DOI:10.11705/j.issn.1672-643X.2014.06.033

基于投影寻踪分类的长江流域水质综合评价模型及其应用模型

熊聘[1]，楼文高[1,2]

(1.上海理工大学 光电信息与计算机工程学院，上海 200093；2.上海商学院，上海 200235)

摘 要：为了研究长江水系的水质情况及发展趋势，根据其103个国控断面2006-2010年8项水质监测指标数据(共计515组样本)和地下水质量评价标准，建立了水质综合评价的投影寻踪(PPC)模型。研究结果表明：长江水系国控断面上整体水质较好，其中Ⅰ类和Ⅱ类水质分别占60%和22.7%；上游水质总体上好于下游水质；化工工业较发达地区的支流和河流水污染情况严重，2006-2009年，水质逐年改善，但2010年水质最差。研究了不同投影窗口半径R值对PPC建模结果的影响规律，理论和实证研究表明，$R=0.1S_z$的较小值方案是不合理的，$r_{max}\leqslant R\leqslant 2p$的较大值方案是错误的，$R$不宜取常数，只有$r_{max}/5\leqslant R\leqslant r_{max}/3$的中间适度值方案是合理的。

关键词：水质综合评价；投影寻踪建模；投影窗口半径R值；长江水系

中图分类号：X234；TP391 **文献标识码**：A **文章编号**：1672-643X(2014)06-0156-07

Water quality comprehensive evaluation and application model based on projection pursuit clustering in Yangtze river basin

XIONG Pin[1], LOU Wengao[1,2]

(1. *School of Optical-Electrical Information and Computer Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*; 2. *Shanghai Business School, Shanghai* 200235, *China*)

Abstract: In order to study the water quality and its changing trend of the Yangtze river(YR) and according to the eight-index measured data on 103 government-control sections of YR from 2006 to 2010 (total 515 samples) and the standards of quality evaluation of groundwater, the paper established the projection pursuit clustering model of comprehensive evaluation of water quality. The results show that the water quality on the control sections is quite good, the water quality of type Ⅰ and Ⅱ take up 60% and 22.7% respectively. The water quality in upperstream is better than that in downstream. The water pollution is quite serious in the Chemical-industry developed areas. From 2006 to 2009, the water quality was gradually improved year by year, and was the worst in 2010. The paper researched the rule of influence of cutoff radius R value (CRRV) on the results of PPC model . The theory analysis show that the scheme of taking the smaller CRRV ($R=0.1S_z$) is absolutely unreasonable, the larger CRRV ($r_{max}\leqslant R\leqslant 2p$) is error, should not be the constant, and only the moderate-suitable value of CRRV ($r_{max}/5\leqslant R\leqslant r_{max}/3$) is reasonable correct.

Key words: water quality comprehensive evaluation; projection pursuit clustering (PPC) modelling; cutoff radius R value; Yangtze River

长江是我国第一、世界第三大河流，也是中华民族的母亲河，分布着全国三分之一的人口、GDP、粮食产量和水资源以及40%的水能资源。长江流域横贯中国大陆中部，跨越东中西3个经济带，连接南北，在我国国民经济的发展中占有非常重要的战略地位。近年来长江流域水质污染日趋严重，长江流域的功能在逐步退化，已引起相关部门和专家的高度重视。为了对长江流域未来的环境规划、水资

收稿日期：2014-07-18； **修回日期**：2014-09-26

基金项目：上海高校知识服务平台“上海商贸服务业知识服务中心”项目(ZF1226)；上海市重点学科“商务经济学”项目

作者简介：熊 聘(1986-)，男，湖北武汉人，硕士研究生，主要研究方向为投影寻踪建模、数据挖掘、数据库和软件工程。

通讯作者：楼文高(1964-)，男，浙江杭州人，博士，教授，主要从事人工智能等数据挖掘技术在环境工程与科学、管理科学与工程等领域的应用。

源利用与管理等提供更加科学的决策依据,对流域的水质进行正确的评价与预测是很有必要的。

目前已有多种水质综合评价方法,如指数评价法、模糊数学法、物元分析法、集对分析法、灰色评价法、Logistic 模型、支持向量机(SVM)、主成分分析(PCA)、人工神经网络(ANN)模型及其 PCA – ANN 组合评价方法等,这些方法都具有各自的优缺点[1-6],如前 5 种方法必须由其他方法确定各评价指标的权重,否则就是等权重,结果的合理性难以保证,而且是以半定量研究为主;PCA 方法不仅需要大量的样本,而且评价结果随样本不同而不同[7],ANN 模型在遵循基本原则的前提下可以取得较好的效果,但建模过程十分繁琐,需要人为确定多种现象和合理参数[5,8-9],存在一定的主观性;SVM 模型也需要人为判定两个合理的参数。与此同时,2000 年以来,由 Friedman 等[10]提出的可用于高维、非线性、非正态分布数据建模的一维投影寻踪分类(PPC)模型在国内已被广泛应用[11-15]。由于涉及到如何确定合理的窗口半径 R 值,应用中也出现了不少问题。为此,本文选取《中国环境年鉴》中长江水系 103 个国控监测断面 2006 –2010 年的主要水质指标监测数据,根据地表水环境质量标准(GB3832 – 2002)[16](简称:水质评价标准)中 DO、COD_{Mn}、Hg 和 NH_3-N 等 8 种主要污染物指标,应用 PPC 建模技术,对长江流域主要省市分界断面及其城市的水质进行可靠性综合评价和变化趋势的预测。同时,研究了由于选取不合理的 R 值将可能导致错误,提出了选取合理 R 值范围的正确方案和原则。

1　投影寻踪水质评价原理

1.1　长江流域水质监测指标及其我国地下水环境质量标准

《中国环境年鉴》(2007 –2011)给出了长江水系国控断面 9 个主要监测指标 2006 –2010 年的年均值数据。因为 I ~ V级水质的 pH 值均为 6 ~9,故我们只选取 DO、COD_{Mn}、BOD_5、NH_3-N、Hg、Pb、挥发酚和石油类共 8 个主要监测指标数据来综合评价长江水系的水质。同时,根据五级水质的指标值及其变化规律或者趋势以及长江水系水质指标实测值的范围确定了水质评价标准中各指标的最大值(10.6, 20, 15, 5, 0.35, 0.15, 1, 1.5)和最小值(0.5, 0.5, 0.5, 0.01, 0.00001, 0.0003, 0.001, 0.001)。

1.2　水质综合评价投影寻踪建模原理简介

根据水质评价标准[16],可用一维 PPC 技术[11-15]建立长江水系水质综合评价模型。PPC 建模基本思想是“使样本投影点整体上尽可能分散,局部上尽可能密集”[10-15],即目标函数为:

$$Q(a) = max(S_z D_z)$$

$$\text{s.t.} \sum_{j=1}^{p} a_j^2 = 1, \quad 1 \geqslant a_j \geqslant -1 \tag{1}$$

样本投影值标准差 S_z 与局部密度值 D_z 为:

$$\begin{cases} S_z = \sqrt{\{\sum_{i=1}^{n} [z(i) - E(z)]^2\}/(n-1)} \\ D_z = \sum_{i=1}^{n} \sum_{k=1}^{n} [R - r_{i,k}] \cdot u[R - r_{i,k}] \end{cases} \tag{2}$$

式中:S_z 越大表示投影点整体上越分散; D_z 越大表示投影点局部越密集。其他符号意义参见文献[10 – 15]。

由(1)和(2)式可知,投影密度窗口半径 R 值显著影响其建模结果 —— 最佳投影向量以及样本投影值,楼文高等[15]在研究了投影密度窗口半径 R 值的本质及其对建模结果的影响后提出了 R 值的合理范围应该为 $r_{max}/5 \leqslant R \leqslant r_{max}/3$。由于(1)式为复杂的高维非线性并含不等式约束的最优化问题,求解十分困难,本文采用基于混沌的人工蜂群算法进行最优化求解。

2　长江水系水质的 PPC 建模结果

为了消除不同水质指标量纲相差很大对建模结果的不利影响,本文首先对各指标值进行极大极小值线性归一化预处理,并且对 DO 进行了正向化处理(即所有指标归一化值越大,PPC 模型的输出值也越大,表示水质越差),从而把各个评价指标的值都转化到[0, 1]范围内。调用笔者采用 Matlab7.0 编制的基于混沌的人工蜂群算法 PPC 程序,得到了 $R = r_{max}/5$ 时的 PPC 建模结果,最佳投影向量 $\alpha_{1\sim8}$ = (0.134, 0.253, 0.314, 0.344, 0.500, 0.287, 0.490, 0.362),各样本的投影值 $z(1) \sim z(7)$(0, 0.155, 0.225, 0.398, 0.667, 1.126, 2.681),样本值标准差 S_z = 0.930,局部密度值 D_z = 8.585,目标函数值 $Q(a)$ = 7.985,投影窗口半径 R = 0.536,最大窗口半径 r_{max} = 2.681。从各个投影值的意义可知,I ~ 劣 V 类水质的 PPC 模型输出值范围分别为(0, 0.155]、(0.155, 0.225]、(0.225, 0.398]、(0.398, 0.667]、(0.667, 1.126]、(1.126, 2.681]和大于 2.681。

《中国环境年鉴》给出了长江水系 103 个国控断面 8 个主要监测指标的值,首先把这些值进行归

一化处理,并代入上述建立的 PPC 模型,就可以得到各个国控断面 2006 – 2010 年的水质综合评价结果 $z(i)$,对照上述各级水质的 PPC 模型输出值范围,就可以很方便地判定各断面每年的综合水质类别。限于篇幅,本文仅列出各省市交界主要国控断面以及苏州、上海等共计 22 个断面的 PPC 模型输出值及其综合评价水质类别(如表 1 所示),图 1 所示是各断面的 PPC 模型输出值与各类水质的区间分布图。

表 1 长江水系主要国控断面 PPC 模型输出值、水质综合评价结果及其类别

断面	地区	断面	三峡	省界	2006		2007		2008		2009		2010	
代码	名称	名称	库区	断面	PPC 值	类别	PPC 值	类别	PPC 值	类别	PPC 值	类别	PPC 值	类别
1	攀枝花	龙洞	上游区	滇 – 川	0.096	Ⅰ	0.089	Ⅰ	0.132	Ⅰ	0.130	Ⅰ	0.122	Ⅰ
2	水富	铁路桥	上游区	滇 – 川	0.221	Ⅱ	0.098	Ⅰ	0.133	Ⅰ	0.096	Ⅰ	0.095	Ⅰ
3	重庆	朱沱	影响区	川 – 渝	0.261	Ⅲ	0.244	Ⅲ	0.240	Ⅲ	0.205	Ⅱ	0.193	Ⅱ
4	重庆	培石	库区	渝 – 鄂	0.193	Ⅱ	0.177	Ⅱ	0.204	Ⅱ	0.208	Ⅱ	0.177	Ⅱ
5	岳阳	城陵矶		湘 – 鄂	0.181	Ⅱ	0.152	Ⅰ	0.141	Ⅰ	0.178	Ⅱ	0.183	Ⅱ
6	九江	姚港		赣 – 鄂	0.129	Ⅰ	0.148	Ⅰ	0.142	Ⅰ	0.134	Ⅰ	0.134	Ⅰ
7	安庆	皖河口		赣 – 皖	0.134	Ⅰ	0.137	Ⅰ	0.129	Ⅰ	0.103	Ⅰ	0.111	Ⅰ
8	南京	江宁河口		皖 – 苏	0.144	Ⅰ	0.180	Ⅱ	0.152	Ⅰ	0.149	Ⅰ	0.154	Ⅰ
9	南通	姚港		苏 – 沪	0.172	Ⅱ	0.180	Ⅱ	0.159	Ⅱ	0.149	Ⅰ	0.143	Ⅰ
10	遂宁	老池	上游区	川 – 渝	0.159	Ⅱ	0.106	Ⅰ	0.176	Ⅱ	0.164	Ⅱ	0.144	Ⅰ
11	广元	八庙沟	上游区	甘 – 川	0.194	Ⅱ	0.073	Ⅰ	0.099	Ⅰ	0.099	Ⅰ	0.114	Ⅰ
12	岳池	赛龙乡	上游区	川 – 渝	0.244	Ⅲ	0.143	Ⅰ	0.202	Ⅱ	0.183	Ⅱ	0.176	Ⅱ
13	铜仁	沿河	上游区	黔 – 渝	0.263	Ⅲ	0.174	Ⅱ	0.179	Ⅱ	0.186	Ⅱ	0.134	Ⅰ
14	重庆	利泽	影响区	川 – 渝	0.153	Ⅰ	0.145	Ⅰ	0.159	Ⅱ	0.114	Ⅰ	0.139	Ⅰ
15	赤水	鲢鱼溪	影响区	黔 – 川	0.214	Ⅱ	0.242	Ⅲ	0.170	Ⅱ	0.169	Ⅱ	0.127	Ⅰ
16	武都	绸子坝		甘 – 川	0.143	Ⅰ	0.137	Ⅰ	0.180	Ⅱ	0.149	Ⅰ	0.123	Ⅰ
17	十堰	羊尾		陕 – 鄂	0.118	Ⅰ	0.095	Ⅰ	0.104	Ⅰ	0.123	Ⅰ	0.219	Ⅱ
18	滁州	汊河		皖 – 苏	0.912	Ⅴ	0.753	Ⅴ	0.683	Ⅴ	0.967	Ⅴ	0.423	Ⅳ
19	南阳	梅湾		豫 – 鄂	0.255	Ⅲ	0.252	Ⅲ	0.215	Ⅱ	0.265	Ⅲ	0.249	Ⅲ
20	南阳	新甸铺		豫 – 鄂	0.545	Ⅳ	0.463	Ⅳ	0.450	Ⅳ	0.353	Ⅲ	0.290	Ⅲ
21	苏州	轻化仓库			0.951	Ⅴ	0.676	Ⅴ	0.503	Ⅳ	0.574	Ⅳ	0.503	Ⅳ
22	上海	吴淞口			0.336	Ⅲ	0.294	Ⅲ	0.322	Ⅲ	0.480	Ⅳ	0.461	Ⅳ

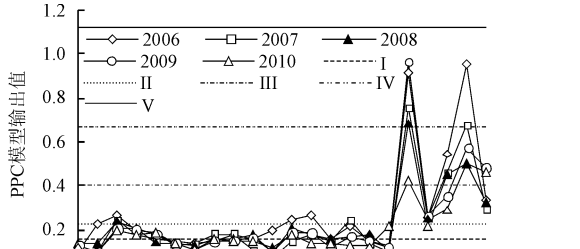

图 1 22 个国控断面及其水质评价标准分界值的 PPC 模型输出值和水质综合评价结果(2006 – 2010 年)

从表 1 可以看出,长江水系的综合水质呈现以下特点:①绝大部分国控断面的综合水质以Ⅰ类和Ⅱ类为主,在 110 个样本(22 个断面、5 年)的评价结果中,Ⅰ类水质 49 个,占 44.5%,Ⅱ类水质 31 个,占 28.2%,合计占 72.7%;②上游的整体综合水质好于下游的水质(如安徽、江苏和上海的整体水质较差);③大城市的水质相对较差(如重庆、苏州、上海的整体水质较差);④在 2006 – 2010 年期间分别有 7、5、4、4 和 2 个断面的综合水质最差,也就是说,长江水系在这些断面上的整体水质在向好的方向变化,沿长江流域各省市或者城市的水污染治理工程取得了初步成效,如铜仁(黔 – 渝断面)、赤水(黔 – 川断面)水质从Ⅲ类变为Ⅰ类,苏州从Ⅴ类变为Ⅳ类,但上海市吴淞口的水质明显变差,从Ⅲ类变为Ⅳ类;⑤长江水系流经某些省市后,综合水质明显变差,如四川、安徽、湖北、江苏和上海等。

3 结果与讨论分析

3.1 水质各评价指标的重要性分析

从 PPC 模型各评价指标的权重可以看出,8 个指标都是重要的,对综合水质都有重要影响,但从影

响程度来看,Hg 的影响最大,其次是挥发酚,他们两者基本相当,然后是石油类、NH_3-N、BOD_5、Pb、COD_{Mn}等,DO 的影响最小,最大权重与最小权重之比达 3.7 倍,差异显著。因此,从有效改善综合水质的角度来看,沿长江各地区应采取有效措施,首先减少 Hg 和挥发酚的排放,并加强控制,尽可能减少石油类、氨氮、BOD_5 和 Pb 等的排放,这样可以起到事半功倍的效果。

3.2　长江水系的整体水质情况及其变化趋势预测

从表 1 和所有 103 个国控断面 2006 – 2010 五年的监测数据(一共 515 个样本)的评价结果可以看出,长江水系的整体水质呈现如下主要特点:①长江水系的综合水质整体上处于较好的状态,以Ⅰ类和Ⅱ类水质为主,其中Ⅰ类水质占 60.0%,Ⅱ类水质占 22.7%,Ⅲ类水质占 10.1%,Ⅳ、Ⅴ和劣Ⅴ类水质各仅占 4.3%、1.5% 和 1.4%;②整体上讲,上游地区(断面)的水质好于下游地区,大城市(断面)的水质比其他区域断面的水质差;③近年来,虽然在不同省市分界的国控断面上综合水质有所好转,表示各省市都十分重视水污染治理工程,并取得了较好的效果,但从所有 103 个国控断面来看,2006 – 2010 年期间分别有 24、16、21、16 和 26 个断面的水质最差,表明 2007 – 2009 年水质较好,这个结果是否表明各省市对水环境污染治理出现了放松的现象,值得沿长江流域各省市环保部门和企业的反思和警惕;④从区域来看,长江在流经贵州省(湘江、横江、清水河和大河等)、四川省(岷江、釜溪河、沱江)、江西省(昌江等)、安徽省(滁河等)、湖北省(唐河、白河等)和江苏省(京杭大运河、外秦淮河等)以及部分支流的水系水质较差,这些区域或支流将成为长江水系水环境污染治理的难点和重点。这些区域由于化工工业(尤其是酿酒业、造纸业、皮革加工业等)非常发达,人口密度又高,水体中 Hg、石油类、挥发酚和氨氮等的含量都较高,从而导致水质大多差于Ⅲ类,有些断面甚至出现了Ⅳ类和Ⅴ类水质,在岷江、釜溪河、螳螂川、大河和横江的部分水域的国控断面甚至出现了劣Ⅴ类水质的情况。

3.3　投影窗口半径 R 值对 PPC 建模结果和长江水系水质评价结果的影响

由公式(1)、(2)知,R 值将显著影响 PPC 模型的最优化结果,而且不同学者提出了选取 R 值的多种方案,如:①Friedman 等提出较小值方案[10-11, 13],取 $R = 0.1\ S_z$ 或者 $R = (0.01 \sim 0.001)S_z$[14];②王顺久等提出较大值方案[12],取 $r_{\max} \leqslant R \leqslant 2p$[12-13],通常取 $R = p$;③楼文高等提出中间适度值方案[15],取 $r_{\max}/5 \leqslant R \leqslant r_{\max}/3$。

上述三种不同 R 值方案在本质上存在较大差异,采用较小值方案时投影窗口内只包含很少的样本点,而较大值方案时投影窗口内包含了所有的样本点,中间适度值方案时窗口内样本点既不太多,也不太少。

为了研究 R 值对 PPC 建模的影响规律,分别取 $R \leqslant 0.001S_z$、$0.002S_z$、$0.01S_z$、$0.1S_z$、$0.25S_z$、$0.5S_z$、$r_{\max}/5$、$r_{\max}/3$、$r_{\max}/2$、$r_{\max}$、$2r_{\max}$ 和 $p(=8)$,PPC 建模结果如表 2 所示。对于本例数据,研究发现如下规律:

(1) R 从 $0.001S_z$ 到 $0.002S_z$ 的区间内,最佳投影向量发生了剧烈的变化,a_1 从 0.306 变为了 – 0.267,即指标性质发生了改变,这个结果虽然难以理解,但确实是真正的全局最优解。

(2) 在 $R \leqslant 0.001S_z$ 和 $R \geqslant r_{\max}$ 的范围内(以及 $R = 8$),结果很稳定,最佳投影向量和各样本投影值相差很小。

(3) 在 $R \in [0.002S_z, 0.25S_z]$ 时,笔者曾尝试多种群智能最优化算法和参数,均较难求得真正的全局最优解,最佳投影向量变化也呈现出没有规律性,多个指标权重出现了小于 0 的情况,其指标性质也是错误的(用下划波浪线表示),$z(1) \sim z(3)$ 的值几乎相等,而且都很小(用下划线表示,下同),在 $R = 0.1S_z$ 时,甚至出现了 $z(1) \sim z(6)$ 几乎都等于 0 的情况,这显然与他们是Ⅰ ~ Ⅴ类不同水质的分界值的要求是相矛盾的。

(4) 在 $r_{\max}/5 \sim r_{\max}/2$ 范围内,不仅所有指标的性质都是正确的,而且指标权重大小的排序也几乎一致,各指标权重的变化正好体现出投影寻踪技术可以从不同视角将样本点投影到低维子空间上,从而揭示出高维数据不同结构特点的精髓。

(5) 在 $R \geqslant r_{\max}/5$ 的范围内,R 值越大,各指标的权重相差越小,即指标的重要性有均衡化的趋势。

由于不少文献 R 取常数,笔者取 $R = 0.0001$(接近于 $0.001S_z$)、0.001、0.01、0.1、0.3(接近于 $r_{\max}/5$)、0.5(接近于 $r_{\max}/3$)、0.7、0.9(接近于 $r_{\max}/2$)、1、1.5、2.9(相当于 $\geqslant r_{\max}$)、5(相当于 $2r_{\max}$)、10、100 分别建立 PPC 模型,结果如表 3 所示。从表 3 可以看出,整体上讲,R 取常数时与上述 3 个方案具有相似的规律和现象,如 R 取很小(如小于 0.0001)和很大(如大于 5)时的结果不但稳定,而且几乎相同,在 $R \in [0.002S_z, 0.25S_z]$ 范围内取常数时,建模结果没有规律,并与 R 在其他范围内时的结果存在明显差异,即使 R 值很接近,取常数的结果与前 3 个方案的结果都明显不同,甚至 R 在 $r_{\max}/5$

~ $r_{max}/2$ 或者 $R \geqslant S_z$ 范围内取常数时，建模结果(各指标的权重和样本投影值)也没有规律性，还出现了部分指标的性质是错误(权重小于0)的情况。因此，笔者建议 PPC 建模时 R 不宜取常数，应该取 $r_{max}/5 \leqslant R \leqslant r_{max}/3$ 或者 $0.5S_z \leqslant R \leqslant S_z$。

那么，上述现象究竟是普遍规律呢，还是仅仅是特殊情况?现分析讨论如下：

(2) 式中，当 $i = k$ 时 $r_{i,k} = r_{k,i} = 0$，则：

$$D_z = nR + 2\sum_{i=1}^{n}\sum_{k=i+1}^{n}(R - r_{i,k})u(R - r_{i,k})$$

当取 $R = 0.1S_z$(较小值方案) 时，通常只有少数几个样本点能满足 $R \geqslant r_{i,k}(i \neq k)$ 条件(假设有 m 个)，则 $D_z = nR + 2\sum_{l=1}^{m}(R - r_l) = (2m + n)R - 2\sum_{l=1}^{m} r_l$，此时目标函数 $Q(a) = \max(S_z D_z) = \max[0.1(2m + n)S_z^2 - 2S_z\sum_{l=1}^{m} r_l]$，主要取决于 S_z 的最大化，同时由于 $S_z\sum_{l=1}^{m} r_l$ 随各评价指标权重的改变而变化，致使最优化收敛过程通常不太稳定(如表 2 和 3 所示)，较难求得真正的全局最优解。如果 R 很小，所有样本点均可能满足 $R < r_{i,k}(i \neq k)$ 条件，则 $Q(a) = \max[0.1nS_z^2]$，即 $Q(a)$ 值只与类间距离 S_z 有关，其最优化过程必定是稳定的。

同理，如果 R 取较大值方案($r_{max} \leqslant R \leqslant 2p$)，则 $R \geqslant r_{max} \geqslant r_{i,k}$，有 $D_z = n^2R - 2\sum_{i=1}^{n}\sum_{k=i+1}^{n} r_{i,k}$，目标函数 $Q(a) = \max[S_z(n^2R - 2\sum_{i=1}^{n}\sum_{k=i+1}^{n})]$。因此，要使 $Q(a)$ 取得最大值，应使 S_z 最大化(即使所有样本点尽可能分散) 和 $r_{i,k}$ 最小化(即使所有样本点尽可能密集)，这两者显然是相互矛盾的。事实上，当 $R \geqslant 2r_{max}$ 就有 $n^2R \gg 2\sum_{i=1}^{n}\sum_{k=i+1}^{n} r_{i,k}$，从而有 $Q(a) = \max(n^2RS_z)$，即只使 S_z 最大化即可。

因此，从上述理论分析和实证结果可知，R 取很小值和较大值($r_{max} \leqslant R \leqslant 2p$) 时，PPC 建模结果几乎相同，但都只实现了使 S_z 最大化的目的，仅仅体现了“使样本投影点整体上尽可能分散”的要求，而没有实现“使样本点局部上尽可能密集”的目标，与 Friedman 等[10]提出的 PPC 建模基本思想是不一致的，也是不合理的。R 取较小值方案时，PPC 建模结果严重受其值大小的影响，但表现出没有规律性；而 R 取中间适度值方案时，目标函数正好体现了 PPC 建模的基本思想，是合理的。

表 2 投影窗口半径 R 取不同值时 PPC 建模结果的对比

最优化结果		$0.001S_z$	$0.002S_z$	$0.01S_z$	$0.1S_z$	$0.25S_z$	$0.5S_z$	$\frac{r_{max}}{5}$	$\frac{r_{max}}{4}$	$\frac{r_{max}}{3}$	S_z	$\frac{r_{max}}{2}$	r_{max}	$2r_{max}$	8
样本投影值	$z(1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	$z(2)$	0.230	0	0	-0.006	0.009	0.132	0.155	0.181	0.189	0.178	0.201	0.231	0.234	0.224
	$z(3)$	0.337	0	0.001	-0.003	0.009	0.189	0.225	0.264	0.277	0.260	0.295	0.339	0.344	0.330
	$z(4)$	0.564	0.145	0	-0.001	0.095	0.358	0.398	0.455	0.475	0.458	0.498	0.561	0.569	0.553
	$z(5)$	0.904	0.307	0.169	-0.012	0.266	0.627	0.667	0.747	0.768	0.752	0.798	0.891	0.903	0.900
	$z(6)$	1.446	0.706	0.399	0.001	0.616	1.101	1.126	1.233	1.257	1.250	1.290	1.416	1.434	1.426
	$z(7)$	2.822	2.088	1.803	1.367	2.093	2.629	2.681	2.753	2.770	2.756	2.789	2.826	2.827	2.821
最佳投影向量系数	a_1	0.306	-0.267	-0.141	0.012	-0.232	0.051	0.134	0.194	0.218	0.179	0.254	0.322	0.327	0.293
	a_2	0.366	0.115	0.016	-0.151	0.143	0.264	0.253	0.287	0.294	0.302	0.304	0.347	0.355	0.360
	a_3	0.350	0.119	0.145	-0.015	0.205	0.305	0.314	0.328	0.327	0.321	0.328	0.349	0.351	0.347
	a_4	0.352	0.397	0.289	0.252	0.282	0.333	0.344	0.350	0.357	0.356	0.362	0.355	0.354	0.353
	a_5	0.339	0.437	0.568	0.739	0.544	0.476	0.500	0.459	0.452	0.443	0.442	0.370	0.356	0.352
	a_6	0.369	0.415	-0.110	-0.037	0.186	0.301	0.287	0.312	0.326	0.335	0.325	0.352	0.358	0.365
	a_7	0.347	0.437	0.568	0.604	0.544	0.481	0.489	0.453	0.445	0.441	0.432	0.370	0.358	0.359
	a_8	0.393	0.437	0.468	-0.037	0.422	0.419	0.362	0.369	0.351	0.379	0.342	0.362	0.369	0.392
目标函数	S_z	0.973	0.760	0.663	0.518	0.761	0.920	0.930	0.952	0.956	0.954	0.960	0.972	0.972	0.973
	D_z	0.068	0.020	0.120	1.740	3.087	6.902	8.585	12.031	18.79	19.87	34.700	92.881	231.28	346.30
	$Q(a)$	0.066	0.015	0.080	0.901	2.350	6.349	7.985	11.450	17.957	18.968	33.313	90.235	224.87	336.89
	r_{max}	2.822	2.088	1.803	1.379	2.093	2.629	2.756	2.753	2.681	2.756	2.760	2.826	2.827	2.821
	R	0.0097	0.0015	0.0066	0.052	0.190	0.460	0.536	0.688	0.923	0.954	1.394	2.826	5.654	8.000

注：带下划波浪线 ~~ 表示指标性质是错误的，带下划线 __ 表示投影值很小且几乎相等。

表 3　投影窗口半径 R 取不同值(常数)时 PPC 建模结果的对比

最优化结果		0.0001	0.001	0.01	0.1	0.3	0.5	0.7	0.9	1	1.5	2.9	5	10	100
样本投影值	$z(1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	
	$z(2)$	0.230	0	0	−0.003	−0.004	−0.015	0	0.024	0.043	0.118	0.196	0.219	0.226	0.229
	$z(3)$	0.337	0	−0.001	0.003	−0.004	−0.020	0	0.035	0.063	0.172	0.287	0.322	0.331	0.337
	$z(4)$	0.564	0.116	0	0	0	0.001	0.054	0.117	0.160	0.325	0.494	0.543	0.556	0.563
	$z(5)$	0.904	0.268	0	−0.005	0	0.030	0.127	0.232	0.298	0.549	0.803	0.874	0.893	0.903
	$z(6)$	1.446	0.644	0.055	0.003	0.043	0.141	0.323	0.496	0.593	0.956	1.314	1.407	1.431	1.444
	$z(7)$	2.822	2.069	1.391	1.367	1.442	1.602	1.845	2.055	2.170	2.547	2.790	2.818	2.821	2.822
最佳投影向量系数	a_1	0.307	−0.330	−0.002	0.003	−0.045	−0.1290	−0.148	−0.122	−0.087	0.060	0.224	0.280	0.297	0.306
	a_2	0.366	0.259	−0.004	−0.142	−0.103	−0.045	0.013	0.070	0.098	0.207	0.322	0.353	0.361	0.366
	a_3	0.350	0.259	0.003	0.027	0.084	0.104	0.140	0.172	0.192	0.264	0.330	0.344	0.348	0.350
	a_4	0.352	0.320	−0.002	0.291	0.252	0.267	0.295	0.313	0.323	0.353	0.358	0.355	0.353	0.352
	a_5	0.339	0.466	0.725	0.757	0.725	0.705	0.676	0.648	0.635	0.557	0.416	0.364	0.348	0.339
	a_6	0.369	0.263	−0.004	−0.081	−0.053	0.010	0.098	0.156	0.179	0.262	0.342	0.361	0.366	0.368
	a_7	0.347	0.521	0.689	0.559	0.622	0.631	0.618	0.601	0.592	0.531	0.416	0.370	0.356	0.348
	a_8	0.393	0.310	−0.015	−0.048	−0.039	0.063	0.154	0.219	0.240	0.314	0.382	0.391	0.393	0.393
目标函数	S_z	0.973	0.754	0.523	0.517	0.543	0.599	0.676	0.740	0.775	0.889	0.964	0.972	0.973	0.973
	D_z	0.001	0.013	0.262	3.580	10.610	16.617	21.795	26.934	29.352	43.043	97.607	199.48	444.25	4854.1
	$Q(a)$	0.001	0.010	0.137	1.851	5.763	9.957	14.723	19.935	22.738	38.267	94.092	193.92	432.23	4723.5
	$r_{\max}$	2.822	2.069	1.392	1.372	1.446	1.622	1.845	2.055	2.170	2.547	2.790	2.818	2.821	2.822
	R	0.0001	0.001	0.01	0.1	0.3	0.5	0.7	0.9	1	1.5	2.9	5	10	100

注:带下划波浪线 表示指标性质是错误的,带下划线 表示投影值很小且几乎相等。

而且,笔者研究还发现,王顺久等[12]和付强等[13]推导得出 $r_{\max} \leqslant R \leqslant 2p$ 取值范围的过程也是不合理的。王顺久等[12]的推导过程可简述为"当 $R \geqslant r_{\max}$ 时,有 $0 \leqslant x_{i,j} \leqslant 1$, $|a| \leqslant 1$,因此, $-p \leqslant z(i) = \sum_{j=1}^{p} a_j x_{i,j} \leqslant p$,且 $r_{\max} \leqslant 2p$。于是可以得出 R 的合理取值范围为 $R_{\max} \leqslant R \leqslant 2p$"。付强等[11]的推导过程基本相同。显然,上述的推导过程至少存在如下错误:从"$0 \leqslant x_{i,j} \leqslant 1$, $|a| \leqslant 1$"并不能推导得出"$-p \leqslant z(i) = \sum_{j=1}^{p} a_j x_{i,j} \leqslant p$"的结论,事实上,由柯西不等式很容易证明[17],只有在 $a_j = 1/\sqrt{p}$ 时 $z(i)$ 才可能取得极大值,并且 $z(i) = \sum_{j=1}^{p} a_j x_{i,j} \leqslant \sum_{j=1}^{p} a_j \leqslant \sqrt{p}$。其次,从"$R \geqslant r_{\max}$ 且 $2p \geqslant r_{\max}$",也不可能得出"$r_{\max} \leqslant R \leqslant 2p$"的结论。当然,$R \geqslant r_{\max}$ 的假设也与 Friedman 等[10]提出选取 R 合理值"应使窗口内的样本点个数既不能太少,以免样本滑动平均时偏差太大,同时又不能随着样本个数 n 的增大而增加太多"的要求相矛盾,因为当 $R \geqslant r_{\max}$ 时,所有样本投影点都在同一个窗口内了。

3.4　R 取中间适度值方案($r_{\max}/5 \leqslant R \leqslant r_{\max}/3$)对长江水系水质综合评价结果的影响

根据表 2 中 $R = r_{\max}/3$ 时的 PPC 建模结果,得到了上述 22 个长江水系主要国控断面的水质综合评价结果。比较 $R = r_{\max}/5$ 和 $R = r_{\max}/3$ 的结果可以发现,权重小于 0.350 的指标的重要性都得到了提高(其中 DO 的提高最为显著,权重从 0.134 提高到 0.218,几乎提高一倍),反之,指标的重要性有所降低,即所有指标的重要性有均衡化的倾向,在 $R = r_{max}/3$ 时,Hg、挥发酚和石油类对水质的影响程度有所降低,其他指标的影响程度则有所提高,从而导致 DO 较低以及 COD_{Mn} 和 Pb 较高水域水质的评价结果变差,反之,Hg 和挥发酚较高水域的水质评价结果有所好转。在 22 个国控断面 110 个样本的评价结果中,20 个样本的水质类别发生了改变,其中 10 个Ⅱ类水质的样本变为了Ⅰ类,5 个Ⅲ类样本变为Ⅱ类,2 个Ⅳ类样本变为Ⅲ类,1 个Ⅲ类样本变为Ⅳ类,2 个Ⅳ类样本变为Ⅴ类。同理,在所有 103 个国控断面的 515 个样本中,47 个样本的水质类别发生了改变,其中 36 个样本从Ⅱ类水质变为了Ⅰ类水质,8 个样本从Ⅲ类水质变为了Ⅱ类水质,2 个样本从Ⅳ类水质变为了Ⅲ类水质,1 个样本从Ⅴ Ⅰ类水质变为了Ⅴ类水质。

从上述 $R = r_{\max}/5$ 和 $R = r_{\max}/3$ 时的水质综合评价结果可以看出,绝大部分样本(占 91%)的水质类别保持不变,但由于 DO 的影响程度显著提高,导

致部分水体的水质评价结果出现了改变,但主要是Ⅰ类和Ⅱ类水质。

4 结 语

本文根据我国地表水环境质量标准[16]和长江水系主要国控断面的 8 项水质监测指标数据,建立了水质综合评价的投影寻踪分类(PPC)模型,得到了 103 个国控断面 2006 – 2010 年共计 515 个样本的水质综合评价结果,长江水系在国控断面上以Ⅰ~Ⅲ类水质为主(Ⅰ类水质占 60%,Ⅱ类水质占 22.7%,Ⅲ类水质占 10.1%),整体水质较好。上游水质总体上要好于下游水质。流经四川、安徽、江苏省后,水质下降明显。在化工工业较发达的部分地区的支流、河流里,水污染情况严重。更为严峻的是,虽然经过长期的治理,2007 – 2009 年的水质比 2006 年有明显的改善,但在 2010 年却又出现了明显的恶化,沿长江流域的当地政府和有关环保部门,必须采取切实有效的水污染治理措施,绝不能有丝毫的放松和懈怠。

投影窗口半径 R 值对 PPC 建模结果有重要影响,理论和实证研究均表明,选取 R 值的较小值方案是不合理的,而较大值方案是错误的,R 取很小和很大值时,建模结果只能体现“使样本投影点整体上尽可能分散”的要求。当 R 取较小值时,各指标权重的稍许变化就可能引起目标函数的改变,最优化过程往往无法求得真正的全局最优解。在通常情况下,R 取常数时也不能得到合理的结果。R 只有取中间适度值方案时,建模结果不仅完全符合 FrIedman 等提出的 PPC 建模基本思想,而且具有很好的规律性,最优化过程也都能求得真正的全局最优解。

参考文献:

[1] 富天乙, 邹志红, 王晓静. 基于多元统计和水质标识指数的辽阳太子河水质评价研究[J]. 环境科学学报, 2014,34(2):473 – 480.

[2] 金菊良, 刘 丽,丁 晶,等. 地下水水质评价的逻辑斯谛曲线模型[J]. 环境污染与防治, 2003,25(1):46 – 48.

[3] 陈润羊, 花 明, 涂安国. 长江流域水质评价的几种方法[J]. 东华理工大学学报(自然科学版), 2008,31(2):146 – 151.

[4] 刘小楠, 崔 巍. 主成分分析法在汾河水质评价中的应用[J]. 中国给水排水, 2009,25(18):105 – 108.

[5] 楼文高, 王延政. 基于 BP 网络的水质综合评价模型及其应用[J]. 环境污染治理技术与设备, 2003,4(8):23 – 26.

[6] 李祚泳, 郭丽婷, 欧阳洁. 水环境质量评价的普适指数公式[J]. 环境科学研究, 2001,14(3):56 – 58.

[7] MacCallum R C, Widaman K F, Zhang S, et al. Sample Size in Factor Analysis [J]. Psychological Methods, 1999, 4 (1):84 – 99.

[8] 楼文高, 乔 龙. 基于神经网络的金融风险预警模型及其实证研究[J]. 金融论坛, 2011,16(11):52 – 61.

[9] StatSoft Inc. Electronic Statistics Textbook [EB/OL]. [2011]. http://www. statsoft. com/textbook.

[10] Friedman J H, Tukey J W. A Projection Pursuit Algorithm for Exploratory Data Analysis [J]. IEEE Transactions on Computers, 1974,23(9):881 – 890.

[11] 张欣莉, 丁 晶, 李祚泳, 等. 投影寻踪新算法在水质评价模型中的应用[J]. 中国环境科学, 2000,20(2):187 – 189.

[12] 王顺久, 张欣莉, 丁 晶, 等. 投影寻踪聚类模型及其应用[J]. 长江科学院院报, 2002,19(6):53 – 55 + 61.

[13] 付 强, 赵小勇. 投影寻踪模型原理及其应用[M]. 北京:科学出版社, 2006.

[14] 封志明, 郑海霞, 刘宝勤. 基于遗传投影寻踪模型的农业水资源利用效率综合评价[J]. 农业工程学报, 2005,21(3):66 – 70.

[15] 楼文高, 乔 龙. 投影寻踪聚类建模理论的新探索与实证研究[EB/OL]. [2013 – 08 – 30]. 数理统计与管理. http://www. cnki. net/kcms/detail/ 11. 2242. O1. 20130830. 1736. 001. html.

[16] 国家环境保护总局. 地表水环境质量标准 GB3838 – 2002[S]. 北京:中国环境出版社,2002.

[17] 龙定源. 柯西不等式在中学数学中的应用[J]. 数学教学通讯:中教版,2000(2):38 – 40.