改进主成分分析与多元回归融合的汉丰湖水质评估及预测

陈昭明[1,2]，王伟[1,3*]，赵迎[1]，徐泽宇[1]

（1. 中国科学院重庆绿色智能技术研究院，重庆　400714；2. 重庆大学，重庆　400044；
3. 重庆理工大学计算机科学与工程学院，重庆　400054）

摘　要：基于 2015—2017 年汉丰湖水质监测数据，采用改进主成分分析和多元回归相融合的评价方法对水环境质量状况进行评价。先对水质主要影响因素采用改进主成分分析作降维处理并计算主成分得分值，再对选定的主成分作多元线性回归处理得到水质预测回归模型，并用于研究区水质的评估预测。结果表明：选出的 4 个主成分因子其累积方差贡献率达到 87.3%，实现了数据结构的简化；同时，改进主成分回归预测值总体上更趋近于实测值，其预测结果的相对误差最大值 <4%，而常规方法预测结果的相对误差最大值接近 10%，体现出该方法所建模型具有较高的预测精度。

关键词：改进主成分分析；多元回归；水质评估；汉丰湖；三峡水库

中图分类号：X524；X824　　**文献标志码**：B　　**文章编号**：1006－2009(2020)04－0015－05

Evaluation and Prediction of Water Quality in Hanfeng Lake Based on Improved Principal Component Analysis and Multivariate Regression Model

CHEN Zhao-ming[1,2], WANG Wei[1,3*], ZHAO Ying[1], XU Ze-yu[1]

(1. *Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing* 400714, *China*; 2. *Chongqing University, Chongqing* 400044, *China*; 3. *College of Computer Science and Engineering, Chongqing University of Technology, Chongqing* 400054, *China*)

Abstract: Based on the water quality monitoring data in Hanfeng Lake from 2015 to 2017, the water environmental quality was evaluated by improved principal component analysis and multivariate regression model. First, the major influence factors were performed to reduce the dimension by improved PCA, the scores of the principal components were calculated. Then, the water quality prediction model was established by multiple linear regressions analysis on the selected principal components. The model was applied in evaluating and predicting the water quality in the study area. The results indicated that the cumulative variance of 4 selected principal component factors reached 87.3%. This simplified the data structure. Meanwhile, the prediction value from improved principal component regression was generally close to the measured value, the maximum relative error was less than 4%. With conventional method, the relative error was close to 10%, showing that the model had high prediction accuracy.

Key words: Improved principal component analysis; Multivariate regression; Water quality evaluation; Hanfeng Lake; Three Gorges Reservoir

三峡水库蓄水运行以来支流水体富营养化趋势加重[1-3]。同时，随着进入后三峡时期，库区经济社会快速发展[4]，工业园区和城镇在水库岸边集聚使库区水环境承受更大的压力，水环境污染问题更加严峻。为促进经济发展与环境保护的深度融合，如何构建客观高效的水质评价体系是当前普

收稿日期：2019－05－05；**修订日期**：2020－05－18

基金项目：国家自然科学基金青年科学基金资助项目(61605205)；国家重大科研仪器研制基金资助项目(51727812)

作者简介：陈昭明(1985—)，男，四川宜宾人，高级工程师，博士研究生，主要从事环境光学监测技术、流域水环境监测及水污染控制技术研究。

*通信作者：王伟 E-mail：wangwei@cqut.edu.cn

遍关注的问题。目前,常用的水质评价方法有模糊综合评价法、综合指数法、人工神经网络法、灰色决策法[5-7]等。由于水环境污染的随机性和模糊性[8-9],以及水质系统的复杂性,导致上述方法具有一定局限性。例如模糊综合评价法无法客观确定复杂指标体系的权重系数,人工神经网络法易陷入局部最优,灰色决策法易使决策值均化导致方案优选困难等[6,10],故需要探寻更加简便有效的评价方法。主成分分析法(PCA)采用降维思想将多个指标缩减为少数几个独立综合指标[11-13],简化了数据结构,提高了分析的直观性,在水质评价中得到广泛应用。然而,传统主成分分析采用零均值标准化方法,破坏了数据样本间的差异性信息,而且采用加权综合得分的评判方法,不能很好地解释分析结果[14]。此外,使用单一方法难以保证评估预测的准确性和全面性。今将主成分分析和多元回归分析相融合,采用改进后的主成分进行多元回归分析得到水质预测模型,并用于汉丰湖的水质评价,以期为水环境保护提供技术支持。

1　研究方法

1.1　主成分分析原理

主成分分析是一种多变量统计方法,采用线性变换构造一组互不相关的新变量,并从中提取少数独立综合变量以降低维数、浓缩信息和简化结构,使分析问题的过程更加直观有效。

1.1.1　主成分分析步骤

首先,对有 n 个样本、每个样本有 m 个观测指标的原始数据矩阵 $\boldsymbol{X}_{nm}$ 作标准化处理,消除量纲的影响;其次,将标准化后的数据作相关分析,参照文献[11-12]求取相关系数矩阵 $\boldsymbol{R}$ 和协方差矩阵 $\boldsymbol{C}$;再次,利用相关系数矩阵 $\boldsymbol{R}$ 或协方差矩阵 $\boldsymbol{C}$ 计算出 n 个特征值 λ_i ($i = 1,2,\cdots,n$),按由大到小顺序排列,并求得各主成分的方差贡献率 G_i 和累积方差贡献率 TG_i [11-12];最后,提取累积方差贡献率 $TG_i \geqslant 85\%$ 的前 s 个特征值对应的主成分 F_i 来降低数据维度,并计算综合评价指数 F 。

$$F = \sum_{i=1}^{s} \alpha_i F_i \ , \ \alpha_i = \lambda_i / \sum_{i=1}^{s} \lambda_i \tag{1}$$

1.1.2　改进主成分分析

传统的主成分分析在标准化处理时抹杀了各指标变异程度的差异信息;又由于数据间的非线性关系,因而其降维效果不理想;采用方差贡献率为权重的综合评价方法不能合理解释分析结果,甚至出现评价结果与事实偏差较大的现象。今参照文献[15]对主成分分析进行改进,其公式为:

$$ZX_i = x_i / \sqrt{\sum_{i=1}^{n} x_{ij}^2} \tag{2}$$

将预处理后的数据矩阵进行主成分分析,提取出主成分,其降维效果明显。

1.2　多元线性回归分析

多元线性回归是用于解释一个因变量 y 与多个自变量 x_i 之间线性关系的分析预测方法,通常采用最小二乘估计可求得回归方程 $E(y)$ 为:

$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{3}$$

同时,还须对求得的回归方程作回归系数显著性检验,主要有相关系数检验、F 检验和 t 检验,以检验自变量与因变量之间的线性关系。今采用改进主成分分析提取的因子作为多元线性回归的自变量建立回归方程,既可减少回归方程中自变量的数量,又由于各因子间的相互独立性从而保证回归方程的稳定性,解决了回归分析的共线性问题。

2　水质评估应用实例

2.1　原始参数选择及数据获取

汉丰湖是三峡库区腹地的一条重要支流,受三峡水位调控、上游来水和调节坝坝底拦水的综合影响,兼具河流、湖泊、回水河湾等多种形态特征,其生态状况对长江水环境具有重大影响。根据该支流的水质特点和监测情况,选定监测指标为透明度(SD)、水温(WT)、pH 值、总氮(TN)、总磷(TP)、溶解氧(DO)、总固体悬浮物(TSS)、氨氮(NH_3-N)、硝态氮(NO_3^--N)、高锰酸盐指数(I_{Mn})等 10 项水质理化指标和叶绿素 a(Chl-a)[16]。检测方法按照《水和废水监测分析方法》(第四版)《环境水质监测质量保证手册》(第二版)进行。按照汉丰湖的分布形态,在湖区共设置 8 个采样点,分别位于南河大丘坝、石龙船大桥、头道河大桥、东湖郡、东河大桥、交汇口、库心和调节坝。2015—2017 年每月下旬采样 1 次,在每个采样点采集水面下 1.0 m 处的水样,各检测指标均设置 3 个平行样,并将 3 次测定结果取平均值作为原始数据。以东湖郡为例,其部分数据见表 1。将该数据按照公式(2)进行标准化处理,作为主成分分析的初始数据,随机抽取标准化后的部分数值见表 2。文中采用 SPSS 22.0、Origin 8.0 和 Excel 2013 进行数据分析和处理。

2.2　主成分因子的选取

主成分分析前须对表 2 数据作 KMO(Kaiser-Meyer-Olkin)和 Bartlett 球型检验,只有当 KMO 检验值 >0.5,Bartlett 检验值 <0.05 时才适合进行主成分分析。采用 SPSS 22.0 处理可得 KMO 检验值为 0.694,Bartlett 检验值为 0.000 3,故表 2 的数据间存在相关性,可进行主成分分析。对表 2 的数据计算得出相关系数矩阵见表 3,特征值及方差贡献率见表 4,初始因子载荷矩阵见表 5。主成分分析的结果显示:前 4 项主成分的贡献率分别为 45.067%、20.405%、11.582%和 10.231%,累积贡献率达到 87.3%,>85%,这 4 个主成分已包含原来 10 个影响因子的绝大部分信息,故选取前 4 个主成分作评价分析。

由表 5 初始因子载荷矩阵中的数据除以表 4 主成分相对应的特征值的平方根,即可得到主成分的载荷值,其值越大说明主成分与该变量的相关性越好。据此可得主成分公式为:

表 1　原始水质数据

Table 1　Raw data of water quality

SD /m	θ(WT) /℃	pH 值	ρ(TN) /(mg·L^{-1})	ρ(TP) /(mg·L^{-1})	ρ(DO) /(mg·L^{-1})	ρ(TSS) /(mg·L^{-1})	ρ(NH_3-N) /(mg·L^{-1})	ρ(NO_3^--N) /(mg·L^{-1})	ρ(I_{Mn}) /(mg·L^{-1})
1.27	20.60	8.38	1.61	0.12	5.63	26.18	0.16	0.57	3.30
1.17	20.15	8.36	1.94	0.16	5.50	16.15	0.20	0.81	4.17
1.10	20.12	8.51	1.40	0.11	6.10	19.08	0.18	0.47	2.92
0.96	21.04	8.35	1.93	0.17	5.67	16.22	0.26	0.81	4.37
0.83	20.95	8.31	1.20	0.18	5.49	22.19	0.17	0.77	4.55
1.22	20.31	8.48	1.80	0.16	5.59	27.96	0.19	0.64	3.92
1.23	21.05	8.23	1.81	0.12	5.73	21.80	0.16	0.69	3.84
1.10	19.81	8.37	1.69	0.13	5.79	18.39	0.19	0.70	3.36

表 2　标准化后的水质数据①

Table 2　Standardized water quality data①

ZX_1	ZX_2	ZX_3	ZX_4	ZX_5	ZX_6	ZX_7	ZX_8	ZX_9	ZX_{10}
0.296	0.258	0.258	0.234	0.209	0.252	0.321	0.219	0.216	0.217
0.273	0.253	0.258	0.282	0.278	0.246	0.198	0.273	0.306	0.274
0.257	0.252	0.262	0.203	0.191	0.273	0.234	0.246	0.178	0.192
0.257	0.235	0.258	0.227	0.278	0.258	0.201	0.328	0.261	0.260
0.224	0.264	0.258	0.280	0.295	0.254	0.199	0.355	0.306	0.287
0.194	0.263	0.256	0.289	0.313	0.246	0.272	0.232	0.291	0.299
0.285	0.255	0.262	0.262	0.278	0.250	0.343	0.260	0.242	0.258
0.257	0.248	0.258	0.246	0.226	0.259	0.225	0.260	0.264	0.221

①ZX_1—ZX_{10}分别代表 SD、WT、pH 值、TN、TP、DO、TSS、NH_3-N、NO_3^--N、I_{Mn}标准化后的数据。

表 3　相关系数矩阵

Table 3　Correlation coefficient matrix

成分	ZX_1	ZX_2	ZX_3	ZX_4	ZX_5	ZX_6	ZX_7	ZX_8	ZX_9	ZX_{10}
ZX_1	1	-0.187	0.206	-0.450	-0.521	0.188	0.362	-0.345	-0.382	-0.574
ZX_2	-0.187	1	0.159	0.444	0.078	-0.031	0.268	-0.417	0.175	0.365
ZX_3	0.206	0.159	1	-0.402	-0.455	0.089	0	0.094	-0.295	-0.249
ZX_4	-0.450	0.444	-0.402	1	0.810	-0.596	0.049	0.050	0.787	0.856
ZX_5	-0.521	0.078	-0.455	0.810	1	-0.556	0.004	0.306	0.709	0.890
ZX_6	0.188	-0.031	0.089	-0.596	-0.556	1	-0.203	-0.201	-0.622	-0.594
ZX_7	0.362	0.268	0	0.049	0.004	-0.203	1	-0.644	-0.299	-0.037
ZX_8	-0.345	-0.417	0.094	0.050	0.306	-0.201	-0.644	1	0.332	0.239
ZX_9	-0.382	0.175	-0.295	0.787	0.709	-0.622	-0.299	0.332	1	0.791
ZX_{10}	-0.574	0.365	-0.249	0.856	0.890	-0.594	-0.037	0.239	0.791	1

表 4　特征值及方差贡献率

Table 4　Eigenvalue and variance contribution rate

成分	特征值	方差贡献率/%	累积贡献率/%
ZX_1	4.507	45.067	45.067
ZX_2	2.041	20.405	65.472
ZX_3	1.158	11.582	77.054
ZX_4	1.023	10.231	87.285
ZX_5	0.567	5.671	92.956
ZX_6	0.333	3.333	96.289
ZX_7	0.179	1.793	98.082
ZX_8	0.096	0.961	99.043
ZX_9	0.067	0.667	99.710
ZX_{10}	0.029	0.290	100.000

表 5　初始因子载荷矩阵

Table 5　Initial factor loading matrix

成分	1	2	3	4
ZX_1	−0.621	0.272	−0.330	0.387
ZX_2	0.250	0.646	0.650	−0.182
ZX_3	−0.397	−0.046	0.677	0.580
ZX_4	0.912	0.281	0.017	−0.061
ZX_5	0.912	0.010	−0.198	−0.022
ZX_6	−0.671	−0.146	0.191	−0.602
ZX_7	−0.137	0.848	−0.263	0.231
ZX_8	0.325	−0.840	0.077	0.262
ZX_9	0.877	−0.100	0.012	0.119
ZX_{10}	0.945	0.110	0.130	0.032

$$F_1 = -0.29ZX_1 + 0.12ZX_2 - 0.19ZX_3 + 0.43ZX_4 + 0.43ZX_5 - 0.32ZX_6 - 0.06ZX_7 + 0.15ZX_8 + 0.41ZX_9 + 0.45ZX_{10} \quad (4)$$

$$F_2 = 0.19ZX_1 + 0.45ZX_2 - 0.03ZX_3 + 0.20ZX_4 + 0.01ZX_5 - 0.10ZX_6 + 0.59ZX_7 - 0.59ZX_8 - 0.07ZX_9 + 0.08ZX_{10} \quad (5)$$

$$F_3 = -0.31ZX_1 + 0.60ZX_2 + 0.63ZX_3 + 0.02ZX_4 - 0.18ZX_5 + 0.18ZX_6 - 0.24ZX_7 + 0.07ZX_8 + 0.01ZX_9 + 0.12ZX_{10} \quad (6)$$

$$F_4 = 0.38ZX_1 - 0.18ZX_2 + 0.57ZX_3 - 0.06ZX_4 - 0.02ZX_5 - 0.60ZX_6 + 0.23ZX_7 + 0.26ZX_8 + 0.12ZX_9 + 0.03ZX_{10} \quad (7)$$

从主成分公式可以看出,与第一主成分密切相关的指标是 ZX_4(TN)、ZX_5(TP)、ZX_9(NO_3^-−N)、ZX_{10}(I_{Mn})等 4 个指标,与碳、氮、磷等元素有关,主要反映水体中的有机物污染状况,可以表征水体的富营养化程度,而且由于第一主成分的方差贡献率达到 45.1%,远大于其他几个主成分的方差贡献率,因而第一主成分对水质的评价起决定性作用;与第二主成分密切相关的指标是 ZX_2(WT)、ZX_7(TSS)和 ZX_8(NH_3−N)3 个指标,主要反映 WT、TSS、NH_3−N 对水体的影响情况,表现出水体外源泥沙及周围环境等对水质的作用;与第三主成分密切相关的指标是 ZX_2(WT)和 ZX_3(pH 值)2 个指标,主要反映 WT 和 pH 值对水体的影响;与第四主成分密切相关的指标是 ZX_1(SD)、ZX_3(pH 值)和 ZX_6(DO)3 个指标,主要反映 SD、pH 值和 DO 对水体的影响。水体中悬浮物质越多,对光的散射和吸收越强,SD 就越小,而 DO 反映了水体的自净能力,该主成分代表了水体的物理状态特征。提取的主成分反映的信息与实际采样调查得出的结果相吻合,均反映出水体受多种水质指标的共同影响而变得较为复杂,氮、磷为主要污染指标,水体呈现富营养化趋势。由此得到的主成分因子为后续的回归分析及评价提供了基础。

2.3　主成分回归计算

Chl-a 浓度的高低可反映水体富营养化状态和水质的动态变化,预测其浓度变化对于水华的预警和水体富营养化防治具有重要意义。回归分析模型作为分析多变量间相互关系的常用方法,适用于 Chl-a 浓度的预测分析。

由式(4)—式(7)分别计算出主成分 F_1—F_4 的得分值,并进行 Chl-a 与 4 个主成分的相关性分析,其 Pearson 系数分别为 0.517、−0.216、0.721、−0.682,具有较好的相关性。以 4 个主成分为自变量,Chl-a 为因变量进行改进主成分回归分析,可得回归模型参数,见表 6。

由表 6 中的信息可得回归模型表达式为:

$$\text{Chl-a}_j = -3.943 + 8.970F_1 + 12.001F_2 + 51.103F_3 - 29.468F_4 \quad (8)$$

该模型中 $n=14$,$r=0.892$。为了与改进主成分回归模型的预测精度对照,采用原始检测数据进行多重线性回归分析得到 Chl-a 的预测模型为:

$$Y = -43.495 - 55.029ZX_1 - 5.473ZX_2 + 212.646ZX_3 + 43.278ZX_4 - 1.722ZX_5 + 31.78ZX_6 - 1.732ZX_7 - 7.386ZX_8 + 3.392ZX_9 - 23.864ZX_{10} \quad (9)$$

该模型中 $n=14$,$r=0.914$,回归系数的显著性水平比式(8)的低。将表 2 中各行数据分别带入式(4)—式(7)可计算出 F_1—F_4 的值,然后带入式

表 6　主成分回归模型参数

Table 6　Parameters of PCA regression model

变量	回归系数	标准误差	t	r	F	显著性
β_0	-3.943	11.261	-0.350	0.892	8.742	0.004
β_1	8.970	4.833	1.856			
β_2	12.001	7.757	1.547			
β_3	51.103	29.424	1.737			
β_4	-29.468	23.862	-1.235			

(8)中可获得改进主成分回归模型的预测值;同理,将表 2 中各行数据带入式(9)可得到多重线性回归模型的预测值。将两种方法得到的结果与实测值相比较,其预测结果对比见图 1。由图 1 可见,改进主成分回归预测值总体上更接近于实测值。经计算,采用改进主成分回归预测的相对误差平均值为 1.52%,最大值为 3.61%,而采用常规的多重线性回归预测的相对误差平均值为 4.92%,最大值为 9.02%。通过改进主成分分析,在降低数据维度和获得数据间的差异信息的同时,有效改善了最小二乘估计在消除变量间多重相关性影响的缺陷,从而有效提高了预测模型的准确度。

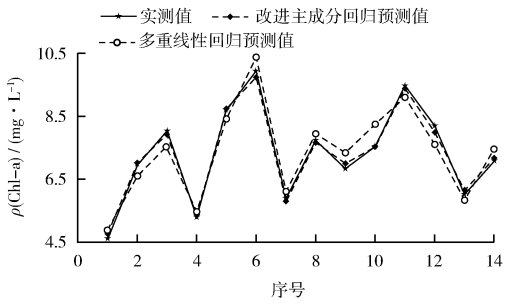

图 1　预测结果对比

Fig. 1　Comparison of prediction results

3　结论

(1) 改进主成分分析方法将原始的多个变量指标压缩为 4 个主成分因子来反映水体的污染程度,实现了数据结构的简化。

(2) 改进主成分分析选出 4 组主成分,第一主成分反映汉丰湖水体中有机物污染状况,可以表征水体的富营养化程度;第二主成分反映 TSS、NH_3-N 对水体的影响,代表外源泥沙环境对水质的作用;第三主成分反映 WT 和 pH 值对水体的影响;第四主成分反映 SD、pH 值和 DO 对水体的影响,代表了水体的物理状态特征。

(3) 将主成分因子作为多元线性回归的自变量构建预测模型,具有较高的预测精度,可快速有效地预判水体状态,便于开展水质监测。后续将进一步优化模型,在水环境质量预测及信息化管理中发挥更好的作用。

[参考文献]

[1] YAN H Y, HUANG Y, WANG G Y, et al. Water eutrophication evaluation based on rough set and petri nets: A case study in Xiangxi River, Three Gorges Reservoir[J]. Ecological Indicators, 2016, 69: 463 - 472.

[2] 赵士波,郭平,李斗果,等. 三峡库区澎溪河营养盐时空变化特征及富营养化评价[J]. 四川环境,2018,37(1):51 - 58.

[3] WU D, YAN H Y, SHANG M S, et al. Water eutrophication evaluation based on semi-supervised classification: A case study in Three Gorges Reservoir[J]. Ecological Indicators, 2017, 81: 362 - 372.

[4] 齐黎明. 三峡库区移民迁建小城镇风貌规划研究——以万州区武陵镇为例[D]. 重庆:重庆大学,2017.

[5] 林海,李阳,李冰,等. 北京市妫水河水质现状评价[J]. 环境监测管理与技术,2019,31(2):40 - 43.

[6] YAN F, QIAO D Y, QIAN B, et al. Improvement of CCME WQI using grey relational method[J]. Journal of Hydrology, 2016, 543: 316 - 323.

[7] MLADENOVIĆ-RANISAVLJEVIĆ I I, TAKIĆ L, NIKOLIĆ D. Water quality assessment based on combined multi-criteria decision-making method with index method[J]. Water Resources Management, 2018, 32(7): 2261 - 2276.

[8] 胡宇飞,余得昭,过龙根,等. 武汉东湖水体异味物质及其与水环境因子相互关系[J]. 湖泊科学,2017,29(1):87 - 94.

[9] 杨浩,张国珍,杨晓妮,等. 基于模糊综合评判法的洮河水环境质量评价[J]. 环境科学与技术,2016,39(S1):380 - 386.

[10] OSTAD-ALI-ASKARI K, SHAYANNEJAD M, GHORBANIZADEH-KHARAZI H. Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran[J]. KSCE Journal of Civil Engineering, 2016, 21(1): 1 - 7.

[11] DALAL S G, SHIRODKAR P V, JAGTAP T G, et al. Evaluation of significant sources influencing the variation of water quality of Kandla creek, Gulf of Katchchh, using PCA[J]. Environmental Monitoring and Assessment, 2010, 163(1/4): 49 - 56.

[12] ARSLAN O. Spatially weighted principal component analysis (PCA) method for water quality analysis[J]. Water Resources, 2013, 40(3): 315 - 324.

[13] 程晨. 基于主成分分析的上海市大气降水中 DL-PCBs 来源初判[J]. 环境监测管理与技术,2018,30(4):18 - 22,45.

[14] 贺密,贾杰,张敏. 主成分分析法在地下水质量评价中的应用[J]. 地下水,2015,37(6):6 - 8.

[15] 刘清园,李永,蒲迅赤,等. 改进的主成分分析法在水库水质评价中的应用研究[J]. 四川环境,2017,36(6):116 - 122.

[16] 杨兵. 三峡前置库汉丰湖试运行年水环境变化特征及控制效果评估[D]. 重庆:西南大学,2017.