聚类分析在秦淮河水质指标相关性研究中的应用

马 振，周 密

（河海大学 水文水资源学院，江苏 南京 210098）

摘 要：随着水质监测尺度和监测网络的扩大，传统的水质指标相关性分析的方法已经不能适应于庞大的水质数据。而采用聚类分析法，在对水质指标进行降维处理的同时，可以筛选出水质相关项。利用SPSS软件计算水质指标相关系数矩阵，并绘制聚类分析树形图，对已知水质数据进行相关性分析，结果发现秦淮河东山站水体总有机碳和高锰酸钾指数、总氮和氨氮具有较强相关性。结合线性回归方程的验证，证明R型聚类分析在庞大数据背景下的水质指标相关性研究中具有较好的效果，可以在水污染治理、水质监测评价中发挥较好的作用。

关键词：水质指标；相关性；聚类分析；线性回归；大数据

中图分类号：X830 文献标识码：A 文章编号：1000-0852(2018)01-0077-04

水环境中水质指标的相关性主要由水中相互联系的组分决定。在一个相对稳定的水域内，污染源及水中的污染物不会发生剧烈变化，则其中各水质指标间的相关性也比较稳定。研究水质指标间的相关性，可以有选择的减少监测项目，优化监测网络，节省监测成本[1]。近年来，不少学者做了关于水质监测数据的合理性分析，及水质参数之间相关性的研究。比如林志贵、徐立中等基于D-S理论，对长江口区域水质数据进行了融合处理[2]；Zhou Feng、Liu Yong等运用多元分析法，探讨了水质空间差异的分析方法[3]；贾利运用线性回归法，对淮河流域水质参数相关性进行了研究[4]。但随着水质监测尺度和监测网络的不断扩大，水质数据变得愈加庞大和复杂[5]，传统的研究水质指标相关性的方法在大量数据中难以奏效。

秦淮河是南京市的母亲河，全长120km[6]，随着生活污水的大量排入，秦淮河的生化需氧量、氨氮、总磷等指标严重超标[7]，破坏了秦淮河的生态环境。目前对秦淮河的研究多集中局部水质指标的监测和污染治理上，对水体内部各污染指标间关系研究较少。因此本文采用多元统计分析中的聚类分析法，以秦淮河东山站为例，对其各污染指标间的关系进行探究，为大数据背景下水质指标的相关性分析提供思路，在目前大数据和机器学习风靡的背景之下，具有重要意义。

1 聚类分析

聚类分析是一种建立分类的多元统计分析方法，它能够将一批样本（或变量）数据根据其诸多特征，按照在性质上的亲疏程度在没有先验知识的情况下进行自动分类，产生多个分类结果；类内部个体特征具有相似性，不同类间个体特征的差异性较大[8]。

常见的聚类方法有层次聚类和K-Means聚类，而层次聚类分析又分为Q型聚类和R型聚类。Q型聚类是对样本进行聚类，使具有相似特征的样本聚在一起，使差异性大的样本分离开来。R型聚类是对变量进行聚类，使差异性大的变量分离开来，具有相似性的变量聚集在一起；在相似变量中选择少数具有代表性的变量参与其他分析，可以减少变量个数，达到变量降维的目的[9]。本文意图通过聚类分析的方法，研究多个水质指标间的相关性，选出关系密切或具有代表性的变量，以达到评价指标降维的目的，因此采用R型聚类分析法。根据其原理，运用SPSS软件进行计算。

首先对原始数据进行预处理，不同水质指标间的量纲或数量级不同，会对聚类分析和相关性研究产生影响。为了使不同水质指标间的数据具有可比性，需对数据进行标准化处理。常见的数据标准化方法有“最小—最大标准化”、“Z-score标准化”和“按小数定

收稿日期：2016-12-02

作者简介：马振（1993-），男，山东菏泽人，硕士研究生，主要研究方向为水质评价与预测。E-mail：myatm2012@126.com

标标准化”等，Z-score 标准化变换后的数据均值为 0，标准差为 1，消去了量纲的影响，故本文采用此法：

$$Z_{ij}=\frac{X_{ij}-\overline{X}_{ij}}{S_j} \quad (1)$$

式中：$i=1,2,3,\cdots,n$（n 为样品数）；$j=1,2,3,\cdots,m$（m 为变量数）；X_{ij} 为原始数据；$\overline{X}_{ij}$ 为第 j 个变量在样品的平均值；S_j 为变量在样品的标准差；Z_{ij} 为标准化数据[10]。

利用标准化后的数据，计算各变量之间的相关系数。对相关系数矩阵逐层分析，不同变量类型下个体距离采用平方欧氏距离计算，个体与小类、小类与小类间距离采用组间平均链锁距离计算，逐步计算至各类对象归为一类，绘制聚类分析树形图。

2 水质指标相关性分析

2.1 研究区概况

本文采用江苏省南京市秦淮河东山水文站 2011~2012 年连续监测的 24 个月的水质数据，选取 pH 值、溶解氧（DO）、高锰酸钾指数（COD_{Mn}）、氨氮（NH_3-N）、总磷（TP）、总氮（TN）、总有机碳（TOC）共 7 项水质指标（约 5 000 条）进行相关性分析。秦淮河是长江下游的一条重要支流，其在江宁东山镇分为两支，一支为秦淮新河，一支流入南京城区。秦淮河江宁东山段起于牛首山河口，止于江宁上坊门桥，全长 8.5km，水功能区划属于工业、景观娱乐用水，水环境功能区目标是Ⅳ类水[11]。

2.2 相关系数的计算

不同的相关系数可以度量不同类型的变量，常用的有皮尔逊简单相关系数、斯皮尔曼等级相关系数和肯德尔秩相关系数等[12]。本文采用度量定距型变量间相关关系的皮尔逊简单相关系数，它的数学定义为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (2)$$

式中：n 为样本量；x_i 和 y_i 分别为两变量的变量值。

表 1 为利用 SPSS 软件计算水质指标的相关系数矩阵。由分析结果知，当显著性水平 α 为 0.01 时，TP 和 COD_{Mn}（相关系数为 0.534）、TN 和 DO（相关系数为 0.706）、TN 和 NH_3-N（相关系数为 0.790）、TOC 和 COD_{Mn}（相关系数为 0.816）、NH_3-N 和 TOC（相关系数为 0.641）、TP 和 TOC（相关系数为 0.723）具有较强的线性关系。

表1 各水质指标相关系数矩阵

Table1 The correlation coefficient matrix of water quality index

样品	pH	DO	COD_{Mn}	NH_3-N	TP	TN	TOC
pH	1.000						
DO	0.328	1.000					
COD_{Mn}	-0.511	-0.134	1.000				
NH_3-N	-0.098	0.498	0.338	1.000			
TP	-0.133	-0.102	0.534	0.417	1.000		
TN	-0.124	0.706	0.221	0.790	0.139	1.000	
TOC	-0.420	0.122	0.816	0.641	0.723	0.475	1.000

2.3 水质指标聚类分析

为了进一步分析 2.2 中七个水质指标相关项的相关性，避免出现假相关，利用相关系数矩阵进行 R 型聚类分析，绘制聚类分析树形图（见图 1）。树形图以躺倒树的形式展现了聚类分析中的每一次类合并的情况，SPSS 自动将各类间的距离映射到 0~25 之间，并将凝聚过程显示在图上[13]。

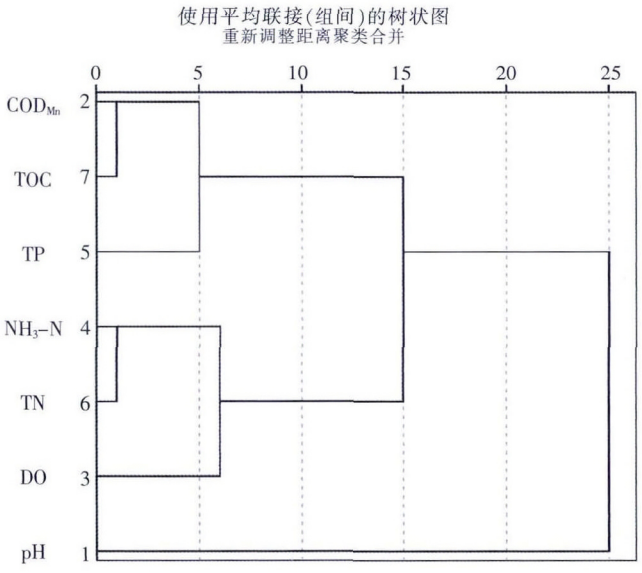

图 1 R 型聚类分析树形图

Fig.1 The tree diagram of R-CA

由图 1 可知，TOC 和 COD_{Mn}、TN 和 NH_3-N 距离最近，率先合并成一类；其次是 TP 与（TOC、COD_{Mn}）合并，它们间的距离大于 TOC 和 COD_{Mn} 距离；然后是 DO 与（TN、NH_3-N）合并；再次是前两个大类合并为一类，此时类间的距离已经比较大；最后所有个体聚成一类。通过聚类过程可知，TOC 和 COD_{Mn}、TN 和 NH_3-N 关系最为密切；结合 2.2 中的相关系数，可知 TOC 和 COD_{Mn}、TN 和 NH_3-N 具有很强的线性相关性。

2.4　线性回归分析

2.4.1　线性回归

为了进一步分析水质指标间的统计关系，考查水质指标间的数量变化规律,现通过回归方程的形式[14],描述和反映 2.3 中得到的相关性最为密切的水质指标的关系,为水质预测提供科学依据。

用 SPSS 软件分别对 TOC 和 COD_{Mn}、TN 和 NH_3–N 进行拟合回归,拟合回归过程见表 2 和图 2。通过表 2 可知，在 $COD_{Mn}=a\times TOC+b$ 关系中,$a=0.555$,$b=1.311$,$R^2=0.666$,符合拟合优度检验;回归系数的显著性检验采用 t 检验法，经计算样本回归系数的 t 值为 6.626,所以回归系数 b 与 0 差异性显著，符合显著性检验。同理,在 $TN=a\times NH_3-N+b$ 关系中,$a=1.413$,$b=3.601$,$R^2=0.622$,符合拟合优度检验;回归系数的显著性检验采用 t 检验法，经计算样本回归系数的 t 值为 6.018,回归系数 b 与 0 差异性显著,符合显著性检验。综上,可得出 TOC 和 COD_{Mn}、TN 和 NH_3–N 的线性回归关系,表达式如下:

$$\begin{cases} COD_{Mn}=0.555\times TOC+1.311 \\ TN=1.413\times NH_3N+3.601 \end{cases} \quad (3)$$

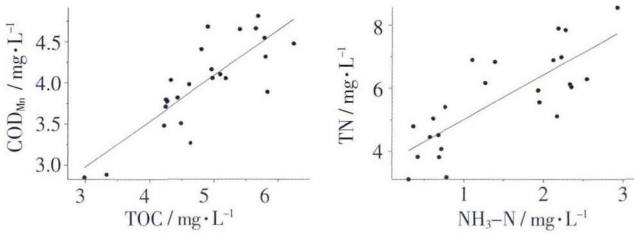

图 2　线性拟合回归图

Fig.2　The linear fitting regression

表2　线性回归过程表

Table2　The linear regression process

项	B	标准误	R	R^2	t	Sig.
TOC	0.555	0.084	0.816	0.666	6.626	0
（常数）	1.311	0.41			3.195	0.004
NH_3–N	1.413	0.235	0.789	0.622	6.018	0
（常数）	3.601	0.39			9.234	0

2.4.2　回归方程检验

利用回归方程，根据总有机碳和氨氮的监测值,分别计算高锰酸钾指数和总氮的浓度,并与实际监测数据进行对比,计算相对误差。分析发现,相对误差在合理范围之内。以 2012 年 12 月 20 日至 2012 年 12 月 31 日数据为例,由表 3 可知,除个别点外,两者的相对误差均在 5%左右,说明两个线性回归方程是可以接受的。

3　结论

(1)通过相关性分析发现,在秦淮河东山水文站监测的水质指标中,TOC 和 CODMn、TN 和 NH3–N 关系密切,具有较强的线性相关关系。

(2)TOC 和 COD_{Mn}、TN 和 NH_3–N 的相关关系得到验证,说明本文提出的通过聚类分析筛选相关项的方法,可以在庞大的数据背景下发挥较好的作用。通过聚类分析,实现了水质指标的降维处理,进而可以优化水质监测项目,用较少的指标即可有效反映区域水质状况,降低监测成本。在进行水质评价时,也可以优化评价过程。

(3)针对秦淮河的污染控制,可以根据本文的方法,对各站水质指标的关系进行系统监测,以实现对秦淮河污染的全方位掌控,为污染治理奠定基础。

表3　监测数据对比表

Table3　The contrast of monitoring data

TOC 监测值 / mg·L⁻¹	COD_{Mn} 监测值 / mg·L⁻¹	COD_{Mn} 计算值 / mg·L⁻¹	相对误差	NH_3–N 监测值 / mg·L⁻¹	TN 监测值 / mg·L⁻¹	TN 计算值 / mg·L⁻¹	相对误差
5.78	4.54	4.52	–0.54%	2.47	6.81	7.14	4.86%
5.64	4.66	4.44	–4.66%	2.65	7.49	7.39	–1.31%
4.96	4.16	4.06	–2.31%	2.59	7.68	7.30	–4.96%
5.10	4.10	4.14	1.02%	2.44	6.71	7.09	5.70%
5.40	4.65	4.31	–7.33%	2.39	7.00	7.03	0.32%
4.90	4.68	4.03	–13.85%	2.27	6.62	6.85	3.47%
4.81	4.41	3.98	–9.65%	2.20	7.52	6.75	–10.27%
4.22	3.48	3.65	4.87%	2.30	7.36	6.89	–6.46%
4.43	3.82	3.77	–1.26%	2.32	6.72	6.92	3.00%
4.61	3.98	3.87	–2.74%	2.06	7.40	6.55	–11.40%
5.18	4.05	4.19	3.29%	2.06	6.48	6.54	0.98%
5.80	4.31	4.53	5.00%	2.17	6.97	6.70	–3.86%

参考文献：

[1] 张旋,王启山,于淼,等. 多元统计分析技术在水质监测中的应用[J]. 中国给水排水, 2010,26(11):120-126. (ZHANG Xuan, WANG Qishan, YU Miao, et al. Application of multivariate statistical techniques to water quality monitoring [J]. China Water & Waste Water, 2010,26(11):120-126. (in Chinese))

[2] 林志贵,徐立中,黄凤辰,等. 基于 D-S 理论的多源水质监测数据融合处理[J]. 计算机工程与应用, 2004,40(10):3-5. (LIN Zhigui, XU Lizhong, HUANG Fengchen, et al. Multi-source water quality monitoring data fusion based on D-S theory [J]. Computer Engineering and Applications, 2004,40(10):3-5. (in Chinese))

[3] ZHOU Feng, LIU Yong, GUO Huaicheng. Application of multivariate statistical methods to water quality assessment of the watercourses in northwestern new territories, Hong Kong [J]. Environmental Monitoring & Assessment, 2007,132(1-3):1-13.

[4] 贾利. 淮河流域水质参数相关性研究[J]. 水资源保护, 2001,(2):48-49. (JIA Li. Study on correlativity of water quality parameters for Huaihe River basin [J]. Water Resources Protection, 2001,(2):48-49. (in Chinese))

[5] 周丰,郭怀成,黄凯,等. 基于多元统计方法的河流水质空间分析[J]. 水科学进展, 2007,18(4):544-551. (ZHOU Feng, GUO Huaicheng, HUANG Kai, et al. Multivariate statistic technique for spatial variation in river water quality [J]. Advances in Water Science, 2007,18(4):544-551. (in Chinese))

[6] 魏玉香,周宁晖,方孝华. 南京市内秦淮河环境综合整治中水质变化趋势回顾[J]. 环境监测管理与技术, 2004,16(6):42-43. (WEI Yuxiang, ZHOU Ninghui, FANG Xiaohua. The trend review of water quality change in the comprehensive treatment of inner Qinhuai River in Nanjing [J]. Environmental Monitoring and Management, 2004,16(6):42-43. (in Chinese))

[7] 高桂枝,卢海龙,陈晨,等. 南京市内秦淮河水葫芦生态修复潜力分析[J]. 安徽农业科学, 2010,38(33):18945-18947. (GAO Guizhi, LU Hailong, CHEN Chen, et al. The analysis on ecological eestoration potential of water hyacinth in inner Qinhuai River in Nanjing [J]. Journal of Anhui Agricultural Sciences, 2010,38(33):18945-18947. (in Chinese))

[8] 何万谦,黄金良. 澳门半岛近岸海域水质时空变异分析[J].环境学报, 2010,31(3):606-611. (HE Wanqian, HUANG Jinliang. Identification of spatio-temporal variation in the seaside water quality along Macau Peninsula [J]. Environmental Science, 2010,31(3):606-611. (in Chinese))

[9] 薛薇. 统计分析与 SPSS 的应用 [M]. 北京: 中国人民大学出版社, 2011:289-292. (XUE Wei. Statistical Analysis and SPSS Application [M]. Beijing: Renmin University of China Press, 2011:289-292. (in Chinese))

[10] 张永波. 地下水环境保护与污染控制 [M]. 北京: 中国环境科学出版社, 2003:56-58. (ZHANG Yongbo. The Groundwater Environment Protection and Pollution Control [M]. Beijing: China Environmental Science Press, 2003:56-58. (in Chinese))

[11] 毛晓文,常虹. 秦淮河南京段水质变化过程及污染控制[J]. 水资源保护, 2014,30(1):74-78. (MAO Xiaowen, CHANG Hong. Water quality variation process in Nanjing reach of Qinhuai River and pollution control measures [J]. Water Resources Protection, 2014,30(1):74-78. (in Chinese))

[12] 张先富. 基于 HSPF 半分布式水文模型的新立城水库流域水环境模拟及预测研究[D]. 长春: 吉林大学, 2015. (ZHANG Xianfu. Study on Water Environment Simulation and Prediction of Xinlicheng Reservoir Basin Based on HSPF Semi-distributed Hydrological Model [D]. Changchun: Jilin University, 2015. (in Chinese))

[13] 刘铭,朱长军,顿珠加措,等. 聚类分析方法在济宁市水质分析中的应用[J]. 河北工程大学学报(自然科学版), 2014,31(2):69-71. (LIU Ming, ZHU Changjun, DUNZHU Jiacuo. Application of cluster analysis method in water quality analysis of Jining City [J]. Journal of Hebei University of Engineering (Natural Science), 2014,31(2): 69-71. (in Chinese))

[14] 顾廷富, 李坤, 包军. 大庆水质自动监测 TOC 与高锰酸盐指数(COD_{Mn})的相关性分析[J]. 环境科学与管理, 2006,31(3):154-155. (GU Tingfu, LI Kun, BAO Jun. Correlation analysis of TOC and COD_{Mn} of automatic water quality monitoring in Daqing [J]. Environmental Science and Management, 2006,31(3):154-155. (in Chinese))

Application of Cluster Analysis in Correlation Study on Water Quality Indexes

MA Zhen, ZHOU Mi

(*College of Hydrology and Water Resources, Hohai University, Nanjing* 210098, *China*)

Abstract: With the enlargement of water quality monitoring scope and network, traditional methods of water quality indexes correlation analysis can no longer be applicable to a large amount of water quality data. But the cluster analysis method can work it. SPSS software was used to calculate the correlation coefficient matrix of water quality index with cluster analysis method. And a tree diagram was drawn to analyze the correlation of water quality data. It is concluded that the total organic carbon and potassium permanganate index, total nitrogen and ammonia nitrogen have strong correlation. Combined with the verification of linear regression, it can be proved that the R cluster analysis has a good effect on the study of water quality index correlation, and it can play an important role in water pollution control or water quality monitoring and evaluation.

Key words: water quality index; correlation; cluster analysis; linear regression; big data