应用红外光谱数据对生油岩评价的尝试

李 鸿 生　曹 慧 缇

（地质部石油地质中心实验室）

红外光谱在石油地质的研究领域中已被广泛应用。E.E.布雷的“芳烃结构分布指数”提出鉴别生油岩的指标，被认为可以比较直接评价生油条件。本文运用数学方法——多维数据处理方法，籍助于电子计算机对红外资料进行系统分析。试图对生油岩的判别和分类进行对比和探讨。我们选择了湘中地区海相碳酸盐岩样品的芳烃红外图谱17个峰值，进行系统聚类分析、因子分析和对应分析。

一、聚类分析

聚类分析就是研究变量之间所存在的相似性以及按其相似程度进行分类的方法，即根据样品的观测值（变量）具体计算样品之间的相似程度，把相近的样品归为同一类别，直到把所有的样品都归到各自类别为止，形成一个由小到大的分类单位，最后把整个分类系统组成一个分类图式以表示所有样品的亲疏关系（1）。对样品进行分类，首先要确定划分类型的数量标准，即能够表示样品间相似程度的统计量，使之有一个度量标准来进行聚类。这里把样品看成M维空间中一个向量。两个样品之间相似程度可用几种方法度量。

（1）欧氏距离：两个样品之间距离为欧氏距离：

$$D_{K1}=\sqrt{\sum_{i=1}^{M}(X_{iK}-X_{i1})^2}$$

k，1为二个样品号。

Xik，Ki1为第K个和第1个样品第i个变量观测值。

为了使Dk1在一定范围内变化，一般用：

$$D_{K1}=\sqrt{\frac{1}{P}\sum_{i=1}^{M}(X_{iK}-X_{i1})^2}$$

其中P为一常数，D_{K1}表示样品X_K和X_1的距离系数，Dk1值越小则二个样品相似程度越大，反之就小。把计算出的D值列成一个N×N的距离系数矩阵。

（2）相似系数：对于任意两个样品X_1和X_K的相似程度可用两个向量夹角余弦（相似系数）Cosθ1k表示，如图1所示。

$$X_K=(X_{1K},\ X_{2K},\ \cdots\cdots X_{MK})'$$

$$X_1=(X_11,\ X_21,\ \cdots\cdots X_{M1})'$$

$$\cos\theta_{1k}=\frac{\sum_{i=1}^{M}X_{i1}\cdot X_{iK}}{\left(\sum_{i=1}^{M}X^2_{iK}\cdot\sum_{i=1}^{M}X^2_{i1}\right)^{\frac{1}{2}}}$$

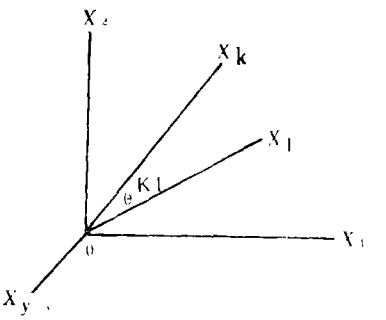

图1　两个样品的相似系数

它表示二个样品向量间夹角余弦 $\cos\theta k1$ 越大（θ 角越小），两个样品性质就越相近。

（3）相关系数：第K个样品与第1个样品相关系数定义为：

$$P_{k1}=\sum_{i=1}^{M}(Xik-\bar{X}k)(Xil-\bar{X}l)\Big/\Big[\sum_{i=1}^{M}(Xik-\bar{X}k)^2\sum_{i=1}^{M}(Xil-\bar{x}l)^2\Big]^{\frac{1}{2}}$$

其中：$\bar{X}k=\frac{1}{M}\sum_{i=1}^{M}Xik$，$\bar{}l=\frac{1}{M}\sum_{i=1}^{M}Xil$

并 P_{Ki} 满足不等式 $-1\leqslant P_{KI}\leqslant+1$。

求得的相似系数和相关系数都可列成N×N的系数距阵，它是一个对称且对角线等于1的矩阵。所以实际上进行聚类的计算处理时只要不包括对角线在内的上三角矩阵（或下三角矩阵）即行(注)。

例如相关系数距阵可表示为：

$$R^{(1)}=\begin{pmatrix} & R_{12}^{(1)} & R_{13}^{(1)}\cdots\cdots R_{1n}^{(1)} \\ & & R_{23}^{(1)}\cdots\cdots R_{2n}^{(1)} \\ & & \cdots\cdots R_{n-1,n}^{(1)} \end{pmatrix}$$

右上角(1)表示为第一次计算的相关系数的标号。当求出样品之间的一种距离系数后就可进行聚类。在距离系数距中阵选取最大值（如相关系数），记下该二个样品的标号，这二个样品就首先组成一类。随后用二个样号中较小号码代表组成这一类（我们视为新样品）的标号，并且求出这个“新样品”的各个变量的数值。

假如是第K和第1号样品首先组成一类，并且K＜1，则把新“样品标号”定为K，且 $x_{iK}^{*}=(x_{1K},x_{2K},\cdots\cdots,x_{MK}^{*})$ 表示

$$x_{iK}^{*}=\frac{x_{1K}^{(1)}+x_{il}^{(1)}}{2}\qquad(i=1,2,\cdots,M)$$

$x_{iK}^{(1)}$ 和 $x_{il}^{(1)}$ 为老样品的各变量值。

x_{iK}^{*} 为新样品的各变量值。

我们选用逐步划分法继续进行分类。将求出的“新样品”和原有样品的相关系数，除去已求得一类的大号老样品，这样组成第二次相关系数矩阵，然后再按照上述办法反复进行计算，进行N—1次，得出样品聚类的结果。

我们对湘中地区40块样品的红外数据，进行了相似系数 $\cos\theta$ 相关系数和欧氏距离系数聚类分析。结果见表1和图2。

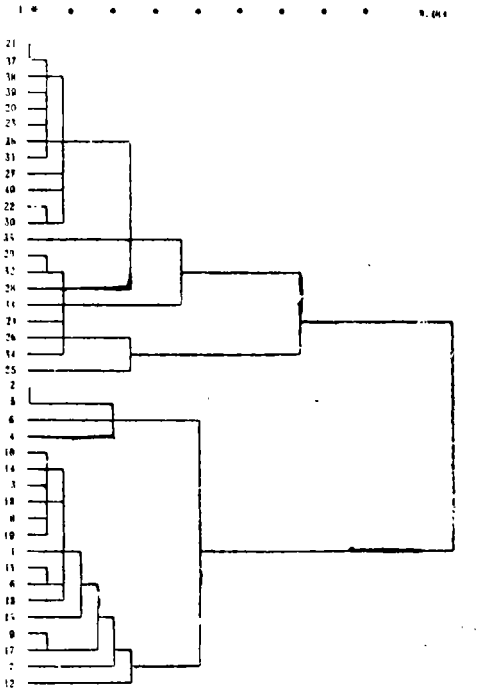

图2 Q型聚类分析—相似系数分枝图

样品：40个　　变量：17

从表(1)中看出：计算结果把样品分成三类：第一类是以生物灰岩为主并包括灰黑色灰岩，主要代表层位是D_2q和C_1y_1组，属较好生油岩，第二类是以泥晶、微晶灰岩为主并包括深灰色泥岩，主要代表层位是D3x

(注)地质科学院：地质学中多元统计方法之一、六（1977年）。

聚类分析分类结果（N＝40块样品，M＝17个峰值） 表1

类别		Ⅰ类 较好生油岩	Ⅱ类 较差生油岩	Ⅲ类较差生油岩
主要岩性特征		生物灰岩、灰黑色灰岩	微晶灰岩、泥灰岩、深灰色灰岩	炭质页岩、泥质灰岩
主要层位		C_1y_1、D_2q	D_3x、D_3s、E	C_1d_1
样号	相似系数	21、22、23、27、28、29、30、31、32、33、35、36、37、38、39、40	8、9、10、11、12、13、14、15、16、17、18、19、1、2、3、4、5、6、7、	24、25、26、34、
	相关系数	20、21、22、23、27、28、29、30、31、32、35、36、37、38、39、40	1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、33*	24、25、26、34、
	欧氏距离	20、21、22、23、25*、26*、27、28、29、30、31、32、35、36、37、38、39、40	1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、33*	24、34、

注：有*者是与相似系数对比属不同类之样品。

和D_3s组，属较差生油岩，第三类是以灰质页岩为主并有泥质灰岩，主要代表层位是C_1d组，属较差生油岩。

利用相似系数、相关系数和欧氏距离系数进行分类，从地质角度来看，相似系数效果更好一些。图2中三类的相似系数为0.80以上，各自相聚成一类，说明同一类的样品相似性是比较好的。例如第33号样品，是较好生油岩，应属第Ⅰ类，用相似系数计算正好划入第Ⅰ类，而用欧氏距离则归入第Ⅱ类；同样，第25、26号样品，是较差生油岩，应属第Ⅲ类，欧氏距离计算划入第Ⅰ类，相似系数正属第Ⅲ类。

二、因子分析

因子分析分成Q型和R型二种，前者是分析样品间的关系，后者是分析变量（指标）间的关系。

1. Q型因子分析

我们把参与聚类分析的40块样品进行了Q型因子分析（有关计算方法从略），计算出前三个主因子特征值及贡献率（表2）。

Q型因子分析选取前三个特征值贡献率 表2

主因子	（λ）特征值	累计百分数%	代表标本号
1	35.40	88.49	33，36
2	2.14	93.85	12，11
3	1.25	96.98	34，24

根据计算，前三个特征值的累计贡献率达96.98%，即所代表的方差已占总方差的97%。换句话说，这三个因子97%地表达了数据中的变差和因素及其整个数据的变化。

根据所选取的三个主因子数进一步计算得到Q型因子分析图（图3）。从图中看出样品明显地分成三大类，与聚类分析结果基本一致。

在第1个主因子上，最大因子载荷为80%和81%，代表样品是33号及36号；第2个主因子最大因子载荷是88%，代表样品12号，其次是77%的11号样品；第3个主因子最大因子载荷为88%和81%，代表样品是34号和24号，现将这几个代表样品的红外光谱数据列于表3。

三个主子的代表样品测定数值表　　表 3

测定值 / 样号 / 红外指标 cm^{-1}	第 1 主因子		第 2 主因子		第 3 主因子	
	33	36	12	11	34	24
2850	0.143	0.646	0.511	0.442	0.793	1.176
2950	0.226	1.084	0.123	0.301	0.993	1.523
1380	0.038	0.155	0.044	0.068	0.595	0.602
1460	0.061	0.358	0.021	0.141	0.771	0.963
720	0.023	0.164	0	0	0	0
1200	0	0	0	0	0.595	0.585
3050	0	0	0	0	0.641	0.664
1600	0	0	0	0.016	0.541	0.446
860—880	0	0	0	0	0.364	0.188
800—810	0	0	0	0	0.481	0.284
740—760	0	0.113	0.011	0.038	0.868	1.097
1720	0.032	0.155	0	0.042	0.595	0.523
1740	0	0	0	0.083	0	0
1660	0	0	0	0	0	0
1040—1170	0	0	0.029	0.047	0	0
1270—1300	0	0	0	0	0	0
3400—3500	0	0	0.043	0.043	0	0

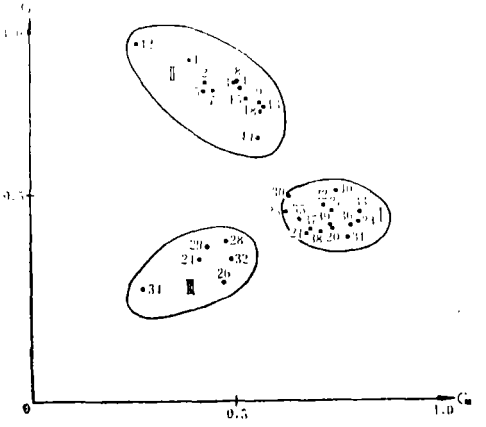

图 3　Q型因子分析图

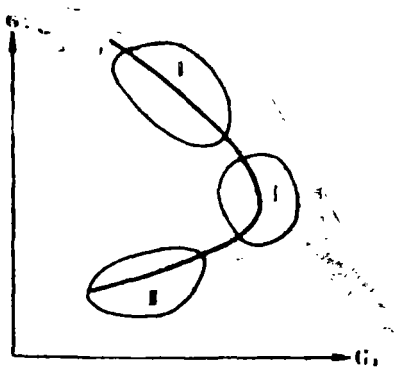

图 4　Q型因子曲线图

表3中代表性样品的数值可以比较清楚地看到：第1主因子是以甲基、亚甲基等为主要成分，含氧基团和芳核结构基本没有反应，这个特征可以认为是母质类型较好的反映；第2个主因子所反映的是以含氧基团为主的成分，甲基、次甲基烷烃基链反映略差一些，其特征反映了母质类型较差；第3主因子则以芳环、芳核结构为主要成分，可作为代表母质类型差並接近煤系的反映。

图4中可以看到这样一根曲线，在一定程度上可以代表样品成熟度趋势曲线。曲线上部为成熟度较低，逐渐形成一个抛物线至中部凸出部分便是成熟度较高的样品聚集，再后，曲线沿着G_1和G_2都降低的方向移动，便是成熟度过高样品聚集的反映。同时，也说明第Ⅱ类聚集的样品数值是以含氧基团的峰值为主要显示，G_2主因子起主要作用；第Ⅰ类则是烷烃，直链烃基团为主要显示，即G_1主因子起主要作用；第Ⅲ类朝芳环、芳核基团峰值移动，G_1和G_2二个主因子都显著降低。对于G_2主因子第Ⅱ类样品均占65%以上，而第Ⅰ类和第Ⅲ类均在60%以下；对于G_1主因子，第Ⅰ类占70%以上，而第Ⅱ、Ⅲ类均在50%以下。

2. R型因子分析：这也是从样品空间找出公因子。以确立因子的最少数目，和Q型因子分析的区别在于它是着重研究变量之间的关系。

根据特征值计算累计贡献率，得出前四个主因子已达到92.21%，把原变量精细地划分成五类（图5、表4）。

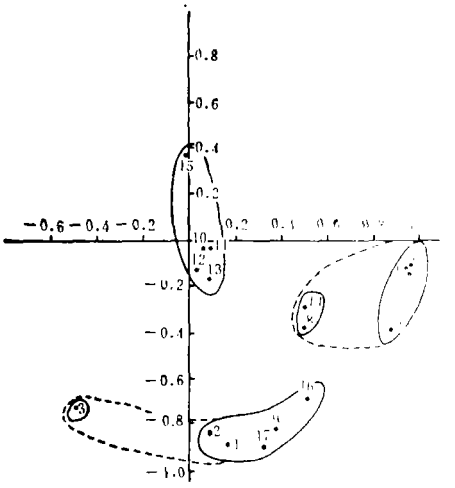

图5 R型因子分析图

红外光谱17个峰分类及相关表（2） 表4

类 别	变 量 号	变量间相关程度	基 团 峰 反 应
Ⅰ	1、2、9、16、17	79—91.5%	烷烃峰CH_3-，-CH_2-为主，CH_3-CH_2- 变形振动 伸展振动
Ⅱ	3	与(Ⅰ)相关达72.5%	-(CH_2)-$\geqslant_4$ 骨架振动
Ⅲ	10、11、12、13、15	80—98.6%	含氧基团C=O,C-O-,OH为主，伸展振动
Ⅳ	4、5、6、7	87.6—96%	芳环，芳核上-C=C-，C-H面外环振动
Ⅴ	8、14	66%—75% 与(Ⅰ)相关达49-50%	芳烃峰C-H-，C-C面外，伸展振动

表中Ⅰ、Ⅱ类是甲基、次甲基及烷烃基反映的峰值，第Ⅲ类是含氧基团等反映的峰值，第Ⅵ、Ⅴ类是以芳烃芳环反映的峰值。同时第Ⅱ类和第Ⅰ类很相关，达到72.5%程度；第Ⅴ类和第Ⅳ类也相关地达到50%。这样，又可看成为三大类，和Q型因子分析的三个主因子的意义得到完全一致的解释，使得对样品的评价和判断更为明确。

对于这一分析，其结果和R型因子分析对变量的分类也是完全一致的（图6）。

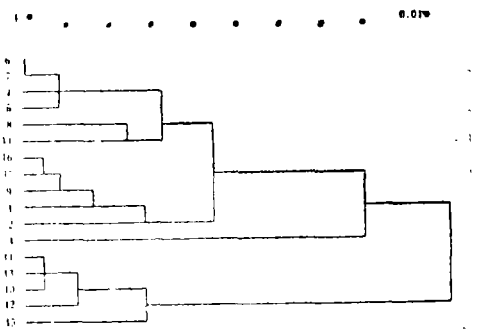

图6　R型聚类分析枝状图

样品：40，　　变量：17

三、对应分析

对应分析是通过R型因子分析的计算，同时得到R型和Q型的二种结果，並统一表示在一张图上。

它的计算是造一个M×N的矩阵Z，首先将原始距阵按行和列求和：

$$\begin{array}{l|l} X_{11}\ X_{12}\cdots X_{1n} & \sum_{j=1}^{N} X_{1j}=X_{1\cdot} \\ X_{21}\ X_{22}\cdots X_{2n} & \sum_{j=1}^{N} X_{2j}=X_{2\cdot} \\ X_{M1}\ X_{M2}\cdots X_{Mn} & \sum_{j=1}^{N} X_{Mj}=X_{M\cdot} \\ \hline \sum_{j=1}^{M} X_{j1}=X_{\cdot 1}\ X_{\cdot 2}\ \cdots\cdots \sum_{i=1}^{M} X_{1M}=X_{\cdot M} & \end{array}$$

得到

$$Z_{M\times n}=\begin{pmatrix} Z_{11}\ Z_{12}\cdots\cdots Z_{1n} \\ Z_{21}\ Z_{22}\cdots\cdots Z_{2n} \\ Z_{M1}\ Z_{M2}\cdots\cdots Z_{Mn} \end{pmatrix}$$

其中

$$Z_{ij}=\frac{X_{ij}-X_{i\cdot}X_{\cdot}/T}{\sqrt{X_{i\cdot}X_{\cdot j}}}\begin{pmatrix} i=1,2,\cdots M \\ j=1,2,\cdots N \end{pmatrix}$$

从矩阵Z出发做R型因子分析，然后再做Q型因子分析。因为这样处理的R型因子分析中的最大特征值也就是Q型的最大特征值，若$\Phi_1,\Phi_2\cdots\Phi_m$，是R型的相应特征值的特征向量，则$Z^T\Phi_1,Z^T\Phi_2,\cdots Z^T\Phi_m$是Q型的相对应的特征向量，各求出R型和Q型因子载荷矩阵值，计算结果可作图7。图中样品与变量间的关系明显地分成三类，这个划分和系统聚类分析完全一致，特别是和相似系数的聚类分析完全一致，同时和因子分析的结果也是基本一致。

现将各类的特征叙述如下。

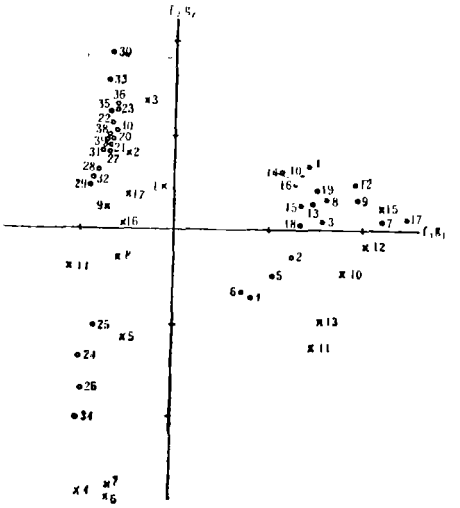

图7 对应分析图
（第1、2主因子座标图解）

· 样品 × 变量（吸收峰编号）

g_1，g_2，表示典型样品，f_1，f_2，代表主因子

第Ⅰ类样品位于对应分析的第二象限，与3号变量（720cm^{-1}）关系密切，同时也与1号（2850cm^{-1}）、2号(2950cm^{-1})、16、17和9号（1380cm^{-1}、1460cm^{-1}及1700—1720cm^{-1}）等几个变量相关。其中720cm^{-1}是石油沥青的主要特征峰，这里C—C骨架振动，—CH_3，—CH_2等基团占主导地位，与芳核、芳环类的峰很远，与含氧基团的峰也很远，这说明母质类型较佳，为腐泥型，且成熟度高，生油条件良好。这些样品的岩性特征主要是含生物灰岩及生物灰岩，大多数样品是属于中泥盆统棋梓桥组，少数是上泥盆统及下石炭统。

第Ⅱ类样品位于第一和第四象限，与10号（1740cm^{-1}）、12号(1040—1170cm^{-1})、11号（1660cm^{-1})、13号(1270—1300cm^{-1})等峰关系密切，说明了这批样品含氧基团值较高，它和甲基、次甲基等类峰的关系较第Ⅰ类远，同时和芳核、芳环类峰的关系也远，其母质类型不如第Ⅰ类好，可能以腐植型为主，其成熟度似较低。从层位上看，它们是湘中上泥盆统与下石炭统的结晶灰岩为主（微晶、泥晶泥灰岩），生油条件较差。

第Ⅲ类样品位于第三象限，它们远离甲基、次甲基等烷基类的代表峰，而与芳烃结构峰关系密切，和4、5、6、7(3050c m^{-1}、1600cm^{-1}、860—880cm^{-1}、800—810c m^{-1})等变量在一起，同时与含氧基团的峰也较远。可以认为其母质类型以腐植型为主，生油条件差，且成熟度有些过高。从岩性来看主要是碳质页岩。

上述三方面的结果与地质情况符合，例如第Ⅰ类样品的生油条件较好，与它所处的岩相条件也是一致，即棋梓桥组是广海低能带的沉积，而第Ⅱ类正好位于下石炭统的近岸滨海沼泽相带。

这项工作初次开展，错误之处，望同志们斧正。

（收稿日期1980年5月10日）

参考文献：

1. 中国科学院地质所：1977年 数学地质引论 地质出版社
2. 董庆年：1977年 红外光谱法 石油化学工业出版社