第9卷 第3期
1997年9月

湖 泊 科 学
JOURNAL OF LAKE SCIENCES

Vol. 9, No. 3
Sep., 1997

255-260 P343.3 X824

太湖北部梅梁湾水域水质因子聚类*

刘元波 高锡芸
(中国科学院南京地理与湖泊研究所,南京 210008)

提要 计算了沿梁溪河河口到太湖湖心断面上10个监测点17个水质因子93组数据的Pearson相关系数和Kendall秩相关系数,进而运用最小距离法进行了因子聚类.正态分布检验和聚类结果表明,采用Kendall秩相关进行因子聚类为宜.结果将诸因子聚为五大类:TDN、TN、CON、NO_3-N、NH_4-N、OH和COD_{Mn}归为一类;TDP、TP、PO_4^{3-}和pH值归为一类;SS和SD归为一类.反映了该水域环境五个方面特性:氮素污染、磷素污染、水体光学性质、水体藻类变化以及氮类污染物时空降解过程.

关键词 太湖,水质,相关,聚类分析,水质因子,湖水

70年代以来,太湖水质下降,富营养化进程加剧,而梅梁湾水域尤为突出.80年代后期,中国科学院南京地理与湖泊研究所对太湖水质进行了综合评价[1].1991年起,太湖湖泊生态系统研究站沿梅梁湾内梁溪河河口到太湖湖心,设立了10个监测点,按月取样,获得大量监测数据,结果表明,沿梁溪河河口到湖心这一断面,TP、TN、COD_{Mn}和电导率(CON)等水质因子呈现出梯度变化[2].本文采用1991—1993年间监测数据,通过对17个水质理化和生物学因子进行聚类分析,来进一步探讨水质因子间关系.这17个因子是:Chl. a、CON、COD_{Mn}、DO、NH_4-N、NO_2-N、NO_3-N、OH、pH值、SD、SS、TDP、TN、TDN、TP、PO_4^{3-}以及T_w.

1 方法与步骤

对变量进行聚类,称为R型聚类分析.进行R型聚类分析,需要计算变量间的相关系数或相似系数.在计算相关(似)系数前,对17个因子数据进行标准化,以消除量纲影响.运用极差标准化方法,按下式计算:

$$x_i = \left| x_i^* - \frac{1}{n}\sum_{i=1}^{n} x_i^* \right| / \{\max\{x_i^*\} - \min\{x_i^*\}\}$$

其中,x_i为变量X的标准化数值;x_i^*为变量X的第i个样本值;n为样本数目;$\max\{x_i^*\}$,$\min\{x_i^*\}$分别为变量X的样本最大和最小值.

1.1 相关系数及其计算

计算因子间相关(似)系数,是进行聚类分析的首要步骤.本文除采用Pearson相关系数外,还采用Kendall秩相关系数.

1.1.1 Pearson相关系数和相似系数 Pearson相关系数常为人们所用.相似系数又称夹角弦.它们分别用下式计算:

* 国家自然科学基金(59506027)和江苏省社会发展研究基金(BS95035)资助项目.
收稿日期:1995-12-27;收到修改稿日期:1996-02-10.

$$r_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}},\quad r_2=\frac{\sum_{i=1}^{n}x_i\cdot y_i}{\sqrt{\sum_{i=1}^{n}x_i^2\cdot\sum_{i=1}^{n}y_i^2}}$$

其中，r_1 为 Pearson 相关系数；r_2 为相似系数；n 为样本数目；x_i、y_i 分别为变量 X、Y 的第 i 个样本值；$\bar{x}$、$\bar{y}$ 为分别为样本的算术平均数.

可以证明，数据经过极差标准化后，Pearson 相关系数和相似系数计算公式等价，因而采用一种计算公式即可. 本文运用 Pearson 相关系数.

1.1.2 秩相关系数 本文采用 Kendall 秩相关系数，其计算基于秩数据[3]. 对两个变量 X 和 Y 的 n 对数据 x_i、y_i($i=1,\cdots,n$)由大到小排序变量 X，求秩 R_{x_i}. 然后对另一组数据中的 R_{y_i} 值，求 C_i(其值为 R_{y_i} 中第 i 个数据之后比第 i 个数大的数据个数). 遇同分值取 0.5.

同分较少时，用下式计算 Kendall 秩相关系数

$$r_1=\left|4\sum C_i-n(n-1)\right|/(n(n-1))$$

同分较多时，对每次涉及到 m 个数据的同分，计 t=m(m−1). 分别对两组数据中同分 t 求和，得 $\sum t_x$ 和 $\sum t_y$. 用下式计算 Kendall 秩相关系数

$$r_2=\left|4\sum C_i-n(n-1)\right|/\sqrt{(n^2-n-\sum t_x)(n^2-n-\sum t_y)}$$

计算 n 个数据变量间的相关系数，可得到 $n\times n$ 阶对称矩阵 R_0.

1.2 用最小距离系统聚类法进行 R 型聚类分析

聚类分析有多种方法，如系统聚类法、分解法、加入法、动态聚类法等等，其中又以系统聚类法用得最多[4]. 在系统聚类诸方法中，最小距离法满足较多的可容性条件，如单调性、最小支撑树结构、最优 $W-K$ 分类等[5]. 本文采用最小距离法对环境因子进行聚类. 步骤如下：

(1) 选择 R_0 中的最大元素(负相关系数采用绝对值)，设此最大元素为 $r_{p,q}$，其中 p、q 分别表示 p 类和 q 类，则将 p 类和 q 类合并成一个新类 $r=\{p,q\}$. 在 R_0 中消去所对应的行与列，并加入由新类 $r=\{p,q\}$ 与剩下的其它未聚合类的系数所组成的一行和一列，得到一个(n−1)×(n−1)阶系数矩阵 R_1.

(2) 由 R_1，重复(1)的作法得到(n−2)×(n−2)阶系数矩阵 R_2. 直到 n 个个体聚为一大类为止.

(3) 合并过程中记下合并时个体的编号及合并时的相关系数，绘制聚类图.

2 结果与讨论

由 17 个水质因子 93 组样本，计算得到 Pearson 积矩相关系数和 Kendall 秩相关系数矩阵，皆为 17×17 阶对称矩阵. 临界相关系数 $r_{5\%}^{93}=0.1868$. 表 1 列出了计算的 Pearson 相关系数(上三角)和 Kendall 秩相关系数(下三角).

2.1 分类结果比较

根据上面计算的 Pearson 相关系数矩阵和 Kendall 秩相关系数矩阵，分别用最小距离法进行聚类，结果见图 1.

表1　17个水环境因子的Pearson相关系数(上三角)和Kendall秩相关系数(下三角)

Tab. 1　Pearson coefficients and Kendall rank coefficients of 17 aquatic environment factors

	Chl-a	CON	COD_{Mn}	DO	NH_4-N	NO_2-N	NO_3-N	OH	pH	SD	SS	TDN	TDP	TN	PO_4^{3-}	TP	T_w
Chl-a		0.19	0.38	-0.19	-0.04	0.17	-0.09	0.09	0.39	-0.03	-0.12	-0.08	-0.18	-0.05	-0.02	-0.06	0.49
CON	0.18		0.66	-0.63	0.80	0.75	0.22	0.76	-0.43	0.11	-0.15	0.83	0.57	0.85	0.49	0.66	-0.05
COD_{Mn}	0.31	0.44		-0.52	0.67	0.68	-0.03	0.74	-0.27	-0.03	0.04	0.56	0.28	0.66	0.25	0.68	-0.14
DO	-0.22	-0.36	-0.22		-0.67	-0.50	0.14	-0.63	0.32	-0.16	0.17	-0.55	-0.35	-0.54	-0.54	-0.51	0.40
NH_4-N	0.02	0.48	0.35	-0.13		0.60	-0.08	0.87	-0.57	-0.07	0.02	0.90	0.54	0.91	0.66	0.78	-0.22
NO_2-N	0.13	0.56	0.42	-0.14	0.51		0.16	0.65	-0.34	0.16	-0.11	0.60	0.27	0.68	0.19	0.47	-0.06
NO_3-N	-0.11	0.20	0.03	0.05	0.06	0.19		-0.13	-0.03	-0.03	0.06	0.26	0.29	0.23	-0.15	-0.09	-0.04
OH	0.13	0.54	0.46	-0.15	0.56	0.50	-0.03		-0.58	0.02	-0.09	0.78	0.48	0.80	0.54	0.75	-0.27
pH	0.27	-0.18	-0.15	0.06	-0.44	0.28	-0.00	-0.35		0.06	-0.10	-0.55	-0.48	-0.54	-0.31	-0.49	0.44
SD	0.07	0.14	-0.03	-0.12	0.13	0.06	-0.07	0.15	0.08		-0.09	-0.04	-0.04	-0.02	0.04	-0.13	0.24
SS	-0.09	0.07	0.08	0.06	-0.11	-0.06	0.05	-0.09	-0.07	-0.09		-0.01	0.04	0.01	0.02	0.13	-0.29
TDN	-0.05	0.55	0.34	-0.17	0.57	0.48	0.34	0.39	-0.38	0.02	-0.01		0.62	0.97	0.52	0.69	0.37
TDP	-0.15	0.40	0.33	-0.03	0.53	0.42	0.14	0.42	-0.43	-0.02	0.04	0.53		0.58	0.42	0.57	-0.28
TN	-0.01	0.57	0.42	-0.17	0.55	0.51	0.36	0.39	-0.33	-0.00	0.01	0.83	0.51		0.50	0.72	-0.28
PO_4^{3-}	-0.09	0.32	0.30	-0.12	0.33	0.37	0.03	0.35	-0.39	-0.05	0.07	0.33	0.47	0.35		0.55	-0.07
TP	-0.02	0.31	0.42	0.03	0.47	0.38	0.00	0.40	-0.39	-0.08	0.11	0.35	0.58	0.57	0.48		-0.38
T_w	0.38	0.06	-0.04	-0.49	-0.23	0.07	-0.00	-0.19	0.35	0.16	-0.17	-0.13	-0.32	-0.13	-0.20	-0.37	

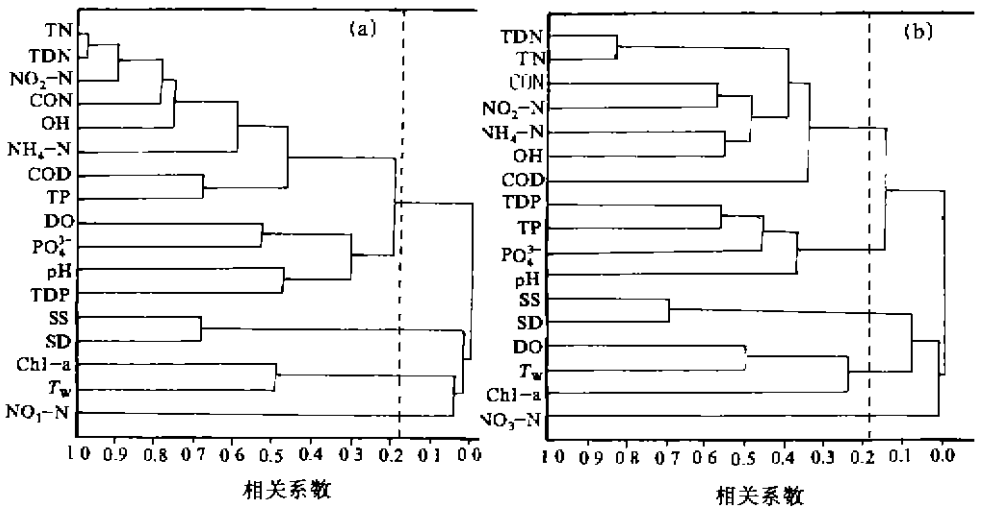

图1　基于Pearson秩矩相关(a)及基于Kendall秩相关(b)的聚类结果

Fig. 1　Cluster analysis of 17 water quality factors based on Pearson coefficient (a) and on Kencall rank coefficient (b)

图 1a 将 17 个因子分为五大类：1) 含氮污染物与 TP、COD_{Mn}、CON、OH；2) 部分含磷污染物与 DO、pH；3) SS 与 SD；4) Chl. a 与 T_w；5) NO_3-N. 图 1b 也分为五大类：1) 含氮污染物及 COD_{Mn}、CON、OH；2) 含磷污染物与 pH；3) SS 与 SD；4) DO、T_w 与 Chl. a；5) NO. N. 两种分类结果有一些相似之处，然而并不相同. 如图 1a 将 TP 归入含氮污染物一类. 虽然无论从 Pearson 相关还是 Kendall 相关系数计算结果来看，TP 与氮类因子(除 NO_3-N 外)都存在较好的相关性，但 TP 为各种形态磷之和，显然应与磷类因子归为一类. 再有，像 COD_{Mn}、pH 值等均为水质综合指标，与诸多环境因子有关，而图 1a 显示 COD_{Mn} 与 TP、pH 与 TDP 关系最密切并归为同一小类. 图 1b 则将 COD_{Mn} 与含氮污染物归为一大类，并与各种形态的氮及 CON 分为一类；同样 pH 值则与各种形态含磷污染物归为一大类. 更加明显的问题是，图 1a 中 DO 与 PO_4^{3-} 直接归为一小类. 事实上，DO 与温度有着更为密切的关系，同时由于藻类光合作用产氧而与叶绿素 a 浓度亦有关，而 DO 与 PO_4^{3-} 并无直接联系. 因此图 1a 的结果显然不甚合理，而图 1b 所揭示的关系才更为恰当.

分别根据 Pearson 相关系数和 Kendall 秩相关系数矩阵，同样运用最小距离法进行聚类分析，得出了两种颇有差异的分类结果. 这是由于两种相关系数间存在一定差别. Pearson 相关系数属于参数相关方法，一般用来研究变量间线性关系. 使用此相关系数的前提条件是：所研究数据应该服从正态分布. 因此在使用 Pearson 积矩相关系数前，应对样本数据作正态分布检验. 但人们在使用此相关系数时，往往忽略了对所使用数据进行分布检验，因而进行聚类分析时可能会导致错误的分类结果. 秩相关系数属非参数方法，其适用性要比前者强得多，不要求研究数据遵从正态分布. 它用来判断数据在 n 维空间中是否具有某种趋势性分布，测量变量间相关或不相关的程度. 下面对所使用数据进行正态分布检验.

2.2 正态分布检验

在正态分布检验诸方法中，采用偏度—峰度检验. 作为经典的参数方法，其检验功效高，并有明确的概率意义. 用 γ_1 和 γ_2 表示总体偏、峰度系数.

双侧偏度系数 t-检验的原假设为：$H_0: \gamma_1=0$　对立假设为：$H_1: \gamma_1\neq 0$

双侧峰度系数 t-检验的原假设为：$H_0: \gamma_2=0$　对立假设为：$H_1: \gamma_2\neq 0$

分别用 g_1 和 g_2 表示样本数为 n 的样本偏度系数和峰度系数，偏度检验和峰度检验的计算统计量分别是：

$$t_1=\frac{|g_1|}{\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}},\quad t_2=\frac{|g_2|}{\sqrt{\frac{24n(n-1)}{(n-3)(n-2)(n+3)(n-5)}}}$$

在 α 显著性水平，对双侧检验，拒绝原假设的条件是：

$$t_1>t^{n-1}(\text{偏度检验}),\qquad t_2>t^{n-1}(\text{峰度检验})$$

对于服从正态分布的总体，两项检验结果都不应当显著.

表 2 为对各项目进行偏度—峰度检验的结果. 显然，在 $\alpha=0.01$ 水平上，只有 pH 值和水温的数据分布服从正态分布. CON、NH_4-N 和 SD 的偏度与峰度统计量，各有一项大于 $t^{n-1}_{0.01}=2.638$，仍判定为不服从正态分布.

正态分布检验表明，17 个项目中，大多数项目的数据并不服从正态分布，因而用 Pearson 相关系数进行计算并不恰当，由此得到的聚类结果也很难反映出因子间的真实关系. 对秩相关

表 2　17 个水质因子的偏度-峰度统计量

Tab. 2　Statistics of Skewness and Kurtosis of 17 water quality factors

水质因子	Chl. a	CON	COD_{Mn}	DO	NH_4-N	NO_2-N	NO_3-N	OH	pH
偏度统计量 r_1	7.2853	5.8015	8.7242	−4.5252	10.4141	7.3526	3.9888	6.7406	−0.7829
峰度统计量 r_2	7.1123	0.7139	15.0319	2.2712	13.7470	6.6001	0.0762	5.5249	1.2284
水质因子	SD	SS	TDN	TDP	TN	PO_4^{3-}	TP	T_w	
偏度统计量 r_1	3.8660	5.6595	6.8892	12.6392	6.7571	26.8094	11.7041	−0.7639	
峰度统计量 r_2	1.2467	4.2070	5.1976	23.9241	4.5712	108.5980	19.2757	−2.4922	

系数而言，就不存在所使用数据应该服从正态分布的限制. 总而言之，根据 Kendall 秩相关得到的聚类结果，才比较恰当地反映出梁溪河河口至湖心断面上诸水质因子间的关系(图 1b). 分类结果显示出沿梁溪河河口到太湖湖心断面上水质环境的重要特性：氮素污染、磷素污染、水体光学特性、水体藻类变化以及氮类污染物时空降解过程. 含氮污染物和含磷污染物是该水域环境污染的两种主要组成成分. COD_{Mn}、CON 和 OH，都与氮类(除 NO_3-N 外)和磷类物质呈现出较高的相关性趋势(见表 1 秩相关系数). 相关系数最低也达 0.30. 从表 1 中可以看出，COD_{Mn}、CON 和 OH 与氮类(除 NO_3-N 外)的相关系数，普遍比磷类物质相关系数要大. 反映在聚类结果中，COD_{Mn}、CON 和 OH 与氮类物质聚为一类. pH 与氮类(除 NO_3-N 外)和磷类物质也都呈现出较好的负相关性(表 1). 相关系数在 −0.44 至 −0.33 之间，但综合分类结果表明，pH 与含磷污染物分为一类. SS 和 SD 相对独立于该水域的其它因子，反映水域藻类情况的 Chl. a，与 T_w 及 DO 的关系相对密切. NO_3-N 自成一类，是由于它并非由污染源直接排出，而是由污染物 NH_4-N 经氧化后生成，从时间上看，其含量随 NH_4-N 排出的时间增长而增加；从空间上看，表现为与距污染源的距离远近有关，而与其余 16 项指标均无直接关系.

3　结语

数理统计方法无论在环境科学还是在生物科学中都得到了广泛应用. 文献[3,6]指出应用中存在着不少问题，其中问题之一是不注意运用各种数理统计方法的前提条件，因而可能导致不正确的结论. 本文运用极差方法将数据标准化，计算了 Pearson 相关系数和 Kendall 秩相关系数，采用最小距离法进行了聚类分析. 结合正态分布检验，说明要注意恰当合理地运用相关方法. 在使用 Pearson 相关系数前须进行正态分布检验. 聚类结果将所采用的水质因子分为五大类，反映了太湖水质环境因子间关系，在一定程度上反映出该水域环境特征. 虽然本文运用聚类分析方法揭示出太湖水质因子间的一些关系，但要全面地揭示因子间的定性定量关系，还需运用其它数学的、物理的、化学甚至生物学乃至多学科的手段和方法，进行更加深入地研究.

参 考 文 献

1　孙顺才，黄漪平主编. 太湖. 北京：海洋出版社，1993. 210−218

2　蔡启铭，高锡芸，陈宇炜等. 太湖水质的动态变化及影响因子的多元分析. 湖泊科学，1995，7(2)：97−106

3　陶　澍编著. 应用数理统计方法. 北京：中国环境科学出版社. 1994

4 方开泰编著. 实用多元统计分析. 上海：华东师范大学出版社，1989. 215－256
5 孙文爽，陈兰祥编. 多元统计分析. 北京：高等教育出版社，1994. 362－367
6 南雄明，刘来福. 生物统计的各种检验方法和使用条件概述. 生态学杂志，1995，14(3)：67－70

FACTOR CLUSTER ANALYSIS OF WATER QUALITY IN MEILIANG BAY, TAIHU LAKE

Liu Yuanbo Gao Xiyun
(*Nanjing Institute of Geugraphy and Limnology, Chinese Acadmey of Sciences, Nanjing* 210008)

Abstract

With the development of regional economy, Taihu Lake, which is situated in the subtropical zone of China, has been becoming a eutrophic lake. Its state of aquatic environment began to be monitored in 1991, under the support of TaiLLER (Taihu Laboratory for Lake Ecosystem Research). From the mouth of the Liangxi River to the centre of Taihu Lake, there set up 10 monitoring sites.

Seventeen factors of water quality are involved in this paper. Pearson correlation and Kendall rank correlation are used to calculate the coefficients of factors and the coefficient matrixes are obtained, respectively. The minimum distance method is adopted to proceed the cluster analysis. The results show that the cluster analysis based on rank correlation coeffee icient is reasonable while that on Pearson correlation coefficient improbable. The normal distribution test indicated that only the data of pH value and water temperature are abided by the normal distribution. It is necessary to proceed the normal distribution test before Pear son correlation is applied.

Seventeen factors are classified into five groups by means of cluster analysis based on Kendall rank correlation coefficient as follows:

(1) total dissolved nitrogen(TDN), total nitrogen(TN), nitrite nitrogen (NO_2-N), ammoniacal nitrogen(NH_4-N), alkalinity(OH), conductivity(CON), and chemical oxygen demond(COD_{Mn}). (2) total dissolved phosphorus(TDP), total phosphorus(TP), phosphate (PO_4^{3-}) and pH value. (3) suspended substance(SS) and transparency(SD). (4) dissolved oxygen(DO), water temperature(T_W) and chlorophyll a(Chl. a). (5) nitrate nitrogen(NO_3-N).

Five groups reflect five respects of this aquatic environment; nitrogen pollution, phosphorus pollution, optical property, algal changes, and temporal and spatial variations of nitrogen pollutant.

Key Words Taihu Lake, water quality, correlation, cluster analysis