

应用 SPSS13 软件对地下水水质进行判断分析模型研究

李 理¹ 赵秋娜²

(1. 东莞新科技研究开发有限公司, 广东 东莞 523087; 2. 河北省唐秦水文水资源勘测局, 河北 唐山 063000)

摘要 :为了能够深层次发现水质指标之间的关系和规律,采用 SPSS13 软件包对地下水中氨氮、亚硝酸盐氮、硝酸盐氮、高锰酸盐指数、总砷、溶解性铁、总锰、氟化物等水质指标进行分析,得出它们之间判断分析模型。该模型对地下水的水质评价简单实用,能够提高水质资料整编工作效率。

关键词 地下水水质综合评价;高级多元统计;SPSS13 软件;判断分析模型

中图分类号 X824 **文献标识码** A **文章编号** 1004-693X(2008)S1-0020-04

在地下水水质评价过程中,有许多参数变量,例如有氨氮、亚硝酸盐氮、硝酸盐氮、高锰酸盐指数、总砷、溶解性铁、总锰、氟化物等数据需要处理。如果对其进行手工处理,不但工作量大,容易产生许多问题,而且不能够深层次地发现水质变量之间的关系和规律。基于上述问题,通过引入高级多元统计分析技术之一——判断分析方法,应用 SPSS13 软件进行运算,得到地下水水质判断分析模型。

1 判别分析模型及处理程序

1.1 地下水水质评价与地下水水质变量反向关系

现代数学和计算机信息技术的发展,使得先进的地下水水质的评价方法得到广泛推广。厉艳君等^[1]介绍了模糊综合评价、灰色系统、BP 人工神经网络、多元线性回归、集对分析、逻辑斯谛曲线、物元可拓等方法模型。罗士心等^[2]论述了浓度标准分类法、污染指数法、分级评价法等常用的评价方法。周丰等^[3]提出了一种基于多元统计分析和 RBFNNs 的水质评价方法。由地下水水质变量按一定算法可以容易地得到地下水水质评价模型,然而地下水水质评价与地下水水质各变量反向关系如何,方向评价模型如何,目前研究的不是太清楚。基于上述问题,本文应用判断分析模型,探讨地下水水质评价与水质变量反向模型。

1.2 判别分析模型介绍

判别分析模型是高级多元统计分析技术。判别分析模型是类别明确的一种分类技术。它根据观测到的某些参数指标对所研究的对象进行分类,得到

判别函数,然后再使用判别函数对未知分类的样品进行分类。判别分析需要有标准,常用的判别分析方法有距离判别、Fisher 的典型判别和 Bayes 判别。距离判别和 Fisher 的典型判别对数据的概率分布无严格要求,而 Bayes 判别则要求数据概率服从多元正态分布。

1.3 判别分析模型算法

朱顺泉^[4]提到应用 SPSS 软件来建立判别分析过程:对于分为 k 组的研究对象 m 个性质可建立 $k-1$ 个典则判别函数(原始自变量的线性组合)和 k 个 Fisher 线性判别函数,然后将各样本的自变量回代到判别函数中,计算其判别分数或者属于各组的概率,根据数值的大小判别样本所属组别,对比样本的原始组别给出出错率。在判断分析模型中,若变量 $x_1, x_2, \dots, x_m$ 影响着判别函数,则可建立判别函数 $z_k = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + \dots + a_mx_m$,其判别分析模型算法解法步骤为:①选择自变量和组变量;②计算各组单变量描述统计量,包括组内均值、组内标准差、总均值、总标准差、各组协方差矩阵、组间相关矩阵,对组间均值相等以及协方差矩阵相等的零假设进行检验;③推导判别系数,给出标准化或未标准化的典型判别函数,并对函数显著性进行检验;④建立 Fisher 线性判别模型;⑤按照一定的规则进行分组;⑥进行样本回判分析,计算错分率;⑦输出运算结果;⑧结合实际情况对结果进行分析。

2 利用 SPSS13 软件包处理地下水水质判别分析模型

宇传华^[5]关于判别分析操作应用的论述非常详

细 ,可以直接利用其介绍的方法 ,很方便地解决复杂数据处理难题。

2.1 SPSS13 软件包统计分析基本步骤

SPSS13 是由 SPSS 公司(WWW.SPSS.COM)出品的大型通用专业统计分析软件。该软件可利用多种数据文件以及数据来源 ,生成统计报表图形 ,进行简单和复杂的统计分析 ,广泛应用于各行业。利用 SPSS13 软件包解决统计分析的基本步骤为 :①输入数据到 SPSS 菜单 ,找到变量浏览菜单定义变量名称等属性 ,在数据浏览菜单输入原始数据 ;②分析前数据准备 ,即数据核查、筛选、转换 ;③选择分析方法和过程分析 ,一般选图形菜单或者分析菜单进行运算分析 ;④选择所要分析的变量和观察个体 ;⑤运行分析过程并浏览结果。

2.2 应用 SPSS13 软件判别分析模型操作

在 SPSS13 菜单中依次单击分析菜单中子菜单分类按钮中的判断按钮 ,弹出判断分析菜单 ,键入分组变量为评分类别 Z ,最小值为 1 ,最大为 4。设氨氮是 X_1 ,亚硝酸盐氮是 X_2 ,硝酸盐氮是 X_3 ,高锰酸盐指数是 X_4 ,总砷是 X_5 ,溶解性铁是 X_6 ,总锰是 X_7 ,氟化物是 X_8 ,全部填入自变量 ,选择逐步逼近

法 ,分别单点击统计按钮 ,出现统计菜单 ,在描述栏中选平均值 ,在函数系数栏中选费舍尔 ;单击方法按钮出现逐步逼近法菜单 ,在方法栏中选中广义方差比最小化方法 ,评价标准栏用 F 值 ,确认值为 3.84 ,剔出值为 2.71 ,显示栏选步骤总结 ;分类按钮 ,设定先验概率为等先验概率 ,显示栏中选择刀切法考核结果表 ,使用协方差阵中组内协方差阵。

2.3 按 SPSS13 格式要求输入原始数据

根据河北省水环境监测中心 2007 年 9 月份的地下水水质监测成果 ,选取其中一部分数据 ,作为 SPSS13 判别函数模型研究数据 ,将这些原始数据输入到 SPSS 处理表格中(表 1)。根据实际情况选 X_1 、 X_2 、 X_3 、 X_4 、 X_5 、 X_6 、 X_7 、 X_8 进行判别分析处理。

表 1 中“ < DL ”表示小于检出限 ,检出限是指产生一个能可靠地被检出的分析信号(或数据)所需要的某元素的最小浓度或含量。本文分析根据规范将“ < DL ”按 1/2 检出限处理 ,评分类别已由水文整编人员评为优、良、好、差 4 种类型。为了评分类别符合 SPSS13 软件运算规则 ,将评分类别的测量值赋予顺序值 ,各值分别为优 = 1 ,良 = 2 ,好 = 3 ,差 = 4。根据以上规范处理 ,得到表 2。

表 1 地下水水质原始数据

测点名称	X_1 / ($\text{mg}\cdot\text{L}^{-1}$)	X_2 / ($\text{mg}\cdot\text{L}^{-1}$)	X_3 / ($\text{mg}\cdot\text{L}^{-1}$)	X_4 / ($\text{mg}\cdot\text{L}^{-1}$)	X_5 / ($\text{mg}\cdot\text{L}^{-1}$)	X_6 / ($\text{mg}\cdot\text{L}^{-1}$)	X_7 / ($\text{mg}\cdot\text{L}^{-1}$)	X_8 / ($\text{mg}\cdot\text{L}^{-1}$)	评分 类别 Z
北戴河	< DL	0.009	20.5	< DL	0.0009	0.04	0.13	0.36	优
铁庄	< DL	0.006	28.6	< DL	0.0007	0.03	0.01	0.89	优
坟坨	< DL	< DL	47.4	< DL	0.0005	< DL	< DL	0.15	差
刘台庄	0.05	0.022	12.2	< DL	0.0026	0.06	0.22	0.34	好
徐新庄	0.28	0.032	21.5	< DL	0.0010	< DL	0.05	0.43	优
赵家港	0.08	0.004	84.2	0.5	0.0014	0.04	0.04	0.37	良
安各庄	0.53	0.016	4.49	0.7	0.0024	0.04	0.69	0.36	差
新庄	0.23	0.016	10.7	1.3	0.0012	0.03	0.11	0.25	良
阎各庄	1.66	< DL	< DL	0.6	0.0137	< DL	0.60	0.38	差
西新庄	0.16	0.021	0.89	< DL	0.0021	< DL	0.70	0.43	差
下马坨	0.65	0.024	5.63	< DL	0.0012	< DL	0.06	0.29	良
汀流河	0.24	0.038	56.9	< DL	0.0007	< DL	0.24	0.27	良
曹家河	0.13	< DL	< DL	0.5	0.0045	< DL	0.07	2.28	良
榛子镇	0.09	0.027	6.89	< DL	0.0010	< DL	0.01	2.11	良

表 2 地下水水质 SPSS13 数据

测点名称	X_1 / ($\text{mg}\cdot\text{L}^{-1}$)	X_2 / ($\text{mg}\cdot\text{L}^{-1}$)	X_3 / ($\text{mg}\cdot\text{L}^{-1}$)	X_4 / ($\text{mg}\cdot\text{L}^{-1}$)	X_5 / ($\text{mg}\cdot\text{L}^{-1}$)	X_6 / ($\text{mg}\cdot\text{L}^{-1}$)	X_7 / ($\text{mg}\cdot\text{L}^{-1}$)	X_8 / ($\text{mg}\cdot\text{L}^{-1}$)	评分 类别 Z
北戴河	0.025	0.009	20.5	0.25	0.0009	0.04	0.13	0.36	1
铁庄	0.025	0.006	28.6	0.25	0.0007	0.03	0.01	0.89	1
坟坨	0.025	0.0015	47.4	0.25	0.0005	0.015	0.005	0.15	4
刘台庄	0.05	0.022	12.2	0.25	0.0026	0.06	0.22	0.34	3
徐新庄	0.28	0.032	21.5	0.25	0.0010	0.015	0.05	0.43	1
赵家港	0.08	0.004	84.2	0.5	0.0014	0.04	0.04	0.37	2
安各庄	0.53	0.016	4.49	0.7	0.0024	0.04	0.69	0.36	4
新庄	0.23	0.016	10.7	1.3	0.0012	0.03	0.11	0.25	2
阎各庄	1.66	0.0015	0.04	0.6	0.0137	0.015	0.60	0.38	4
西新庄	0.16	0.021	0.89	0.25	0.0021	0.015	0.70	0.43	4
下马坨	0.65	0.024	5.63	0.25	0.0012	0.015	0.06	0.29	2
汀流河	0.24	0.038	56.9	0.25	0.0007	0.015	0.24	0.27	2
曹家河	0.13	0.0015	0.04	0.5	0.0045	0.015	0.07	2.28	2
榛子镇	0.09	0.027	6.89	0.25	0.0010	0.015	0.01	2.11	2

2.4 运行结果

运行结果见表 3 ~ 13 及说明。

表 3 分析案例过程化总结

无权重事件		数量	百分比/%
有效		14	93.3
被排除	丢失或者超出组编码范围	0	0
	至少一个丢失判别变量	0	0
	兼具丢失或者超出组编码范围和至少一个丢失判别变量	1	6.7
	总计	15	100.0

表 4 组统计					
Z 评分 类别	平均值	标准方差	有效数量		
			无权重	有权重	
优	X_1	0.110	0.147	3	3.000
	X_2	0.016	0.014	3	3.000
	X_3	23.533	4.416	3	3.000
	X_4	0.250	0.000	3	3.000
	X_5	0.001	0.001	3	3.000
	X_6	0.024	0.020	3	3.000
	X_7	0.063	0.061	3	3.000
	X_8	0.560	0.288	3	3.000
良	X_1	0.237	0.214	6	6.000
	X_2	0.018	0.014	6	6.000
	X_3	27.39	34.695	6	6.000
	X_4	0.508	0.407	6	6.000
	X_5	0.002	0.001	6	6.000
	X_6	0.013	0.018	6	6.000
	X_7	0.088	0.081	6	6.000
	X_8	0.928	0.983	6	6.000
好	X_1	0.050	数据不足	1	1.000
	X_2	0.022	数据不足	1	1.000
	X_3	12.200	数据不足	1	1.000
	X_4	0.250	数据不足	1	1.000
	X_5	0.003	数据不足	1	1.000
	X_6	0.060	数据不足	1	1.000
	X_7	0.220	数据不足	1	1.000
	X_8	0.340	数据不足	1	1.000
差	X_1	0.594	0.742	4	4.000
	X_2	0.010	0.010	4	4.000
	X_3	13.205	22.878	4	4.000
	X_4	0.450	0.235	4	4.000
	X_5	0.005	0.006	4	4.000
	X_6	0.011	0.019	4	4.000
	X_7	0.499	0.332	4	4.000
	X_8	0.330	0.124	4	4.000
总 计	X_1	0.298	0.435	14	14.000
	X_2	0.016	0.012	14	14.000
	X_3	21.427	25.131	14	14.000
	X_4	0.418	0.299	14	14.000
	X_5	0.002	0.003	14	14.000
	X_6	0.018	0.021	14	14.000
	X_7	0.230	0.257	14	14.000
	X_8	0.636	0.682	14	14.000

变量选入历史的步骤为 1 步 ,相关变量为 X_7 ,公差为 1.00 ,剔除 F 值为 4.367。

表 5 变量被剔除历史

步	公差	最小公差	确认的 F 值	广义方差比最小化	
0	X_1	1.000	1.000	0.935	0.781
	X_2	1.000	1.000	0.421	0.888
	X_3	1.000	1.000	0.255	0.929
	X_4	1.000	1.000	0.558	0.857
	X_5	1.000	1.000	0.857	0.795
	X_6	1.000	1.000	2.147	0.608
	X_7	1.000	1.000	4.367	0.433
	X_8	1.000	1.000	0.655	0.836
1	X_1	0.878	0.878	0.175	0.409
	X_2	0.852	0.852	0.961	0.328
	X_3	0.876	0.876	0.095	0.420
	X_4	0.956	0.956	0.556	0.365
	X_5	0.919	0.919	0.064	0.424
	X_6	0.960	0.960	2.136	0.253
	X_7	0.995	0.995	0.329	0.390
	X_8	0.995	0.995	0.329	0.390

表 6 广义方差比最小化

步	变量的数量	方差比最小化	自由度 1	自由度 2	自由度 3	精确 F 值			
						统计量	自由度 1	自由度 2	检验的显著性
1	1	0.433	1	3	10.0	4.367	3	10.0	0.033

特征值。函数 1 的特征值为 1.310 ,变量的百分数为 100.0% ,累计变量 100.0% ,标准相关 0.753。

广义方差比最小化。测试函数 1 的广义方差比最小化为 0.433 ,卡方 8.879 ,自由度为 3 ,检验的显著性为 0.033。

标准规范判别分析函数系数。函数 1 与 X_7 函数系数是 1.000。

表 7 结构矩阵

分类	函数 1	分类	函数 1
X_7	1.000	X_8 (不使用)	0.284
X_8 (不使用)	0.384	X_6 (不使用)	0.210
X_6 (不使用)	- 0.351	X_5 (不使用)	0.200
X_5 (不使用)	0.349	X_4 (不使用)	- 0.070

表 8 未标准化判别函数系数

Z 分类类别	函数 1
X_7	5.187
常数	- 1.087

表 9 同一组函数

Z 分类类别	函数 1
优	- 0.759
良	- 0.629
好	0.054
差	1.500

表 10 分类处理过程总结

已处理		15
被排除	丢失或超出组编码范围	0
	至少一个判别变量丢失	1
输出使用		14

表 11 成组优先概率

Z 评分类别	优先权	分析中已使用事件	
		无权重	有权重
优	0.250	3	3.000
良	0.250	6	6.000
好	0.250	1	1.000
差	0.250	4	4.000
总计	1	14	14.000

表 12 Fisher 线性判别分类系数

项目	Z 评分类别			
	优	良	好	差
X_7	1.704	2.377	5.919	13.419
常数	-1.440	-1.491	-2.037	-4.733

表 13 分类结果

Z 评分类别		原始							
		计数				百分比/%			
		优	良	好	差	优	良	好	差
预测组成员	优	2	4	0	1	66.7	66.6	0.0	25.0
	良	1	1	0	0	33.3	16.7	0.0	0.0
	好	0	1	1	0	0.0	16.7	100.0	0.0
	差	0	0	0	3	0.0	0.0	0.0	75.0
总计		3	6	1	4	100.0	100.0	100.0	100.0

Z 评分类别		交叉确认							
		计数				百分比/%			
		优	良	好	差	优	良	好	差
预测组成员	优	1	4	0	1	33.3	66.6	0.0	25.0
	良	2	1	1	0	66.7	16.7	100.0	0.0
	好	0	1	0	0	0.0	16.7	0.0	0.0
	差	0	0	0	3	0.0	0.0	0.0	75.0
总计		3	6	1	4	100.0	100.0	100.0	100.0

3 结 论

在评分类别模型化过程中 ,令优 = Z_1 ,良 = Z_2 ,好 = Z_3 ,差 = Z_4 , X_1 、 X_2 、 X_3 、 X_4 、 X_5 、 X_6 、 X_7 、 X_8 与评价类别 Z 多元关系可以简化为 X_7 与评分类别 Z 的一元关系模型。未标准化判别模型 : $Z = -1.078 + 5.153X_7$ 。Fisher 线性判别模型 $Z_1 = -1.440 + 1.704X_7$, $Z_2 = -1.491 + 2.377X_7$, $Z_3 = -2.037 + 5.919X_7$, $Z_4 = -4.733 + 13.419X_7$ 。

实践证明 ,影响水质评分类别的参数存在一定的地域性和时间差异 ,而在一定的时空区域 ,并非所有地下水水质参数都影响水质评分类别 ,仅有少数污染项目参数起主要作用 ,其他多数污染项目参数可以完全忽略不计。通过判断分析模型 ,可以迅速发现这些主要污染项目关键参数。应用判断分析模

型 ,可以快速解决水质分析数据处理难题 ,极大地提高地下水水质资料整编效率。

参考文献 :

[1] 厉艳君 ,杨木壮 .地下水水质综合评价方法综述[J].地下水 ,2007 ,29(5) :19-24 .
[2] 罗士心 ,毛红梅 ,陶守耀 ,等 .水质评价方法综述[J].水资源研究 ,2002 ,23(3) :15-19 .
[3] 周丰 ,郭怀成 ,刘永 ,等 .基于多元统计分析和 RBFNNs 的水质评价方法[J].环境科学学报 ,2007 ,27(5) :846-853 .
[4] 朱顺泉 .管理科学研究方法统计与运筹优化应用[M].北京 :清华大学出版社 ,2007 :130-134 .
[5] 宇传华 .SPSS 与统计分析[M].北京 :电子工业出版社 ,2007 :472-476 .

(收稿日期 2008-03-17 编辑 徐娟)

(上接第 12 页)

pH 值存在优化点 ,对于 COD 的去除 ,pH 值在 9.0 ~ 9.5 之间处理效果最佳 ,而对于 $\text{NH}_3\text{-N}$,pH 值在 9.5 ~ 10.0 之间处理效果最好 ;

根据以上结论 ,针对目前潜流人工湿地氨氮去除效果偏低的问题 ,在实际运用中 ,可考虑在污水中投加适当的碱类 ,使得污水的净化处于一个良好的碱性环境 ,从而提高潜流人工湿地对 $\text{NH}_3\text{-N}$ 的处理效果。

参考文献 :

[1] 徐丽花 ,周琪 .不同填料人工湿地处理系统净化能力研究[J].上海环境科学 ,2002 ,21(10) :603-605 .
[2] BEN U .Storm water :best management practices and detention for water quality[J].Drainage and CSO Management ,1993 :382-389 .
[3] 国家环保总局 .水和废水监测分析方法[M].2 版 .北京 :中国环境科学出版社 ,1989 :230-351 .
[4] 叶正芳 ,倪晋仁 .污水处理的固定化微生物与游离微生物性能的比较[J].应用基础与工程科学学报 ,2002 ,10(4) :327 .
[5] 戴树桂 .环境化学[M].北京 :高等教育出版社 ,1997 :139-140 .
[6] 陈曦 ,崔莉凤 ,杜兵 ,等 .温度和 pH 值对厌氧氨氧化微生物活性的影响分析[J].北京工商大学学报 ,2006 ,24(3) :7-8 .
[7] 尹军 ,崔玉波 .人工湿地污水处理技术[M].北京 :化学工业出版社 ,2006 :71-72 .

(收稿日期 2008-06-07 编辑 高渭文)