多元统计在水资源利用方面的应用*

孙 景 艳
(重庆师范大学 数学与计算机科学学院, 重庆 400047)

摘 要:运用多元统计分析中的聚类分析和判别分析,对我国 30 个省、市、自治区(除西藏,台湾外)的水资源利用情况进行分类排序,并对结果进行了检验。以期为我国供水部门估计供水量与政府决策提供一些参考数据。

关键词:聚类分析 判别分析 维尔克斯(Wilks)统计量 水资源 利用
中图分类号:P333.6 **文献标识码**:A **文章编号**:1672-6693(2007)02-0067-04

Multi-elemental Statistical Analysis on the Water Utilization in Various Regions of China

SUN Jing-yan
(College of Mathematics and Computer Science, Chongqing Normal University, Chongqing 400047, China)

Abstract: This paper gives an order and a classification for the water utilization in various regions of China, respectively by utilizing the analysis of cluster and the discrimination analyses. Simultaneously, the outcome passes the outstanding test. It provides some scientific bases not only for the comprehensive estimation of water demands but also for the government management.

Key words: cluster analysis; discrimination analysis; Wilks statistic data; water resource; utilization

我国是贫水国家,虽然水资源总量丰富,但人均占有量很低。21 世纪将是我国经济高速发展时期,水资源短缺,固然有其自然的原因,也与气候变化有关,但人为的浪费,用水不合理也是很重要的因素。本文运用多元统计方法,选取一些具有代表性的指标,对全国 30 个省市自治区的水的利用状况进行了综合分析,希望为各地区合理、优化、节约用水提供一些科学依据。

1 数据来源与指标选取

由于各地受人口密度,经济结构,作物组成,节水水平,水资源条件等多种因素的相互影响,用水指标值有很大差别。本文引用的数据摘自文献[1]中具有代表性的 5 个方面的 10 个指标。(原始数据略)。

(1) 工业方面:x_6——万元工业增加值用水量(m^3),x_8——工业年用水量(亿 m^3);

(2) 农业方面:x_9——农业年用水量(亿 m^3);

(3) 生活方面:x_2——人均用水量(m^3),x_4 城镇生活人均生活用水量(L/d),x_5——农村居民人均生活用水量(L/d),x_7——生活年用水量(亿 m^3);

(4) 环境方面:x_{10}——生态年用水量(亿 m^3);

(5) 经济方面:x_1——人均国内生产总值(万元),x_3——万元国内生产总值用水量(m^3)。

2 多元统计分析

2.1 聚类分析

聚类分析的内容非常丰富,本文采用 Q 型的系统聚类法进行分析。由于数据存在量纲和数量级的差别,在聚类之前先进行标准化处理,计算样品之间的距离采用欧氏距离的平方,类与类之间的距离采用类平均法(具体过程利用 SPSS12[3] 在计算机上完成,结果见图 1)

系统聚类的结果能够给出 n 个样品自成一类到全部样品聚为一类,这个过程中所有结果都有,根据

* 收稿日期:2006-06-15 修回日期:2006-12-20
作者简介:孙景艳(1978-),女,河北衡水人,硕士研究生,研究方向为随机经济系统分析。

我国区域划分和图1可知,根据在水资源利用方面的10个指标,可将全国30个省市划分为三类较好,具体为:

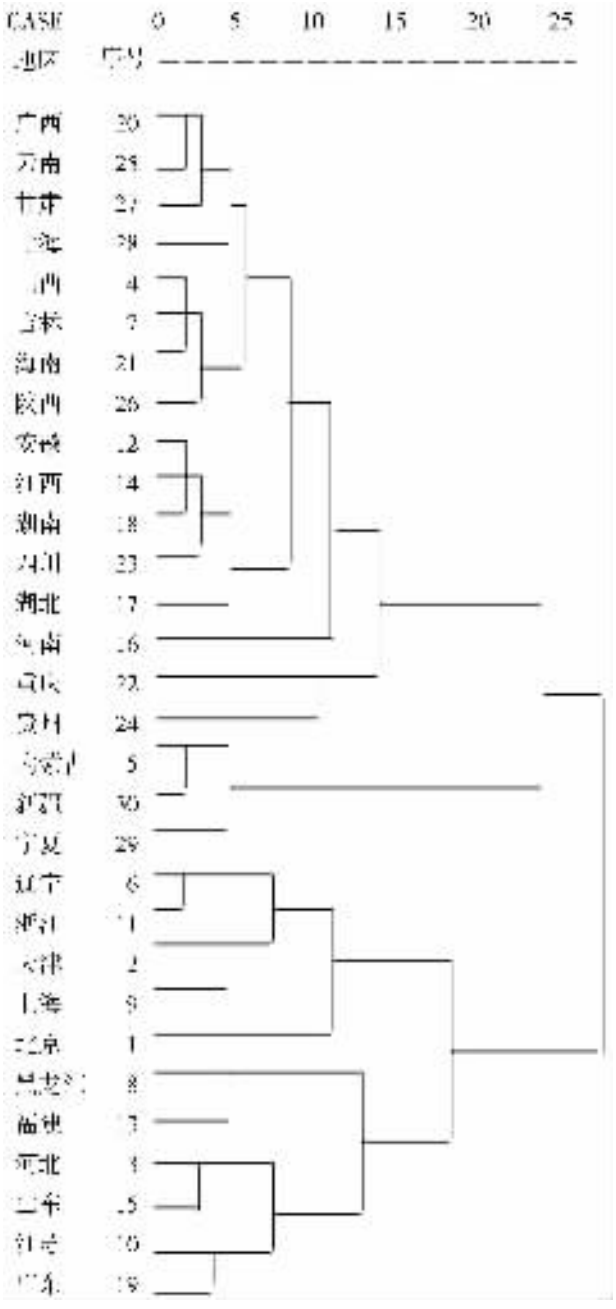

图1 聚类分析结果

第1类:北京,天津,河北,辽宁,上海,江苏,浙江,福建,山东,广东,黑龙江;

第2类:山西,吉林,安徽,江西,河南,湖北,湖南,广西,海南,重庆,四川,贵州,云南,陕西,甘肃,青海;

第3类:内蒙古,宁夏,新疆。

从三类来看,水供应量最少的是新疆等西部偏远地区,其次是广大的中部地区,水资源消耗量最大的是东部发达地区。上述分类是否合理,下面将运用判别分析的方法进行检验。

2.2 判别分析

判别分析是已知研究对象分成若干类型,并已取得各种类型的一批已知样品的观测数据,在此基础上根据某些准则建立判别方程,然后将样品的属性代入判别方程,对样品进行判别分类。

判别分析的类型很多,本文运用应用比较广泛,对分布、方差没有什么要求的费歇尔(Fisher)判别(又称典型判别)方法(本例采用SPSS12[3]软件)对上述聚类分析的情况做判别,其具体步骤[5]如下。

(1)计算类内离差平方和阵 W 和类间离差平方和阵 B。

(2)给出 $W^{-1}B$ 的特征根(即典型判别方程的特征值)以及方差贡献率(见表1)。特征根数取变量数及类别减1中的较小值,本例为3类,变量数10,因此特征根个数为2,其中第一个特征根为3.4,方差贡献率为53.5%,能够解释所有信息的53.5%,而这二个特征根累计方差贡献率100%,说明这两个典型方程能反映所有原始信息。

表1 特征根

典型方程	特征根	方差贡献率/%	累计方差贡献率/%
1	3.4	53.5	53.5
2	2.951	46.5	100

(3)计算 $W^{-1}B$ 的2个特征根所对应的特征向量,即得出两个典型判别方程对应的系数(见表2),由表2得

$$y_1 = 1.549x_1 + 0.001x_2 + 0.003x_3 - 0.003x_4 - 0.018x_5 - 0.006x_6 + 0.017x_7 + 0.007x_8 + 0.002x_9 - 0.139x_{10} - 1.745$$

$$y_2 = 1.619x_1 - 0.001x_2 - 0.001x_3 - 0.004x_4 + 0.008x_5 - 0.001x_6 + 0.014x_7 - 0.003x_8 + 0.007x_9 + 0.004x_{10} - 1.425$$

由此可计算出各地区的两判别因子得分(见表3)。

(4)判别方程的有效性检验。建立的判别方程是否显著,即对下面的识别样品归类的准确性如何,是否可信,需要检验。维尔克斯统计量表达为:类内离差平方和矩阵的行列式与总离差平方和矩阵行列式的比值。从表4可得,相伴概率均为0,远远小于置信水平0.01,从而得两个判别方程的判别能力都很显著。

(5)样品判别归类。由两个显著的判别方程构成一个判别空间,将样品数据代入2个判别方程进

表 2　典型判别方程系数矩阵

判别方程	人均产值	人均用水	万元用水	城镇生活	农村生活	万元工业	生活用水	工业用水	农业用水	生态用水	常数项
1	1.549	0.001	0.003	-0.003	-0.018	-0.006	0.017	0.007	0.002	-0.139	-1.745
2	1.619	-0.001	-0.001	-0.004	0.008	-0.001	0.014	-0.003	0.007	0.004	-1.425

表 3　样品判归结果

原始地区序号	聚类分析类别	判别分析类别	第一判别方程得分	第二判别方程得分
1	1	1	0.521	2.784
2	1	1	1.361	2.94
3	1	1	0.335	1.058
4	2	2	-0.662	-0.297
5	3	1**	2.041	-0.593
6	1	1	0.274	1.088
7	2	2	-0.708	-0.527
8	1	1	1.283	0.395
9	1	1	1.958	4.221
10	1	1	0.768	2.205
11	1	1	-0.768	2.094
12	2	2	-1.466	-0.603
13	1	1	-0.278	0.901
14	2	2	-1.867	-1.251
15	1	1	0.993	2.103
16	2	2	-0.56	0.35
17	2	2	-0.497	-0.418
18	2	2	-2.402	-0.527
19	1	1	0.641	2.825
20	2	2	-1.593	-0.898
21	2	2	-1.463	-1.711
22	2	2	-2.678	-1.121
23	2	2	-1.217	-0.213
24	2	2	-2.738	-2.15
25	2	2	-0.795	-0.795
26	2	2	-1.139	-0.707
27	2	2	-0.148	-1.985
28	2	2	-0.655	-2.213
29	3	3	4.625	-3.98
30	3	3	6.833	-2.975

表 4　典型方程显著性检验

判别方程	维尔克斯统计量	卡方值	自由度	相伴概率
1	0.058	64.251	20	0.000
2	0.253	30.914	9	0.000

行计算,各类的判别方程取值在该 2 维空间中构成 3 组点集群,每组点集群都有一个重心点,对应的向量分别是:$\bar{Y}_1 = (0.645, 2.056)$;$\bar{Y}_2 = (-1.287, -0.942)$;$\bar{Y}_3 = (4.5, -2.516)$。样品判别归类时,将待判样品点代入 2 个判别方程,得 2 个判别值,可将之看作 2 维空间中的一个点,例如以第一个样品北京作为待判样品,则对应的向量为 $Y_1 = (0.521, 2.784)$,考察样品点北京与各类重心欧氏距离的平方,距离平方分别为

$$l_1 = (0.521 - 0.645)^2 + (2.784 - 2.056)^2 = 0.761$$

$$l_2 = (-1.287 - 0.521)^2 + (-0.942 - 2.784)^2 = 17.149$$

$$l_3 = (0.521 - 4.5)^2 + [2.784 - (-2.516)]^2 = 43.922$$

由此得北京到第 1 类的重心距离最短,所以北京判归到第 1 类。所有样品的判归结果见表 3。

(6)对判别效果作检验。如果各个总体的均值向量在统计上没有显著差异,作判别分析意义不大。所以下面对均值向量做检验,三个类别的均值向量如下。

$\hat{u}^1 = (2.192, 442.636, 221.455, 218.273, 88.636, 129, 27.909, 59.227, 123.409, 3.864)'$

$\hat{u}^2 = (0.806, 365.563, 469.188, 204.438, 64.625, 283, 17.463, 34.731, 95.456, 1.169)'$

$\hat{u}^3 = (1.014, 1503.333, 1466, 167.333, 35.333, 127.667, 8.333, 7.2, 225, 6.967)'$

第 1 类组内离差平方和　$S_1 = \sum_{i=1}^{11} (x_i - \hat{u}^1)(x_i - \hat{u}^1)'$

第 2 类组内离差平方和　$S_2 = \sum_{i=1}^{16} (x_i - \hat{u}^2)(x_i - \hat{u}^2)'$

第 3 类组内离差平方和　$S_3 = \sum_{i=1}^{3} (x_i - \hat{u}^3)(x_i - \hat{u}^3)'$

组内总离差平方和　$S = S_1 + S_2 + S_3$

组间总离差平方和　$B = \sum_{i=1}^{3} n_i (\hat{u}^i - \hat{u})(\hat{u}^i - \hat{u})'$

($\hat{u}$ 为所有样品的均值向量),统计量 $\Lambda = \frac{|S|}{|S+B|} \sim \Lambda(p, n-k, k-1) = \Lambda(10, 27, 2)$,代入数据(根据文献[4]在计算机中完成计算)得　$\Lambda = \frac{3.282 \times 10^{47}}{9.748 \times 10^{49}} \approx 0.003\ 37$。$\Lambda_{0.01}(10, 27, 2) = \Lambda_{0.01}(2, 10, 19) = 0.022$[2],$0.003\ 37 < 0.022$,所以三组均值有显著差异,即说明上述判别分析有效,也即检验了聚类分析的结果是合理的。

3　讨　论

(1)图 1 中的序号 5 样品内蒙古,聚类分析为第

3类,而判别分析为第1类,这与上面采取的方法有关,也与内蒙古的数据有关,其数据介于1类和3类之间,差别不显著。

(2)从图1的分类结果可以看出,第1类几乎全是沿海或东部的经济发达省市,水供应充足,人口稠密,高科技产业和服务业较发达,用水较合理,投资环境最好[6],文化水平也较发达[5];而第3类均为偏远地区,水供应量严重不足,人口稀疏,气候干燥,经济、文化发展也比较落后,投资环境最差的省市[6];中部16个省级行政区大多以重工业为主,经济、文化处于中等水平,水消耗量也较大,有一定的节水空间[7-9]。这些分析统计与实际情况基本一致,所以从这些分析可以得到,一个地区首先要保证水资源供应充足,利用合理,投资环境才可能好,从而才能吸引更多外资,最终经济、文化才能较快发展,人民生活水平得到较大提高。

参考文献:

[1] 中华人民共和国水利部,中国水资源公报[M]. 北京:中国水利水电出版社,2004.
[2] 张尧庭,方开泰. 多元统计分析引论[M]. 北京:科学出版社,1982.
[3] 章文波,陈红艳. 实用数据统计分析及SPSS 12.0应用[M]. 北京:人民邮电出版社,2006.
[4] 李世奇,杜慧琴. Maple计算机代数系统应用及程序设计[M]. 重庆:重庆大学出版社.,1999.
[5] 于秀林,任雪松. 多元统计分析[M]. 北京:中国统计出版社,2003.
[6] 张雅清. 中国各省市投资环境的统计分析[J]. 重庆师范大学学报(自然科学版),2006,23(1):67-70.
[7] 陈忠,任雪梅,周心琴,等. 重庆市降水量的时空分布[J]. 四川师范学院学报(自然科学版),2003,24(2):171-76.
[8] 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):17-20.
[9] 陈晔,赵纯勇,魏兴萍. 重庆市沙坪区水资源用量趋势及保护措施[J]. 重庆师范大学学报(自然科学版),2005,22(1):53-56.

(责任编辑 李若溪)