

基于多元统计分析的水环境质量评价及趋势分析

蒋同斌¹, 李继玲²

(1 淮阴工学院 数理学院, 江苏 淮安 223001;
2 江苏经贸职业技术学院 信息技术系, 江苏 南京 211168)

[摘要] 用因子分析的方法, 找出影响水质的主成分; 用聚类分析的方法将样本水质分类, 结合分类结果对各样本水质进行评价; 用秩相关系数检验法作污染趋势分析. 作为上述方法的应用, 对淮河淮安段 2006~2007 年的水质进行分析评价并对其作污染趋势分析, 结果表明: 影响淮河水质的因素依次为 NH_4-N 、 TN 、 COD_{Cr} 、 COD_{Mn} 及 DO 等; 而淮河淮安段的水质在逐渐好转.

[关键词] 因子分析, 聚类分析, 方差分析, 水质评价, 秩相关检验

[中图分类号] X143 [文献标识码] A [文章编号] 1672-1292(2010)04-0052-05

Water Environmental Quality Evaluation and Trend Analysis
Based on Multivariate Statistical Analysis

Jiang Tongbin¹, Li Jiling²

(1 School of Mathematics and Physics Huaiyin Institute of Technology, Huai'an 223001, China
2 Department of Information Technology, Jiangsu Vocational and Technical Institute, Nanjing 211168, China)

Abstract Through factor analysis, the principal components which mainly affect the water quality are obtained. The water quality of samples are classified by using cluster analysis method, the water quality of each samples are evaluated based on the classification results. Water pollution trend is analysed with rank correlation coefficient test. By applying the above method, water quality of Huaihe river in Huai'an area from 2006 to 2007 is evaluated, and water pollution trend is analyzed. The research result demonstrate that the factors affecting water quality of Huaihe River are NH_4-N 、 TN 、 COD_{Cr} 、 COD_{Mn} & DO in order. Water quality of Huaihe river in Huai'an area is becoming good.

Key words factor analysis, cluster analysis, variance analysis, water quality evaluation, rank correlation

水生态系统不仅提供了维持人类生活和生产活动的基础产品, 还具有维持自然生态系统结构、生态过程与区域生态环境的功能. 所以水环境的质量备受关注. 水质评价就是以定量的方式直观表征水环境的质量状况. 目前, 水质评价的方法有综合指数法^[1]、模糊综合评判法^[2]、五元联系数^[3]、判别分析方法^[4]、神经网络法^[7]、指数模型^[6]、灰色聚类分析^[7]、灰色关联分析法^[8]等. 本文探讨用主成分分析法^[9], 找出影响水质的主成分, 用聚类分析法^[10]将样本水质分类, 并用秩相关系数检验法^[10]作污染趋势分析. 本文的计算由统计分析软件 SPSS 完成.

1 建立统计分析模型

1.1 因子分析的数学模型

因子分析起源于 20 世纪初, 它是由 K Pearson 和 C Spearman 等为定义和测定智力所作的统计分析. 因子分析是主成分分析的推广和发展, 是用少数几个因子来描述许多因素之间的联系, 以较少因子反映原料大部分信息的统计学方法.

设有 p 维可观测随机向量 $X = (X_1, X_2, \dots, X_p)'$, 其均值 $\mu = (\mu_1, \mu_2, \dots, \mu_p)'$, 协方差矩阵 $\Sigma = (\sigma_{ij})$,

收稿日期: 2010-06-12
基金项目: 淮安市科技发展规划项目 (HAS07014).
通讯联系人: 蒋同斌, 副教授, 研究方向: 概率与统计应用. E-mail: jiangtongbin2006@163.com

其因子分析的一般模型为:

$$\begin{cases} x_1 = u_{11}y_1 + u_{12}y_2 + \dots u_{1p}y_p + \varepsilon_1 \\ x_2 = u_{21}y_1 + u_{22}y_2 + \dots u_{2p}y_p + \varepsilon_2 \\ \dots \\ x_p = u_{p1}y_1 + u_{p2}y_2 + \dots u_{pp}y_p + \varepsilon_p \end{cases}$$

矩阵形式为 $X = UY + \varepsilon$ 其中, Y 为因子变量; U 为因子载荷矩阵, u_{ij} 为因子载荷, 是第 i 个原有变量在第 j 个因子变量上的负荷, 即 x_i 在第 j 个因子变量上的相对重要性, u_{ij} 绝对值越大, Y_j 和 x_i 关系越强; ε 为特殊因子, 表示原有变量不能被因子变量所解释的部分.

因子分析的基本步骤为:

- (1) 确定原若干变量是否适合于因子分析, 即对原变量作相关性分析;
- (2) 构造因子变量, 常用极大似然法、最小二乘法等, 本文用主成分分析法;
- (3) 计算因子变量的得分.

主成分分析是将分量相关的原始变量, 借助于一个正交变换, 转化为不相关的新变量, 并以方差作为信息量的测度, 对新变量进行降维, 取累计贡献率大的若干成分作为主成分. 这些主成分能够反映原始变量的绝大部分信息, 它们通常表示为原始变量的某种线性组合. 为使主成分所含信息互不重叠, 应要求它们之间互不相关.

定义 设 $X = (X_1, X_2, \dots, X_p)'$ 为 p 维随机向量. 称 $Y_i = a_i'X$ 为 X 的第 i 主成分 ($i = 1, 2, \dots, p$), 如果:

- (1) $a_i'a_i = 1$ ($i = 1, 2, \dots, p$);
- (2) 当 $i > 1$ 时, $a_i'\Sigma a_j = 0$ ($j = 1, 2, \dots, i-1$);
- (3) $\text{var}(Y_i) = a_i'\Sigma a_i = \max_{a_i'a_i=1} \text{var}(a_i'X)$.

其中, Σ 及 $\text{var}(Y)$ 分别为 $X = (X_1, X_2, \dots, X_p)'$ 的协方差矩阵及 Y 的方差.

主成分分析的步骤如下:

设 $X = (X_1, X_2, \dots, X_p)$ 为 p 维样本数据.

- (1) 数据的标准化, $x_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j}$, $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$. 其中, n 为样本点数, p 为原变量数, $x_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$, $S_j = \left[\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - x_j)^2 \right]^{\frac{1}{2}}$;
- (2) 计算数据 $(x_{ij})_{n \times p}$ 的协方差矩阵 R , 当数据标准化后, R 即为相关系数矩阵;
- (3) 求 R 的特征值 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$ 称 $\lambda_k / \sum_{i=1}^p \lambda_i$ 为成分 Y_k 的贡献率, $\sum_{k=1}^m \lambda_k / \sum_{i=1}^p \lambda_i$ 为成分 Y_1, \dots, Y_m ($m < p$) 的累计贡献率, 取 m 使 $\sum_{k=1}^m \lambda_k / \sum_{i=1}^p \lambda_i$ 达到一定值 (一般取 80%);
- (4) 求 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$ 对应的特征向量 a_1, a_2, \dots, a_m , 它们标准正交; 则 $Y_i = a_i'X$ 为 X 的第 i 主成分 ($i = 1, 2, \dots, m \leq p$);
- (5) 计算主成分得分, 一般方法有回归法、Bartlette法、Anderson-Rubin法等. 将主成分表示为原变量的线性组合, 并代入样本数据, 计算出相应的主成分得分.

1.2 聚类分析的基本原理和方法

聚类分析是根据事物间的不同特征、亲疏程度和相似性等关系进行分类的一种方法. 聚类分析方法主要有两种, 快速聚类分析方法 (K-Means Cluster Analysis) 和层次聚类分析方法 (Hierarchical Cluster Analysis). 聚类分析的形式也分为两种, Q型聚类 (对样本进行分类) 和 R型聚类 (对研究对象的观察变量进行分类). 本文介绍并使用 Q型聚类.

快速聚类分析的一般过程为:

- (1) 对样本指标量化, 并标准化处理, 使样本指标成为具有可比性的数据;
- (2) 根据实际解决问题的需要, 确定聚类成多少类 (例如 k 类);
- (3) 确定 k 个类的初始类中心点;
- (4) 计算所有样本点到 k 个类中心点的欧氏距离, 按照 k 个类中心点距离最短原则, 形成一个新的 k

类,完成一次迭代过程;

(5) 重新确定类中心点: 计算各类中变量值的均值,并以均值点作为新的类中心点;

(6) 重复 (4)、(5) 的计算过程,直到达到指定的迭代次数或终止迭代的判断要求为止;也可按初始类中心点分类,仅作一次迭代计算. 本文作一次迭代进行水环境质量评价.

1.3 秩相关检验的基本原理

秩相关检验法是用秩相关系数检验二元定序变量间线性相关程度的一种方法. 秩相关系数的计算公式为: $r = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$. 其中, n 为样本容量; d_i 为两种秩评定下第 i 个体的两种秩次之差, $-1 \leq r \leq 1$. 将 r 的绝对值与秩相关系数统计表中的临界值 w 进行比较,若 $r > w$ 则表明变化趋势有显著意义;若 $r > w$ 且是负数,则表明为下降趋势.

2 应用实例

2.1 水质样本

《地表水环境质量标准》(GB3838—2002)规定的评价因子共为 24 项,然而在不同时期、不同地点选取的评价因子也有所不同. 根据具体情况,对淮河水质取评价因子为: 高锰酸盐指数 COD_{Mn} 、化学需氧量 COD_{Cr} 、生化需氧量 BOD_5 、总氮含量 TN 、总含磷量 TP 及 $NH_4 - N$ 等. 本文取 2006~2007 年淮河淮安段水质月监测数据均值作为样本,监测数据见表 1.

表 1 2006~2007 年淮河水质监测数据
Table 1 Huaihe River water quality monitor data from year 2006 to 2007

年份	因子	月											
		1	2	3	4	5	6	7	8	9	10	11	12
2006	DO	9.77	10.73	7.50	8.13	8.03	7.83	7.47	7.03	6.67	7.73	6.50	8.27
	COD_{Mn}	3.17	3.67	4.77	4.37	3.93	4.07	3.60	3.10	3.67	3.07	3.83	3.97
	BOD_5	3.17	3.23	3.53	3.73	3.87	3.77	3.30	3.33	3.17	3.40	3.17	3.10
	$NH_4 - N$	1.23	1.46	1.33	1.45	1.09	0.97	1.46	0.95	0.99	0.95	0.95	0.88
	COD_{Cr}	13.67	12.67	10.67	10.67	13.33	12.67	12.33	12.33	12.33	12.67	11.33	12.33
	TN	4.08	3.65	3.67	4.57	2.71	2.24	2.61	2.53	2.61	2.55	2.89	2.27
	TP	0.20	0.26	0.19	0.20	0.21	0.19	1.80	0.21	0.23	0.23	0.24	0.16
2007	DO	10.88	11.50	9.15	9.28	7.99	7.50	5.37	6.22	6.03	7.28	6.56	9.70
	COD_{Mn}	3.84	4.20	5.07	4.35	4.18	4.30	5.37	4.45	4.25	3.25	4.28	4.08
	BOD_5	2.82	2.78	2.47	3.53	2.42	3.22	2.85	3.47	2.72	2.70	2.69	2.77
	$NH_4 - N$	1.19	1.32	1.79	1.29	0.82	0.81	0.97	0.77	0.72	0.30	0.58	0.50
	COD_{Cr}	15.80	15.00	21.80	17.83	14.00	18.50	18.60	15.33	15.80	14.83	15.73	14.33
	TN	2.97	3.25	3.68	3.36	2.89	1.97	2.98	2.06	2.70	2.35	1.66	1.87
	TP	0.08	0.18	0.16	0.17	0.12	0.18	0.15	0.16	0.14	0.11	0.10	0.14

注: 资料来源于淮安市环境监测中心站.

2.2 淮河水质的主成分分析

用软件 SPSS for Windows^[10]对表 1 数据进行分析,得到的结果如表 2~表 4 所示.

表 2 主成分提取结果
Table 2 Principal components obtained

成分	1	2	3	4	5	6	7
特征值	2.185	1.906	1.11	0.936	0.48	0.291	0.093
贡献率	31.209	27.228	15.85	13.374	6.858	4.151	1.33
累计贡献率	31.209	58.437	74.287	87.662	94.52	98.67	100

表 2 第二行为协方差阵的 6 个特征值,由第三行知,成分 1、2、3、4 的贡献率分别为 31.209%、27.228%、15.850%、13.374%,累计贡献率 $\sum_{k=1}^4 \lambda_k / \sum_{i=1}^7 \lambda_i$ 为 87.662%,即第一、二、三、四成分包括了原监测变量 87.662% 的信息,所以取成分 1、2、3、4 作为主成分.

表 3 成分载荷矩阵
Table 3 Component matrix

成分		DO	COD _{Mn}	BOD ₅	NH ₄ -N	COD _{Cr}	TN	TP
	1	0.561	-0.059	0.425	0.901	-0.286	0.834	0.308
	2	0.264	0.729	-0.563	0.285	0.823	0.28	-0.39
	3	-0.678	0.57	0.366	0.194	0.082	-0.013	0.382
	4	0.131	-0.172	-0.419	0.126	0.225	-0.23	0.771

设 x_i 为第 i 监测变量, y_i 为第 i 主成分, 由表 3 可知: 主成分 1 与 NH_4-N 及 TN 有较强相关性; 主成分 2 与 COD_{Mn} 及 COD_{Cr} 有较强相关性; 主成分 3 则与 DO 有较强相关性. 由此, 影响淮河水质的因素依次为 NH_4-N 、 TN 、 COD_{Cr} 、 COD_{Mn} 及 DO 等. 且

$$\begin{cases} x_1 = 0.561y_1 + 0.264y_2 - 0.678y_3 + 0.131y_4 \\ \dots \\ x_7 = 0.308y_1 - 0.390y_2 + 0.382y_3 + 0.771y_4 \end{cases}.$$

2.3 淮河水质的评价

根据《地表水环境质量标准》评价因子溶解氧 DO 、化学需氧量 COD_{Cr} 、生化需氧量 BOD_5 、总含磷量 TP 、高锰酸盐指数 COD_{Mn} 、总氮含量 TN 及 NH_4-N 等的质量标准见表 4

表 4 地表水质分类标准
Table 4 Classification standard of surface water quality (mg/L)

水质标准	评价指标							
	DO ≥	COD _{Mn} ≤	BOD ₅ ≤	NH ₄ - N ≤	COD _{Cr} ≤	TN ≤	TP ≤	
	I	7.5	2	3	0.15	15	0.2	0.02
	II	6	4	3	0.5	15	0.5	0.1
	III	5	6	4	1.0	15	1.0	0.1
	IV	3	8	6	1.5	20	1.5	0.2
	V	2	10	10	2.0	25	2.0	0.2

用公式 $d_{ij} = \frac{c_{ij}}{c_j}$ 将表 1 的数据标准化, 其中 d_{ij} 、 c_{ij} 、 c_j 分别为监测数据的标准化数据、原数据及对应因子分类标准数据均值. 取各类水质标准的区间中点, 将其标准化作为聚类分析的类中心点 (见表 5), 用 1.2 的原理和方法, 对淮河淮安段 2006~2007 年水质进行分类, 结果见表 6

表 5 地表水质分类的类中心点及其标准化
Table 5 Cenetr point of surface water quality classification and standardization

评价指标	水质分类的类中心点							类中心点的标准化						
	DO	COD _{Mn}	BOD ₅	NH ₄ -N	COD _{Cr}	TN	TP	DO	COD _{Mn}	BOD ₅	NH ₄ -N	COD _{Cr}	TN	TP
水质标准	I	7.75	1.00	1.50	0.08	7.50	0.10	0.01	1.46	0.2	0.36	0.09	0.48	0.10
	II	6.75	3.0	3.00	0.33	15.00	0.35	0.06	1.27	0.6	0.71	0.39	0.97	0.58
	III	5.50	5.0	3.50	0.75	15.00	0.75	0.10	1.04	1	0.83	0.90	0.97	0.96
	IV	4.00	7.0	5.00	1.25	17.50	1.25	0.15	0.75	1.4	1.19	1.51	1.13	1.50
	V	2.50	9.0	8.00	1.75	22.50	1.75	0.20	0.47	1.8	1.90	2.11	1.45	1.92

表 6 淮河淮安段 2006~2007 年水质分类结果

Table 6 The results of water quality classification of Huahe River in Huai'an area from year 2006 to 2007

月 水质 年	1	2	3	4	5	6	7	8	9	10	11	12
2006	V	V	V	V	IV	IV	V	IV	IV	IV	IV	IV
2007	IV	IV	V	IV	IV	IV	IV	IV	IV	III	III	III

2.4 水质的污染趋势分析

用 SPSS 对表 1 的水质样本进行秩相关 (Spearman) 检验, 结果如表 7

2.5 结论

根据表 2 及表 3 影响淮河水质的因素依次为 NH_4-N 、 TN 、 COD_{Cr} 、 COD_{Mn} 及 DO 等; 由表 6 知, 淮河淮安段的水质 2006 年有 5 个月为 V 级, 2007 年只有一个月为 V 级, 且 2007 年的最后 3 个月为 III 级, 说明淮河淮

安段的水质在逐渐好转. 由表 7 知, 污染物 DO 的相伴概率 $0.12 > 0.01$, 其他污染物的浓度随时间变化的趋势都不明显, 且其中有 5 个监测因子的浓度与时间呈负相关, 即淮河淮安段的水质状况在逐月好转.

表 7 水质污染的秩相关检验

Table 7 Rank correlation coefficient test of water pollution

Spearm an							Sig (2- tailed)						
DO	COD _{Mn}	BOD ₅	NH ₄ -N	COD _{Cr}	TN	TP	DO	COD _{Mn}	BOD ₅	NH ₄ -N	COD _{Cr}	TN	TP
- 0.33	0.36	- 0.63	- 0.71	0.68	- 0.53	- 0.77	0.12	0.08	0.00	0.00	0.00	0.01	0.00

3 结语

水环境是一种由多介质组成的多元体系, 涉及到大量的污染因素和变量, 各变量分别从某一方面反映了水环境的质量, 但由于污染因素具有高度的复杂性、随机性和综合性, 依据全部评价因子的监测数据对水环境的质量作出综合评价, 即使在目前可以借助计算机进行大量数据处理的条件下, 仍难以实现. 因子分析法能较全面、客观地反映水资源各项指标的综合污染程度, 可客观地找出影响水质的主要因素, 不需要主观地确定各监测变量的权重, 是一种比较理想的方法. 而统计分析软件 SPSS 的应用, 使该方法具有可推广性.

[参考文献] (References)

[1] 王文强. 综合指数法在地下水水质评价中的应用 [J]. 水利科技与经济, 2008(1): 54-55.
Wang Wenqiang. Application of aggregative index number method in groundwater quality valuation [J]. Water Conservancy Science and Technology and Economy, 2008(1): 54-55. (in Chinese)

[2] 刘荣珍, 赵军. 模糊评价模型在长江水质评价中的应用 [J]. 兰州交通大学学报: 自然科学版, 2007(6): 50-52.
Liu Rongzhen, Zhao Jun. Application of the fuzzy evaluation model to water quality of the Yangtze River [J]. Journal of Lanzhou Jiaotong University: Natural Science Edition, 2007(6): 50-52. (in Chinese)

[3] 陈丽燕, 付强, 魏丽丽, 等. 五元联系数在湖泊水质综合评价中的应用 [J]. 环境科学研究, 2008(3): 82-86.
Chen Liyan, Fu Qiang, Wei Lili, et al. Application of five-element connection number to the quality assessment of eutrophication in lakes [J]. Research of Environmental Sciences, 2008(3): 82-86. (in Chinese)

[4] 辛欣, 卢文喜, 李海杰, 等. 判别分析方法在水质评价中的应用 [J]. 环境科学与技术, 2008(1): 113-115.
Xin Xin, Lu Wenxi, Li Haijie, et al. Application of discriminant analysis method in water quality assessment [J]. Environmental Science and Technology, 2008(1): 113-115. (in Chinese)

[5] 刘金生, 周焕银, 刘金辉, 等. 基于 BP 神经网络的抚河水环境质量评价研究 [J]. 东华理工大学学报: 自然科学版, 2008(1): 85-88.
Liu Jinsheng, Zhou Huanyin, Liu Jinhui, et al. Study on water environmental quality assessment of Fu River based on the BP neural network [J]. Journal of East China Institute of Technology: Natural Science Edition, 2008(1): 85-88. (in Chinese)

[6] 董永权. 基于指数模型的长江水质的评价和预测 [J]. 唐山师范学院学报, 2008(2): 22-25.
Dong Yongquan. Evaluation and forecast of Changjiang river's water contamination with exponential model [J]. Journal of Tangshan Teachers College, 2008(2): 22-25. (in Chinese)

[7] 王海玲, 刘廷玺, 王敏, 等. 达拉特旗地表水水质灰色聚类分析评价 [J]. 内蒙古农业大学学报: 自然科学版, 2006(1): 90-93.
Wang Hailing, Liu Tingxi, Wang Min, et al. Surface water quality assessment using grey clustering method in Dalate banner [J]. Journal of Inner Mongolia Agricultural University: Natural Science Edition, 2006(1): 90-93. (in Chinese)

[8] 李红艳, 魏永霞, 安瑞强. 运用灰色关联分析法评价建三江垦区地下水水质的研究 [J]. 东北农业大学学报, 2007(2): 243-246.
Li Hongyan, Wei Yongxia, An Ruiqiang. Applying grey relation method to evaluate groundwater quality in Jiansanjiang land reclamation area [J]. Journal of Northeast Agricultural University, 2007(2): 243-246. (in Chinese)

[9] 高惠璇. 应用多元统计分析 [M]. 北京: 北京大学出版社, 2006.
Gao Huixuan. Applied Multivariate Statistical Analysis [M]. Beijing: Peking University Press, 2006. (in Chinese)

[10] 余建英, 何旭宏. 数据统计分析与 SPSS 应用 [M]. 北京: 人民邮电出版社, 2003.
Yu Jianying, He Xuhong. Data Statistical Analyses and Application of SPSS [M]. Beijing: Post and Telecommunications Press, 2003. (in Chinese)

[责任编辑: 严海琳]