

# 主成分-聚类复合模型在水环境管理中的应用 ——以松花江吉林段为例

张 妍<sup>1</sup>, 尚金城<sup>2</sup>, 于相毅<sup>2</sup>

(1. 北京师范大学环境学院环境模拟与污染控制国家重点实验室, 北京 100875; 2. 东北师范大学城市与环境科学学院, 吉林 长春 130024)

**摘要:** 以松花江吉林段 5 年的监测资料为依据, 采用主成分分析与聚类分析的复合模型, 定量诊断出水污染因子对松花江吉林段的污染贡献率, 合理划分出各类水环境管理区域, 并提出相应的水环境管理措施。结果表明: 松花江吉林段主要污染物是挥发酚、COD 和氨氮, 表现为有机污染和非点源污染; 在水环境管理分区中, 丰满 - 龙潭桥区域属于轻污染, 为水源保护江段, 龙潭桥 - 哨口区域属于重污染, 为水环境控制江段, 哨口 - 白旗区域属于中度污染状况, 为水环境自净江段。这一工作的开展对于实施河流总氮总量控制, 实现松花江吉林段水域的科学化分类管理具有一定的参考价值。

**关 键 词:** 主成分分析; 聚类分析; 水环境管理; 松花江

**中图分类号:** X522      **文献标识码:** A      **文章编号:** 1001-6791(2005)04-0592-04

水环境研究是一个复杂的多因子体系, 无论是地表水还是地下水, 极易受到各种因素的影响和污染。尤其是流域水环境, 对其水资源的开发利用, 往往会引起下游河段、湖泊的水质变化, 城市水体富营养化等环境问题。因此必须加强流域中各段水环境的管理, 合理划分流域水环境管理区, 这对于保护水源、改善环境、维持生态平衡、促进经济发展具有重要意义<sup>[1,2]</sup>。

松花江是我国七大江河之一, 是吉林、黑龙江两省的重要水资源。而松花江流经的吉林市是一个以化工、电力、造纸、化纤、制药为主的化工城。随着松花江流域经济的高速发展, 特别是沿江工业的发展, 使松花江水质污染日趋严重<sup>[3,4]</sup>。松花江吉林江段是从丰满大坝(北纬 43°40′、东经 126°40′)至舒兰县法特乡黄鱼圈屯西北处(北纬 44°32′、东经 126°25′), 全长 126 km。采样监测江段是从丰满大坝到舒兰白旗镇白旗渡口, 全长 112 km。共设置 7 个采样监测断面, 分别为丰满、龙潭桥、哈达湾、哨口、九站、安达、白旗。

## 1 复合模型

主成分分析及聚类分析方法常被应用于水环境管理研究中。流域主要污染物的辨识、地下水水质脆弱性指标体系的确定等常采用主成分分析方法<sup>[5,6]</sup>, 而王双合、郭晓君等分别用聚类法分析地表水水质污染的类型、程度及趋势<sup>[7,8]</sup>, 实现了分析过程与分析结果的科学真实性。常见的水环境管理复合模型有模糊聚类评价法、多元统计方法等<sup>[9,10]</sup>, 而本文针对已有水环境管理复合模型在应用中存在的主观随意性问题, 提出将主因子分析与聚类分析方法集成, 研究水环境管理问题。

由于水的弥散作用, 水质的变化是一个连续、渐变的复杂过程, 采用主成分分析和聚类分析相结合的模型进行水质评价与分区是水环境管理研究的一大进步, 它克服了过去评价方法的不足。主成分分析(Principal Component Analysis), 又称多元分析, 它是将分散在一组变量上的信息集中到某几个综合指标(主成分)上的探索性统计分析方法, 以便利用主成分描述数据集内部结构, 实际上也起着数据降维的作用。聚类分析(Cluster Analysis)是根据事物本身的特性研究个体分类的方法, 依据同类中个体相似性和不同类中个体差异性对观测量

收稿日期: 2004-04-07; 修订日期: 2004-06-30

作者简介: 张 妍(1973-), 女, 吉林镇赉人, 博士, 主要从事区域环境规划与管理等研究。

通讯作者: 尚金城, E-mail: shangjc@nenu.edu.cn

进行聚类。

设污染物监测指标为  $x_1, x_2, \dots, x_n$ , 用  $x_{ki}$  表示第  $k$  个监测断面的第  $i$  个分析指标的数据, 建立数据矩阵。按特征值大于 1 的原则, 提取出相互独立的  $m$  个主成分, 从而将信息重复的指标剔除。每个主成分概括原始指标信息的程度用其贡献率表示, 以给定的阈值(本文为 85%)为确定主成分个数的依据, 当前  $i$  个主成分的累积贡献率大于等于阈值时, 表明这  $i$  个主成分已充分概括了大多数原始指标的信息, 就以这  $i$  个主成分作为新的分析指标。这样, 原始分析指标由多个转变为少数几个, 分析指标数量的减少并未使信息损失多少<sup>[5]</sup>。利用主成分就可以方便、直观的揭示流域的主要污染物, 然后再通过对主成分的聚类分析, 将水域划分成几个区域<sup>[8]</sup>, 每个区域的污染程度与主要污染物直观的显现出来, 就可以有针对性的对每个区域进行水环境管理, 提出相应的减缓与控制水污染的措施, 避免了一概而论。

本文利用复合模型的结果, 参考实际环境条件, 得出水环境管理分区, 提高了水质评价与分区的客观性<sup>[11,12]</sup>。并尝试将复合模型应用于松花江吉林江段水环境质量评价与分区中。

## 2 应用实例

根据松花江吉林江段污染状况与污染源情况, 从众多污染物中选取 6 类污染物作为分析指标, 分别是: 油、氨氮、高锰酸盐指数、挥发酚、5 日生化需氧量、悬浮物。统计出各指标 5 年的平均值, 并进行标准化处理, 作为分析的原始数据。

### 2.1 主成分分析

主成分分析的主要目的是将具有相近的因子载荷的各个变量置于一个公因子之下, 当初始因子不能典型的代表变量的含义时, 应对因子载荷矩阵实行旋转, 使因子载荷量向两个极端发展(图 1), 以便对因子的意义做出更合理的解释, 输出结果列入表 1、表 2。运用主成分分析方法, 提取出两个主成分(累积贡献率为 86.983%)作为松花江吉林段水环境污染物的分析指标。

表 1 因子的特征值

Table 1 Eigenvalue of factors

特征值	因子旋转前			因子旋转后		
	特征根	贡献率	累积贡献率	特征根	贡献率	累积贡献率
第一主成分	4.047	67.896	67.896	3.114	51.906	51.906
第二主成分	1.145	19.086	86.983	2.105	35.076	86.983

表 2 旋转因子载荷矩阵

Table 2 Rotated component matrix

主成分	挥发酚	生化需氧量	油	高锰酸盐指数	悬浮物	氨氮
第一主成分	0.926	0.850	0.825	0.715	- 3.17E-02	0.585
第二主成分	0.218	0.427	- 5.98E-02	0.682	0.967	0.687

由表 1 可以看出, 从 6 种污染物中提取两个主成分就已提供了原有信息的 86.983%, 满足了因子选取的原则(>85%), 说明提取的两个主成分就可以很好的反映原始环境信息。

由表 2 和图 1 可以看出, 第一主成分主要是由挥发酚、生化需氧量、油和高锰酸盐指数组成, 表现为水体的有机污染, 生化需量和高锰酸钾指数都有很高的载荷量, 分别是 0.850 和 0.715; 第二主成分由悬浮物、氨氮组成, 因子载荷量分别为 0.967 和 0.687, 表现为以氮污染物为主要特征, 主要来源于农业施用化肥农药引起的非点源污染。

### 2.2 聚类分析

根据对两个主成分的聚类分析, 对监测断面进行分类。聚类方法采用最远邻法, 对等间隔测度的变量使用欧式距离平方作为类间距离, 为使问题简化、直观, 更加突出主导因子及制图方便, 取第一和第二主成分作为坐标轴, 计算 7 个监测断面在这两个坐标轴上的数值, 并做成二维坐标图(图 2)。

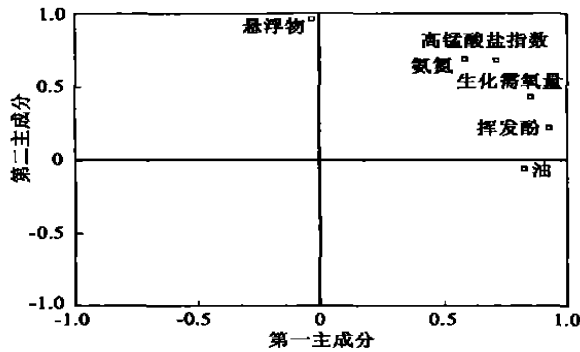


图1 旋转因子载荷图

Fig. 1 Rotated principal component

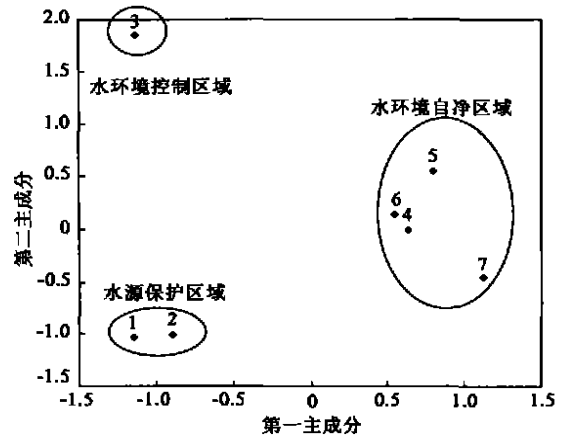


图2 7个监测断面的聚类二维坐标图

Fig. 2 Clustering coordinate chart of seven monitoring sections

由二维坐标图可以看出, 7个监测断面分为3类的类间距比较大, 说明各类的特点比较突出, 经过实际调查, 此种分类与现实情况基本相符。丰满和龙潭桥监测断面为一类, 属于轻污染; 九站、安达、哨口和白旗4个为一类, 属于中度污染; 而哈达湾监测断面自成一类, 为重污染状况。

### 2.3 结果分析

通过对主成分进行聚类分析, 可将松花江吉林江段划分为3个水环境管理区域。丰满-龙潭桥区域属于轻污染, 为水源保护江段, 龙潭桥-哨口区域属于重污染, 为水环境控制江段, 哨口-白旗区域属于中度污染状况, 为水环境自净江段。

**第一类: 水源保护区域。**这一部分水质较好, 沿江没有大的工业污染源, 只有少量的生活污水排入江中。对这个区域的水环境管理, 可以实行生活污水的集中化, 将直接排入江中的生活污水经过排污管道, 统一集中于建成的吉林市污水处理厂中进行处理。同时要避免在经济开发过程中, 在此区域增添新的直接入江的污水排放源。这样, 才能减少此区域COD的含量, 保证水质的良好状态, 为下游用水提供一个相对安全的环境。

**第二类: 水环境控制区域。**这部分是松花江吉林江段的重污染区, 吉林市区内各类污染源的污水几乎都在这个区域排入江中。主要工业污染源有吉化染料厂、吉林纸业有限公司、吉林化纤股份有限公司、吉林新源玉米开发公司、吉林天河酒精公司等。为更好的了解此区域的污染情况, 设置了多个监测断面, 监测数据经过处理显示, 在哈达湾监测断面处形成了一个明显的污染带。对于这个区域的水环境管理, 就必须首先减少直接入江的污水排放源, 将各个主要污染源的排污干线进行汇总, 科学的减少排污点; 其次, 加强对各个污染大户企业的监督与管理, 督促企业进行技术改造。有条件的企业, 一定要在内部建立污水处理厂, 条件不成熟的企业, 尽可能将其排污管线连接至吉林市污水处理厂, 减少直接排入江中的、未经处理的工业污水量, 同时, 对于污染较为严重的小企业, 一定要实现“关、停、并、转”; 最后, 要充分发挥吉林市污水处理厂的作用, 除加强对生活污水的处理外, 还应加强对于工业废水的处理力度。

**第三类: 水环境自净区域。**这部分间隔距离较长, 江面开阔, 水流速缓慢, 污染物的混合比较均匀, 并且沿江大企业污染源的排放也较少。对于这个区域的水环境管理, 要减少污染源的数量, 充分发挥水体自净的作用。

## 3 讨 论

(1) 以主成分分析方法提取的两个主成分, 其累积贡献率达到85%以上, 也就是说, 两个主成分几乎包含了原始数据所有的信息量。因此, 以这两个主成分来对监测断面进行分类, 具有一定的科学性与可靠性。与其他分类方式比较, 其精确度也比较高。根据因子的得分矩阵, 还可以构造线性回归函数, 对水质进行预测。

(2) 对于流域水环境管理分区的问题,往往是按照区域内环境组合特征与流域水环境的特征来进行划分,具有一定的主观因素。同时由于影响水环境的因素众多,使区划者面对海量的监测数据与错综复杂的影响因子,往往无从下手。主成分分析与聚类分析复合模型,就成为了区划者解决此类问题的首选工具。主成分分析可以对原始变量因子进行简化并提取主成分,找出影响环境的主要因子,并通过主成分进行聚类分析,对流域的监测断面进行分类。此法科学严谨,并通过海量数据的计算机处理,提高了水环境管理区划的科学性与可靠性。对于环境区划者,具有一定的实用参考价值。

#### 参考文献:

- [1] 王树功,孙新庚.小东江流域管理的思考[J].环境与开发,2000,15(4):50-51.
- [2] 郑亚西.流域管理模式与水污染防治[J].四川师范大学学报(自然科学版),2003,26(4):417-420.
- [3] 马逊风,刘伟,于力.第二松花江吉林江段水质状况分析[J].松辽学刊(自然科学版),2000,2(1):39-41.
- [4] 包存宽,张敏,尚金城.吉林省松花江中下游污染物排放总量控制[J].城市环境与城市生态,2000,13(3):28-30.
- [5] 姚桂基.青海湟水河主要污染物主成分分析[J].青海环境,2002,12(3):100-102.
- [6] 雷静,张思聪.唐山市平原区地下水脆弱性评价研究[J].环境科学学报,2003,23(1):94-99.
- [7] 王双合.用灰色聚类法对石羊河下游水环境质量的综合评价[J].甘肃水利水电技术,2000,(1):59-63.
- [8] 郭晓君.南通六县(市)城镇地表水水质的聚类分析[J].南通工学院学报,2003,2(4):22-24.
- [9] 王静.湟水水环境监测断面优化设置研究[J].青海环境,2002,12(1):27-30.
- [10] Roberto A, Daniela G, et al. Pollution of a river basin and its evolution with time studied by multivariate statistical analysis[J]. Analytica chimica ACTA, 1995, 310:15-25.
- [11] 张祖陆,彭利民,孙庆义.南四湖水水质污染综合评价及水质分区[J].地理学与国土研究,1998,14(4):30-33.
- [12] 胡立平,邢久生.江西省水功能区划分与水质达标分析[J].江西水利科技,2003,29(3):154-157.

## Application of principal component-cluster analysis complex model to water environment management :Case study in Songhua River in Jilin section as an example

ZHANG Yan<sup>1</sup>, SHANG Jin-cheng<sup>2</sup>, YU Xiang-yi<sup>2</sup>

(1. School of Environment, Beijing Normal University, State Key Joint Laboratories of Environmental Simulation and Pollution Control, Beijing 100875, China; 2. Department of Environmental Science and Engineering, Northeast Normal University, Changchun 130024, China)

**Abstract:** Based on the monitoring data from Jilin section in Songhua River, the contribution rate of water pollution factors to Jilin section in Songhua River is calculated quantitatively, and all kinds of water environmentally managed areas are divided rationally with the complex model of the principal component-cluster analysis. According to the assessment results, the corresponding measures of water environment management are put forward. The results show that the main pollutions in Jilin section in Songhua River are volatile phenol, Chemical Oxygen Demand (COD) and Ammonia-nitrogen ( $\text{NH}_3\text{-N}$ ), which have their pollution characteristics of organic Pollution and non-point Source Pollution. The results also show that in the seven selected monitoring sections (i. e., Fengman, Longtanqiao, Hadawan, Shaokou, Jiuzhan, Anda and Baiqi), the water quality in the sections between Fengman and Longtanqiao, Longtanqiao and Shaokou, Shaokou and Baiqi is lightly, heavily and moderately polluted respectively, which means that the water resources in these sections should be preserved, controlled and exerted respectively. The development of the study has certain reference value to gross control of the regional water environment, and can implement the scientificall classified management of the water areas of Songhua River in Jilin Province.

**Key words:** principal component analysis; cluster analysis; water environment management; Songhua River