第11卷第4期　　石　油　实　验　地　质　　Vol.11, No.4
1989年12月　　EXPERIMENTAL PETROLEUM GEOLOGY　　Dec., 1989

化石孢粉的有序集群

童国榜　　林锦璇　　陈祖荫

（地质矿产部水文地质与工程地质研究所）　（北京工业大学）

本文使用三种有序集群方法（最优分割法、有序点群分析法、有序图论方法）对化石孢粉的分带问题进行处理。实际数据选自河北肃开10孔等处。处理结果表明三种方法均能有效地模拟古植被的演变，但又各有其特色。

一、引　　言

随着地质工作的发展，化石孢粉分析的应用日益广泛，工作内容也更为丰富。对于大量的化石孢粉信息，数量分析是一种重要的研究方法。

近年来，我们对黄淮海地区的井下第四纪孢粉资料使用多种数量方法进行了研究。所用的方法包括点群分析（中国科学院地质研究所，1977；童国榜等，1988）、因子分析(王开发等，1982；I.C.Rrentlce，1980)、对应分析（童国榜等，1983；A.D.Gordon，1982）、时间序列分析（D.G.Green，1983）等等。本文以河北省肃宁县肃开10孔的孢粉资料为例，报告三种有序集群方法的原理及其地质效果。

所谓有序集群（聚类）是指符合以下要求的集群方法：对于一批样品 X_1，X_2，……，X_n，经集群后所得的每个类必须采取以下形式

$$\{X_i, X_{i+1}, \cdots, X_j\}, \quad 1 \leqslant i \leqslant j \leqslant n,$$

即同一类中各样品的下标必须彼此邻接。显然，这一方法对于井下孢粉分析是有用的。本文中所报告的方法除传统的最优分割法外，还有有序点群分析方法和有序的图论集群方法。

二、地层概述及孢粉类型的选择

肃开10孔位于河北平原西部偏北。该区第四纪厚度大，地层发育完整，为一套松散的粉、细砂与粘性土互层。上更新统岩性较细，以浅黄、黄灰色粉砂质亚砂土为主，夹有粉、细砂及淤泥质粘土层。下更新统岩性较粗，以粉、细砂为主，夹浅棕色粘土质粉砂或砂质粘土。

古地磁极性：布容正极性世位于140m以上，松山负极性世位于140—450m，再下是高斯正极性世，详见图一。

该孔孢粉化石丰富。全井共鉴定孢粉3821粒，分属96个科、属，其中绝大部分是华北地区的现生植物类型（该孔孢粉由柯曼红鉴定）。

依据各类孢粉出现的数量多少、相互关系及生态习性，我们选取了18个孢粉类型来反映孢粉谱的变化。被选孢粉类型为松属(Pinus)、云杉属(Picea)+冷杉属（Abies)、桦属(Betula)、栎属(Quercus)、榆科（Ulmus）、麻黄属（Epherlra）、蒿属（Artemisia）、藜科（Chenopodiacacu）、禾本科（Gramineae）、榛属（Corylus）、香蒲属（Typha）+眼子菜属（Potamogeton）、大戟科（Euphobiacrae）、莎草科（Cyperacone）、毛茛科（Ranunculacoae）、蓼科（Polygo）、石松属（Lycopodium）+卷柏属（Selaginella）、水龙骨科（Poıypoɑıaceae）、水蕨（Ceratopteris）。

三、有序集群方法

1.最优分割法（Fisher）（福永圭之介，1978；A.D.Gordon，1982；S.Björck、P.Möller，1987）

本方法是广为使用的方法。定义类{X_i，X_{i+1}，……，X_j}的直径为

$$d(i,j)=\sum_{l=i}^{j}(X_l-\bar{X}_{ij})^T(X_l-\bar{X}_{ij}),$$

其中每个样品X_l视为一个列向量，而类均值

$$\bar{X}_{ij}=\sum_{l=i}^{j}X_l/(j-i+1).$$

集群原则是使误差函数e〔p(n，g)〕达到最小，其中n为样品总数，g为类数。有递推公式：

$$e〔p(n,g)〕=\min_{g\leqslant j\leqslant n}\{e〔p(j-1,g-1)〕+d(j,n)\}.$$

当类数g确定时，本方法可以求得（在离差平方和的意义下的）全局最优解。在类数未确定时，可以选取多个g值并根据较稳定的分界线确定分类方案。

2.有序点群分析（Björck.S.，Möller，P.，1987；D.G.Green，1983）

对常用的点群分析〔系统聚类（Björck.S.,Möller.P，1987）〕加以以下的限制：每次合并时，被合并的两样品序号必须相邻，便得到有序点群分析方法。

算法的基本原则是每次合并最近的两类并遵循有序原则。点间距离采用欧氏距离。类间距离有以下几种定义方式：

（1）最短距离　对于类w_i和w_{i+1}，定义

$$d(w_i,w_{i+1})=\min_{\substack{X_t\in w_i\\X_m\in w_{i+1}}}d(X_t,X_m),$$

后一个d表示点（样品）间的欧氏距离，下同。

（2）最长距离　定义

$$d(w_i,w_{i+1})=\max_{\substack{X_t\in w_i\\X_m\in w_{i+1}}}d(X_t,X_m).$$

（3）平均距离　定义$d^2(w_i,w_{i+1})$为两类中各样品点两两距离平方的平均

数，即

$$d^2(w_i, w_{i+1}) = \frac{1}{n_i n_{i+1}} \sum_{\substack{X_t \in w_i \\ X_m \in w_{i+1}}} d^2(X_t, X_m),$$

其中n_i，n_{i+1}分别为w_i，w_{i+1}中样品个数。

（4）离差平方和增量　假设将w_i，w_{i+1}合并得到新的类w_j。w_i的离差平方和定义为

$$S_i = \sum_{X_t \in w_i} (X_t - \bar{X}_i)^T (X_t - \bar{X}_i),$$

$\bar{X}_i$是w_i的均值。类似地可以定义S_{i+1}与S_j。定义作上述合并后的离差平方和增量为

$$d(w_i, w_{i+1}) = S_j - S_i - S_{i+1},$$

它可以作为一种距离量度。

（5）重心距离　定义

$$d(w_i, w_{i+1}) = d(\bar{X}_i, \bar{X}_{i+1}).$$

（6）间隙距离　定义

$$d(w_i, w_{i+1}) = d(X_t, X_{t+1}), \quad X_t \in w_i, \ X_{t+1} \in w_{i+1}$$

即恰好分别处于两类中的两个相邻点的距离。

3.有序的图论集群方法

Koontz（D.G.Green, 1983）提出了利用密度和树的图论集群方法。我们在此基础上提出的有序集群方法如下。

确定参数R（o）和步长h。在算法的第k步，以每个样品点X_i为中心，R（k）为半径作超球，称为X_i的R（k）邻域。计算邻域中相邻样品点序列X_a，X_{a+1}，……，X_i，……，X_b（$d(X_l, X_i) \leqslant R(k)$，$l = a, a+1, \cdots\cdots, b$；$d(X_{a-1}, X_i) > R(k)$，$d(X_{b+1}, X_i) > R(k)$）的点数$N_i$（不包括$X_i$自身），并令

$$g_{ij} = (N_j - N_i)/d_{ij},$$

其中d_{ij}表示X_i，X_j的距离。然后对每个X_i做以下判断：

（1）若$N_i = 0$，则X_i是一有向树的根（孤立点）。

（2）若$N_i > 0$，则求X_i两个邻点中使g_{ij}达到最大的一个，记作X_l，即

$$g_{il} = \max_{j = i-1, i+1} g_{ij}.$$

①若$g_{il} < 0$，则X_i是根（密度最大的点）。

②若$g_{il} > 0$，则X_i的上一节点（“父亲”）是X_l。

③若$g_{il} = 0$，则考察X_{i-1}，X_{i+1}中使$g_{ij} = 0$而不是X_i的下一节点（“儿子”）的所有点。记这些点的集合为π_i。若$\pi_i = \phi$（空集），则X_i是根；若$\pi_i \neq \phi$，则X_i的上一节点为X_p：

$$d_{ip} = \min_{X_j \in \pi_i} d_{ij},$$

即π_i中到X_i最近的点。

由此可以得到一批有向树，每个树对应于一类。树根是密度最大的点，以下各节点的密度依次减少。当全体样品点已并为一类时计算停止，否则，令R(k+1)=R(k)+h，重新分类。

四、结果分析

三种方法的结果均如图1所示。各种算法的结果比较一致，它们都将全孔样品分成8—10个孢粉组成的类群，用阿拉伯数字在图上标出。

1.最优分割法

当类数超过4以后，分界线较为稳定。最终分出8个孢粉类群，由上至下的特征如下：

(1)榆属(Ulmus)-桦属(Betula)-蒿属(Artemisia)。

(2)松属(Pinus)-榆属(Ulmus)。

(3)松属(Pinus)-藜科(Chenopodiaceae)-蒿属(Artemisia)。

(4)松属-榆属-禾本科(Gramineae)。

(5)松属-藜科-禾本科。

(6)云杉属(Picea)-松属。

(7)桦属-藜科。

(8)榆属-松属-云杉属。

全新世对应于第1，2类，晚更新世对应于第3类，中更新世对应于第4类，早更新世对应于第5，6类，上更新世则对应于7，8类。将其与地质解释结果比较，误识样品共7个，误识率$\hat{\varepsilon}$之值为0.1428。

2.有序点群分析

在使用不同的类间距离时，分类效果有所差别。综合黄淮海地区七个钻孔的计算结果表明，使用重心距离、离差平方和增量及最长距离时效果较好，间隙距离及最短距离效果一般尚可，平均距离则效果较差，如图1所示。

使用各种不同的类间距离时一般也可将全孔划分为8个孢粉类群。一般说来，不同方法所得的各序列间的重心位置可能有较大的差异，这一特点从孢粉样品计算结果也可看到(W.D.Fisher, 1958)。对于肃开10孔，以重心距离法和离差平方和增量法所得的分段结果最佳。与最优分割法的结果相比，主要差别在于本方法把上一方法结果中第五类的部分样品划归到第四类。这一事实与地质综合解释的分层界线更为一致。误识样品仅有3个，误识率$\hat{\varepsilon}=0.0816$。

3.有序的图论集群法

步长超过10以后，出现了九条较稳定的分类界线，将全孔分为10个孢粉类群。由图

1可见，分类结果与前两种方法仍然大体相似，只是对原有的两个类群划分较细，即原第四类（Pinus-Ulmus-Gramineae）的下部又分出一类（Ulmus-Gramineae-Typha），原第五类（Pinus-Chenopodiaceae-Gramineae）的顶部又分出一类（Pinus-Ulmus-Chenopodiaceae，木本树种增多）。

本方法是一种基于寻峰（凝聚点）的方法。使用这一方法，在孢粉组合变化不很明显的井段仍能区分出小的峰值，并提供细分的依据。但是，应注意在步长较小时异常样品点的影响。

综合全区共四个钻孔的计算结果，本区地层共可划分为五个孢粉带，含十个孢粉亚带。分带结果见图1的最右一列。带Ⅰ为针阔混交林草原植被类型，早期针叶树种含量较高，晚期阔叶树种突出，属全新世早、中期。带Ⅱ为针叶林草原或草原植被类型，属于晚更新世。带Ⅲ为针阔混交林植被类型，早期阔叶、落叶的榆、栎树种频繁出现，晚期松属占优势，且出现云杉树种，属中更新世。带Ⅳ为针阔混交林草原植被类型，早期针叶树种占优势，云杉大量出现；中期以阔叶落叶树种榆、桦相对丰富；晚期针叶树种复又增多，且出现云杉。本期属早更新世。带Ⅴ为针阔混交林草原或疏树草原植被类型，早期以松、桦为主，并出现云杉，晚期草木植物显著增加，属上新世晚期。由图1可见，前述三种有序聚类方法都能较好地揭示古植被演变的过程。

上述结果与古地磁也具有可比性。带Ⅳ的底位于松山反向世的底界附近，一般相差1—8个样品。带Ⅲ的底位于布容正极性世的底界附近，一般相差1—5个样品。考虑到时间地层学的资料，带Ⅲ底部为中更新世，距今73万年。带Ⅳ底为早更新世，距今248万年。

对四口钻孔的集群误识率在表1中列出。表中A，B，C分别表示最优分割法、有序点群分析法和有序图论方法，第i带（i=Ⅰ，Ⅱ，……，Ⅴ）的误识情况用一个分数标明，分母为该带样品总数，分子为该带样品中被错分到其他带的个数。

表1　各孔样品误识情况

孔号		肃开10孔			XK 63孔			沧13孔			玉11孔		
算法		A	B	C	A	B	C	A	B	C	A	B	C
各带误识情况	Ⅰ	1/3	0/3	0/3	0/2	0/2	0/2	0/5	0/5	0/5	1/6	1/6	1/6
	Ⅱ	0/5	1/5	1/5	3/6	0/6	0/6	0/3	0/3	0/3	0/3	0/3	0/3
	Ⅲ	5/9	1/9	1/9	2/7	0/7	2/7	1/4	1/4	1/4	4/14	0/14	0/14
	Ⅴ	1/20	1/20	2/20	1/9	0/9	0/9	8/28	8/28	6/28	1/12	3/12	3/12
	Ⅳ	0/12	0/12	0/12	0/4	0/4	0/4	4/27	4/27	0/27	0/5	1/5	1/5
误识样品总数		7	3	4	6	0	2	13	15	7	6	5	5
误识率		0.143	0.082	0.082	0.214	0.000	0.071	0.209	0.224	0.105	0.105	0.125	0.125

五、结　　论

1.三种有序集群方法均能有效地模拟古植被的演变。

2.最优分割法和有序点群分析利于反映较长周期的变化现象，其中有序点群分析的类间距离除离差平方和增量（它与最优分割的目标函数有着内在联系）外，重心距离和最长距离效果也较好。但是，在其他一些试验中，我们利用最短距离也曾得到过很好的效果。

3.有序的图论方法利于刻划周期较短的变化现象。但使用本方法时应注意排除步长较短时出现的假象。

4.Ⅳ，Ⅴ两孢粉带间的类群界线稳定性最好。这表明该界线上、下孢粉组合特征差别最为显著。我们建议以该界线作为第四纪与晚第三纪的分界。这一建议与以松山底界（距今248万年）为Q/N界线的意见一致。

总之，对孢粉资料进行有序集群有利于恢复古植被、讨论古气候、古环境的变迁，以及论证生物地层的划分与对比。

（收稿日期：1986年10月23日）

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。

〔2〕王开发等，1982，东海沉积物中孢粉组合的因子分析，海洋学报。

〔3〕王碧泉等，1984，用聚类分析法研究强震的孕震过程，地震学报。

〔4〕王碧泉等，1986，研究强震孕震过程的几种有序集群方法，地震学报。

〔5〕方开泰，1982，有序样品的一些聚类方法，应用数学学报。

〔6〕童国榜等，1983，河北平原第四纪孢粉组合及其地质意义，海洋地质与第四纪地质。

〔7〕童国榜等，1988，华北平原第四纪孢粉的数学地质分析，植物学报。

〔8〕福永圭之介，1978，统计图形识别导论，陶笃纯译，科学出版社。

〔9〕A. D. Gordon, 1982, Numarical methods in guatarnary Palaeoecology, V. Simultaneous graphical representation of the levels and taxa in a pollen diagram, Review of Palaeobotany and Palynology.

〔10〕Björck, S., Möller, P., 1987, Late weichselian environmental history in southeastern Sweden during the deglaciation of the scandinavian ice sheet.

〔11〕D.G.Green, 1983, Interactive pollen time series analysis, Pollen et Spores.

〔12〕I.C.Prentice, 1980, Multidimensional scaling as a research tool in quaternary palynology, A review of theory and methods, Review of Palaeobotany and Palynology.

〔13〕W.L.G.Koontz,1976, A graph-theoretic approach to nonparmetric cluster analysis,IEEE Trans. Computers.

〔14〕W.D.Fisher, 1958, On grouping for maximum homoseneity, J.Am. statist.Assoc.

ORDERED COLONY OF FOSSIL SPORE—POLLEN

Tong Guobang **Lin Jinxuan**

(Research Institute of Hydrogeology and Engineering Geology, MGMR)

Chen Zuyin

(Beijing Industry University)

Abstract

Using three kinds of ordered colonizing methods (optimum decollation, ordered point group, and ordered graph theory), the authors dealt with the zoning problem for spore-pollen. Acurate data were taken from Sukai NO. 10 well in Hebei. The results show that the three methods are efficient in moldling the evolution of palaeo-vegatation, but each has its peculiarities.

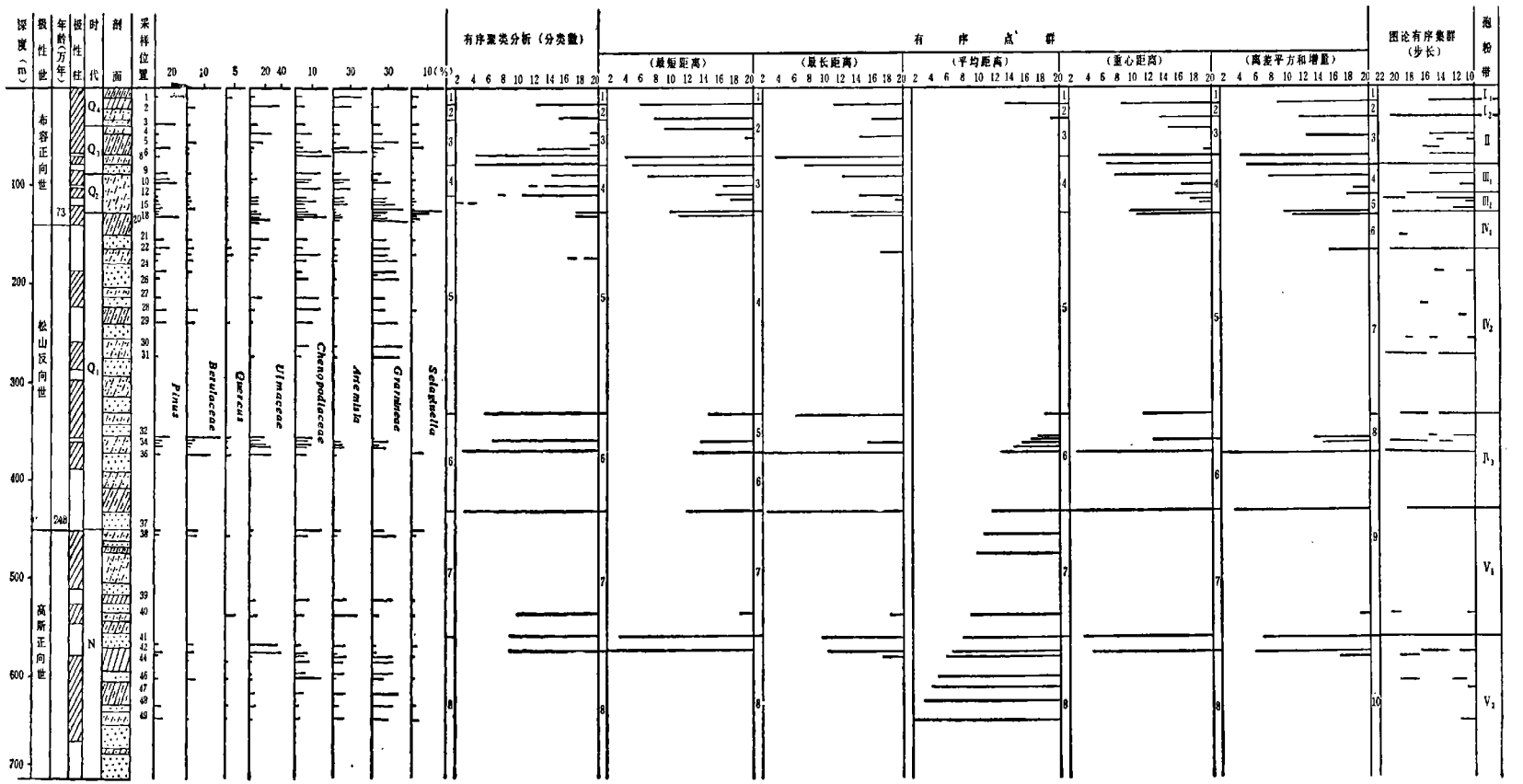

图1 肃开10孔地质性质及集群结果图