第10卷 第4期 海 洋 与 湖 沼 Vol. 10, No. 4
1979 年 10 月 OCEANOLOGIA ET LIMNOLOGIA SINICA Oct., 1979

应用 Q 因子进行样品分类的 A-B 相关法*

范 守 志
(中国科学院海洋研究所)

海洋地质和海洋生物的研究方法是十分相似的：采集标本—测定参数—资料综合对比。往往要对许多标本进行多种参数的测定，结果获得了数量可观、结构复杂的数据阵(data matrix)。

在数学地质和统计生物学中，多参数样品群的对比和分类研究的任务就是要通过对数据阵的处理来确定各对标本在这些参数的组合状态上(或各对参数在这些标本中的分布状态上)的一致性或相似程度，以便对这些标本(或参数)分群划类，进而区划环境、阐明各区的统计特征，并为进一步研究标本(或参数)与生成环境之间的动态关系提供基础。这就是所谓的Q型分析(或R型分析)。

为此，首先是用 n 维向量这个数学概念来描述单个样品中所测的 n 种成份的组合状态，而用某种相似性统计量的数值来度量一对样品在这 n 个参数的组合状态方面的相似程度。至于一批样品的统计分群，则相应于某种数理统计过程。虽然不同的研究者们设计了各种相似性统计量，但在统计分群的方法上，被广泛采用的只有两种：群分析-聚合分类法(枝状图法)[1,2]和因子分析法(座标图法)[1,3,4]。由于这两种方法在应用时不必对区域的学科背景预作判断，因而获得了广泛的应用[5,6,7,8,9]。但其缺点是计算量大，原因也许就在于在统计的过程中抛开了区域的已知特点。其实，在海洋沉积物的研究中，往往有若干特点是可供利用的。因而，数据处理过程的大大简化不仅有必要而且也是可能的。此外，因子分析法并没有提供在载荷因子座标图上识别亚群存在与否的依据。

本文提出的 A-B 相关法同样也是一种二维相关图示法，计算量小，易于在所谓的 A-B 相关图上识别亚群的存在与否。它采用 Q 因子(即比例相似系数 $\cos\theta$)为相似性统计量，方法的实质是 A、B 双端元相关对比。

文中将先对 Q 因子的有关性质加以概括，然后介绍方法本身。至于与本方法有关的 A、B 端元的确定、在 A-B 相关图上判别同群、异群及识别亚群存在与否等问题，作者将另文介绍。

一、基本原理

1. 用 n 维向量表示样品的组成

设对样品 A 分析了 n 种成分，例如：n 种矿物，n 种生物，n 种化学元素，等等。每种成分的含量各为 X_1，X_2，……X_n。诸 X_i 可以是百分含量也可以是绝对含量的数值。

* 中国科学院海洋研究所调查研究报告第 469 号；本文承我所张兆瑾同志审阅并提供宝贵意见，李清同志绘图，谨致谢意。
本刊编辑部收到稿件日期：1978 年 6 月 27 日。

下标 $i=1, 2, \cdots\cdots n$，依约定顺序代表所测 n 种成分；起初随意约定，而后贯彻始终。

于是测量结果可以表示为一个 n 维向量：

$$\mathbf{X}=(X_1, X_2, \cdots\cdots X_n),$$

称之为样品 A 的、该 n 种成分的组合向量。其中，诸 X_i 称为分量。

每一个样品都有自己的组合向量。

量

$$X=\sqrt{\sum_{i=1}^{n} X_i^2} \tag{1}$$

称为向量 $\mathbf{X}$ 的长度。

向量

$$\mathbf{x}=(x_1, x_2, \cdots\cdots x_n)$$

称为样品 A 的归一化组合向量，其中

$$x_i=\frac{X_i}{X}, \quad (i=1, 2, \cdots\cdots n) \tag{2}$$

称为归一化分量。

显然，诸 x_i 满足归一化条件：

$$\sum_{i=1}^{n} x_i^2=1。 \tag{3}$$

由于

$$\frac{x_1}{X_1}=\frac{x_2}{X_2}=\cdots\cdots=\frac{x_n}{X_n}=\frac{1}{X},$$

所以在表示样品的相对组成时，归一化组合向量 $\mathbf{x}$ 与向量 $\mathbf{X}$ 是等价的。

2. 样品间组成相似性的定性描述

现在考察一对样品 A 和 B。

设其组合状态各为：

$$\mathbf{X}=(X_1, X_2, \cdots\cdots X_n),$$
$$\mathbf{Y}=(Y_1, Y_2, \cdots\cdots Y_n);$$

或者，

$$\mathbf{x}=(x_1, x_2, \cdots\cdots x_n),$$
$$\mathbf{y}=(y_1, y_2, \cdots\cdots y_n);$$

则下述定义可以给出 A 和 B 组成相似性的定性描述。

(定义一)若

$$\left.\begin{aligned} X_1&=Y_1,\\ X_2&=Y_2,\\ &\cdots\cdots\\ &\cdots\cdots\\ X_n&=Y_n, \end{aligned}\right\} \tag{4}$$

则称样品 A 和 B 是相同的。

(定义二)若

$$\frac{X_1}{Y_1}=\frac{X_2}{Y_2}=\cdots\cdots=\frac{X_n}{Y_n}, \tag{5}$$

则称样品 A 和 B 是一致的。

不难证明，条件 (5) 等价于：

$$\left.\begin{aligned} x_1&=y_1,\\ x_2&=y_2,\\ &\cdots\cdots\\ x_n&=y_n, \end{aligned}\right\} \tag{6}$$

若记

$$k_i=\frac{y_i}{x_i},\quad (i=1,2,\cdots\cdots n) \tag{7}$$

则条件 (5) 即 (6) 又等价于

$$k_1=k_2=\cdots\cdots=k_n=1。 \tag{8}$$

(定义三)若当 $X_i\neq 0$ 时 $Y_i=0$，且当 $Y_i\neq 0$ 时 $X_i=0$，

则样品 A 和 B 称为无关。

依 (7) 式，样品 A 和 B 无关的条件等价于

$$\left.\begin{aligned} &k_i=0, \quad \text{当 } k_i<1;\\ &k_i=\infty, \quad \text{当 } k_i>1。 \end{aligned}\right\} \quad (i=1,2,\cdots\cdots n) \tag{9}$$

或

(定义四)若诸 $k_i\approx 1$，则称样品 A 和 B 相似性强；反之，

若 $k_i\approx 0$ (当 $k_i<1$)，或 $k_i\gg 1$ (当 $k_i>1$)，则称样品 A 和 B 相似性弱。

3. Q 因子——比例相似性统计量

由下式 (10) 定义的因子 Q[4]，其数值可以作为样品 A 和 B 在所测 n 种成分的百分组成上相似程度的指标：

$$\left.\begin{aligned} Q&=x_1y_1+x_2y_2+\cdots\cdots+x_ny_n,\\ Q&=\cos\theta=\mathbf{x}\cdot\mathbf{y}=\frac{\mathbf{X}\cdot\mathbf{Y}}{XY}。 \end{aligned}\right\} \tag{10}$$

即

其理由可由下述定理来说明。

(定理一)样品 A 和 B 一致的必要而充分的条件是

$$Q=1。$$

(定理二)样品 A 和 B 无关的必要而充分的条件是

$$Q=0。$$

(定理三)若样品 A 与 B 既非一致又非无关，

则必

$$0<Q<1,$$

且 A 与 B 相似性越强，Q 值越近于 1；

A 与 B 相似性越弱，Q 值越近于 0。

(证明从略)

依式(10),角度

$$\theta = \cos^{-1}Q$$

是 n 维向量 $\mathbf{x}$ 和 $\mathbf{y}$ 之间的夹角,称为样品 A 和 B 之间的相关角。显然,相关角越近于零,两个样品的相似性越强;反之,相关角越近于 90°,两个样品越近于无关。

二、A-B 相关法

设在某海区采集了一批标本,共有 N 个(图 1)。

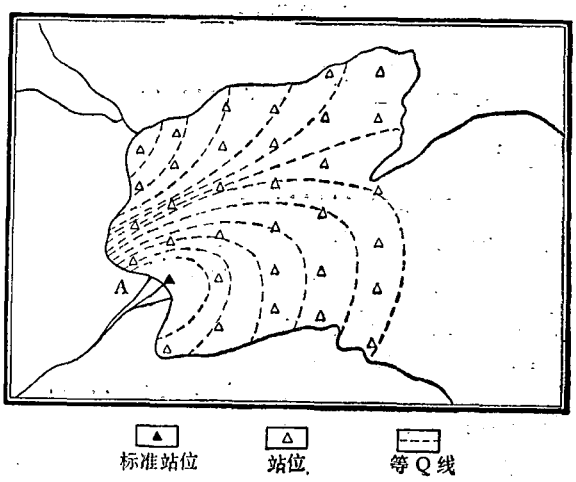

图 1 等 Q 图

▲标准站位 △站位 ——等 Q 线

如果依群分析-聚合分类法或因子分析法来分群,那么第一步需将任何一对标本间的 Q 值(或其它相似性统计量)都计算出来。这样一共要计及

$$C_n^2 = \frac{1}{2}N(N-1)$$

个 Q 值。

在所研究的海区具有已知的学科背景时,不必如此。

方法的各步如下。

1. 选择坐标系

即依研究内容,选择有对比价值的 n 种成分的含量作为组合向量的各分量。第 i 个样品 S_i 的组合向量设为

$$\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots\cdots X_{in}) \quad i = 1, 2, \cdots\cdots N。$$

2. 选择参考向量 A

依地区的学科背景,选择某站位为标准站位,以该站位上的标本的组合向量 $\mathbf{A}$ 作为参考向量,其余 $N-1$ 个标本的组合状态分别用来与 $\mathbf{A}$ 进行对比,而不是在任何一对标本之间都进行一次对比。

例如,在有河流注入的海区中,为研究河流挟带的泥砂在海区中分布的态势,可以选

择河口处的站位 A（图 1）作为标准站位[1]。

3. 计算诸 $Q_{A,i}$ 值

$$Q_{A,i}=\frac{\sum_{k=1}^{n}X_{ik}A_k}{\sqrt{\sum_{k=1}^{n}X_{ik}^2}\sqrt{\sum_{k=1}^{n}A_k^2}},\tag{11}$$

以及

$$Q_{A,i}=\cos^{-1}Q_{A,i}\tag{12}$$

注意：$0\leqslant\theta_{A,i}\leqslant 90°$。

4. 标绘等 Q 图

在站位图上标注各站位的 $Q_{A,i}$ 值。将 $Q_{A,i}$ 值相同的站位联成线，依 Q 值急剧变化的位置对所研究的海域进行区划。

5. 作 A-B 相关图

首先可以作 A 相关图。取直角坐标系 xoy，作单位圆并取第一象限（图 2）。

规定以 ox 轴为量取 $\theta_{A,i}$ 角的起始轴，于是向量 $\mathbf{A}$ 即位于 x 轴上。其余的每个样品，分别用圆周上的点来代表，点的位置是：

$$\angle S_iOA=\theta_{A,i}\tag{13}$$

如此，在圆周上共有 N 个点。离 A 最远的一个记为 B，它代表本区中与 A 最不相关的那个标本，即

$$Q_{A,B}=\mathrm{Min}\,Q_{A,i},\quad i=1,2,\cdots\cdots N。\tag{14}$$

于是又获得另一端元样品 B。

计算诸 $Q_{B,i}$，$i=1,2,3,\cdots\cdots N$。

于是又可标注等 $Q_{B,i}$ 站位图，并可作如下的 A-B 相关图（图 3）。

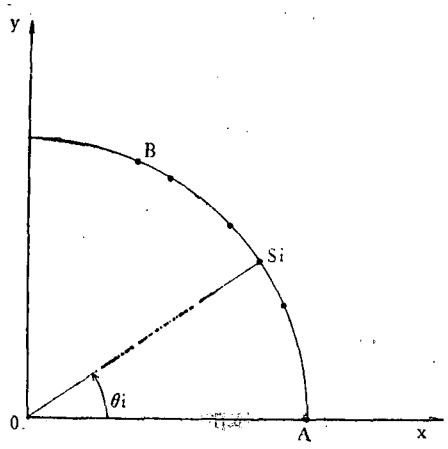

图 2　A 相关图

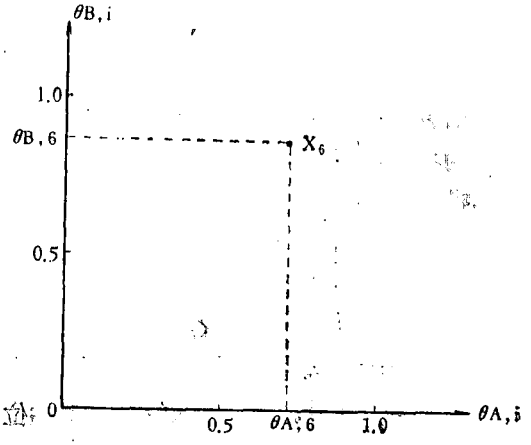

图 3　A-B 相关图

1）产生 A 向量的数学方法，将另文介绍。

每一个样品都依其与 A、B 两个端元样品之间的 Q 值，即依 $Q_{A,i}$ 值和 $Q_{B,i}$ 值用 A-B 相关图上的一个点来表示。如图 3 中点 X_6 代表第 6 号样品。

这样的点可称为相点（phase point）。

一般来讲，有了上述图件，分群划区的工作就可以进行了。

三、 例

实际的应用将另文介绍。这里举一个例子（取自文献[1]第六章），用以说明 A-B 相关图的有效性。

采自某地区的 27 块白云岩标本需要分类。由于标本中物质颗粒细小，进行了化学分析。共分析了十三种成分，数据见表 1。用因子分析法 Q 型（即主因素分析法）对上述资料处理的最后结果见表 2、表 3 及图 4。这 27 块样品被分成三类，样品 1—15 为一类，样

表 1 各标本化学分析原始资料

站号 \ 含量成分 \ 样品号	1	2	3	4	5	6	7	8	9	10	11	12	13
	SiO_2	Fe_2O_3	FeO	Al_2O_3	TiO_2	MnO	CaO	MgO	P_2O_5	$S_{全}$	V_2O_5	$C_{有}$	X
H 119	1	39.81	0.29	2.69	10.13	0.64	0.09	11.02	7.77	0.18	1.44	0.008	0.06	0.033
117	2	37.36	0.42	3.42	9.58	0.56	0.12	12.76	8.69	0.16	2.03	0.009	0.09	0.020
116	3	35.28	0.60	4.12	10.33	0.56	0.12	12.70	8.83	0.24	4.43	0.007	0.18	0.008
121	4	36.95	0.08	2.52	7.67	0.16	0.18	19.34	7.78	0.20	0.63	0.005	0.10	0.006
120	5	31.26	0.11	3.30	7.23	0.40	0.13	15.78	10.69	0.16	1.67	0.009	0.32	0.015
101	6	31.00	0.18	2.61	7.45	0.44	0.08	14.73	10.81	0.16	1.45	0.009	0.54	0.005
126	7	30.55	0.37	2.41	5.35	0.24	0.11	22.08	7.00	0.04	0.12	0.005	0.54	0.005
209	8	42.09	0.75	7.69	5.28	0.32	0.12	12.63	5.67	0.04	3.63	0.008	0.32	0.018
69	9	40.16	0.39	2.24	8.76	0.48	0.16	12.38	9.10	0.16	0.54	0.012	0.41	0.012
293	10	33.07	0.42	5.65	10.17	1.44	0.28	11.26	8.57	0.20	4.57	0.021	0.06	0.096
143	11	36.83	0.60	1.90	8.60	0.44	0.12	13.74	9.12	0.24	0.93	0.011	0.18	-0.578
145	12	29.59	0.10	3.67	6.35	0.24	0.16	16.50	10.74	0.12	1.48	0.028	0.06	0.298
68	13	40.93	0.57	3.17	10.78	0.72	0.12	10.83	7.55	0.16	1.91	0.000	0.12	0.625
131	14	40.23	0.44	4.24	8.67	0.56	0.12	11.14	7.79	0.20	2.19	0.012	0.18	0.139
115	15	37.94	0.49	5.82	8.60	0.48	0.08	11.66	7.37	0.16	4.05	0.007	0.08	0.056
118	16	23.72	0.00	2.61	4.79	0.28	0.12	20.72	13.08	0.18	0.77	0.009	0.06	0.005
122	17	22.58	0.02	3.17	1.95	0.16	0.22	23.56	12.29	0.10	1.06	0.004	0.78	0.004
154	18	16.50	0.00	4.74	1.42	0.08	0.00	26.22	12.63	0.22	1.40	0.004	0.05	0.008
109	19	6.50	0.00	2.35	1.08	0.04	0.72	31.84	14.73	0.08	0.19	0.003	0.08	0.003
237	20	27.70	0.21	1.28	0.86	0.00	0.48	21.54	13.46	0.28	0.08	0.006	0.10	0.017
144	21	24.47	0.07	2.20	3.29	0.08	0.24	23.92	10.95	0.10	0.47	0.014	0.39	0.014
99	22	23.04	0.08	2.11	1.73	0.08	0.14	22.69	15.36	0.04	0.54	0.003	0.59	0.004
141	23	21.77	0.00	1.96	2.45	0.16	0.24	23.20	14.04	0.08	0.35	0.001	7.08	0.018
147	24	18.30	0.00	2.65	3.59	0.12	0.20	22.91	14.52	0.12	0.75	0.030	0.59	0.005
183	25	16.38	0.52	1.62	1.11	0.00	0.40	28.28	12.69	0.04	0.19	0.010	0.05	0.379
280	26	8.10	0.00	3.15	1.19	0.08	0.80	27.54	15.74	0.16	1.45	0.008	0.03	0.144
187	27	5.30	1.37	1.18	1.69	0.11	0.638	31.11	12.78	0.10	0.832	痕量	0.014	0.554

表 2　前三个特征值

主　因　素	特　征　值	特征值累计	累计百分比%
1	23.927	23.927	88.8
2	2.869	26.796	99.4
3	0.069	26.865	99.6

表 3　各标本的权系数

样 品 号	λ_1 的权系数	λ_2 的权系数	样 品 号	λ_1 的权系数	λ_2 的权系数
1	0.950	0.307	15	0.940	0.331
2	0.930	0.366	16	0.705	0.706
3	0.921	0.377	17	0.642	0.764
4	0.870	0.483	18	0.492	0.866
5	0.862	0.504	19	0.185	0.980
6	0.871	0.487	20	0.721	0.679
7	0.963	0.376	21	0.678	0.732
8	0.938	0.316	22	0.640	0.761
9	0.936	0.345	23	0.613	0.766
10	0.923	0.363	24	0.571	0.817
11	0.916	0.395	25	0.458	0.886
12	0.840	0.540	26	0.257	0.962
13	0.955	0.292	27	0.160	0.980
14	0.950	0.311			

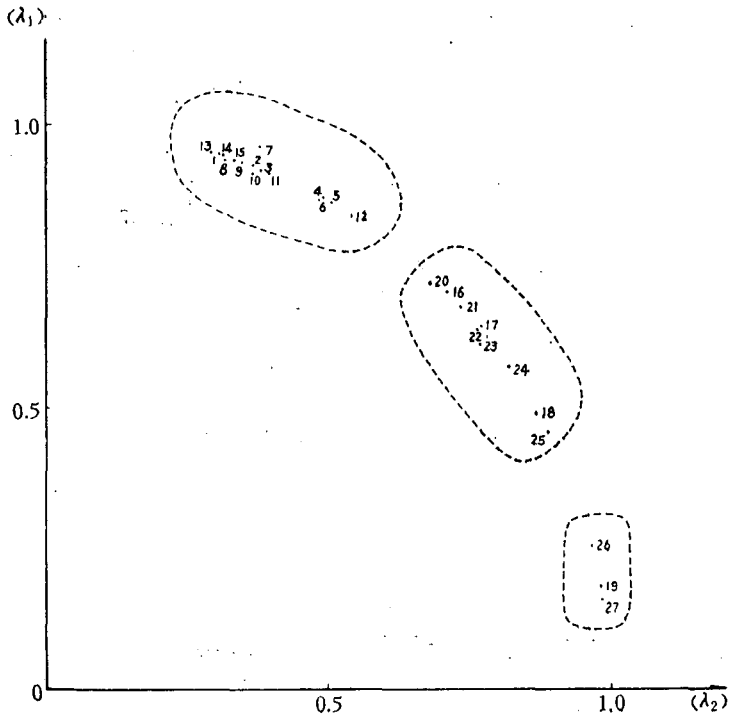

图 4　样品依权系数分类图

品 19、26 和 27 为一类，其余九块样品为第三类(过渡类型)。

现以本文所述方法进行分类。

即以样品 27 为标准样品，于是可算出诸 $Q_{27,i}$ 之值，列入表 4。依此又作出图 5。

图 5 表明，样品 13 为另一端元；计算诸 $Q_{13,i}$ 之值，也列入表 4。依表 4 点绘 Q 因子分类图，即 A-B 相关图(图 6)。

表 4 各标本的 Q 因子

样品号 i	$Q_{27,i}$	$Q_{13,i}$	样品号 i	$Q_{27,i}$	$Q_{13,i}$
1	0.456	0.999	15	0.478	0.993
2	0.511	0.997	16	0.729	0.799
3	0.524	0.990	17	0.849	0.835
4	0.619	0.974	18	0.929	0.721
5	0.633	0.972	19	0.997	0.466
6	0.617	0.975	20	0.772	0.885
7	0.717	0.932	21	0.828	0.862
8	0.458	0.983	22	0.840	0.828
9	0.484	0.988	23	0.841	0.809
10	0.510	0.986	24	0.891	0.785
11	0.537	0.993	25	0.945	0.698
12	0.664	0.962	26	0.985	0.527
13	0.444	1.000	27	1.000	0.444
14	0.458	0.998			

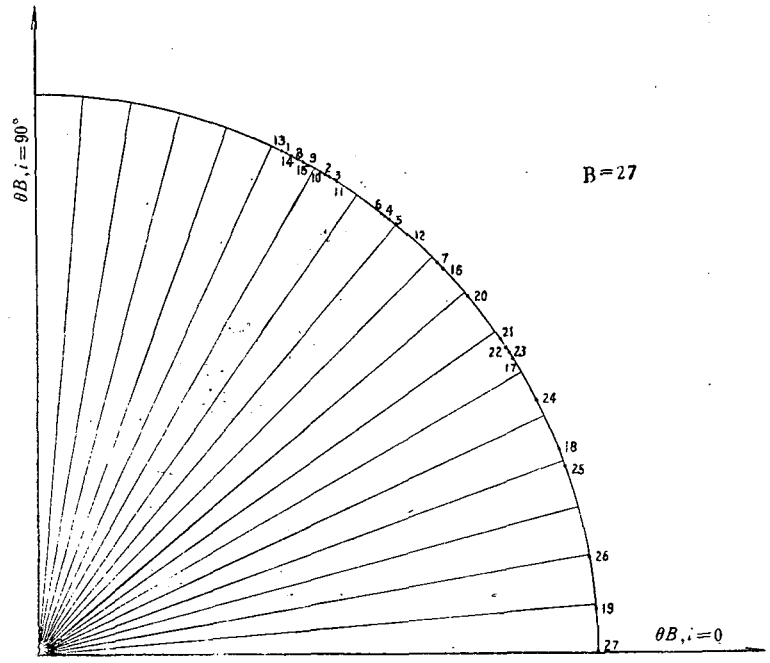

图 5 相关角分布图

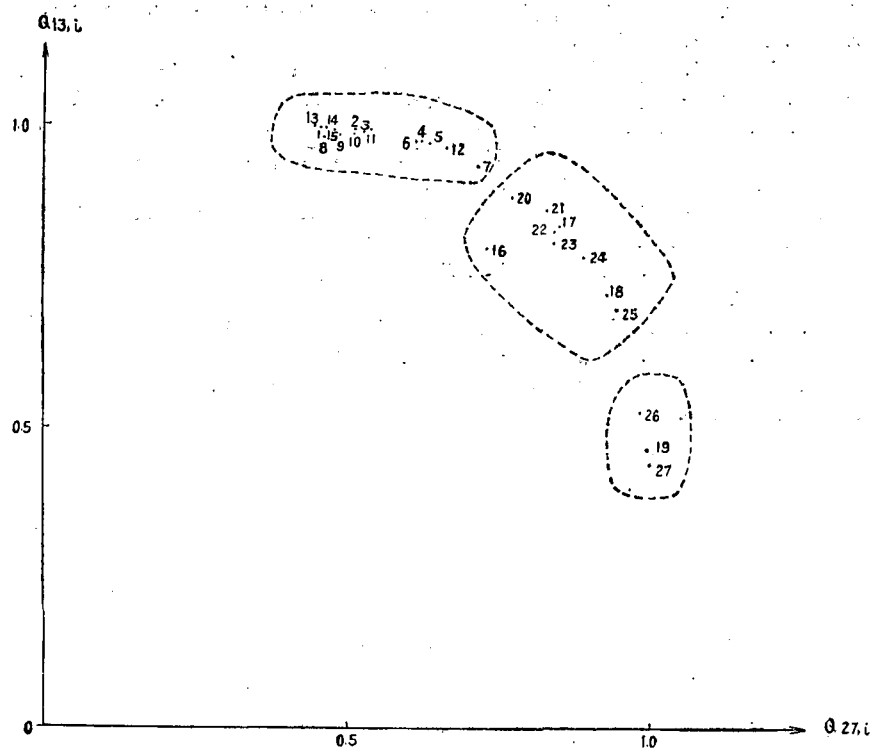

图 4　A-B 相关图

很清楚，图 6 给出的分类结果是与图 4 一致的。

应当指出，在本例中，标准样品 A 端元（即 27 号样品）是由主因素分析法处理的结果提供的。在实际应用中，应依所研究海区的具体特征来选定 A 样品。

本方法的优点是仅计及 $2N$ 个 Q 值，比需要至少要计及 $\frac{1}{2}N(N-1)$ 个 Q 值的枝状图法来说是简捷多了；比起主因素分析法来说，就更为简捷。

下表也许可以说明这点。

表 5　计算量对比

样 品 数 N	A-B 相 关 法	枝 状 图 法
50	100	1225
100	200	4950
200	400	19900

至于本方法的有效性，有待于今后的实践来检验。

参　考　文　献

[1] 中国科学院地质研究所，1977。数学地质引论。地质出版社。第四章，第六章。

[2] McCammon, R., 1968. The dendrograph: a new tool for correlation. *Geol. Soc. Amer. Bull.* 79(11):1663—1670.

[3] Imbrie, J. and T. H. Van Andel, 1964. Vector analysis of heavy mineral data. *Geol. Soc. Amer. Bull.* 75(11):1131—1156.
[4] Klovan, J. E., 1975. R- and Q-mode factor analysis, in: Concepts in geostatistics, ed. by R. McCammon. Springer-Verlag, Berlin-Heidelberg-New York. Chapter 2 pp. 21—67.
[5] Stephenson, W. and W. T. Williams, 1971. A study of the benthos of soft bottoms, Sek Harbour, New Guinea, using numerical analysis. *Aust. J. mar. Freshwat. Res.* 22(1):11—34.
[6] Hughes, R. N. and M. L. H. Thomas, 1971. The classification and ordination of shallow-water benthic samples from Prince Edward Island, Canada. *J. exp. mar. Biol. Ecol.* 7(1):1—39.
[7] Johnson, K. R. and A. D. Albani, 1973. Biotopes of recent benthonic foraminifera in Pitt Water, Broken Bay, N. S. W. (Australia). *Palaeogeography, Palaeoclimatology, Palaeoecology* 14, pp. 265—276.
[8] Albani, A. D. and K. R. Johnson, 1976. Resolution of firaminiferal biotopes in Broken Bay, N. S. W. *Jour. Geol. Soc. Aust.* 22(4):435—446.
[9] Albani, A. D., 1974. Sedimentary environments in Broken Bay, N. S. W. *Jour. Geol. Soc. Aust.* 21(3):279—290.

THE A-B CORRELATION METHOD FOR SAMPLE CLASSIFICATION BY USING THE Q FACTOR*

Fang Shouzhi

(*Institute of Oceanology, Academia Sinica*)

ABSTRACT

In this paper, a simpler statistical method for sample classification is given. In this method each of the N samples is compared with the end-member samples A and B. The selection of the end-member samples from the N samples is based on a consideration of some specific features of the area investigated or is based on a formula which will be presented in a forthcoming paper. The classification of N samples may be done on the A-B relation diagram after the values of all the factors $Q_{A,i}$ and $Q_{B,i}$ ($i = 1, 2, \ldots, N$) are calculated.

* Contribution No. 469 from the Institute of Oceanology, Academia Sinica.