基于 SPSS 的山东枣庄小沙河水质分析

吴智洋,张志强,谢宝元,朱　悦,康　璇
(北京林业大学/教育部水土保持与荒漠化防治国家重点实验室,北京 100083)

摘要:利用多元统计的因子分析方法对采自于山东枣庄小沙河 7 个断面的 17 份水样数据进行分析,将 4 个评价指标转化为有机质污染因子和有毒有害污染因子,通过因子模型和因子得分表达式刻画小沙河的水质污染情况,得出第一、二断面水质污染最为严重。

关键词:水质评价;因子分析;小沙河
中图分类号:X824　　**文献标识码**:A　　**文章编号**:1002-2767(2011)02-0048-04

水环境污染在全球普遍存在,而我国的水环境污染尤为严重。近 30 年来,经济飞速发展造成大量污染物被排放到水环境中,天然水体受到了不同程度的污染。水环境中的污染物质日益增多,污染物成分也越来越复杂,水环境的不断恶化使原本就匮乏的水资源更加紧张,造成了严重的水质性缺水。对于南水北调工程,如在通水之前未能完成治污任务,则不仅难以实现调水的目的,而且必将引发大面积的水污染,把"南水北调"变成"南污北调",贻害无穷。目前,小沙河作为南四湖流域和南水北调工程的一部分,水质污染问题已经成为社会各界关注的焦点。

南四湖水质是否达标直接关系到南水北调东线调水水质,影响调水工程的顺利实施。南四湖流域的工农业生产,使入湖河流生态系统遭到破坏,河流水质下降,是造成南四湖水质恶化的原因之一。北京林业大学水体污染控制与治理科技重大专项课题组于 2009 年 9 月和 11 月对小沙河的水质情况进行了取样,并针对所取样的数据在 SPSS 中进行分析处理,探索有价值的信息和规律。

1　因子分析理论

1.1　因子分析方法综述

水质分析中,在取得大量实测有代表性的数据之后,面临的是要从中提取出有用信息及对信息进行加工,也即数据信息的提取和理论归纳分析,这已是当今信息科学的一个新热点,也是多元统计理论充分发挥作用的机遇[1]。

因子分析方法最早是 1904 年由德国心理学家 C. Spearman 提出的,1957 年由美国 W. C. Krumbein 引进地质学,用于研究沉积学,1962 年 J. Imbrie 和 E. G. Purdy[2] 加以完善和发展。因子分析是将多个实测变量转换为少数几个不相关的综合指标的多元统计方法,也就是在较少损失原始数据信息的前提下,用少量的因子去代替原始的变量,从而达到对原始变量的分类,揭露原始变量之间的内在联系[3]。

1.2　因子分析模块在水质分析领域的应用

李波[4]在文章"洪泽湖水质的因子分析"中将水质监测数据与空间数据结合,建立了具有时空特征的洪泽湖水质数据库。通过相关性分析判断参数的重要度,完成数据筛选,然后进行因子分析,研究主要因子的施工变化规律,总结各分量的水质净化差异,最后提出了富营养化指数的模拟方程。

何兴江[5]在《基于 SPSS 的城市区域地下水变异的因子分析过程》一文中建立了城市建设引发城市区域的地下水缓慢变异的模型,调用因子分析过程,对城市区域地下水场变异起重要作用的组分进行提取,以起重要作用的组分最少组合来作为评价地下水变异的指标体系。

朱万森等[6]应用因子分析法对地面水质污染状况的研究中,应用因子分析法对 11 条河道的地面水质分析数据进行了信息掘取,研究了各河道污染程度的排序、主因子、主要污染因子的组合及污染源的分析,其科研成果可以为河流水质污染治理提供重要的理论指导。

2　数据分析

2.1　水质样本

采用南四湖地区小沙河流域 7 个断面的 17

收稿日期:2010-11-01
基金项目:国家水体污染控制与治理科技重大专项资助项目(2009ZX07210-009)
第一作者简介:吴智洋(1986-),男,黑龙江省鸡西市人,在读硕士,从事水土保持与河流生态研究。E-mail:yang12355930@163.com。

个水质监测点的取样数据进行基于 SPSS 的分析过程。水质分析主要做了 COD、氨氮、总氮和总磷[7]的分析(见表 1)。

表 1　描述性统计

水质分析	样本数	最小值	最大值	均值	标准差
化学需氧量(COD)	17	24.45	39.40	31.0188	4.40526
氨氮(NH_3-N)	17	0.13	1.55	0.3935	0.43475
总氮(TN)	17	17.27	21.26	18.8065	1.01493
总磷(TP)	17	0.38	0.80	0.5906	0.11470

2.2　数据分析过程

2.2.1　KMO 抽样适度测定值与 Bartlett 球形检验值　Bartlett 球度检验的概率值为 0.000,即假设被拒绝,也就是说可以认为相关系数矩阵与单位矩阵有显著差异。同时,KMO 值为 0.539,根据度量标准该值大于 0.5,因此,原数据可用于因子分析。

2.2.2　数据标准化　数据标准化后的情况见表 2,表 3 和表 4。

表 2　K 均值和球形检验

凯瑟-梅耶样本采集频率		0.539
巴特利特球体检验	约方	30.741
	自由度	6.000
	概率	0.000

表 3　描述性统计

指标	均值	标准差	样本数
总磷(TP)	0.0000000	1.00000000	17
总氮(TN)	0.0000000	1.00000000	17
氨氮(NH_3-N)	0.0000000	1.00000000	17
化学需氧量(COD)	0.0000000	1.00000000	17

表 4　相关矩阵

指标		化学需氧量(COD)	氨氮(NH_3-N)	总氮(TN)	总磷(TP)
相关度	化学需氧量(COD)	1.000	0.726	−0.527	−0.506
	氨氮(NH_3-N)	0.726	1.000	−0.130	−0.776
	总氮(TN)	−0.527	−0.130	1.000	0.076
	总磷(TP)	−0.506	−0.776	0.076	1.000
概率	化学需氧量(COD)		0.000	0.015	0.019
	氨氮(NH_3-N)	0.000		0.309	0.000
	总氮(TN)	0.015	0.309		0.386
	总磷(TP)	0.019	0.000	0.386	

2.2.3　因子分析的共同度　表 5 是因子分析的共同度。其中第二列显示的是初始共同度,全部为 1;第三列是提取特征根的共同度,可以看到总氮、氨氮 2 个指标的共同度非常高,这 2 个变量的信息保留得很完整,而 COD 和总磷的共同度相对低一些,信息有所缺失。

表 5　公因子方差

指标	初始值	提取值
化学需氧量(COD)	1.000	0.859
氨氮(NH_3-N)	1.000	0.915
总氮(TN)	1.000	0.935
总磷(TP)	1.000	0.836

2.2.4　因子分析的总方差解释　由表 6 可知,第一个因子的特征根为 2.466,解释了 4 个原始变量总方差的 61.646%;第二个因子的特征根为 1.079,解释了 4 个原始变量总方差的 26.986%,累计方差贡献率为 88.631%。

表 6　总方差解释

成分	初始特征值			平方载荷的提取量			平方载荷的旋转量		
	总和	方差百分率/%	累计百分率/%	总和	方差百分率/%	累计百分率/%	总和	方差百分率/%	累计百分率/%
1	2.466	61.646	61.646	2.466	61.646	61.646	2.176	54.399	54.399
2	1.079	26.986	88.631	1.079	26.986	88.631	1.369	34.232	88.631
3	0.332	8.305	96.936						
4	0.123	3.064	100.00						

另外,从表 6 中可以看出,只有这 2 个变量的特征值大于 1,且只有这 2 个因子被提取和旋转,其累计解释总方差百分比和初始解的前 2 个变量相同,但经旋转后的因子重新分配各个因子的解释原始变量的方差,使得因子的方差更接近,也更易于解释。

2.2.5　因子分析的碎石图　图 1 呈现的是因子分析碎石图,利用此图可以帮助确定最优的因子数量。此图呈下降趋势,前 2 个因子特征值较大,连线相对陡峭,后面的几个因子特征值连线趋于

平缓。表明前 2 个因子的信息覆盖面大，因此，取前 2 个作为综合因子来表征这 4 个变量。

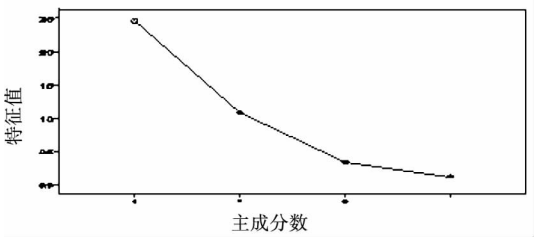

图 1　散点图

表 7　成分矩阵

指标	成分	
	1	2
化学需氧量(COD)	0.888	−0.265
氨氮(NH_3-N)	0.908	0.300
总氮(TN)	−0.442	0.860
总磷(TP)	−0.810	−0.423

表 8　旋转后的成分矩阵

指标	成分	
	1	2
化学需氧量(COD)	0.669	−0.642
氨氮(NH_3-N)	0.945	−0.148
总氮(TN)	0.000	0.967
总磷(TP)	−0.914	−0.006

表 7、表 8 分别显示的是旋转前以及旋转后的因子载荷矩阵，给出了每个变量在 2 个因子上的载荷。和没有旋转相比，因子的含义更容易解释。可以看出，大部分变量如 COD、氨氮、总磷在第一个因子上的载荷都比较高，即与第一个因子的相关程度较高，反映了水中有机物质对水体的影响；总氮在第二个因子上的载荷较高，即与第二个因子的相关程度较高，反映了水中有毒有害物质对水体的影响。

设 COD 为 X_1，氨氮为 X_2，总氮为 X_3，总磷 X_4，写出各变量与因子的关系式为：

$X_1=0.669F_1-0.642F_2$

$X_2=0.945F_1-0.148F_2$

$X_3=0F_1+0.967F_2$

$X_4=-0.914F_1-0.006F_2$

2.2.6　*旋转空间的因子图*　图 2 是旋转空间的因子图，可以看做是旋转后的载荷矩阵的图形表示，从图中再次验证了前面旋转后的载荷矩阵对因子的解释。

2.2.7　*因子得分系数矩阵及因子得分图*

$F_1=0.208X_1+0.455X_2+0.205X_3-0.472X_4$

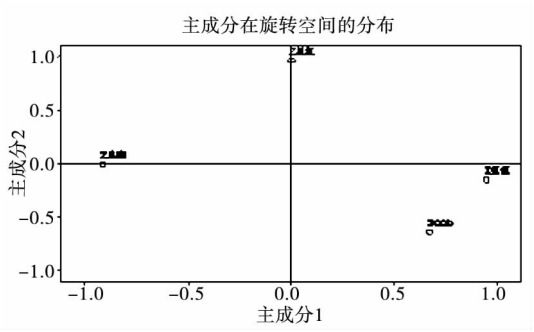

图 2　旋转空间里的成分图

$F_2=-0.383X_1+0.079X_2+0.791X_3-0.198X_4$

表 9　得分系数矩阵

指标	成分	
	1	2
化学需氧量(COD)	0.208	−0.383
氨氮(NH_3-N)	0.455	0.079
总氮(TN)	0.205	0.791
总磷(TP)	−0.472	−0.198

根据因子得分系数矩阵得到以下算式：

其中，第一个因子表示有机质因子，第二个因子表示有毒有害因子。

2.2.8　*综合得分 F 的计算*　以各因子的方差贡献率占两个因子总方差贡献率的比重作为权重进行加权汇总，以得出各水质监测点水质情况的综合得分 F[8]，即 $F=(0.616\ 46F_1+0.269\ 86F_2)/0.886\ 31$。

有第一因子 F_1 看，1～3 和 10～16 号水样受有机污染物影响较重，从第二因子 F_2 上看，4～9 和 16、17 号水样受有毒有害物质污染严重，从综合得分 F 看，1～5 号水样污染情况最严重。

3　结论

COD、氨氮和总磷是表示各种水中还原性物质多少的指标。水中的还原性物质有各种有机物、亚硝酸盐、硫化物和亚铁盐等，并且主要的是有机物。因此，第一因子可以作为衡量水中有机物质含量多少的指标。因子得分系数越大，说明水体受有机物的污染越严重。

有机氮是反映水中蛋白质、氨基酸和尿素等含氮有机化合物总量的一个水质指标。总氮(英文缩写 TN)则是一个包括从有机氮到硝酸氮等全部含量的水质指标。第二因子总氮的因子得分系数达到了 0.791，说明水质受总氮的影响显著。可以说，第二因子是衡量水中含氮有机化合物多

少,及化学毒害物质多少的重要标准。

表 10　综合得分情况比较

监测点	COD	氨氮	总氮	总磷	F1	F2	F
1	39.40	1.55	18.87	0.42	12.57056	−0.12474	8.705289
2	36.08	1.16	18.68	0.40	11.67304	0.96968	8.414257
3	37.21	1.15	18.12	0.38	11.79817	0.0971	8.23561
4	27.32	0.13	21.26	0.60	9.81681	6.24457	8.72926
5	24.45	0.20	20.89	0.60	9.17585	7.05664	8.530705
6	27.92	0.17	19.01	0.58	9.508	4.24214	7.90478
7	25.81	0.24	18.82	0.60	9.05258	4.90155	7.788794
8	27.25	0.19	19.15	0.60	9.397	4.60711	7.9387
9	24.83	0.23	19.38	0.62	8.94955	5.7151	7.964839
10	31.92	0.24	17.27	0.54	10.03403	1.34725	7.389228
11	32.75	0.20	18.07	0.58	10.33359	1.65108	7.690092
12	32.45	0.19	17.65	0.58	10.18054	1.43297	7.517231
13	34.57	0.25	18.63	0.58	10.8497	1.40093	7.9729
14	31.47	0.23	18.02	0.68	10.02355	2.08434	7.606365
15	33.89	0.18	18.63	0.74	10.60089	1.62416	7.867812
16	30.79	0.20	18.91	0.74	10.02259	3.03452	7.895004
17	29.21	0.18	18.35	0.80	9.54173	3.18324	7.605831

1～5 号水样来自 1 号和 2 号断面,而 1～5 号水样综合得分 F 最大,可见 1 号和 2 号断面水质污染严重,应加强附近污染源的监控。

参考文献:

[1] 王岚,王亚平,许春雪,等. 多元校正分析在水环境地球化学领域中的应用[J]. 地质通报,2009(2):78-79.

[2] 胡键颖,冯泰编. 实用统计学[M]. 北京:北京大学出版社,1997.

[3] 张崇甫,陈述云,胡希铃. 统计分析方法及其应用[M]. 重庆:重庆大学出版社,1995.

[4] 李波,濮培民,韩爱民. 洪泽湖水质的因子分析[J]. 中国环境科学,2003,23(1)69-73.

[5] 何兴江,张信贵,易念平,等. 基于 SPSS 的城市区域地下水变异 Factor Analysis 过程[J]. 地质与勘探,2006,1(1):93-96.

[6] 朱万森,陈红光,刘志荣,等. 应用因子分析法对地面水质污染状况的研究[J]. 复旦大学学报,2003,42(3):72-74.

[7] 王苏明,王亚平. 水分析技术进展[J]. 岩矿测试,1998,17(3):165-167.

[8] 何晓群. 多元统计分析[M]. 2 版. 北京:中国人民大学出版社,2009:223-225.

Water Quality Analysis of Xiaosha River by SPSS in Zaozhuang City of Shandong Province

WU Zhi-yang, ZHANG Zhi-qiang, XIE Bao-yuan, ZHU Yue, KANG Xuan

(Beijing Forestry University/State Key Laboratory on Soil and Water Conservation of Ministry of Education, Beijing 100083)

Abstract: Using multivariate statistical method of factor, the data of 17 water samples from 7 sections were analyzed. Conversing the four evaluations into organic pollutants and toxic pollutants, the water quality of pollution in Xiaosha River was characterized by expression of factor models and factor scores. The results showed that water pollution of first and second sections were the most serious.

Key words: water quality assessment; factor analysis; Xiaosha River