

太湖水质指标主成分分析

冯 健,崔广柏,张红举

(河海大学水资源环境学院,江苏 南京 210098)

摘要:用主成分分析的方法研究了太湖各水质监测站监测的水质指标,从原始数据出发提取了占总方差 80% 以上的两个主成分,对此作出了合理的解释:第一主成分主要是由水体中的 N、P 和有机物所引起的,而第二个主成分主要体现了水体的清澈程度。得到了水质监测站的水质分类图,从水质分类图中明显看出各站水质的污染情况及其污染原因,将大量抽象的数据变为形象的图表,降低分析的难度。

关键词:主成分分析法;水质指标;线性回归;太湖;水质监测

中图分类号:X524 文献标识码:A 文章编号:1004-693X(2004)03-0049-02

近年来随着经济的不断发展,太湖水质不断恶化,太湖流域面临着水质型缺水。为了监测太湖的水质,太湖流域管理局在太湖中建立了 20 多个水质监测站。由于站点和水质指标多,从中难以综合判断各站的水质情况。基于上述原因,本文提出主成分分析方法,对太湖各处水质进行综合性比较与评价,以减少人为因素的影响,使得到的结论更加科学、合理,为进一步改善太湖水环境提供科学依据。

1 主成分分析法的基本思想

下面以二维情况为例,对照线性回归问题,说明主成分分析的基本思想。设有 N 个观测点 (x_{i1}, x_{i2}) , $i = 1, \dots, N$ 散布如图 1,线性回归的问题是要找 1 条对 N 个点 (x_{i1}, x_{i2}) 的拟合直线 $\hat{x}_2 = a + bx_1$,使离差平方和

$$Q = \sum_{i=1}^N (x_{i2} - \hat{x}_{i2})^2 = \sum_{i=1}^N (x_{i2} - a - bx_{i1})^2$$

最小。

主成分的基本思想是,先对 N 个点 (x_{i1}, x_{i2}) 绘出 1 条“最佳”拟合直线,使得这 N 个点 to 直线的垂直距离的平方和最小,并称此直线为第一主成分。然后,再求与第一主成分相互独立(或者说垂直)的,且与 N 个点 (x_{i1}, x_{i2}) 垂直距离平方和最小的第二主成分,如图 1 2 所示。

对于具有 N 个观测点的情况,可仿照二维的情况类推。如要求第 k 个主成分,必须使它与前 $k - 1$ 个主成分不相关,且使它与各观测点的垂直距离平方和为最小。如此继续,直到求出 m 个主成分。

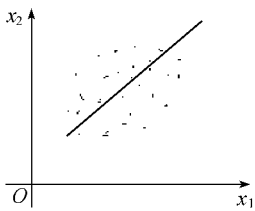


图 1 线性回归示意图

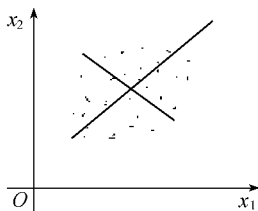


图 2 主成分示意图

从上述变换中不难看出主成分分析是把原来 N 维随机变量的总方差分解成 k 个互不相关的方差和,并且使第一主成分方差最大。这个方差在总方差中比率越大,它的贡献越大,其“综合”原变量能力愈强。在实用中,笔者采用前几个主成分,多数研究者认为当前几个主成分的方差占总方差的 80% 以上即可。

2 数据准备及应用软件

本文采用太湖环境监测中心 2001 年 12 月、2002 年 1 月、2002 年 2 月的太湖中的 22 个站的采样数据(站点位置如图 3 所示),选取的研究指标有 DO、TP、TN、NH₃-N、BOD₅、pH 值、COD_{Mn}、悬浮物(SS)、透明度(SD),从而构成了 22 × 9 的原始矩阵。

采用 SPSS(Statistical Package for Social Science)软件进行分析。SPSS 是由美国 SPSS 公司自 20 世纪 80 年代初开发的大型统计学软件包。它集数据整理、分析过程、结果输出等功能于一身,使用 Windows 的窗口方式展示各种管理数据和分析方法的功能,清晰、直观,易学易用,涵盖面广。

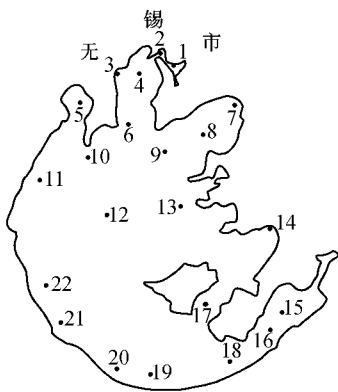


图3 太湖水质监测点位示意图

3 应用与分析

下面以 2001 年 12 月为代表说明数据的分析过程。所选的 9 个水环境指标组成原始数据矩阵,为了去除水环境指标的量纲对计算的影响,将原始矩阵标准化,然后对这些标准化后的数据作主成分分析,得到了两个主成分,分别为第一主成分 G_1 和第二主成分 G_2 。其中 G_1 占总方差的 63.71%, G_2 占总方差的 18.99%,前两个主成分共占总方差的 82.7%,故用 G_1, G_2 取代 9 个水质指标是可行的。

分别用 $DO_i, TP_i, TN_i, NH_3-N_i, BOD_i, pH_i, COD_i, SS_i, SD_i$ 代表第 i 个站的 9 个水质指标,则由计算可得第 i 个站的第一、二主成分分别表示为:

$$G_1(i) = -0.971G'_{DO_i} + 0.942G'_{TP_i} + 0.928G'_{TN_i} + 0.905G'_{NH_3-N_i} + 0.879G'_{BOD_i} - 0.845G'_{pH_i} + 0.838G'_{COD_i} - 0.164G'_{SS_i} - 0.092G'_{SD_i}$$

$$G_2(i) = 0.075G''_{DO_i} + 0.178G''_{TP_i} - 0.148G''_{TN_i} - 0.205G''_{NH_3-N_i} + 0.025G''_{BOD_i} - 0.022G''_{pH_i} + 0.356G''_{COD_i} + 0.903G''_{SS_i} - 0.816G''_{SD_i}$$

可以看出 G_1 在 $DO, TP, TN, NH_3-N, BOD_5, pH$ 值、 COD_{Mn} 指标上有高载荷变量,说明 $G_1(i)$ 体现的是它们的影响,即第一主成分主要是由水体中的 N、P 有机物所引起的,而 $G_2(i)$ 在 SS, SD 变量上有高载荷,它们体现的是水的清澈程度,应用 $G_1(i)$ 和 $G_2(i)$ 的对应关系,将各个站的数据代入就得到了 22 组 $(G_1(i), G_2(i))$ 数据,见表 1。22 个站分布于平面的不同位置,如图 4 所示。

在分析图 4 时,应明确它的含义:因为 G_1 占总方差的 63.71%, G_2 占总方差仅为 18.99%,所以横轴代表了主要的水体污染,而纵轴代表的水体污染比横轴要小得多。 G_1 主要体现了 N、P 等营养物质对水体的污染值:其中第 2 号站 G_1 的值最大,说明此处水体中 N、P 等营养物质含量过高,导致水生

表 1 各水质监测站 G_1, G_2 得分

站号	G_1	G_2	站号	G_1	G_2
1	1.3842	-1.4522	12	-0.5381	0.7238
2	3.6301	0.3901	13	-0.5236	0.4996
3	0.9163	-0.3678	14	-0.5081	-0.1248
4	-0.4343	-0.0556	15	-0.6671	-2.1479
5	1.1962	-0.2403	16	-0.6154	-0.9065
6	-0.4851	-0.0464	17	-0.5770	-0.5874
7	-0.4472	-0.9249	18	-0.1141	3.1083
8	-0.4832	0.5899	19	-0.0943	0.8663
9	-0.5143	0.5377	20	0.0636	0.0204
10	-0.3495	0.1275	21	-0.4212	0.0692
11	0.0374	0.1035	22	-0.4548	-0.1824

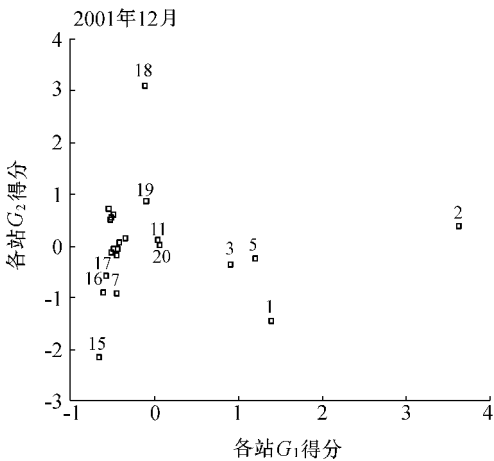


图 4 2001 年 12 月各水质监测站主成分得分

物的大量繁殖和水中溶解氧的降低,其次污染严重的是第 1、3、5 号站,它们的污染原因与 2 号站相同,主要由 N、P 等营养物质引起。而 18 号的 G_2 混浊因子比其他的站高出很多,说明水体的 SS 或 SD 过高。在实测数据中,查到该站样的 SS 高达 308,而 SD 却只有 10,说明可能当时这个区域正在下雨,土壤侵蚀严重,将大量的泥沙冲入体中,造成水体混浊。

用同样的方法可以得到 2002 年 1 月和 2002 年 2 月各站的主成分点图如图 5、6 所示。

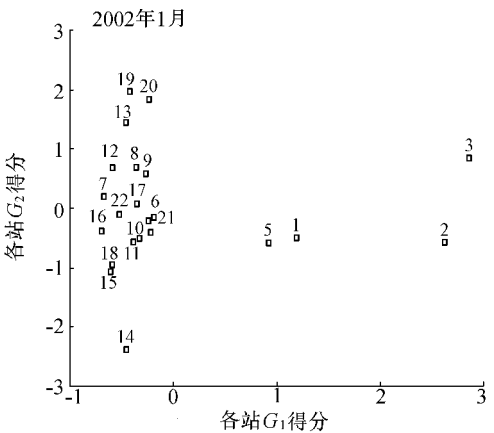


图 5 2002 年 1 月各水质监测站主成分得分

(下转第 63 页)

法都不尽相同,造成同一时期、同一断面上监测数据有较大出入,而有些河流却出现水质监测空白的局面。因此,建议统一布设监测站网,合理布局,增设新的断面,逐步使监测手段和资料整编规范化,使汉江水质监测工作逐步走向协调发展的轨道。

d. 提高全民族的水环境保护意识,依靠广大群众,自觉地保护和监督水环境,广泛开展多种形式的宣传教育活动,对重点河段、水源的水质状况,定期进行监测与评价,并发布水质公告。

(收稿日期 2002-12-17 编辑 胡新宇)

(上接第 50 页)

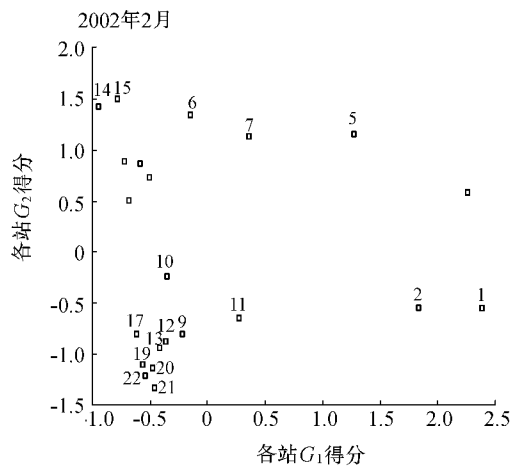


图 6 2002 年 2 月各水质监测站主成分得分

比较以上 2001 年 12 月、2002 年 1 月、2002 年 2 月的各站的主成分点图,可以看出虽然各点水样的水质指标值在变化,但污染的大体程度是不变的,1,2,3,5 号站是水污染严重的地区,它们分别是五里湖、梅园、闸江口、竺山湖。这 4 处位于太湖北部,与

无锡临近,无锡是太湖污染最严重的地区,可见应用的主成分分析法与实际相符。

4 结 语

研究人员可以用 SPSS、SAS 等商业化的统计软件,按照以上方法对每月的太湖水质进行计算分析,从各站点水质的主成分图中可明显看出各站水质的污染情况及其污染原因,将大量的抽象数据变为形象图表,降低分析难度,减少分析人员的工作量。

参考文献：

[1] 李硕,许萌芽.主成分聚类分析法在宁夏水文分区中的应用[J].水文,2002,22(2):44~46.
[2] 张巍,王学军,江耀慈,等.太湖水质指标相关性与富营养化特征分析[J].环境污染与防治,2000(1):50~53.
[3] 张运林,秦伯强.太湖水体富营养化的演变及研究进展[J].上海环境科学,2001,20(6):263~265.
(收稿日期 2003-03-10 编辑 高渭文)

(上接第 57 页)

表 4 2010 年与 2020 年汽车整车及主要零部件用水定额

产品名称	2010 年		2020 年	
	万元产值定额 /m ³	单位产品定额 /m ³	万元产值定额 /m ³	单位产品定额 /m ³
东风载重卡车	7.6	54.6	4.6	33.1
东风轻卡	8.5	33.2	4.3	16.8
神龙系列轿车	0.64	8.7	0.43	5.9
卡车发动机	2.6	3.3	1.6	2.1
卡车底盘	27.2	7.8	14.6	4.7
卡车车身	2.8	3.2	1.7	2.0
卡车车厢	5.5	12.6	3.3	7.6
卡车车架	4.0	1.6	2.5	1.0
卡车总装配 (装配整车)	0.36	2.4	0.22	1.5

5 结 论

a. 神龙汽车公司是东汽车公司与法国雪铁龙公

司合资的企业,汽车万元产值取水量、用水定额和耗水率大大低于载重车公司和轻卡公司,也低于目前国内同类产品生产用水定额,代表了目前国内汽车生产的先进水平。

b. 通过分析、预测东汽公司的汽车及主要汽车零部件用水定额,可为十堰市及湖北省水资源规划、管理和实施计划用水、节约用水工作提供重要依据。

c. 东汽公司节水潜力较大,主要是提高生产用水的重复利用率,杜绝输水管道跑、冒、滴、漏现象;提高工艺水平与技术水平。

参考文献：

[1] 李智慧,徐卫东,王力,等.水平衡测试技术[M].太原:山西科学技术出版社,1998.
[2] 李滨江,郭姚生.佳木斯市水资源评价中需用水系统若干参数的确定[J].水文,1999,19(2).
(收稿日期 2003-05-16 编辑 胡新宇)