主成分分析方法在水库水质综合评价中的应用

库路巴依[1]，白云鹏[1,2]，王玲[1]
[1]河海大学水文水资源学院，南京（210098）
[2]河北省张家口市水文水资源勘测局，河北张家口（075000）
E-mail：kulubay6@yahoo.com

摘　要：主成分分析方法是一种将多维因子纳入同一系统进行定性、定量化研究，理论比较完善的多元统计分析方法。该方法应用于水环境质量评价中，对客观、准确、全面的评价水环境质量有很好的实用性。本文采用 SPSS 统计软件，运用主成分分析方法对张家口市内三个典型水库水质进行综合评价、年际变化分析和季节性变化分析，确定影响水库水质的主要指标、掌握其变化规律。对张家口市区域水库的水资源保护研究与决策提供技术依据。

关键词：主成分分析法；水质评价；变化规律

1. 引言

水质评价是水环境质量评价的主要内容,用数学模型方法进行这方面的定量化研究,国内目前常见的方法主要有:简单指数法、分级加权评分法、普通概率统计法、模糊数学法等等,每种方法各有其优点和缺陷[1]。水质综合评价的难点在于:水质系统是由多维因子(各种污染物含量指标变量)组成的复杂系统,因子间具有不同程度的相关性,每一因子从某一方面反映了水质质量,但依据它们作综合评价有一定难度。主成分分析方法正是一种将多维因子纳入同一系统进行定量化研究、理论比较完善的多元统计分析方法,且在解决很多实际问题时已取得较好效果[2]。

张家口市地理位置特殊,位于首都北京上游,流入官厅水库、密云水库的地表水是北京的主要饮用水源，加强对张家口区域内地表水水质状况的评价和研究有很重要的实际意义。本文选取张家口区域内的三座典型水库，选用张家口水环境监测中心监测的 17 项污染指标的 10 年监测资料，运用主成分分析方法对典型水库水质进行综合评价、年际变化分析和季节性变化分析。

2. 水质评价的主成分分析法

2.1 主成分分析法的基本原理

主成分分析(Principal Components Analysis，PCA)也称为主分量分析，是一种通过降维来简化数据结构的方法：如何把多个变量（指标）化为少数几个综合变量（综合指标），而这几个综合变量可以反映原来多个变量的大部分信息。为了使这些综合变量所含的信息互不重叠，应要求它们之间互不相关。用少量的几个综合变量代替原来的许多变量是有实际意义的。由这几个综合变量出发还有可能得到一个总的指标，按此总指标来排序、分类，问题就可能简单多了[3, 4]。主成分分析法是多元分析法的一种，设法将原来众多具有相关性的指标重新组合成一组新的相互无关的综合指标来代替原来指标。

3. 分析水库基本概况和资料的选取

3.1 分析区域基本概况

张家口市位于河北省西北部，东经113°50′至116°30′，北纬39°30′至42°10′之间。北和西北与内蒙古自治区毗连，西南与山西省接壤，东南和南部与北京市及河北省保定市相邻，面积36965km^2。张家口市地势较高，自西北向东南倾斜，地形复杂，可划分为坝上、坝下两

个不同的地貌单元。境内河流分属三个流域，即：内陆河流域、滦河流域及海河流域。主要有五个水系，即：内陆河水系、滦河水系及海河流域的永定河水系、潮白河水系和大清河水系[5]。

建国以来，水利建设取得了较大成绩。全市建成各类水库98座，其中大型水库2座，中型水库7座，小型水库89座。考虑地理位置的因素，并结合当地实际，本文主要选取位于滦河源头附近的闪电河水库、位于作为北京主要水源地的密云水库上游的云洲水库和位于河北和陕西交界处的壶流河水库三座典型水库做为分析对象进行分析。三座水库位置见图1。

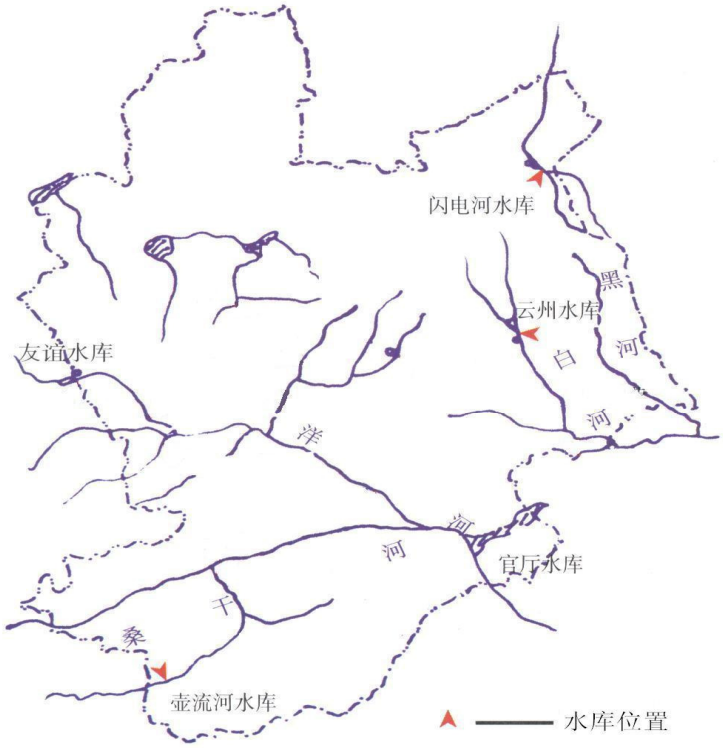


图1 张家口市分析区域水库位置

3.2 资料的选取

根据张家口水环境监测中心的资料，选取闪电河水库、云洲水库和壶流河水库三座典型水库17项污染指标的10年监测资料作为分析资料，三座水库监测频次一般为每年的4月、8月和10月。监测项目分别为：PH值、悬浮物、钙离子、镁离子、钾离子、钠离子、氯离子、硫酸根、重碳酸根、矿化度、溶解氧、氨氮、亚硝酸盐氮、高锰酸盐指数、溶解性铁、氟化物、总磷，所有监测项目张家口水环境监测监测均按照1998年水利部颁布的《水环境监测规范》（SL219-98）标准进行分析测得。

3.3 缺失值数据的处理

在所收集的监测数据中，个别项目有低于方法检出限的情况出现，主成分分析中把这种情况称为缺失值，为了便于统计分析，结合实际工作，对低于监测限的情况，在实际水质报告时，按照未检出处理，但是在本次分析应用 SPSS 统计软件处理，必须有监测数据，在实际工作，统计年平均值时，对于未检出项目，按照未检出项目的检出限的 1/2 处理，所以本次分析对于未检出项目统一按照该项目方法检出限的 1/2 来代替。

3.4 原始数据标准化

主成分综合评价方法的关键是求主成分,其工具是协方差矩阵。由于协方差矩阵易受指标的量纲和数量级的影响,经常要对原始数据进行标准化处理,标准化使协方差矩阵变成了相关系数矩阵。但在消除量钢和数量级影响的同时,也消除了各指标变异程度上的差异信息，

标准化处理公式为[6-7]：

$$x_{ij}^{*} = \frac{x_{ij}}{\overline{x}_j} \quad （其中 \overline{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}）$$

式中 i 代表某测站的监测时段； j 代表某指标的第 j 行。

4. 水库水质的主成分分析

4.1 确定主成分个数

主成分个数主要由特征值和累计方差贡献率确定。取所有特征值大于某一确定值的成分为主成分。本文确定该值为 0.6[8]。根据累计贡献率达到的百分比值确定，本文取累计贡献率达到 85%以上，其含义是此前 L 个主成分所包含的信息占原始变量包含的总信息的 85%，其余 M-L 个新变量对方差影响很小，我们认为可以接受，取前 L 个成分为主成分[9]。结果见表 1。

根据表1结果，确定8个主成分，由原来17个项目缩减为8个主成分，占据了所有监测信息的86.7%。

表1 总方差解释表

主成分	初始特征根			提取初始特征根		
	特征根	方差贡献率（%）	方差累计贡献率（%）	特征根	方差贡献率（%）	方差累计贡献率（%）
1	6.681	39.299	39.299	6.681	39.299	39.299
2	1.716	10.095	49.394	1.716	10.095	49.394
3	1.487	8.748	58.142	1.487	8.748	58.142
4	1.379	8.111	66.253	1.379	8.111	66.253
5	1.039	6.111	72.364	1.039	6.111	72.364
6	0.968	5.696	78.059	0.968	5.696	78.059
7	0.755	4.444	82.503	0.755	4.444	82.503
8	0.718	4.221	86.724	0.718	4.221	86.724
9	0.644	3.79	90.514			
10	0.549	3.232	93.746			
11	0.38	2.235	95.981			
12	0.255	1.498	97.479			
13	0.179	1.055	98.533			
14	0.129	0.761	99.294			
15	0.069	0.403	99.697			
16	0.032	0.191	99.888			
17	0.019	0.112	100			

4.2 主成分综合得分计算

每个污染物种类初始因子荷载系数表示与主成分的相关程度，正值表示正相关，负值表示负相关，其绝对值越接近 1，表示相关程度越高。表 2 列出了所选取的 8 个主成分中的荷载。

表2 初始因子荷载矩阵表

监测项目	主成分							
	F1	F2	F3	F4	F5	F6	F7	F8
PH 值	0.450	-0.583	0.016	-0.002	-0.166	-0.079	0.247	0.130
悬浮物	-0.229	-0.063	-0.482	0.254	0.238	0.440	0.601	-0.150
钙离子	0.012	0.700	0.443	-0.193	0.130	0.045	0.175	-0.309
镁离子	0.946	-0.059	-0.159	0.063	-0.103	-0.068	0.047	0.107
钾离子	0.678	0.257	0.218	-0.030	-0.126	0.307	-0.075	-0.078
钠离子	0.961	-0.099	-0.099	0.007	-0.007	0.043	-0.050	-0.011
氯离子	0.960	-0.079	-0.047	-0.035	-0.001	0.057	-0.058	0.029
硫酸根	0.891	0.007	0.100	0.013	0.001	0.016	0.057	0.009
重碳酸根	0.841	0.339	0.026	0.011	-0.055	-0.071	0.095	-0.123
矿化度	0.945	0.091	-0.031	0.036	-0.013	0.021	0.036	0.010
溶解氧	-0.264	0.279	-0.146	-0.729	0.039	-0.074	0.193	0.200
氨氮	0.369	0.296	-0.299	-0.172	0.567	0.009	-0.083	0.466
亚硝酸盐氮	0.001	0.457	-0.211	0.657	0.076	0.187	-0.246	0.093
高锰酸盐指数	-0.381	-0.358	0.535	0.024	0.353	0.424	-0.147	0.128
溶解性铁	-0.206	0.287	0.494	0.291	-0.384	0.049	0.328	0.511
氟化物	0.651	-0.293	0.424	-0.179	0.264	0.226	0.059	-0.038
总磷	0.125	-0.056	0.310	0.399	0.488	-0.619	0.197	-0.055

根据总方差解释表 1 与因子荷载表 2，并结合研究区域的实际情况对各主成分含义进行分析。

第一主成分即 F1，镁离子、钾离子、钠离子、氯离子、硫酸根、重碳酸根、矿化度共 7 个指标占荷载最大。F1 占据了所有信息的 39.3%，该主成分主要反映分析水库中常规离子和水的矿化度。第二主成分即 F2，PH 值、钙离子共 2 个指标所占荷载最大。主要反映水库水质的酸碱条件的变化。第三主成分即 F3，高锰酸盐指数、氟化物共 2 个指标所占荷载最大，主要与水库富营养变化有关。第四主成分即 F4，包含亚硝酸盐氮、溶解氧共 2 个指标的主要信息，该主成分也主要反映水库富营养化的信息。第五主成分即 F5，氨氮指标荷载最大，该主成分反映三座水库上游的面源污染，即农民大量施用化肥，由于降雨冲刷带来地表水污染。第六主成分即 F6，总磷指标标荷载最大。也主要反映水库的富营养变化，在水库的富营养中磷也是一个主要指标。第七主成分即 F7，悬浮物指标荷载最大，主要反映上游水土流失造成的地表水悬浮物变化。第八主成分即 F8，溶解性铁指标荷载最大，反映水库金属污染的主要信息。

各主成分表达式系数用初始因子荷载矩阵第 i 列向量除以 λ_i 开平方就得到第 i 个主成分的系数向量[8，10-11]。具体到本文分析结果，根据初始因子荷载矩阵表，利用公式 $A_1=B_1/SQR(6.681)$（B_1 为第一主成分因子荷载矩阵系数，6.681 为第一主成分特征值)可以计算出第一主成分系数，同理可以确定 8 个主成分系数，各主成分得分系数见表 3。

表3 主成分系数表

监测项目	F1 系数	F2 系数	F3 系数	F4 系数	F5 系数	F6 系数	F7 系数	F8 系数
PH 值	0.1741	-0.4451	0.0131	-0.0017	-0.1629	-0.0803	0.2843	0.1534
悬浮物	-0.0886	-0.0481	-0.3953	0.2163	0.2335	0.4472	0.6917	-0.177
钙离子	0.0046	0.5344	0.3633	-0.1644	0.1275	0.0457	0.2014	-0.3647
镁离子	0.366	-0.045	-0.1304	0.0536	-0.101	-0.0691	0.0541	0.1263
钾离子	0.2623	0.1962	0.1788	-0.0255	-0.1236	0.312	-0.0863	-0.0921
钠离子	0.3718	-0.0756	-0.0812	0.006	-0.0069	0.0437	-0.0575	-0.013
氯离子	0.3714	-0.0603	-0.0385	-0.0298	-0.001	0.0579	-0.0668	0.0342
硫酸根	0.3447	0.0053	0.082	0.0111	0.001	0.0163	0.0656	0.0106
重碳酸根	0.3254	0.2588	0.0213	0.0094	-0.054	-0.0722	0.1093	-0.1452
矿化度	0.3656	0.0695	-0.0254	0.0307	-0.0128	0.0213	0.0414	0.0118
溶解氧	-0.1021	0.213	-0.1197	-0.6208	0.0383	-0.0752	0.2221	0.236
氨氮	0.1428	0.226	-0.2452	-0.1465	0.5563	0.0091	-0.0955	0.55
亚硝酸盐氮	0.0004	0.3489	-0.173	0.5595	0.0746	0.1901	-0.2831	0.1098
高锰酸盐指数	-0.1474	-0.2733	0.4387	0.0204	0.3463	0.431	-0.1692	0.1511
溶解性铁	-0.0797	0.2191	0.4051	0.2478	-0.3767	0.0498	0.3775	0.6031
氟化物	0.2519	-0.2237	0.3477	-0.1524	0.259	0.2297	0.0679	-0.0448
总磷	0.0484	-0.0427	0.2542	0.3398	0.4788	-0.6291	0.2267	-0.0649

据表 3 可以写出综合主成分的表达式。 根据主成分表达式，可以计算出各项目 8 个主成分的得分，各主成分得分与各主成分的方差贡献率相乘再求和，即可得到各站点各测次的主成分的综合得分。主成分综合得分计算公式为[12]：

$$F_{综} == 0.39299F_1 + 0.10095F_2 + 0.08748F_3 + 0.08111F_4 + 0.06111F_5 + 0.05696F_6 + 0.04444F_7 + 0.04221F_8$$

根据各主成分表达式和综合主成分表达式可以计算出三座水库95年至2003年每一年的8个主成分得分和主成分的综合得分，各站点1996年至2005年综合得分见表4。

表4 水库1995年至2005年主成分综合得分

站　　点	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
闪电河水库	1.035	0.954	0.973	1.115	1.151	1.061	1.238	1.513	1.3	1.063
壶流河水库	1.606	1.827	1.779	1.932	2.214	2.59	2.394	2.788	2.716	2.953
云州水库	1.003	1.213	1.08	1.019	1.031	1.245	1.232	1.508	1.212	1.167

4.3 主成分水质综合分析和年际变化分析

4.3.1主成分水质综合分析

根据主成分综合得分可以对分析水体进行不同角度的水质评价和分析，本文主要根据其综合得分对水库进行水质综合分析、年际变化分析和季节变化变化分析。据表 4，绘制水库水质综合得分年际变化折线图。见图 2.

据图 2 结果显示壶流河水库总体水质要差于另外两个水库水质状况。闪电河水库和云州水库水质状况基本相似，水质状况好于壶流河水库。分析结果符合客观实际，壶流河水库位于河北山西交界，上游一直有点源污染，水质本底值高于闪电河水库和云州水库，而闪电

河水库和云州水库上游一直没有工业污染源，主要是面源污染，水质状况一直较好。图中显示壶流河水库水质有变坏趋势。其主要原因是，近年来上游矿业迅猛发展，尾矿排水经常有疏漏现象，对壶流河水库水质造成一定程度的污染，所以水质状况有污染上升趋势。闪电河水库和云州水库两座水库水质变化一直较为平稳，2003 年以后有变好趋势。

4.3.2 主成分水质季节变化变化分析

水库储水主要来自于降雨和上游基流流入，以地表水的形式存在，所以其水质状况变化受季节的影响很大，即上游的来水量和来水水质状况不同，则水库的水质会相应的进行变化。结合当地实际，并分别选取三个水库每年的枯水期、丰水期、平水期的4月、8月、10月监测成果的主成分综合得分，进行季节性水质变化分析。三座水库4月、8月、10月多年平均主成分得分见表5。

表5 水库不同季节多年平均主成分综合得分表

站名	4 月多年平均	8 月多年平均	10 月多年平均
闪电河水库	0.9954	1.2717	1.1433
壶流河水库	2.4857	2.0891	2.1547
云州水库	1.0982	1.3263	1.088

根据图3显示，闪电河水库和云州水库丰水期水质状况要差于其他两个时期，其主要原因是，两座水库上游主要污染是面源污染，除丰水期外基流很小，上一年沉积的污染物在枯水期得不到稀释，丰水期遇大的降雨，将河道沉积的污染物冲入水库，导致水库水质变坏，丰水期过后，水中污染不断得到稀释，水质状况有所好转，平水期过后，河道水量减小污染物会继续沉积。但是壶流河水库出现丰水期水质变好，其主要原因是上游有相对固定的点源污染水质流入水库，河道沉积污染较少，丰水期降雨和产流使得水质得到稀释，水质状况有所好转，到平水期水质不断转好，平水期过后，污染得不到稀释，水质状况又会变坏。总之，水库水质状况发生季节性的变化一是由于上游基流大小和主要污染源有关，二是与季节降雨和产流有直接的关系。

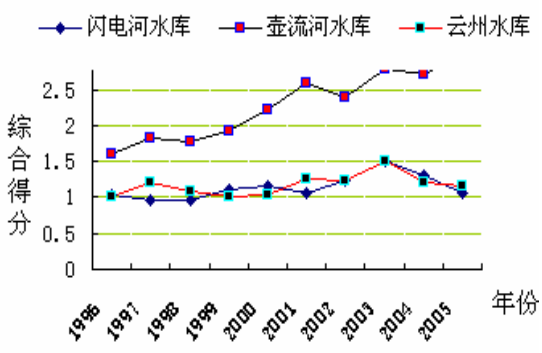


图2 水库综合得分年际变化折线图

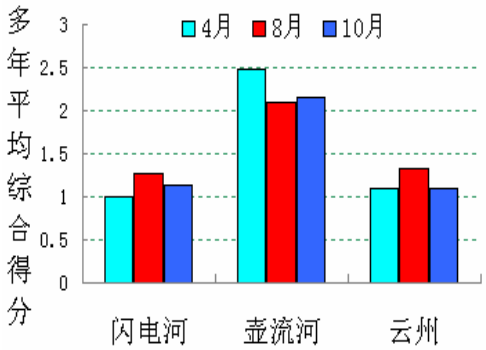


图3 水库不同季节主成分综合得分柱状图

5. 结论

文章通过对主成分分析方法的阐述，并基于 SPSS 软件应用大量的监测资料，对张家口区域内的三座典型水库进行主成分分析，得出的结果符合客观实际，并根据所得结果进行个更加深入的分析研究，得出了优于其他分析方法的分析结论， 对当地生产实践和科学管理

提供的技术依据。

该方法在水质综合评价中体现的主要特点：一，是主成分分析法处理分析数据客观性强，对数据的前处理和合理性分析有其独特性。二，是用主成分分析的方法还可以指导实际工作人员删除那些次要的指标,对我们筛选指标提供了很好的科学依据。三，是主成分分析优于传统的水质评价，评价结果兼顾水中离子的大部分信息，能综合反映水质状况。比如：单因子评价遵循悲观原则，一个污染因子超标整个水体定为劣质水。传统的综合污染指数法一般依据国家规定评价标准进行比照，部分离子如：钙、镁、钾钠等项目没有评价标准，那就或略了水中部分离子对污染的的贡献。四，是在水库综合评价中，体现出该方法的优越性，区别于传统的湖库只考虑富营养化的评价，或略了其他离子的污染。

参考文献

[1] 李惠明,尚广平.水质现状评价数学模型综合研究[J].中国环境科学,1991,11(5):356-360.
[2] 王晓鹏 河流水质综合评价之主成分分析方法 数理统计与管理 2001(40) 49-52.)
[3] 叶宗裕.主成分综合评价方法存在的问题和改进.统计与信息论坛，2004（2）29-31
[4] 侯文.对应用主成分法进行综合评价的探讨.2006（2）211-214
[5] 张家口市第二次水资源评价报告，2005
[6] 王晓鹏，曹广超. 水环境污染状况评价变量加权主成分分析方法.理统计与管理，2005（6）：1-5.
[7] 孟生旺.用主成分分析法进行多指标评价应注意的问题.统计研究 1992:86-87
[8] 高惠璇.《SAS 系统软件使用手册》.北京：中国统计出版社，1997.
[9] 米红，张文璋编著，《实用现代统计分析方法与SPSS 应用》 2002.
[10] 王文博 ,陈秀芝 .多指标综合评价中主成分分析和因子分析方法的比较 统计与信息论坛.2006（5）19-22.
[11] 林海明 ,张文霖 .主成分分析和因子分.析的异同和ＳＰＳＳ软件 .统计研究 ,
[12] 苏金明等。统计软件 SPSS for Windows 实用指南[M].北京:电子工业出版社，

Application of principal component analysis on the evaluation of the water quality of the reservoir

Kulubay[1], Bai Yunpeng [1,2], Wang Ling[1]
[1] College of Hydrology and water Resourses, Hohai University, Nanjing (210098)
[2]Zhangjiakou Branch of Hebei Provincial Hydrology & Water Resources Survey Bureau, Zhangjiakou, Hebei (075000)

Abstract
The principal components analysis is a multivariate statistical tool which brings the multi-dimensional factor into the identical system to carry on the qualitative and quantitative research. And the theory is quite perfect .This method applies in the appraisal of the water environment quality, and has a very good usability in appraising water environment quality objectively, accurately, and comprehensively. By SPSS statistics software, this article uses principal components analysis method to carry on the synthesis assessment, the year border change analysis and the seasonal variability analysis to the water quality of three typical reservoirs in Zhangjiakou, determining the major component which affects the reservoir water quality, identifying its changing rule. And also it provides the technical basis to the research and the decision-making on the conservation of water resources in Zhangjiakou's reservoir.
Keywords: principal components analysis； the assessment of water quality；seasonal variability

作者简介：库路巴依，男，1980 年生，研究生，主要研究方向是水文与生态环境。