主成分分析在河流水质综合评价中的应用

刘德林,刘贤赵
(烟台师范大学地理与资源管理学院,山东 烟台 264025)

摘 要:采用主成分分析法(PCA),对大沽夹河流域水质进行了定量化综合评价。结果表明:流域水质具有明显的区域差异,在 14 个典型监测断面中,福山水闸下和新夹河桥 2 个监测断面水质污染较为严重,宫家岛等 4 个断面水质较好,其余断面水质良好。就全流域而言,水质污染程度不是很严重,基本满足功能区的要求。

关键词:烟台市;水质;综合评价;PCA 分析
中图分类号:P343.1 **文献标识码**:A **文章编号**:1005-3409(2006)03-0124-02

Application of Principal Component Analysis to the Comprehensive Evaluation of Water Quality in River

LIU De lin, LIU Xian zhao
(*College of Geography and Resource Management, Yantai Normal University, Yantai, Shandong* 264025, *China*)

Abstract: The water quality in Dagujia River basin was assessed by using the PCA method. The results showed that the water quality pollution of the whole basin is not serious and can achieve the water quality standard of this region basically. The water quality of all the 14 representative sections were all right except the sections both water gate of Fushan and bridge of Xinjiahe River which were polluted seriously.

Key words: Yantai city; water quality; comprehensive evaluation; PCA method

水质评价是水环境质量评价的主要内容之一,它为水资源合理开发利用和水体污染的综合防治提供科学依据,是国民经济和人类社会健康、持续发展的重要工作之一。目前,常用的水质评价方法主要有简单指数法、分级加权平均法、综合污染指数法、模糊数学法、普通概率统计法等,上述方法虽然都有一定的数学基础和理论依据,但水质系统是一个由多因子构成的复杂系统,水质评价受诸多指标因子的影响,从而使上述方法在进行水质评价时表现出一定的局限性[1]。对于涉及多因素的水质评价主要不是分别考虑各因素的作用效果,而应是在最少人为因素影响的情况下,正确分析影响水质的各因素之间相互作用,从而得出反映各因素特征信息的综合评价结果。主成分分析方法(PCA 法)正是一种将多维因子纳入同一系统中进行定量化研究、理论比较完善的多元统计分析方法,在解决很多实际问题时取得了较好的效果[2,3]。鉴于此,本文采用 PCA 分析法,对烟台市大沽夹河流域进行水质综合评价,以期为该区水资源的合理开发利用和水体污染综合防治提供一定的决策依据。

1 水质综合评价的 PCA 分析法

PCA 分析是在确保系统原有数据信息量丢失最小的原则下,在各个变量相关关系研究的基础上,将多个变量的信息压缩为几个能反映原问题特征的综合变量指标,并据此特征信息指标对系统进行综合分析,从而避免了人为地确定各指标权重的主观随意性,具有降维、简化变量之优点[4,5]。

设共有 n 个待评水体样本,每个样本有 p 个指标变量,则构成一个 $n \times p$ 阶的水质数据矩阵:

$$X = \begin{matrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{matrix} \quad (1)$$

经降维处理,p 个指标变量可综合成 m 个新指标 F_1, F_2, …, F_m,则 x_i 可由 F_m 表示为:

$$X = L F + \varepsilon \quad (2)$$

其中,

$$X = (x_1, x_2, \cdots, x_p)^T \quad (3)$$

$$L = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p1} & l_{p2} & \cdots & l_{pm} \end{bmatrix} \quad (4)$$

$$F = (F_1, F_2, \cdots, F_m) \quad (5)$$

$$\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_p) \quad (6)$$

上述模型应用于水质综合评价时,可根据精度分析要求(通常累计贡献率 $E \geq 80\%$),在 p 个指标变量中合理选取 m 个综合指标($m < p$),略去线性表达式中的特殊因子(ε),从而达到数据降维的目的。其中,F 为综合指标的向量集合,据此可进行相关分析。L 为原变量上的载荷值,体现了原指

收稿日期:2005-10-03
基金项目:山东省自然基金项目(Q02E03)资助
作者简介:刘德林(1979-),男,山东潍坊人,在读硕士研究生,主要从事水土资源高效利用方面的研究工作。

标变量与综合指标变量的相关程度[6]。具体计算步骤见参考文献[4]。

2　实例分析 ——大沽夹河流域典型断面水质综合评价

2.1　样本点(断面)选取和监测指标的确定

大沽夹河为烟台市区供水的主要水源地,对其水质状况做出科学、准确的评价具有重大意义。在遵循科学性和可比性原则的基础上,结合流域监测断面实际观测情况,选取流域内具有代表性的 14 个监测断面、水质评价常用的 7 项指标及指标在断面上的具体观测值(表 1),采用 PCA 分析法对河流水质进行综合评价。

表 1　2003 年大沽夹河水质监测指标统计值

指标 断面	x_1 (pH 值)	x_2 (溶 解氧)	x_3 (化学 需氧量)	x_4 (高锰 酸钾指数)	x_5 (生化 需氧量)	x_6 (氨氮)	x_7 (石油类)
1 庵里水库入口	8.41	8.66	13.92	5.40	2.33	0.301	0.005
2 庵里水库出口	8.34	8.76	14.02	5.20	2.20	0.333	0.005
3 南　桥	8.44	8.70	6.51	2.39	2.34	0.025	0.005
4 门楼水库入口	8.33	8.91	10.33	2.88	1.00	0.077	0.013
5 门楼水库出口	8.33	9.02	9.50	2.92	0.97	0.075	0.005
6 荆子埠	7.84	8.71	8.19	1.41	2.23	0.025	0.005
7 大沙埠	7.31	6.27	9.42	4.44	3.36	0.936	0.002
8 福山水闸下	8.14	6.29	15.30	7.41	4.74	3.931	0.006
9 仉　村	7.17	5.59	7.18	3.16	2.68	0.808	0.002
10 回　里	8.18	9.36	9.83	2.81	1.17	0.071	0.009
11 套　口	8.34	9.27	11.00	3.59	1.43	0.088	0.021
12 东陌堂桥	8.32	9.29	10.17	2.62	1.13	0.079	0.005
13 宫家岛	8.52	9.92	12.50	4.82	2.78	0.253	0.036
14 新夹河桥	8.37	10.98	17.17	5.45	3.01	0.260	0.110
15 全流域	8.15	8.62	11.06	3.87	2.19	0.510	0.010

2.2　结果与分析

表 2　特征值及主成份贡献率与累积贡献率

主成份	特征值	贡献率/%	累计贡献率/%	主成份	特征值	贡献率/%	累计贡献率/%
Z_1	3.1057	44.37	44.37	Z_5	0.1424	2.03	99.02
Z_2	2.6896	38.42	82.79	Z_6	0.0567	0.81	99.83
Z_3	0.7377	10.54	93.33	Z_7	0.0119	0.17	100.00
Z_4	0.1424	3.66	96.99				

根据表 1 所列数据资料,通过对原始数据进行标准化处理,采用 DPS 软件(v5.12)进行计算得到各指标的特征值、主成份的贡献率和累积贡献率(表 2)。累积贡献率说明主成分所包含全部指标信息的百分比。从表 2 可知,主成分分量 Z_1、Z_2 是由七个原始变量 x_1、x_2、x_3、x_4、x_5、x_6 和 x_7 通过 PCA 分析得到的一组新变量,以 82.79%的累积贡献率(概率)替代了原变量系统,充分地反映了原始数据的主要信息。因此,可以利用主成份 Z_1 和 Z_2 对大沽夹河流域断面水质污染状况进行可比性研究。

从各评价指标在主成分中的载荷值可以看出(表 3),第一主成份 Z_1 在化学需氧量(x_3)、高锰酸钾指数(x_4)、生化需氧量(x_5)和氨氮(x_6)上具有很大的载荷,载荷值变化在 0.624 8～0.913 3 之间。在这些指标中高锰酸钾指数代表的是河流中的有机污染[7],氨氮表示引起湖泊富营养化的营养元素污染状况,化学需氧量和生化需氧量为综合类污染物[8]。上述指标均为增长较快的污染物,因此,可将第一主成分看成潜力因子,随着该区人口增长和经济的发展,这一因子的增长会比较显著。第二主成份 Z_2 代表了 pH 值(x_1)、溶解氧(x_2)和石油类污染物(x_7)等的污染,这些指标污染程度在一定时间内变化幅度不会不大。

表 3　主成分载荷值

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
第一主成分(Z_1)	-0.1926	-0.4669	0.6248	0.8989	0.9113	0.8903	0.1713
第二主成分(Z_2)	0.8024	0.8692	0.7252	-0.3302	-0.1197	-0.2376	0.7645

另外,根据特征向量矩阵

$$U=\begin{bmatrix} -0.109 & 0.489 & 0.602 & 0.411 & -0.106 & -0.303 & -0.338 \\ -0.265 & 0.530 & -0.023 & 0.186 & -0.027 & 0.468 & 0.627 \\ 0.355 & 0.442 & 0.066 & -0.497 & 0.115 & 0.480 & -0.428 \\ 0.510 & 0.201 & 0.710 & 0.366 & -0.296 & -0.458 & 0.489 \\ 0.517 & -0.073 & -0.224 & 0.519 & -0.561 & 0.283 & -0.117 \\ 0.505 & -0.145 & 0.262 & 0.340 & 0.699 & 0.093 & 0.206 \\ 0.097 & 0.466 & -0.696 & 0.174 & 0.291 & -0.400 & -0.188 \end{bmatrix} \quad (7)$$

和与之对应的特征值(表 2),可得到各指标与主成份 Z_1 与 Z_2 的线性关系:

$$Z_1 = -0.109x_1 - 0.265x_2 + 0.355x_3 + 0.510x_4 + 0.517x_5 + 0.505x_6 + 0.097x_7 \quad (8)$$

$$Z_2 = -0.489x_1 - 0.530x_2 + 0.442x_3 + 0.201x_4 + 0.073x_5 + 0.145x_6 + 0.466x_7 \quad (9)$$

表 4　大沽夹河流域各监测断面水质综合评判结果(2003)

项目断面	第一主成 分得分 F_1	第二主成 分得分 F_2	水质污染综 合得分 F	污染综合排 名(重轻)	污染程度 分级结果
1	0.660	0.842	0.616	4	中
2	0.554	0.785	0.548	5	中
3	-1.426	-0.597	-0.862	13	轻
4	-1.443	0.224	-0.554	9	轻
5	-1.598	0.007	-0.706	12	轻
6	-1.445	-1.216	-1.108	15	轻
7	1.411	-2.493	-0.332	8	中
8	5.225	-0.633	2.075	2	重
9	0.448	-3.372	-1.097	14	轻
10	-1.501	0.039	-0.651	11	轻
11	-0.932	0.678	-0.153	7	轻
12	-1.581	0.146	-0.645	10	轻
13	0.341	1.685	0.799	3	中
14	1.360	3.978	2.132	1	重
15	-0.073	-0.075	-0.061	6	中

根据主成份 Z_1 和 Z_2 与相应的贡献率之积的和,计算各监测断面和全流域的水质污染综合得分(表 4),给予水质污染程度的定量化描述,得分越大,表明污染程度越严重,由此可对样点就污染程度进行分级。由表 4 可知,大沽夹河 14 个监测断面,有 12 个水质较好。其中,荆子埠等 8 个断面水质良好,宫家岛等 4 个断面水质较好。新夹河桥和福山水闸下断面水质污染较为严重,其主要原因是大沽夹河下游为补充地下水而新增设 4 处橡胶坝,导致下游基本处于断流状态,河水流动性较差而致使氨氮超标。总体而言,大沽夹河流域污染程度不是很大,基本满足功能区的要求。

3　结论与讨论

河流水质系统是一个由多因子构成的复杂系统,其综合评价的数量化指标很多,采用 PCA 法对其进行综合评价,可以在原始数据信息量丢失最小的情况下,减少评价指标,同时客观的确定权重,避免主观随意性。结果表明,大沽夹河流域水质具有明显的区域差异,在流域内 14 个典型监测断面中,河流水质较好的断面达 85.7%,主要分布在内夹河仉村以上河段及外夹河大沙埠以上河段,只有福山水闸下和新夹河桥 2 个监测断面水质污染较为严重。就全流域而言,水质污染程度不是很大,基本满足功能区的要求。尽管如此,继续控制水污染源的数量,加大水质污染严重区域的水环境保护和治理仍是紧迫任务。

(下转第 128 页)

表2 叶片电导率 μv/cm

草种	胁迫程度	天数			
		0	2	4	6
匍匐翦股颖	对照	210	242	262	343
	轻度	187	245	316	345
	中度	270	346	450	708
	重度	840	2460	3990	4928
无芒雀麦	对照	160	282	262	290
	轻度	190	206	224	242
	中度	247	347	524	735
	重度	358	876	739	1380
高羊茅	对照	187	198	227	230
	轻度	138	185	162	266
	中度	197	195	276	428
	重度	433	925	1296	2400
紫羊茅	对照	205	212	227	225
	轻度	130	199	215	213
	中度	230	340	460	516
	重度	276	866	1337	2950
多年生黑麦草	对照	178	181	205	210
	轻度	185	200	220	272
	中度	209	243	320	350
	重度	666	678	1084	2590

表3 干旱胁迫下5种冷季型草坪草各指标得分比较

品种名称	植株高度	根系生长	叶片相对含水量	叶片电导率	总计
匍匐翦股颖	1	1	1	1	4
无芒雀麦	3	2	4	2	11
高羊茅	4	4	3	3	14
紫羊茅	2	3	2	4	11
多年生黑麦草	5	5	5	5	20

3 结论与讨论

干旱胁迫下,不同草种的植株高度随胁迫程度加强而减小。胁迫程度越高,植株生长速率越小。不同冷季型草的根系长度随水分胁迫的影响各不相同,其中多年生黑麦草对水分胁迫的反映最不敏感,其次为高羊茅和紫羊茅,它们在重度水分胁迫时,根的生长量才明显减弱,无芒雀麦在中度水分胁迫时根的生长量明显减小,匍匐翦股颖的根系发育最差,轻度水分胁迫时,根的生长即明显受抑制。

随着水分胁迫的加强,冷季型草坪草叶片含水量逐渐下降,而电导率升高。胁迫使细胞膜受到伤害,削弱其选择透性,细胞液外渗液增加,即表现为电导率增大。干旱对细胞膜的伤害随着胁迫程度的提高而加大,且胁迫程度越高电导率就越大[10]。抗性强的品种在干旱条件下叶片相对含水量降低幅度小,电导率上升较慢,表明植物受伤害程度较轻。不同胁迫水平效应间电导率差异呈极显著,多年生黑麦草电导率增加最少,匍匐翦股颖电导率增加最多。

综合上述抗旱指标对5种冷季型草坪草苗期水分胁迫下的变化规律,赋分综合评价,其幼苗抗旱能力由强到弱依次为:多年生黑麦草>高羊茅>无芒雀麦>紫羊茅>匍匐翦股颖。

参考文献:

[1] 庞鸿宾. 节水农业工程技术[M]. 郑州:河南科学技术出版社, 2000. 40 - 63.

[2] Huang B R, Fry J D. Root anatomical physiological and Morphological responses to drought stress for tall fescue cultivars[J]. Crop Sci., 1998(38):1017 - 1022.

[3] Bonos S A, Murphy J A. Growth response and performance of Kentucky bluegrass under summer stress[J]. Crop Sci., 1999(39):770 - 774.

[4] 徐炳成, 山仑, 黄占彬. 草坪草对干旱胁迫的反应及适应性研究进展[J]. 中国草地, 2001,23(2):55 - 61.

[5] 山仑, 陈陪元. 旱地农业生理生态基础[M]. 北京:科学出版社, 1998. 100 - 121.

[6] Carrow R N. Drought avoidance characteristics of diverse tall fescue cultivars[J]. Crop Sci., 1996(36):371 - 377.

[7] 李敏. 草坪地被植物的引种[A]. 见:陈佐忠. 面向21世纪的中国草坪科学与草坪业[M]. 北京:中国农业大学出版社, 1998. 106 - 111.

[8] 华东师范大学生物系植物生理教研组. 植物组织含水量测定, 植物生理学试验指导[M]. 北京:高等教育出版社, 1980. 2 - 5.

[9] 西北农大生理教研室. 植物生理大试验指导[M]. 西安:陕西农业科学出版社,1985. 1 - 2,42 - 44,47 - 48,92 - 94,148 - 154.

[10] 李培英. 优良草坪草研究[J]. 北京:北京农业出版社, 2001. 132 - 145.

(上接第125页)

参考文献:

[1] 李惠明,尚广平. 水质现状评价数学模型综合研究[J]. 中国环境科学,1991,11(5):356 - 360.

[2] 傅湘,纪昌明. 区域水资源承载能力综合评价 - 主成分分析法的应用[J]. 长江流域资源与环境,1995,8(2):168 - 173.

[3] 董菁,张毅,张佐,等. 基于主成分分析法的城市交通路口相关性分析[J]. 西南交通大学学报,2003,38(6):619 - 622.

[4] 王红芬. 计量地理学概论[M]. 济南:山东教育出版社,2001. 142 - 144.

[5] 任广平. 因子分析及其在河网水质综合评价中的应用研究[J]. 环境污染治理技术与设备,2005,6(4):91 - 94.

[6] Richard A. Johnson, Dean W. Wichern. 实用多元统计分析[M]. 北京:清华大学出版社,2001. 388 - 425.

[7] 张路,范成新,秦伯强,等. 太湖宜粟河水系沉积物中多环芳烃来源解析[J]. 地球化学,2003,32(2): 124 - 130.

[8] 程晓如,方正,薛英文. 东湖西南湖区水质监测与评价[J]. 武汉大学学报(工学版),2001,34(5): 96 - 100.