the procedures of developing the PPDC model are described as follows.

2.1 Data Standardization

In order to eliminate the effect of different ranges of values of cluster factors, the initial data are standardized before it is used in the PPDC model. And the standardization formula used in this study is

$$x_{ij} = \left(x_{ij}^0 - x_{j\min}^0\right) \Big/ \left(x_{j\max}^0 - x_{j\min}^0\right) \tag{1}$$

where $x_{j\max}^0$, $x_{j\min}^0$ are the initial maximum and minimum of the j^{th} factor respectively.

2.2 Linear Projection

Essentially, projection is used to observe data characteristic from all angles. The main purpose of projection pursuit is to find the hidden structure from high-dimensional data sets by searching through all their low-dimensional projections (Cui 1997). If $\overrightarrow{a} = \left(a_1, a_2, \cdots, a_j, \ldots, a_m\right)^T$ is a m-dimensional unit vector and z_{i} is the projected characteristic value of x_{ij}, linear projection can be described as,

$$z_i = \sum_{j=1}^{m} a_j x_{ij} \tag{2}$$

where $\overrightarrow{a}$ is projection axis vector, and it is also called projection direction vector.

2.3 Projection Index

Cluster analysis is a tool for exploratory data analysis that tries to find the intrinsic structure of data by organizing patterns into groups or clusters (Har-even and Brailovsky 1995). In the PPDC model, a new projection index is generated on the basis of dynamic cluster principle (Ren and Wang 1999).

Define $s(z_i, z_k)$ $(k = 1, \cdots, n)$ as the absolute value of distance between the projected characteristic value z_i and z_k, namely $s(z_i, z_k) = |z_i - z_k|$.

Let $\Omega = \{z_1, z_2, \cdots, z_n\}$, and define

$$ss\left(\overrightarrow{a}\right) = \sum_{z_i, z_k \in \Omega} s(z_i, z_k) \tag{3}$$

Then, we assume that the all samples are classified as $p(2 \le p < n)$ clusters. Θ_h $(h = 1, 2, \cdots, p)$ is the projected characteristic value space of cluster h, which contains all the projected characteristic value of cluster h, and thus

$$\Theta_h = \{z_i | d(A_h - z_i) \le d(A_t - z_i), \forall\, t = 1, 2 \cdots, p, t \ne h\} \tag{4}$$

where $d(A_h - z_i) = |z_i - A_h|$ and $d(A_t - z_i) = |z_i - A_t|$, A_h and A_t is the initial cluster core of both cluster h and cluster t, respectively. In practice, we use the average projected characteristic value of clusters as new cluster core to conduct the iteration until the criterion is met (Ren and Wang 1999). We also define

$$d_h\left(\overrightarrow{a}\right) = \sum_{z_i, z_k \in \Theta_h} s(z_i, z_k) \tag{5}$$

and

$$dd\left(\overrightarrow{a}\right) = \sum_{h=1}^{p} d_h\left(\overrightarrow{a}\right) \tag{6}$$

Finally, the new projection index $QQ(\vec{a})$ in the PPDC model is as follow

$$QQ(\vec{a}) = ss(\vec{a}) - dd(\vec{a}) \tag{7}$$

The projection index $QQ(\vec{a})$ measures the degree to which the data points in the projection are both concentrated locally $(ss(\vec{a})\text{small})$ while, at the same time, expanded globally $(dd(\vec{a})\text{large})$ (Friedman and Tukey 1974).

2.4 Model Optimization

According to the above analysis, it can be found that the PPDC model can be expressed by

$$\begin{cases} \max QQ(\vec{a}) \\ \|\vec{a}\| = 1 \end{cases} \tag{8}$$

From formula (8), it is shown that the PPDC model reflects an optimum problem. As Friedman pointed out, the efficiency of PP strongly relies on the ability of the optimization

Table 1 Data points and cluster results

No.	Province or city	Index				Cluster results		
		(1)	(2)	(3)	(4)	p=3	p=4	p=5
1	Beijing	0.25	0.26	243	372.727	3	4	5
2	Tianjin	0.11	0.06	129	163.043	3	4	5
3	Hebei	1.67	1.46	126	378.171	3	4	5
4	Shanxi	1.15	0.95	92	484.848	3	4	5
5	Inner Mongolia	3.71	2.48	44	2,304.545	3	4	5
6	Liaoning	3.25	1.06	250	898.293	2	4	4
7	Heilongjiang	6.47	2.69	166	2,155.556	2	4	4
8	Shanghai	0.19	0.12	435	201.493	2	4	4
9	Jiangsu	2.49	1.15	319	470.333	2	4	4
10	Zhejiang	8.85	2.13	881	2,071.594	1	3	2
11	Anhui	6.17	1.67	485	1,161.235	2	3	3
12	Fujian	11.68	3.06	963	3,758.842	1	2	2
13	Jiangxi	14.16	3.23	852	3,636.829	1	2	2
14	Shandong	2.64	1.54	219	389.082	2	4	4
15	Henan	3.11	1.99	244	460.497	2	4	4
16	Hubei	9.46	2.91	528	1,758.065	2	3	3
17	Hunan	16.20	3.75	768	2,599.042	1	2	2
18	Guangdong	18.01	4.67	1021	2,814.241	1	2	2
19	Guangxi	18.80	3.98	791	4,292.237	1	2	2
20	Hainan	3.10	0.79	927	4,647.059	1	3	3
21	Sichuan	31.31	8.02	552	3,917.500	1	2	2
22	Guizhou	10.35	2.59	588	3,080.357	1	3	3
23	Yunnan	22.21	7.38	579	5,798.956	1	2	2
24	Xizang	44.82	10.94	373	203,727.273	1	1	1
25	Shaanxi	4.20	1.65	215	1,323.353	2	4	4
26	Gansu	2.73	1.33	69	1,186.147	3	4	5
27	Qinghai	6.23	2.58	87	13,608.696	3	4	4
28	Ningxia	0.085	0.16	19	208.333	3	4	5
29	Xinjiang	7.93	5.80	54	5,696.774	2	3	4

algorithm to find substantive optima of the projection index among a forest of dummy optima caused by sampling fluctuations (Friedman 1987). Therefore, an efficient algorithm is one of the key issues of the PPDC model.

Holland (1975) has put forward GA as an optimization method, which apply the concept on the artificial survival of the fittest coupled with a structured information exchange using randomized genetic operators taken from the nature (Chau and Albermani 2003). GA is globally oriented in searching and thus useful in optimizing the large and complex problems, especially where the objective functions are ill-defined (Cheng et al. 2006). GA has been widely applied to solve water resources system optimization such as rainfall–runoff model calibration (Cheng et al. 2002, 2005, 2006), Nash model parameter estimation (Dong 2007), water quality assessment (Zhang et al. 2000), flow and water quality model calibration (Chau 2002), complex reservoir system optimal operation (Cheng et al. 2007), and so on. Also, GA has been successful used in other fields (Chau 2004; Chau and Albermani 2003; Yu et al. 2005). GA is just used to estimate the right projection direction in this study (Wang et al. 2006).

(1) Encoding. The encode methods can be classified as binary encoding and real number encoding. In order to avoid frequent binary-to-decimal conversion and reduce the computational cost, real number encoding is used in this study. If $\overrightarrow{g}_0^h$ (Let $g \in [0, 1]$; $h = 1, \cdots, N$; N is the number of forerunner population) is the hth group m-dimensions random real number, we are going to encode $\overrightarrow{a}^h = -1 + 2 \cdot \overrightarrow{g}_0^h$, where $\overrightarrow{g}_0^h$ are called forerunner individuals.

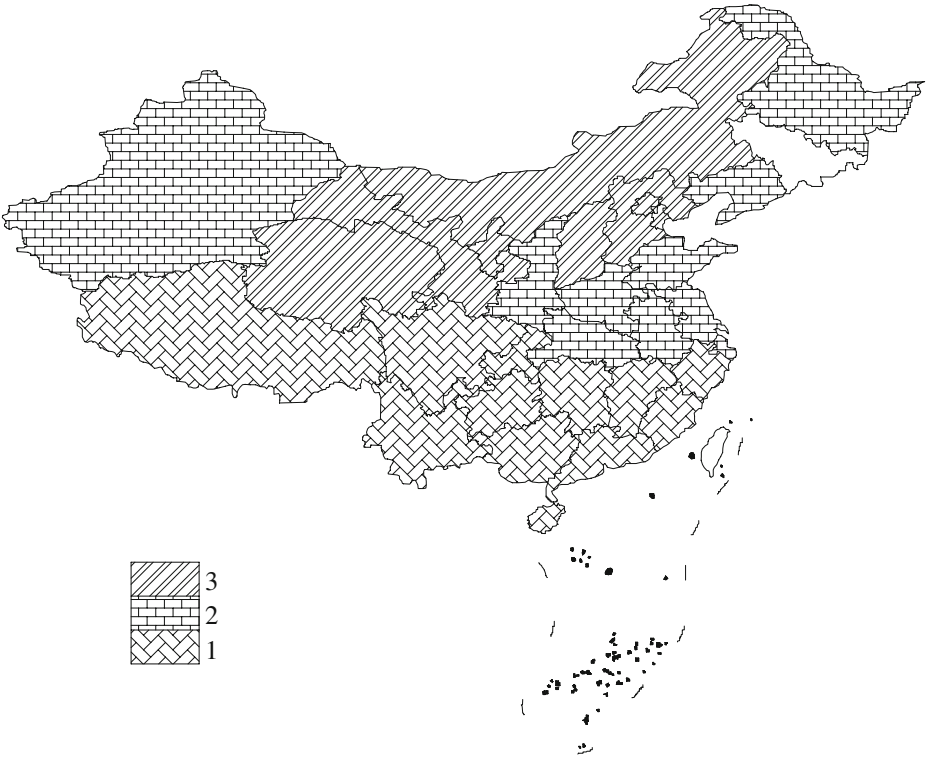

Fig. 1 Schematic diagram of regional partition of water resources in China (Three clusters)

(2) Fitness evaluation. According to formula (1) to (7), $QQ\left(\overrightarrow{a}^h\right)$ are calculated. Also, individual fitness can be evaluated according to $QQ\left(\overrightarrow{a}^h\right)$. The larger the values of $QQ\left(\overrightarrow{a}^h\right)$ are, the better the individual fitness will be.

(3) Selection. Let $y_h = 1 - \frac{1}{QQ\left(\overrightarrow{a}^h\right) \cdot QQ\left(\overrightarrow{a}^h\right) + 0.0001}$, where y_h are selection probabilities.

The number "0.0001" is introduced here to avoid zero in denominator. According to the values of y_h, we can select the hth individual in order to produce the first generation of individuals $\overrightarrow{g}^h_1$.

$$\begin{cases} \overrightarrow{g}^h_1 = \overrightarrow{g}^h_0, y_h \geq \mu_1 \\ \overrightarrow{g}^h_1 = \overrightarrow{\mu}_2, y_h < \mu_1 \end{cases}, \text{where } \mu_1 \text{ is a random numbers, } \overrightarrow{\mu}_2 \textit{ is a } m - \text{dimensions random series.}$$

(4) Crossover and Recombination. To guarantee the multi-variety of individuals, the second generation of individuals $\overrightarrow{g}^h_2$ is going to be generated through the random linear combination as the following:

$$\begin{cases} \overrightarrow{g}^h_2 = \mu_4 \overrightarrow{g}^{h_1}_1 + (1-\mu_4)\overrightarrow{g}^{h_2}_1, \mu_3 < 0.5 \\ \overrightarrow{g}^h_2 = \mu_5 \overrightarrow{g}^{h_1}_1 + (1-\mu_5)\overrightarrow{g}^{h_2}_1, \mu_3 \geq 0.5 \end{cases}, \text{where } h_1, h_2 \leq N, \text{ and } \mu_3, \mu_4, \mu_5 \text{ are random numbers.}$$

(5) Mutation. For forerunner individual $\overrightarrow{g}^h_0$, the smaller the values of $QQ\left(\overrightarrow{a}^h\right)$ are, the smaller the selection probabilities will be, whereas, the larger the mutation

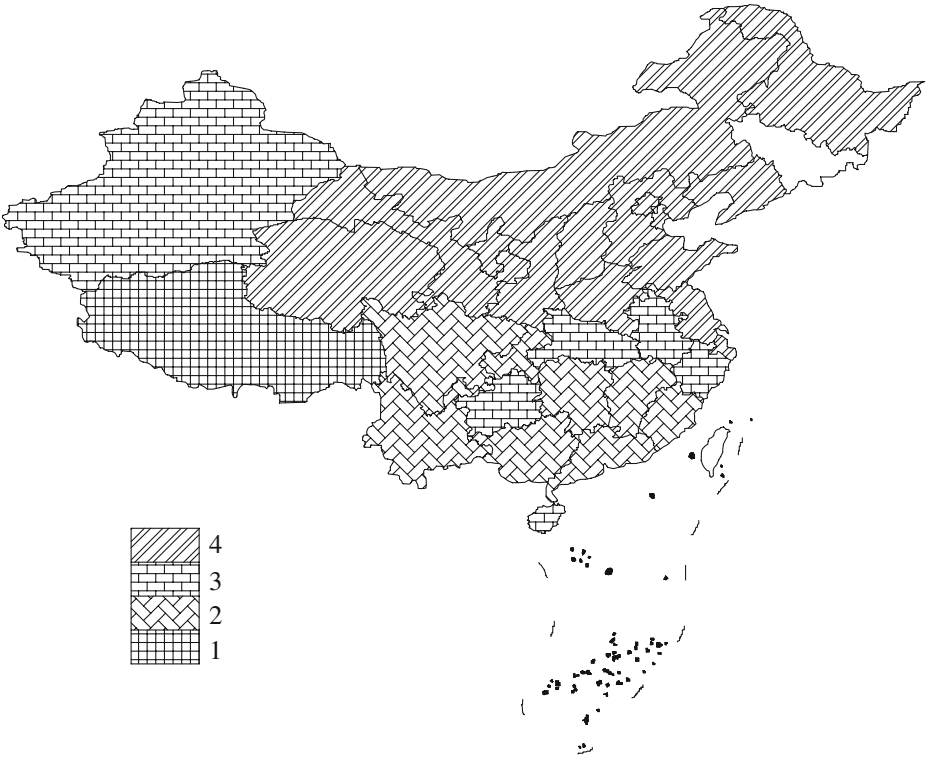

Fig. 2 Schematic diagram of regional partition of water resources in China (Four clusters)

probabilities will be. Let $y_h^{'} = 1 - y_h$, where $y_h^{'}$ are called mutation probabilities. Therefore, the third generation of individuals $\overrightarrow{g}_3^h$ can be produced as the following:

$$\begin{cases} \overrightarrow{g}_3^h = \overrightarrow{\mu}_7, y_h^{'} \geq \mu_6 \\ \overrightarrow{g}_3^h = \overrightarrow{g}_0^h, y_h^{'} < \mu_6 \end{cases}, \text{where } \mu_6 \text{ is a random number, } \overrightarrow{\mu}_7 \text{ is a } m - \text{dimensions random series.}$$

(6) Evolutionary iteration. Consider $\overrightarrow{g}_1^h$, $\overrightarrow{g}_2^h$, and $\overrightarrow{g}_3^h$ as a new population, repeat the processes from step 2 until computational error meets the criteria, and get the right projection direction $\overrightarrow{a}$ at long last.

3 Application

The PPDC model is used to conduct regional partition of water resources for 29 administrative regions in China. The index system includes four factors: (1) annual total river runoff ($10^9 m^3$), (2) annual gross amount of groundwater ($10^9 m^3$), (3) average annual water producing modulus ($10^9 m^3$ hm $^{-2}$) and (4) average annual per capita water resource (m^3 person^{-1}). The factor statistic values are shown in Table 1 (Liu and Fu 2000).

In order to comparative analysis, water resources can be partitioned by three cases, namely three clusters, four clusters and five clusters.

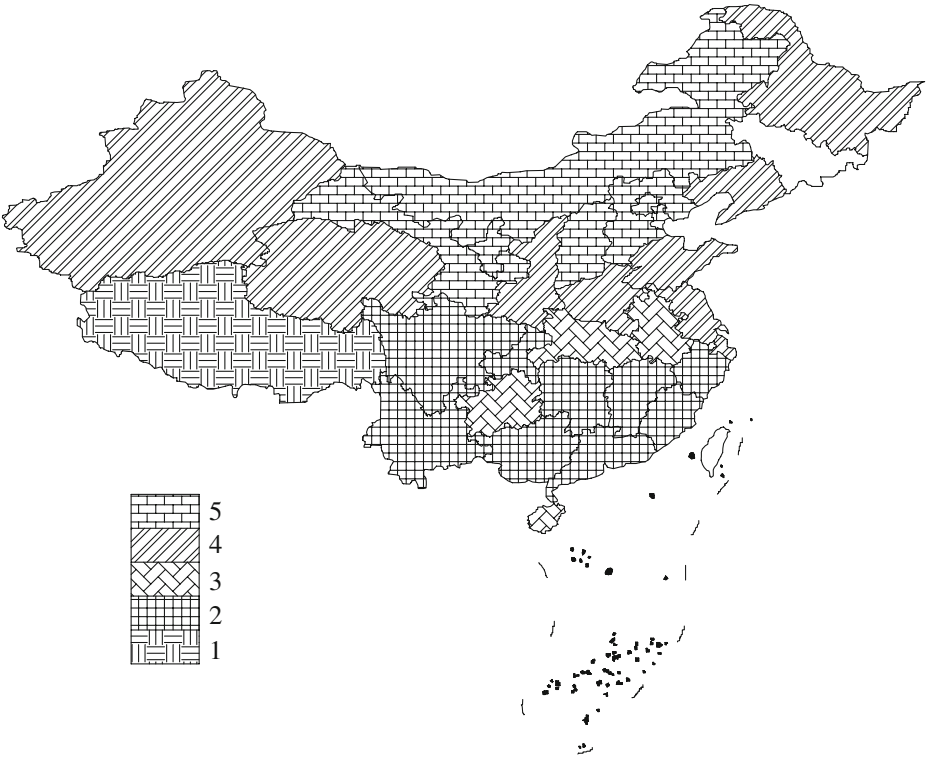

Fig. 3 Schematic diagram of regional partition of water resources in China (Five clusters)

Let $m = 4$, $n = 29$, $p = 3$, 4 and 5. We can get the right projection direction on the basis of the PPDC model as follows,

$$\vec{a} = \begin{cases} (0.4507, 0.3882, 0.7948, 0.1203)^T, & \text{for} \quad p = 3 \\ (0.5582, 0.5647, 0.5635, 0.2281)^T, & \text{for} \quad p = 4 \\ (0.4936, 0.4400, 0.7273, 0.1840)^T, & \text{for} \quad p = 5 \end{cases}$$

The cluster results are also shown in Table 1. The schematic diagram of water resources partition of China is shown in Figs. 1, 2 and 3.

The main results of this case study are as follows: (1) water resources in the southern part of China is richer than that in the northern part of China. Xizang, Guangdong and Sichuan rank the first three regions in richest regions of water resources in China; (2) the most serious in water resources shortage exist in north China and Gansu Panhandle. Ningxia is the most serious shortage area, while Tianjin, Shanxi, and Gansu are followed, respectively; (3) the cluster results reflect the actual situation of water resources of China. Many rivers such as Yangtze River, Ya-lu-tsang-pu River, Nujiang-Salween River, Lancangjiang-Mekong River, and Pearl River just rise or run through in the southern part of China, as a consequence, there are abundant water resources. South-to-North Water Transfer Project, which is going on right now, is one of the efficient ways to improve the water resources distribution of north China; (4) the PPDC model is a new method for regional partition of water resources.

4 Conclusions

Based on projection pursuit and dynamic cluster, the PPDC model is proposed. The PPDC model makes the cluster results more objective and easier to operate in practice than that of PPC model because there is no parameter calibration. The study demonstrates that the PPDC model is: (1) a new cluster method on the basis of projection pursuit principle; (2) a powerful tool for multi-factor cluster analysis; (3) able to output the cluster results directly; (4) practicable for regional partition of water resources. However, there are still some problems that need to have further investigation for the PPDC model, e.g., (1) the mathematical theory of the PPDC model needs to be researched, (2) the feasibility of the PPDC model needs to be tested and verified with more complicated high-dimension cases, and (3) the application of the PPDC model in multi-factor assessment or other fields needs to be improved.

Acknowledgements This work is part of the Program of China Meteorological Administration (CCSF2007-23), Institute of Plateau Meteorology of China Meteorological Administration (BROP200701, LPM2005014 and PMP2006005) and Sichuan Meteorological Bureau (2006-2). The constructive comments and suggestions from the editor and anonymous reviewers, which resulted in a significant improvement of the manuscript, are gratefully appreciated. The opinions expressed here are those of the authors and not those of other individuals or organizations.

References

Chau KW (2002) Calibration of flow and water quality modeling using genetic algorithm. Lecture Notes in Artificial Intelligence 2557:720

Chau KW (2004) A two-stage dynamic model on allocation of construction facilities with genetic algorithm. Autom Constr 13(4):481–490

Chau KW, Albermani F (2003) Knowledge-based system on optimum design of liquid retaining structures with genetic algorithms. J Struct Eng, ASCE 129(10):1312–1321
Cheng CT, Ou CP, Chau KW (2002) Combining a fuzzy optimal model with a genetic algorithm to solve multiobjective rainfall–runoff model calibration. J Hydrol 268(1–4):72–86
Cheng CT, Wu XY, Chau KW (2005) Multiple criteria rainfall–runoff model calibration using a parallel genetic algorithm in a cluster of computer. Hydrol Sci J 50(6):1069–1087
Cheng CT, Zhao MY, Chau KW, Wu XY (2006) Using genetic algorithm and TOPSIS for Xinanjiang model calibration with a single procedure. J Hydrol 316(1–4):129–140
Cheng CT, Wang WC, Xu DM, Chau KW (2007) Optimizing hydropower reservoir operation using hybrid genetic algorithm and chaos. Water Resources Management. DOI 10.1007/s11269-007-9200-1
Cui HJ (1997) The laws of the iterated logarithm for two kinds of PP statistics. Stat Probab Lett 32(3):235–243
Dong SH (2007) Genetic algorithm based parameter estimation of Nash model. Water Resources Management. DOI 10.1007/s11269-007-9208-6
Friedman JH (1987) Exploratory projection pursuit. J Am Stat Assoc 82:249–266
Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. IEEE Trans Comput C-23:881–890
Friedman JH, Stuetzle W (1981) Projection pursuit regression. J Am Stat Assoc 76:817–823
Friedman JH, Stuetzle W, Schroeder A (1984) Projection pursuit density estimation. J Am Stat Assoc 79:599–608
Hall P (1989) On polynomial-based projection indices for exploratory projection pursuit. Ann Stat 17 (2):589–605
Har-even M, Brailovsky VL (1995) Probabilistic validation approach for clustering. Patter Recogn Lett 16 (11):1189–1196
Holland JH (1975) Adaptation in natural and artificial system. University of Michigan Press, Ann Arbor, Michigan
Hwang JN, Lay SR, Maechler M, Martin RD, Schimert J (1994) Regression modeling in back-propagation and projection pursuit learning. IEEE Trans Neural Netw 5(3):342–353
Lin W, Tian Z, He F (2003) On improving unsupervised restoration of image with PPWLN. Journal of Northwestern Polytechnical University 21(3):344–347 (in Chinese)
Liu CM, Fu GB (2000) Water world today. Tsinghua University Press, Beijing (in Chinese)
Ren RE, Wang HW (1999) Multi-dimensional statistics data analysis – theory, method and practice. National Defence Industry Press, Beijing (in Chinese)
Wang SJ, Zhang XL, Ding J et al (2002) Projection pursuit cluster model and its application. Journal of Yangtze River Scientific Research Institute 19(6):53–55 (in Chinese)
Wang SJ, Zhang XL, Yang ZF, Ding J, Shen ZY (2006) Projection pursuit cluster model based on genetic algorithm and its application in Karstic water pollution evaluation. Int J Environ Pollut 28(3–4):253–260
Yu B, Cheng CT, Yang ZZ, Chau KW (2005) Application of PGA on optimization of distribution of shopping centers. Lecture Notes in Artificial Intelligence 3673:576–586
Zhang XL, Ding J, Li ZY, Jin JL (2000) Application of new projection pursuit algorithm in assessing water quality. China Environ Sci 20(2):187–189 (in Chinese)