

武汉东湖浮游植物水华的多元分析*

蔡庆华

(中国科学院水生生物研究所, 武汉)

提 要

应用多元分析中的 fuzzy 聚类分析、Fisher 判别分析和逐步回归分析,对武汉东湖 1983 至 1985 年浮游植物水华进行一系列分析处理,得到其 I、II 两站的水华判别函数,分析了判断水华形成的主要指标: I 站为初级生产量和氨氮浓度, II 站为温度、硝酸盐浓度、叶绿素 a 浓度和 COD,此外,本文还就 I、II 两站浮游植物的两个因子(chl a 和初级生产量)分别对不同形态的营养元(氨氮、亚硝酸盐、硝酸盐、总氮、磷酸盐、总磷和硅酸盐)进行了回归分析;结果指出,现在东湖浮游植物不是以磷为限制因子而是氮限制。

关键词 fuzzy 聚类分析, Fisher 判别分析, 逐步回归分析, 水华

水华的形成与发展是生态系统内藻类与环境多种因素作用的结果,它与营养物浓度、光照、温度、透明度、藻类丰度浓度等各方面化学、物理、生物、地理指标有密切的联系^[11-14]。若孤立地以某项指标单一分析(统计学上称为一元分析)来反映此现象的发展状态及其与环境因素相互关系的规律性是不理想的。而且,不同藻类水华的开始时间,发展速度以及爆发程度等都存在着很大的差异性,各国学者至今尚未对水华进行数学分析,目前所见的文献仅限于生物学上的描述。

进入 70 年代以来,随着生物数学和计算机科学的日益发展和广泛应用,一种被称为多元分析的统计方法正在逐步地应用于生物学生态学研究之中。多元分析(multivariate analysis)又称做多变量分析或多指标分析,可以看作是普通的一元统计分析的拓广,它能帮助解决一元分析中无法回答的一些复杂问题,更好地揭示客观事物的本质。

本工作的目的,在于引用多元分析中的 fuzzy 聚类分析, Fisher 判别分析和逐步回归分析等,揭示判断武汉东湖浮游植物水华形成的主要指标并用于预报。

数据的来源和处理

本文所引用的数据,来自本所湖泊生态组 1983 年 1 月—1985 年 12 月在武汉东湖两个采样站(水果湖湾为 I 站,郭郑湖心为 II 站)获得常规采集(每月一次)数据。考虑到目

* 本文在导师刘建康教授亲切指导下完成,承梁彦龄、张尧庭、俞敏娟教授和黄祥飞副教授审阅指导,王建、沈国华、张水元、刘衢霞、华俐等同事提供数据和帮助,蔡丽复墨,特此深表谢意。

国家自然科学基金资助项目,编号 3860600。

1987 年 9 月 2 日收到。

前尚未对形成水华的主要因子作出结论,因此尽可能地考虑诸项指标:浮游植物指标中的叶绿素 a (chl a, $\mu\text{g/L}$)和毛初级生产量 (P_G , gO_2/L),水环境指标中的水温($^{\circ}\text{C}$)和透明度 (tran, m) 以及化学指标中的 pH 值,溶解氧 (DO , mg/L) 化学耗氧量 (COD , mg/L) 和各种形态的氮、磷、硅等营养元素: 氨氮 (NH_4N , mg/L)、亚硝酸盐 (NO_2N , mg/L)、硝酸盐 (NO_3N , mg/L)、总氮 (TN, mg/L)、磷酸盐 (PO_4P , mg/L)、总磷 (TP, mg/L) 和硅酸盐 (Si, mg/L) 等 14 个因子。

在取得了 I、II 两站各 3 年 (36 个月) 的样本之后,首先要判别其中哪些样本有水华生成哪些样本没有。一般来说,东湖在每年的 4、5 月份,鱼腥藻有一次水华生成,而 8、9 月份则是微囊藻水华的生成时间,但由于每年具体的环境因子不同,水华的生成月份也可能推迟或提前。鉴于当前对水华还没有一个确切的定义,以感官为根据判断是否出现水华乃是一个软性指标(或称为 fuzzy 指标),也同样是一个多因素指标,只有综合地考虑上述诸项因子,才能确切地判断水华的生成与否。在此基础上,用 Fisher 线性判别分析便可以得出一个定量指标(称之为水华判别函数),用这个指标便可以定量地反映水华的出现与否。

数学方法简述

任何一门学科都要通过分类来建立概念,也要通过分类来发现和总结规律,分类是建立和识别模型的重要基础和手段。所谓分类,就是将事物的总体分成若干个类,使总体中每一事物处在一个且仅在一个类中,而聚类分析就是根据实际情况,按某一个标准来鉴别事物之间接近程度,并把彼此接近的事物归为一类。常用的方法主要有系统聚类法,分割聚类法和 fuzzy 聚类法等,本文使用的是 fuzzy 聚类法。

令 n 表示各站所采集的样本数(本文中 $n = 36$), m 表示每次采样所采集的项目数(本文中 $m = 14$, 又称为变量数),而 x_i 表示第 i 项采集指标集, x_{ij} 表示第 i 个样本的第 j 项指标 ($i = 1, 2, \dots, n, j = 1, 2, \dots, m$)。因此, $X = (x_{ij})_{n \times m}$ 就是样本矩阵。fuzzy 聚类分析就是在矩阵 X 的基础上定义相似矩阵,经过一系列的 fuzzy 运算得到等价矩阵的过程^[9]。

假设经过 fuzzy 聚类分析发现有 n_1 个样本属于有水华出现,而 n_2 个样本无水华出现 ($n_1 + n_2 \leq n$)。根据这个结果,使用 Fisher 线性判别分析,求出系数 a_j ($j = 1, 2, \dots, m$), 建立一个判别函数 $f(x) = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m$ 这里 a_0 为一个常数,其负值就是临界判别点。在此基础上,对任何一次采集的一组数据集 $X_0 = (x_1, x_2, \dots, x_m)$ 代入判别函数 $f(x)$, 若 $f(x_0) \geq 0$, 则判断为有水华出现,若 $f(x_0) < 0$, 则判断为无水华出现。

武汉东湖浮游植物水华多元统计分析的步骤是这样的:

① 原始数据读入

$$x_{ij}, i = 1, 2, \dots, n, j = 1, 2, \dots, m$$

② 进行 fuzzy 聚类分析^[9]

③ 将有水华出现的样本排列为

$$x_{kjl}, k = 1, 2, \dots, n_1, j = 1, 2, \dots, m$$

将无水华出现的样本排列为

$$x_{k'j2}, k' = 1, 2, \dots, n_2, j = 1, 2, \dots, m$$

而 $n' = n_1 + n_2 \leq n$

④ 计算各指标的两类均值:

$$\bar{x}_{j1} = \frac{1}{n_1} \sum_{k=1}^{n_1} x_{kj1} \quad \bar{x}_{j2} = \frac{1}{n_2} \sum_{k=1}^{n_2} x_{kj2}$$

计算类内离均差平方和:

$$\omega_{ij} = \sum_{g=1}^2 \left(\sum_{k=1}^{n_g} x_{kig} \cdot x_{kjg} - n_g \cdot \bar{x}_{ig} \cdot \bar{x}_{jg} \right) \quad i, j = 1, 2, \dots, m$$

⑤ 解方程组,求判别函数的系数

$$\sum_{j=1}^m \omega_{ij} \cdot a_j = (n - 2) \cdot (\bar{x}_{i1} - \bar{x}_{i2}) \quad i = 1, 2, \dots, m$$

确定临界判别值(两类分界点)

$$c = \frac{1}{2} (\bar{z}_1 + \bar{z}_2)$$

$$\text{其中 } \bar{z}_1 = \sum_{j=1}^m a_j \cdot \bar{x}_{j1} \quad \bar{z}_2 = \sum_{j=1}^m a_j \cdot \bar{x}_{j2}$$

建立判别函数

$$f(x) = a_0 + \sum_{j=1}^m a_j \cdot x_j$$

其中 $a_0 = -c$

⑥ 检验显著性

$$D^2 = \sum_{j=1}^m a_j \cdot (\bar{x}_{j1} - \bar{x}_{j2}) \quad F = \frac{(n_1 + n_2 - m - 1) \cdot n_1 \cdot n_2}{m \cdot (n_1 + n_2 - 2) \cdot (n_1 + n_2)} \cdot D^2$$

查表 $F_\alpha(m, n_1 + n_2 - m - 1) = F_0$

若 $F \geq F_0$ 则显著,否则不显著。

⑦ 将原始分类数据回代,检验符合率

设 a = 实际为组 1 而判断为组 1 之数

b = 实际为组 1 而判断为组 2 之数

c = 实际为组 2 而判断为组 1 之数

d = 实际为组 2 而判断为组 2 之数

组 1 和组 2 的判断正确率分别为:

$$a/(a + b) \text{ 和 } d/(c + d)$$

而总的符合率则为

$$(a + d)/(a + b + c + d)$$

结 果

以上步骤用 BASIC 语言编制程序,在 IBM PC/XT 微型计算机上实现,得出如下结果:

1. 1983 年 1 月至 1985 年 12 月的 36 个月中,东湖 I、II 两站浮游植物水华的动态 fuzzy 聚类如图 1、图 2。其中小数表示相似系数,整数表示相应月份(表 1)。

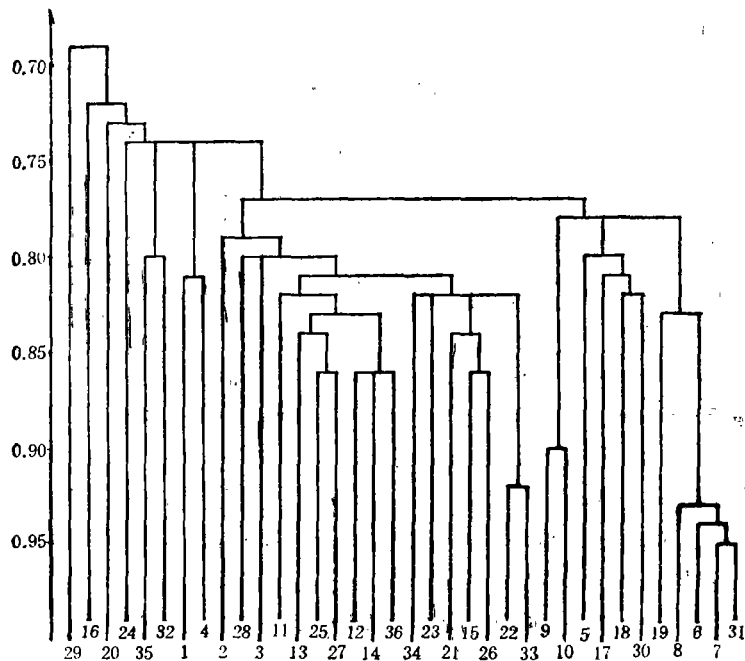


图 1 I 站 1983 年至 1985 年水华的动态 fuzzy 聚类

Fig. 1 Fuzzy cluster dendrogram of water-bloom at station I from 1983 to 1985

表 1 月份与序号的对应关系

Tab. 1 Correspondent relationship of month and number

Month Number	(1983)	1	2	3	4	5	6	7	8	9	10	11	12
		1	2	3	4	5	6	7	8	9	10	11	12
Month Number	(1984)	1	2	3	4	5	6	7	8	9	10	11	12
		13	14	15	16	17	18	19	20	21	22	23	24
Month Number	(1985)	1	2	3	4	5	6	7	8	9	10	11	12
		25	26	27	28	29	30	31	32	33	34	35	36

2. 取 $\lambda_1 = 0.77$, $\lambda_2 = 0.72$ 为截集水平。对 I 站来说,在 $\lambda_1 = 0.77$ 的水平上,水华的情况可以分成 8 类: $\{5-10, 17-19, 30, 31\}$, $\{2, 3, 11-15, 21-23, 25-28, 33, 34, 36\}$, $\{1, 4\}$, $\{32, 35\}$, $\{24\}$, $\{20\}$, $\{16\}$ 和 $\{29\}$, 其中第 1 类有水华出现,第 2 类无水华出现,后 6 类在综合考虑 14 项因子的基础上单独成类,在本文的水华计算问题中不予考虑。同

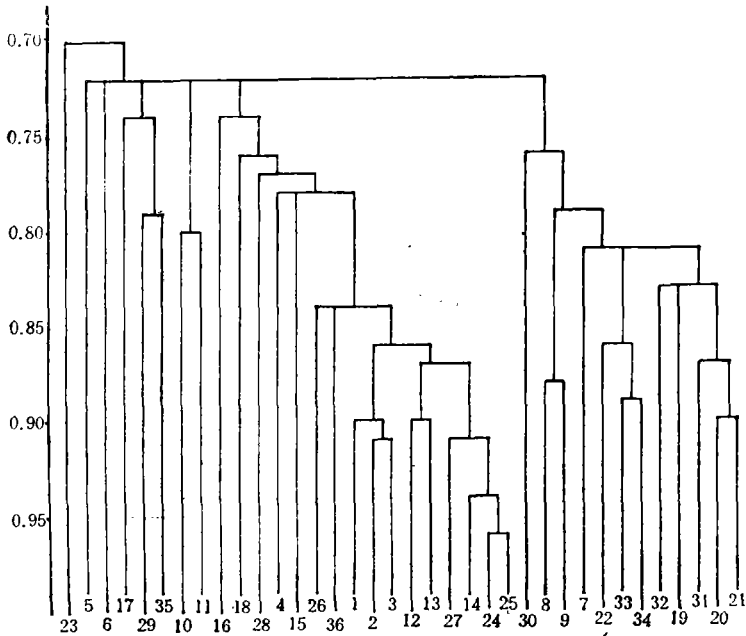


图2 II站1983年至1985年水华的动态 fuzzy 聚类

Fig. 2 Fuzzy cluster dendrogram of water-bloom at station II from 1983 to 1985

样,对II站来说,在 $\lambda_2 = 0.72$ 的水平上,水华情况可以分成7类: {7—9, 19—22, 30—34}, {1—4, 12—16, 18, 24—28, 36}, {10, 11}, {17, 29, 35}, {5}, {6}和{23},其中第1类为有水华出现,第2类无水华出现,后5类不予考虑。

也就是说, I站有水华出现的月份是1983年5—10月, 1984年5—7月和1985年6, 7月共11个月,无水华出现的月份是1983年2, 3, 11, 12月, 1984年1—3, 9—11月和1985年1—4, 9, 10, 12月共17个月,而1983年的1, 4月, 1984年的4, 8, 12月以及1985年的5, 8, 11月在综合考虑14项因子的基础上既不能判定为有水华,也不能说无水华,不予考虑。II站有水华出现的月份是1983年7—9月, 1984年7—10月和1985年6—10月共12个月,无水华出现的月份是1983年1—4, 12月, 1984年1—4, 6, 12月和1985年1—4, 12月共16个月,而1983年5, 6, 10, 11月, 1984年5, 11月和1985年的5, 11月,不能判定是否有水华出现,不予考虑。因此,对I站来说, $n' = 28$, $n_1 = 11$, $n_2 = 17$,对II站来说, $n' = 28$, $n_1 = 12$, $n_2 = 16$ 。这里, n' 就是Fisher判别分析中的样本数,不妨仍以 n 记之。

3. 进行Fisher判别分析,分别得到I、II两站的水华判别函数:

I站: 水华判别函数 $f(x)$

$$\begin{aligned} f(x) = & -60.95 - 0.026chl\ a + 1.91P_G + 0.315temp + 6.784tran + 6.714pH \\ & - 1.272DO - 1.054COD - 4.225NH_4N + 39.057NO_2N \\ & - 4.373NO_3N - 0.517TN - 4.771PO_4P - 0.181TP + 0.339Si \end{aligned}$$

$$D^2 = 25.4755 \quad F = 6.0764 ***$$

符合率 100% (表 2(a))

II 站: 水华判别函数 $f(x)$

$$\begin{aligned} f(x) = & -81.664 + 0.591chl\ a + 1.802P_G + 0.564temp + 3.31tran + 8.706pH \\ & - 2.22DO + 0.816COD - 0.753NH_4N - 151.849NO_2N \\ & - 35.64NO_3N - 3.159TN + 19.82PO_4P - 1.414TP + 0.347Si \end{aligned}$$

表 2 Fisher 判别结果: I 站 (a), II 站 (b); 逐步判别结果: I 站 (c), II 站 (d)

Tab. 2 Result of Fisher discrimination at station I(a) and II (b), and stepwise discrimination at station I(c) and II(d)

	G ₁	G ₂	Total	Correct
G ₁	11	0	11	100%
G ₂	0	17	17	100%
Total	11	17	28	100%

(a) (c)

	G ₁	G ₂	Total	Correct
G ₁	12	0	12	100%
G ₂	0	16	16	100%
Total	12	16	28	100%

(b) (d)

表 3 chl a 和 P_G 对温度和透明度的相关分析

Tab. 3 Correlation analysis of chl a and P_G on transparency and temperature

站 Station	方程 Equation	r	F	Prob. level
I	chl a = 59.08 - 33.92tran	0.53	13.38	0.00085
	chl a = 0.14 + 1.55temp	0.44	8.35	0.00666
	p _G = 9.297 - 4.57tran	0.55	14.89	0.00049
	p _G = -0.43 + 0.31temp	0.68	28.92	0.00000
	p _G = 3.45 + 0.06chl a	0.46	9.33	0.00437
II	chl a = 31.87 - 9.36tran	0.56	15.37	0.00041
	chl a = -1.34 + 0.99temp	0.54	13.90	0.00070
	p _G = 8.14 - 2.26tran	0.64	23.36	0.00003
	p _G = -0.99 + 0.30temp	0.77	50.76	0.00000
	p _G = 2.83 + 0.099chl a	0.47	9.62	0.00385

表 4 Chla 和 P_G 对营养元的相关分析

Tab. 4 Correlation analysis of chl a and P_G on nutrients

Station	Equation	r	F	Prob. level
I	ln(chl a) = 3.88 - 0.82NH ₄ N	0.37	5.30	0.0276
	ln(chl a) = 3.15 - 10.55NO ₂ N	0.82	70.96	0.00000
	ln(chl a) = 3.37 - 5.37NO ₃ N	0.40	6.65	0.01442
	p _G = 9.39 - 2.66NH ₄ N	0.56	15.58	0.00038
	p _G = 6.67 - 10.07NO ₃ N	0.36	4.99	0.03214
II	chl a = 25.24 - 34.12NH ₄ N	0.38	5.69	0.02277
	chl a = 25.51 - 55.00NO ₃ N	0.53	13.50	0.00082
	p _G = 6.30 - 7.27NH ₄ N	0.38	5.84	0.02117
	p _G = 6.49 - 12.52NO ₃ N	0.58	16.85	0.00024
	p _G = 1.41 + 1.70TN	0.44	8.21	0.00710

$$D^2 = 54.8302 \quad F = 13.4282 ***$$

符合率 100% (表 2(b))。

讨 论

上述所取用的 14 个参数之中,参数与参数之间并不是互相独立的,如东湖 I、II 两站 chl a 的含量与毛产量 P_G 存在明显正相关,透明度,水温对毛产量都有不同程度的影响^[2]。表 3 列出了东湖 I、II 两站 chl a, P_G 分别与透明度,水温以及 chl a 与 P_G 的相关方程,相关系数和概率水平。进一步分析时,抛去 14 个参数中彼此相关的对水华的判别分析无显著意义的参数,用尽可能少的彼此独立无关(或相关不甚显著)的参数来描述水华的判别过程。为此,选用多元逐步判别分析。

由于考虑的仅仅是水华的出现与否,即所谓的二类判别分析问题;二类判别问题除上述 Fisher 意义下的线性判别分析外,还可以用二值回归方法来进行分析,即假想要得到的回归方程的因变量 y 仅取两个不同的值,这里,因变量 y 的含义就是待判别的属性,即有水华出现的样本其因变量 $y = 1$, 而无水华出现的样本 $y = -1$, 自变量仍取 x_{k1} 和 x_{k2} (定义同前),用多元线性回归分析方法,求出回归系数 b_i 和常数项 b_0 , 从而建立起用以判别两类的多元回归方程: $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_mx_m$

当 $y \geq 0$ 时,判断有水华出现,否则判断无水华。数学上已经证明, Fisher 判别分析和二值回归判别分析在效果上是等价的,即不论用什么方法,得出的结果都是一样的。如果用多元逐步回归分析方法对上述问题进行处理,则得出的结果与多元逐步判别分析的结果等价。

在使用逐步回归分析对东湖水华进行逐步判别时,为避免一个自变量都不能选入方程,取显著水平为 10%, 另外,在全部 14 个变量中,最后选入方程者不应该超过 7, 所以,取临界水平为 $F_{0.1}(1, n - 7 - 1)$, 得到如下结果:

$$\text{I 站: } n_1 = 11, n_2 = 17, n = n_1 + n_2 = 28, f_0 = F_{0.1}(1, 28 - 7 - 1) = 2.97$$

回归方程(水华判别函数)为

$$y = -0.41 + 0.14p_G - 0.39\text{NH}_4\text{N}$$

$$R^2 = 0.8887, \quad F = 46.95 ***$$

符合率 100% (表 2(c))。

$$\text{II 站: } n_1 = 12, n_2 = 16, n = n_1 + n_2 = 28, f_0 = F_{0.1}(1, 28 - 7 - 1) = 2.97$$

回归方程(水华判别函数)为

$$y = -1.23 + 0.016\text{chl a} + 0.048\text{temp} + 0.067\text{COD} - 1.979\text{NO}_3\text{N}$$

$$R^2 = 0.9535, \quad F = 57.60 ***$$

符合率 100% (表 2(d))。

上述结果表明,对 I 站来说,只要测出毛初级生产量和氨氮 (NH_4N) 浓度就可以判定是否有水华出现,对 II 站来说,判断水华形成的主要因子为温度 (temp.), 硝酸盐 (NO_3N)、叶绿素 a (chl a) 和化学耗氧量 (COD)。

另外,对 I、II 两站浮游植物两个指标(chl a, p_G) 与水体中营养元 (NH_4N , NO_2N ,

表 5 chl a 和 P_G 对几种营养元的逐步回归分析Tab. 5 Stepwise regression analysis of chl a and P_G on several nutrients

站 Station	因变量 Dependent variable	营 养 元 Nutrients							R ²	F	P-value
		常数 Const.	NH ₄ N	NO ₂ N	NO ₃ N	TN	PO ₄ P	TP	SI		
I	chl a	40.92			-83.66					5.94	0.0202
	t	5.72			2.44						
	显著水平 ²⁾	0.00			0.0202						
	P _G	10.09	-2.42		-7.22					10.00	0.0004
	t	8.14	3.64		1.83						
	显著水平 ²⁾	0.00	0.0009		0.0761						
II	chl a	33.38			-60.63	-3.86				7.97	0.0015
	t	5.17			3.97	1.42					
	显著水平 ²⁾	0.00			0.0004	0.1641					
	P _G	4.75		-25.12	-10.62	1.05				9.06	0.0002
	t	3.56		1.47	3.63	1.98					
	显著水平 ²⁾	0.0012		0.1505	0.001	0.0568					

注: 1) coefficient; 2) significant level

NO_3N , TN , PO_4P , TP , 和硅酸盐 Si) 之间的关系分别作进一步分析发现: 浮游植物生物量(以 chl a 代表) 和生产量只与氮相关而与磷和硅酸盐无关; 浮游植物对硝酸盐的依赖大于对氨氮的依赖。表 4 列出了其相关方程, 相关系数和概率水平。考虑到各种形态的营养元之间有一定的联系性, 简单的相关系数并不能反应出总体的营养对浮游植物的贡献, 因此再应用多元的逐步回归分析对浮游植物与各种形态的营养元之间的关系进行分析处理(表 5)。

I 站: chl a 仅与 NO_3N 相关, p_G 则与 NH_4N 主要相关, 与 NO_3N 的相关在 90% 的水平上亦显著; II 站: chl a 与 NO_3N 极相关, 与 TN 相关不甚显著, 但作为一个因子又不可缺少, p_G 与 NO_3N 极相关, 与 TN 在 90% 的水平上相关显著, 与 NO_2N 相关不甚显著。但不管是 I 站还是 II 站, 浮游植物均与磷不相关, 这说明, 从 1983—1985 年的观测数据来看, 武汉东湖的浮游植物并不是通常所认为的磷限制, 而应该说是氮限制。至于其机制到底是什么, 则有待于进一步的研究。

参 考 文 献

- [1] 王 林、张晓卫, 1983. 微型计算机算法与程序。上海科学技术文献出版社。
- [2] 王 骥、沈国华, 1981. 武汉东湖浮游植物的初级生产力及其与若干生态因素的关系。水生生物学集刊, 7(3): 295—311。
- [3] 王学仁, 1985. 地质数据的多变量统计分析。科学出版社。
- [4] 冯士雍, 1985. 回归分析方法。科学出版社。
- [5] 冯德益等, 1983. 模糊数学方法与应用。地震出版社。
- [6] 汪培庄, 1983. 模糊集合论及其应用。上海科学技术出版社。
- [7] 张水元、刘衡霞、黄耀桐, 1984. 武汉东湖营养物氮磷的主要来源。海洋与湖沼, 15(3): 203—213。
- [8] 黄正南, 1986. 医用多因素分析及计算机程序。湖南科学技术出版社。
- [9] 蔡庆华, 1988. 东湖生态系统污染状况的 Fuzzy 聚类分析。水生生物学报, 12(3): 193—198。
- [10] 田中豊、垂水共之、脇本和昌, 1984. パソコン統計解析ハンドブック, II, 多変量解析編。共立出版株式会社。
- [11] Reynolds, C. S. and Walsby, A. E., 1975. Water-blooms. *Biological Reviews*, 50(4): 437—481.
- [12] Shapirc, J., 1973. Blue-green algae: why they become dominant. *Science*, 179(4071): 382—384.
- [13] Smith, V. H., 1983. Low nitrogen to phosphorus ratios favor dominance by blue-green algae in lake phytoplankton. *Science*, 221(4611): 669—671.
- [14] Wetzel, R. G., 1983. *Limnology*, 2nd ed. Saunders College Publishing Company. Philadelphia, London, Toronto.

MULTIVARIATE ANALYSIS OF PLANKTONIC BLOOMS IN LAKE DONGHU, WUHAN

Cai Qinghua

(*Institute of Hydrobiology, Academia Sinica, Wuhan*)

Abstract

By using multivariate analysis for 14 parameters obtained in 1983 to 1985 at the two sampling stations (I, II) in Lake Donghu, Wuhan, the relationships between water-blooms and relevant ecological parameters were analysed, and the discriminant function of water-blooms at each station was yielded. The principal factor in relation to water-blooms are gross primary production and concentration of ammonia-N at station I, and water temperature, nitrate concentration, chlorophyll a and COD at station II. Regression analysis of phytoplankton indexes (chl a and pG) against several nutrients (ammonia-N, nitrite, nitrate, total-N, phosphate, total-P and silicate) demonstrate that the principal limiting factor for phytoplankton in Lake Donghu at present is nitrogen instead of phosphorus.

Key words Fuzzy cluster analysis, Fisher discriminant analysis, Stepwise regression analysis, Water-blooms