

# 地球化学数据

化探单元素异常统计内容

编写人:刘红杰

QQ:498236930

- 1、异常 ID ID
- 2、样品个数 N
- 3、异常面积 S
- 4、样品最大值 Max
- 5、样品最小值 Min
- 6、异常下限 T

7、算术平均值  $\bar{X} = \frac{\sum_{i=1}^n Xi}{n}$

8、几何平均值  $\overline{X}_g = \frac{1}{n} \sum_{i=1}^n \log Xi$

9、标准离差  $S_0 = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$

10、异常衬度  $A_c = \frac{\bar{X}}{T}$

11、异常规模  $A_d = S \times (\bar{X} - T)$

12、异常 NAP 值  $NAP = A_c \times S$

浓集克拉克值 (C) 计算公式:  $C = \frac{\bar{X}}{\text{某元素的克拉克值}}$

变化系数 (Cv) 计算公式:  $Cv = \frac{S}{\bar{X}}$

致矿系数 (Z) 计算公式:  $Z = Cv(\text{全区}) + 10 \times Cv(\text{剔高值后}) + 100 \times \text{高值比例} + C$ , 高值是大于 3 倍的标准离差。

化探背景分析

$$\text{中位数: } Me = X_{50} + \frac{H(50 - F_{50})}{f_{50}}$$

$$\text{偏度: } R_1 = \frac{\frac{1}{n} \sum f_i (X_i - \bar{X})^3}{\left( \sqrt{\frac{1}{n} \sum f_i (X_i - \bar{X})^2} \right)^3} \cdot \sqrt{\frac{n}{24}}$$

$$\text{峰度: } R_2 = \left( \frac{\frac{1}{n} \sum f_i (X_i - \bar{X})^4}{\left( \sqrt{\frac{1}{n} \sum f_i (X_i - \bar{X})^2} \right)^4} - 3 \right) \cdot \sqrt{\frac{n}{96}}$$

$$\text{正态检验: } \lambda = \max |F(x) - F_1(x)| \cdot \sqrt{\sum f_i}$$

Xi: 组中值或含量值;  $f_i$ : Xi 所对应的频数; H: 组距; X50: 包括累计频率 50% 在内的所在组的组下限;  
F<sub>50</sub>: 累计频率 50% 所在组之前的累计频率;  $f_{50}$ : 包括累计频率 50% 所在组的组频率; F(X): 为经验累计频率; F<sub>1</sub>(X): 为理论累计频率。

## R 型聚类分析

$$X'_{ij} = \frac{X_{ij} - X_i}{S_i}$$

$$\text{其中: } \bar{X} = \frac{1}{n} \sum_{j=1}^n X_{ij}; \quad S_i = \sqrt{\frac{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{n-1}}$$

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} \cdot S_{kk}}} = \frac{\sum_{i=1}^n ((X_{ji} - \bar{X}_j)(X_{ki} - \bar{X}_k))}{\sqrt{\sum_{i=1}^n (X_{ji} - \bar{X}_j)^2 \cdot \sum_{i=1}^n (X_{ki} - \bar{X}_k)^2}}$$

式中:  $r_{kj}$  为第 j 个变量和第 k 个变量的相关系数;

$X_{ji}$  为第 j 个变量第 i 个样品的观测值;

$\bar{X}_j$  与  $\bar{X}_k$  为第 j 个和第 k 个变量的平均值。

## Q 型聚类分析

$$X'_{ij} = \frac{X_{ij} - X_{i(\min)}}{X_{i(\max)} - X_{i(\min)}} \quad (i=1, 2, \dots, P; j=1, 2, \dots, n)$$

其中: P, n 分别为变量数和样品数;

$X_{i(\max)}$  及  $X_{i(\min)}$  分别为数据中第 i 个指标的极大值与极小值

$$D_{jk} = \sqrt{\sum_{i=1}^P ((X_{ij} - X_{ik})^2 / P)} \quad (j, k=1, 2, \dots, n; j \neq k)$$

式中：D<sub>jk</sub> 为第 j 个样品与第 k 个样品的距离系数；

X<sub>ij</sub> 为第 i 个变量第 j 个样品的观测值。

## 因子分析

数学原理：设有一批含 p 个变量，n 个样品的观测数据，如果其变量为 X<sub>1</sub>、X<sub>2</sub>、……、X<sub>p</sub>，它们的综合变量记为 F<sub>1</sub>、F<sub>2</sub>、……、F<sub>m</sub> (m ≤ p)，其数学表达式为

$$\begin{cases} F_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p \\ F_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p \\ \dots \\ F_m = a_{m1}X_1 + a_{m2}X_2 + \dots + a_{mp}X_p \end{cases}$$

要求  $a_{11}^2 + a_{12}^2 + \dots + a_{1p}^2 = 1$  (k=1, 2, …, m)

系数  $\{a_{ij}\}$  由下列原则决定：

- 1、F<sub>i</sub> 与 F<sub>j</sub> (ij, i, j=1, 2, …, m) 互相无关
- 2、F<sub>1</sub> 是 X<sub>1</sub>、X<sub>2</sub>、……、X<sub>p</sub> 的一切线性组合中方差最大；  
F<sub>2</sub> 是与 F<sub>1</sub> 不相关的 X<sub>1</sub>、X<sub>2</sub>、……、X<sub>p</sub> 所有线性组合中方差最大的；  
……；  
F<sub>m</sub> 是与 F<sub>1</sub>、F<sub>2</sub>、……、F<sub>m-1</sub> 都不相关的 X<sub>1</sub>、X<sub>2</sub>、……、X<sub>p</sub> 所有线性组合中方差最大。

回归系数：

$$b = \frac{\sum XY - \bar{Y} \sum X}{\sum X^2 - \bar{X} \cdot \sum X} = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{X^2 - (\bar{X})^2}$$

标准差：

$$\sigma_X = \sqrt{X^2 - (\bar{X})^2}$$

$$\sigma_Y = \sqrt{Y^2 - (\bar{Y})^2}$$

$$\text{相关系数: } r = \frac{\overline{XY} - \bar{X} \cdot \bar{Y}}{\sigma_X \cdot \sigma_Y}$$

$$r = b \times \frac{\sigma_X}{\sigma_Y} \quad b = r \times \frac{\sigma_Y}{\sigma_X}$$

## 回归分析

1、多元线性回归分析研究某一变量与多个变量之间的线性关系（某两变量之间的非线性关系有时也可以转化为线性关系）。这是以大量收集到的观测数据为基础，找出相关变量

之间的内部规律性，以定量形式建立一个因变量与另一个自变量（或另几个自变量）之间关系的数学表达式，从而可以根据一个或几个变量的观测值来预测（预报）另一个变量的估计值，并能从多个指标变量中找出对所研究的问题起重要作用的某些指标。

2、多元逐步回归分析是以大量收集到的观测数据为基础，建立某一个变量与另一个变量（或几个变量）之间关系的数学表达式。逐步回归是在多元回归基础是派生出来的一种算法，它能从众多的变量（或预先尽可能多地考虑一些变量）中自动挑选重要变量（指标或因子），并确定其数学表达式的一种统计方法。

3、正交化回归分析与线性回归不同，它不一定建立随机变量与全体变量之间的关系，而是一种具有挑选因子的方案。线性回归和逐步回归仅考虑因变量（一个或多个）与自变量（多外）之间的关系，而不考虑自变量本身之间的关系，因而信息有重复的可能，而降低某一独立因素的相对权系数。这种重复也因具体问题的不同，对回归效果就有一定的影响。正交化回归不仅考虑自变量与因变量之间的关系，而且还要考虑自变量本身之间的相互影响。即是将自变量因子进行正交化，排除自变量之间的相互影响，得到一组新的正交化了的因子，再建立新因子与因变量之间的回归方程，并还原到非正交化、非正规化的回归方程，作为建立预报模型之用。

## 对应分析

对应分析是在 R-型与 Q-型因子分析基础上发展起来和一种多元统计方法，它把 R-型和 Q-型因子分析结合起来，综合考虑变量之间、样品之间及变量与样品之间的关系。它揭示了 R-型与 Q-型分析之间的两重性，以较少重要的几个公共因子的综合指标去研究对象在成因上或空间上的联系，应用载荷平面投影图进行地质解释与推断，在地质学的应用可以包括：对矿床成因的解释；成岩（成矿）的物质来源、作用方式、作用因素；含矿岩体的预测、评价；地球化学的研究等。根据 R-型与 Q-型分析具有的对偶性，由 R-型分析结果很容易地得 Q-型分析结果，它可以克服由于样品容量大而对 Q-型因子分析所带来计算上的困难，把变量和样品同时反映到同一坐标轴（因子轴）的一张图形上，更便于地质解释与推断。

## 相关分析

1、典型相关分析是研究两组地质指标（或变量）之间相关关系的一种统计方法。它揭示了两个因素“集团”之间的内部联系，而两个因素“集团”的内容和变量数目又可以不同，这就决定了它在解决地质问题上的许多特点，在地质上用来研究二组地质特征之间的关系。用以研究地质成因和地质体对比。如研究某种矿物成分与其围岩成分的关系；研究古生物群同古地理环境之间依存关系；两条区域剖面的化学成分、岩石成分的对比等等。典型相关分析的目的，是在两组众多地质变量中分别寻找若干对（每对为若干个变数的线性组合）有代表性、且具有相关关系的综合指标，通过研究两组综合指标间的关系来反映两组地质变量间的主要作用因素和作用形式。

2、秩相关分析是研究一组观测值与另外一组观测序列之间的相关关系，在矿床统计预测中，研究各种控矿地质因素和找矿标志与矿化的相关关系，查明这样的因素与标志最有利的数值范围，即查明找矿有利的统计标志。

两个观测序列的秩相关系数：

$$P = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

其中 d 为对比序列的序差；n 为对比序列的等级数。

3、相关频数比值法：筛选因子（变量）的目的在于从大量的预报因子中选出与预报量

相关较好，而因子之间相关较差，即因子独立性强的若干较优因子，由它们组成数学模型使预报效果达到最佳。

计算公式化：第  $i$  个因子的相关频数比

$$m_i = \frac{n_i}{n'_i}$$

$n_i$ : 变量  $X_i$  报对的频数，指预报量  $Y=1$  时  $X_i=1$ ， $Y=0$  时  $X_i=0$

$n'_i$ : 因子间相关频数，指  $X_i$  因子在各样本个体中的报错时，相应样本个体其它因子重复报错的总频数。

### 特征分析

特征分析是一种对定性描述资料的多元统计方法，其理论简单，通过对已知模型区内各变量间相互关系的定量考察，来确定每个变量的权系数，用于预测的原则是类比法。可综合处理各种数据，尤其适用于多类型定性描述资料的综合处理。在地质找矿过程中常用于预测与评价找矿远景区，亦可用来挑选控矿的重要变量。

方法原理：从乘积矩阵  $Z'Z$  ( $Z$  为原始数据矩阵) 和匹配矩阵  $T$  出发，采用主分量分析的雅可比法 (Jacobi) 寻找最大特征值 ( $\lambda_1$ ) 及其相应的一组特征向量 ( $b_{ij}$ )，来确定各标志的权系数，建立模型与变量线性关系，同时得到预测区的综合指标——关联值，提供定量评价、预测。

### 点聚图 0-1 法

$$\text{区分率} = \frac{N - \text{判错的样品数}}{N} \times 100\%$$

### 隶属度和贴近度

1、根据在集合上的隶属函数，按隶属原则识别对象，判定其属于哪一个类型；2、根据各弗晰两两间的贴近度，按择近原则，确定哪两个弗晰集最贴近。

#### 1、隶属度

设有  $n$  种类型， $m$  种指标，则第  $i$  种类型在第  $j$  种指标上的隶属函数为：

$$A_{ij} = \begin{cases} 0 & X \leq a_{ij}^{(1)} - b_{ij} \\ 1 - \left( \frac{X - a_{ij}^{(1)}}{b_{ij}} \right)^2 & a_{ij}^{(1)} - b_{ij} < X < a_{ij}^{(1)} \\ 1 & a_{ij}^{(1)} \leq X \leq a_{ij}^{(2)} \\ 1 - \left( \frac{X - a_{ij}^{(2)}}{b_{ij}} \right)^2 & a_{ij}^{(2)} < X < a_{ij}^{(2)} + b_{ij} \\ 0 & a_{ij}^{(2)} + b_{ij} \leq X \end{cases}$$

其中  $a_{ij}^{(1)}$  和  $a_{ij}^{(2)}$  分别是第  $i$  类元素第  $j$  种指标的最小值和最大值； $b_{ij}^2 = 2\sigma_{ij}$ ，而  $\sigma_{ij}^2$  是第  $i$  类元素第  $j$  种指标上的方差，给定一具体对象  $X^*$ ，设它的  $m$  个指标为  $X_1^*$ 、 $X_2^*$ 、 $\dots$ 、 $X_m^*$ ，令

$$S_i = \min A_{ij}(X_j^*) \quad 1 \leq j \leq m$$

$$\text{又若 } S_{i_0} = \max(S_i) \quad 1 \leq i \leq n$$

则认为此对象属第  $i_0$  类  $A_{i_0}$ 。

## 2、贴近度

已知有  $n$  种类型  $(A_1, A_2, \dots, A_n)$ ，它们都有  $m$  种指标，均为正态型弗晰变量，相应的参数分别为  $(a_{ij}^{(1)}, a_{ij}^{(2)}, b_{ij})$  ( $i=1, 2, \dots, n; j=1, 2, \dots, m$ )，其中  $a_{ij}^{(1)} = \min(X_{ij})$ ， $a_{ij}^{(2)} = \max(X_{ij})$ ， $b_{ij}^2 = 2\sigma_{ij}^2$ ， $\sigma_{ij}^2$  是  $X_{ij}$  的方差，待判断对象  $B$  的  $m$  个指标分别是具有参数  $(a_j, b_j)$  ( $j=1, 2, \dots, m$ ) 的正态型弗晰变量，则  $B$  与各类型的贴近度为：

$$(A_{ij}, B) = \begin{cases} 0 & a_j \leq a_{ij}^{(1)} - (b_j + b_{ij}) \\ 1 - \frac{1}{2} \left( \frac{a_j - a_{ij}^{(1)}}{b_j + b_{ij}} \right)^2 & a_{ij}^{(1)} - (b_j + b_{ij}) < a_j < a_{ij}^{(1)} \\ 1 & a_{ij}^{(1)} \leq a_j \leq a_{ij}^{(2)} \\ 1 - \frac{1}{2} \left( \frac{a_j - a_{ij}^{(2)}}{b_j + b_{ij}} \right)^2 & a_{ij}^{(2)} < a_j < a_{ij}^{(2)} + (b_j + b_{ij}) \\ 0 & a_{ij}^{(2)} + (b_j + b_{ij}) \leq a_j \end{cases}$$

$$\text{又记 } S_i = \min(A_{ij}, B)$$

$$\text{若有 } S_{i_0} = \max(S_i) \quad 1 \leq i \leq n$$

则按贴近原则，可以认为此  $B$  与  $A_{i_0}$  最贴近。

## 信息量预测

信息量算法是通过计算各地质因素和找矿标志所提供的找矿信息量，定量地评价各地质因素和标志对指导找矿的作用，借以选择与矿化关系密切的变量；同时根据每个单元中各标志信息量总和的大小，评价每个单元相对的找矿意义，用以对找矿远景区进行预测。

标志  $X_j$  指示有矿的信息量：

$$I_j = \lg \frac{N_j/N}{S_j/S}$$

式中： $N_j$ —具有标志  $X_j$  含矿单元数； $N$ —研究区内的含矿单元数； $S_j$ —具有标志  $X_j$  的单元数； $N$ —研究区内单元数。

有用信息量：

$$\Delta I^+ = K \sum_{j=1}^n I_j$$

式中： $K$ —有用信息水平（一般取 0.75）； $n$ —信息量为正值的标志个数；

预测单元找矿信息量临界值：

$$I_r = \frac{1}{r} \sum_{i=1}^r I_j \quad (r \text{ 为有矿单元个数})$$

### 原生晕元素分带序列

令原始矩阵数据为  $X=[X_{ij}]$ ；其中  $X_{ij}$  为线金属量数值； $i$  为标高或水平点距序号  $i=1, 2, \dots, n$ ； $j$  为元素的种类序号  $j=1, 2, \dots, m$ 。

1、确定每个元素的规格化系数  $K_j$ ，对  $X$  矩阵进行规格化变换

从每个元素中挑选出最大值  $K_j$  得  $M$  个最大值，从这  $M$  个最大值中，以最大数量级的  $K_j$  为  $10^0$ ，次大数量级的  $K_j$  取值为  $10^1$ ，……，其余以此类推，

得规格化系数矩阵

$$K = \begin{pmatrix} K_1 & & & 0 \\ & K_2 & & \\ & & \ddots & \\ 0 & & & K_m \end{pmatrix} \quad \text{得规格化数据阵：} A=XK=[a_{ij}]$$

2、确定分带指数

$$\text{由 } A \text{ 计算出 } b_{ii} = \frac{1}{\sum_{j=1}^m a_{ij}} \quad (i=1, 2, \dots, n)$$

$$\text{得 } B = \begin{pmatrix} b_{11} & & & 0 \\ & b_{22} & & \\ & & \ddots & \\ 0 & & & b_{nn} \end{pmatrix} \quad \text{所以分带指数矩阵为：} D=BA=[d_{ij}]$$

3、计算变化梯度

$$\text{公式 } \Delta G_j = \sum_{i=k+1}^n \frac{d_{kj \max}}{d_{ij}} - \sum_{i=1}^{k-1} \frac{d_{kj \max}}{d_{ij}}$$

式中： $d_{kj \max}$  表示第  $j$  个元素的最大分带指数。

4、确定分带序列的方法如下

1) 在同一行里即同一水平（或标高）只出现一个元素的最大分带指数时，不用考虑变化梯度的大小，按标高自上而下排列分带序列。

2) 在同一水平（或标高）出现两个或两个以上的最大分带指数时，可由变化梯度的大小顺序和标高自上而下排列分带序列。

### GRD 数据变换

标准化变换：

$$X'_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}$$

其中  $X_{ij}$  是原始观测值， $\bar{X}_j$  是第  $j$  个变量的算术平均值， $S_j$  是第  $j$  个变量的标准差， $i$  为样本数， $j$  为变量数。



正规化变换:

$$X'_{ij} = \frac{X_{ij} - X_{j\min}}{X_{j\max} - X_{j\min}}$$

其中  $X_{ij}$  是原始观测值,  $X_{j\min}$  是第  $j$  个变量的最小值,  $X_{j\max}$  是第  $j$  个变量的最大值,  $i$  为样本数,  $j$  为变量数。

平均值计量变换:

$$X'_{ij} = \frac{X_{ij}}{\overline{X_j}}$$

反正弦变换:

$$X'_i = \arcsin(X_i)$$

反余弦变换:

$$X'_i = \arccos(X_i)$$

对数变换:

$$X'_i = \log(X_i + C)$$

其中  $C$  为常数。

解析延拓

$$\Delta g(0,0-h) = \sum_{i=1}^{10} k(r_i, h) \Delta g(r_i)$$

其中:  $k(r_i, h)$  第  $i$  环上延(或下延) $h$  时的环系数;  $\Delta g(r_i)$  第  $i$  环上点的平均值;

$i=1 \sim 10$  半径分别为: 0, 0.5, 1, 2, 5, 8, 13, 25, 50, 136

八方位化极

$$gm(m) = f \times \frac{d}{j} \times \sum_{m=0}^{10} \sum_{n=0}^7 \alpha_n \beta_n Z_{mn} = f \times \frac{d}{j} \left[ \frac{0.3}{Za(0)} + \sum_{m=0}^{10} \sum_{n=0}^7 \alpha_n \beta_n Z_{mn} \right]$$

其中:  $Za(0)$ —计算点实测值

$\alpha$ —求  $Z$  的方位系数,  $n=0 \cdots 7$

0 方位与测线重合, 顺时针方向等分为八个方位

$\beta$ —环半径系数

$Z_{mn}$ —八个方位与环交叉点的场值、由插值求得

$g_m(m)$ —伪重力值, 单位  $mg/l$

$Z_m(m)$ —由  $\Delta T$  或  $Z_a$  求得的垂直磁化情下的垂直分量

$j$ —磁化强度 (CGSM)

$\sigma$  --- 剩余密度 ( $g/cm^3$ )

$f$ —万有引力常数  $6.67 \times 10^{-8}$  CGSM