

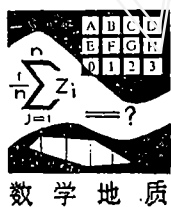
加权中位数法的基本理论、方法及其应用

侯景儒 吴雨沛 向永生

(北京科技大学地质系)

物化探及其他地质数据常呈现非正态分布特征,例如数据中存在特异值或在研究区有几种矿化类型时,致使数据呈现出长尾巴分布。为了提高数据处理的质量和矿石储量计算的精度,本文介绍一种地质统计学方法——加权中位数法。文中讨论了方法的理论,重点研究了方法步骤,包括距离权、丛集权和综合权的计算方法,最后给出一个计算金矿床矿石储量的实例。

关键词: 非参数地质统计学, 特异值, 中位数, 距离权, 丛集权



在处理地质、物化探数据或估计矿产储量时,一些因素在不同程度上影响着计算精度,例如数据中特异值(实际存在的特高观测值或特低观测值)

的出现,使估计结果偏高或偏低。又如在研究区或矿床中存在几种不同类型的矿化,也会影响估计的精度。为了解决这些问题,地质学家和统计学家做了大量研究,提出了一些方法,其主要方法是将所研究的数据进行某种变换,使其分布尽量符合所用经典统计方法的要求(使之呈正态分布),但当对数据经过某种变换仍达不到要求或被研究数据来自不同母体时,其计算结果仍存在着不可忽视的偏差。为此,本文将讨论另一种方法——非参数统计方法。

所谓非参数统计方法(non-Parametric statistics)又称无分布法或自由分布法(free-Parametric),该法无需假设数据来自具有某种特定分布的母体(如正态分布、对数正态分布等),也无需对原始数据进行某种变换。

经典统计学常用两个参数表征一个总体的特征:一个是分布的中心位置(或期望值及平均值),另一个是偏离该中心位置的离散的量。在非参数统计方法中常用的较为稳健的中心位置估计量有:顺序统计量;顺序—秩统计量以及指示数(Indicator)等。而用于估计中心位置的较为稳健的方法有 α -截尾均值法、加权均值法和中位数法。本文重点讨论的加权中位数法也是一种稳健的估计中心位置的非参数地质统计学方法。该法在不必要去掉特异值或在几种矿化同时存在的条件下来处理各种数据或进行储量计算,而且可在一定风险条件下,给出某一待估点或待估块段的最佳估计量及其空间分布。

加权中位数法的基本原理

按照定义,中位数是使分布曲线二等分的点,分布的一半位于中位数以上,分布的另一半位于其下,在连续分布中,中位数是单个的值,在离散分布中,中位数是某个区间的任意值,因而,中位数是使估计绝对离差和最小的估计量,也就是说,中位数受特

异值的影响较小。

在对称分布(如正态分布)中,中位数与均值重合,这时,均值的标准差为 $\sigma/\sqrt{n}$ 。

(σ 为样本估计方差, n 为样品数),中位数的估计标准差为 $1.253\sigma/\sqrt{n}$,显然,在这种情况下用中位数作为中心位置的估计量比均值差,但对于非对称分布而言,均值受特异值影响很大,而中位数则很小,这时用中位数作为中心位置的估计量则比较理想。

众所周知,地质学所研究的地质变量(或地质特征)既具随机性又具结构性(相关性),而且变量的空间位置又十分重要。因此,当我们利用加权中位数法进行估计时,必须既要考虑样品的大小、形状、信息样与待估点(或块段)之间的空间关系,又要考虑信息样之间的空间关系,也就是说,在赋予每一信息样品一权系数时,必须考虑上述因素,把这种考虑分别用所谓“距离权”和“丛集权”表示,同时,为了使估计保证无偏,必须使

$$\sum_{i=1}^n W_i = 1 \quad (1)$$

(1)式中, W_i 是第 i 个样品的权系数; n 是估计时所用的信息样品数。

设 $Z(x)$ 是服从二阶平稳假设的定义于点支撑上的区域化变量,即

$$E[Z(x)] = m \quad (2)$$

$$E[Z(x+h) \cdot Z(x)] - m^2 = C(h) \quad (3)$$

设中心点在 x_0 点上的待估域 $V(x_0)$ 的平均品位是 Z_V

$$Z_V = \int_{V(x_0)} Z(x) dx \quad (4)$$

Z_V 可用 $V(x_0)$ 周围(估计邻域)一组信息样品 Z_i ($i=1,2,\dots,n$)借助于加权中位数法进行估计,即 Z_V 的估计量 Z_V^* 就是这 n 个信息样观测值的加权中位数。它在品位—权分布曲线上为中位权所对应的数值(品位),其表达式为

$$Z_V^* = \text{med}[F_W(Z_i)] \quad (5)$$

(5)式中,med表示取中位数;

$F_W(Z_i)$ 表示区域化变量 Z_i 的权累积分布函数。

再集中讨论权函数,尤其是距离权与丛集权的计算问题。

1. 距离权

主要的距离权函数的计算式有:

$$\text{距离倒数权} \quad WD(i) = 1/d_i \quad (6)$$

$$\text{距离平方倒数权} \quad WD(i) = (1/d_i)^2 \quad (7)$$

$$\text{距离负幂指数权} \quad WD(i) = e^{-Kd_i^\alpha} \quad (8)$$

$$\text{距离商权} \quad WD(i) = (d_{\min}/d_i)^K \quad (9)$$

(6)、(7)、(8)、(9)式中, $WD(i)$ 为距离权;

d_i 为第 i 个信息样到待估点(或块段中心点)的距离;

$d_{\min}$ 为诸 d_i 中最小者;

α, K 均为待定参数(视具体情况)。

以上各种权函数表明,距待估点不同距离的信息样在估计待估点(或域)平均品位时的贡献大小,距离近者贡献大,反之,贡献小。经过实际比较证明,在许多情况下,距离商权比其他距离权函数效果更好。

从距离商权的计算公式(9)可知,当 $d_{\min}=0$,即某信息样位于待估域中心点时,除了该信息样的权因子为1外,其他信息样的权因子均为0,从而丢弃了其他信息样品的贡献,为此,提出如下的距离权函数

$$WD(i) = (d_{\min} + \Delta/d_i)^K \quad (10)$$

当 $d_{\min} \neq 0$ 时, $\Delta=0$;

当 $d_{\min}=0$ 时, Δ 为次近点到待估域中心点的距离,所谓次近点是指除待估域中心点上的信息点之外的距待估域最近的信息点。设待估域为一矩形,在二维空间中两个边长为 LX 与 LY ,当 $d_{\min} > LX$ 或 $d_{\min} > LY$ 时,则

$$d_{min} + \Delta = \sqrt{\left(\frac{LX}{2}\right)^2 + \left(\frac{LY}{2}\right)^2} \quad (11)$$

(10) 式中的 K 视估计空间而定, 在一维条件下, $K=1$, 在二维情况下, $K=2$ 。

2. 从集权

当信息样与待估域之间的距离相等, 且信息样均匀分布时, 各信息样的距离权 (即贡献) 相等, 如果某些信息样丛聚时, 其中的部分信息样提供的信息量重复, 致使待估域的平均值的估计量产生偏差, 丛集权就是要充分考虑丛集的样品对待估值的贡献、信息样之间的空间关系以及信息样与待估域之间的空间关系, 通过丛集权而达到无偏估计。

设研究域的面积 (或体积) 为 S , 信息样总数为 N , 则域 S 内的平均采样密度为 N/S , 当这 N 个信息样在域 S 内均匀分布时, 它的频率分布曲线为

$$F = \frac{1}{S} \left(1 - \frac{1}{N}\right) a_i + \frac{1}{N} \quad (12)$$

当某一群样品的分布密度比平均采样密度大时, 它的密度分布函数就与 F 之间存在一定差异, 这种实际分布偏离均匀分布的程度可作为衡量样品丛集程度的标准, 通常可用 Kolmogorov 统计量来估计这种偏离程度。

对于单个信息样来说, 估计它的丛集权的关键就是确定它周围一定范围内的样品实际分布偏离均匀分布的程度, 这种偏离程度与信息样之间的距离有关, 换言之, 它与需计算丛集权的某一信息样周围一定面积 (二维) 或体积 (三维) 范围内的信息样品数有关。因此, 必须先计算各信息样 i 与需计算丛集权的某信息样之间的距离 d_{ip} , 并按空间维数将 d_{ip} 转化为面积 S_{ip} 或体积 V_{ip} , 在二维情况下

$$S_{ip} = \pi d_{ip}^2 \quad (13)$$

在三维情况下

$$V_{ip} = \frac{4}{3} \pi d_{ip}^3 \quad (14)$$

我们可以把估计邻域 (为估计待估点或待估域而搜索信息样的范围) 划分成若干等分, 统计出每一区间内的信息样总数, 并计算出频率, 然后按照面积或体积递增的顺序排列, 画出累积频率直方图 F^* , 在同一坐标系上画出均匀分布的累积频率直方图 F , 逐点求出两条曲线之间的差值为

$$D_{(ai)} = F^*(ai) - F(ai) \quad (15)$$

从中确定一个最大的正偏差

$$D_{(ai)}^+ = \sup[D_{(ai)}] = \sup[F^*(ai) - F(ai)] \quad (16)$$

该待估样品的丛集权就定义为最大正偏差上的理论频率 $F(ai)$ 与实际频率 $F^*(ai)$ 之比

$$W_{(ip)} = F(ai) / F^*(ai) \quad (17)$$

如果无正偏差, 该比值所在的位置就定义在 $ai=0$ 上, 此时

$$F^*(0) = F(0) = 1/N \quad (18)$$

$$W_{(ip)} = F(0) / F^*(0) = 1 \quad (19)$$

3. 加权中位数法的估计方差

令 $Z(x)$ 为服从二阶平稳的区域化变量, 待估域 V 的平均值为 Z_v , Z_v 的估计值为 Z_v^* , 其估计方差可用下式表示

$$\sigma E^2 = 2 \sum_{i=1}^n \lambda_i \overline{\gamma}(V, v_i) - \overline{\gamma}(V, V) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \overline{\gamma}(v_i, v_j) \quad (20)$$

式中, v_i, v_j 为估计邻域内的信息点 (或域);

$\overline{\gamma}(V, v_i)$ 为待估域与信息域之间的变异函数的平均值;

$\overline{\gamma}(V, V)$ 为待估域 V 自身的变异函数平均值;

$\overline{\gamma}(v_i, v_j)$ 为信息域 v_i 与信息域 v_j 之间的变异函数的平均值;

λ_i, λ_j 为信息域 v_i 及信息域 v_j 各

自的权系数；

n 为信息样的总数。

(20) 式中的平均变异函数 $\bar{\gamma}(h)$ 用中心化协方差 $\bar{C}(h)$ 表示时, (20) 式变为

$$\sigma_k^2 = \bar{C}(V, V) + \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \bar{C}(v_i, v_j) - 2 \sum_{i=1}^n \lambda_i \bar{C}(V, v_i) \quad (21)$$

加权中位数法的方法步骤

应用加权中位数法进行物、化探数据处理或储量估计时的计算步骤简述如下(图1):

1. 对原始数据进行统计分析, 研究样品的分布特征, 绘制样品的分布曲线。
2. 计算变异函数, 进行结构分析, 确定变异函数理论模型。

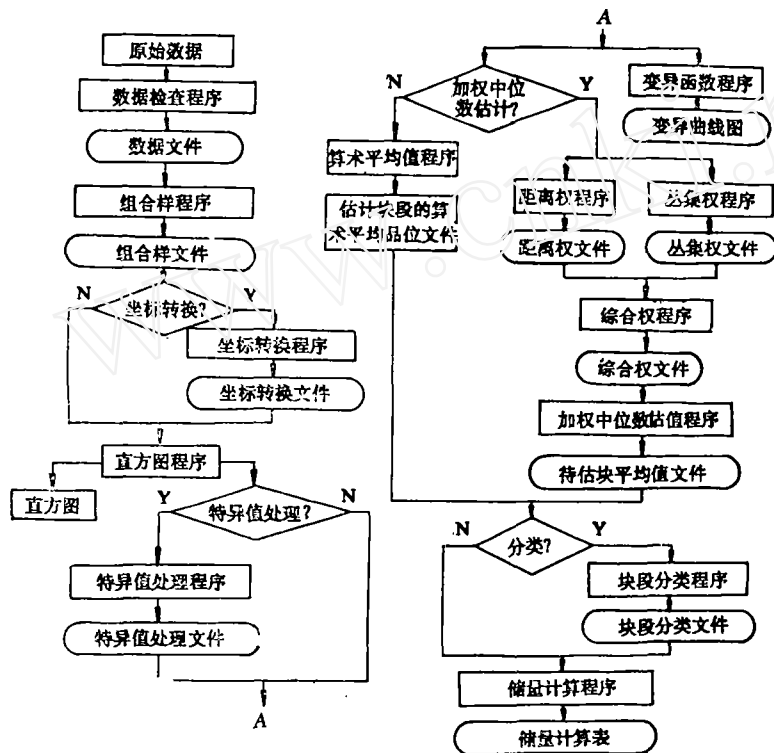


图 1 二维加权中位数法估计流程

3. 计算距离权值。

设待估域 V 的 $LX = LY = 6\text{m}$, 即 V 为正方形, 它的中心点 R 的坐标为 $x_R = 21\text{m}$, $y_R = 117\text{m}$, 为了对 V 进行平均品位的估计, 我们应用了估计邻域内 16 个信息样品, (图 2、表 1) 这些信息样的坐标及观测值 (品位) 如表 1 所示。对每一信息样按下式

$$d_{ip} = \sqrt{(x_i - x_R)^2 + (y_i - y_R)^2}$$

计算信息样 $i (i=1, 2, \dots, 16)$ 到 R 点的距离 (表 1) 求出这些距离中的最小者 $d_{\min} = 3.497$, 再按公式 (10) 求出每个信息样的

距离权 (表 1)。

4. 计算丛集权值。

用图 2 和表 1 所列的数据来计算 1 号信息样的丛集权, 其具体计算方法如下。

(1) 从观测值的变异函数曲线得到该观测值的变程 $a = 6\text{m}$, 因此求信息样丛集权的极限距离 $H = 6\text{m}$, 这样, 计算 1 号样丛集权极限距离范围内的信息样号为 4, 5, 8, 9, 10, 13 (图 2、表 2)。

(2) 计算极限面积 A_i , $A_i = \Pi \times 6^2 = 113.09\text{m}^2$, 将 A_i 分成 5 等分 (5 个区间),

信息样的坐标、观测值(品位)以及计算的 d_{ip} 及 $WC(i)$

表 1

样品号	1	2	3	4	5	6	7	8
x 坐标	19.42	19.36	21.26	18.42	24.12	22.26	18.36	23.54
y 坐标	120.12	113.50	113.00	120.12	119.80	113.00	113.20	120.04
观测值(品位)	0.92	4.21	1.05	0.90	36.50	4.31	0.86	6.71
距离 d_{ip}	3.497	3.865	4.008	4.018	4.192	4.194	4.352	4.571
距离权 $WC(i)$	1.00	0.671	0.579	0.575	0.484	0.483	0.417	0.342

样品号	9	10	11	12	13	14	15	16
x 坐标	17.42	25.12	17.36	24.54	16.42	16.36	26.12	25.54
y 坐标	120.04	119.76	113.58	113.14	119.98	113.64	119.76	113.10
观测值(品位)	0.68	8.44	0.63	3.16	0.45	0.45	10.60	1.17
距离 d_{ip}	4.697	4.959	4.995	5.227	5.464	5.729	5.788	5.985
距离权 $WC(i)$	0.307	0.247	0.240	0.199	0.168	0.139	0.133	0.116

注: 由于待估块段 $Lx = Ly = 6.0m$, 则 $d_{min} = 3.497 < Lx = Ly$ 故 $\Delta = 0 (K=2)$ 。

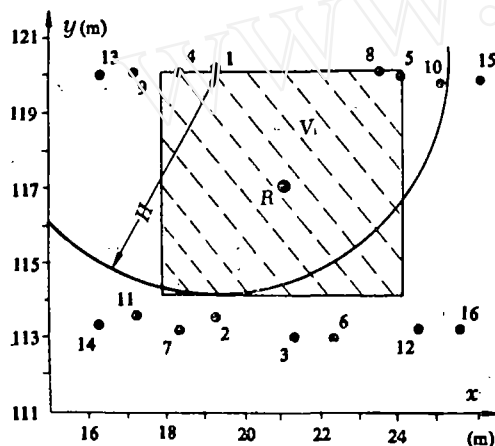


图 2 计算距离权和丛集权

H 为计算丛集权的极限距离

各信息样到 1 号样距离及相应的面积

表 2

信息样号	4	5	8	9	10	13	1
x 坐标	18.42	24.12	23.54	17.42	25.12	16.42	19.42
y 坐标	120.12	119.80	120.04	120.04	119.76	119.98	120.12
到1号样号距离	1.00	4.76	4.12	2.00	5.70	3.00	0.00
变换成的面积	3.14	71.14	53.30	12.56	102.02	28.31	0.00

其分界点分别为: $22.62m^2$; $45.24m^2$; $67.86m^2$; $90.48m^2$ 和 $113.09m^2$ 。

(3) 计算极限面积内各信息样至 1 号信息样之间的距离, 并将其变换成面积 (表 2)。

(4) 统计落入每一区间内信息样点的数目 i , 计算其试验频率, 然后, 按面积(或体积) a_i 递增的顺序排列, 计算其试验累积频率 F^* , 其计算公式为

$$F^* = i / N$$

上式中, N 为研究区内的样品总数; i 为样品序号 (或累计样品数)。在实例中, 整个研究区共计有样品 463 个 ($N = 463$), 这样, 当 $i = 1$ 时, $F^* = 1/463 = 2.159 \times 10^{-3}$; 当 $i = 2$ 时, $F^* = 2/463 = 4.319 \times 10^{-3}$; 当 $i = 3$ 时, $F^* = 3/463 = 6.479 \times 10^{-3}$; 当 $i = 4$ 时, $F^* = 8.639 \times 10^{-3}$; ...。从表 2 可知, 落入 $0 \sim 22.62m^2$ 区间的样品数为 3, 所以对应的 $F^* = 3/463 = 6.479 \times 10^{-3}$, 落入 $22.62 \sim 45.24m^2$ 区间的样品数为 4, 对应的 $F^* = 4/463 = 8.639$, 其余类推 (表 3)。

(5) 计算均匀分布时的理论累积频率 F , 令研究区总面积为 S (实例中 $S = 126 \times 78m^2$), 令每小区间的面积为 ΔS (本例中 ΔS 分别为 $22.62m^2$, $45.24m^2$, $67.86m^2$, $90.48m^2$, $113.09m^2$), 设样品总数为 N , (本例中 $N = 463$), 则

试验频率与理论频率比较 表 3

区 间	~22.62	~45.24	~67.86	~90.48	~113.09
试验频率	6.48×10^{-3}	8.64×10^{-3}	10.80×10^{-3}	12.96×10^{-3}	15.12×10^{-3}
理论频率	4.46×10^{-3}	6.76×10^{-3}	9.06×10^{-3}	11.37×10^{-3}	13.67×10^{-3}
二者差	2.02×10^{-3}	1.88×10^{-3}	1.74×10^{-3}	1.59×10^{-3}	1.45×10^{-3}

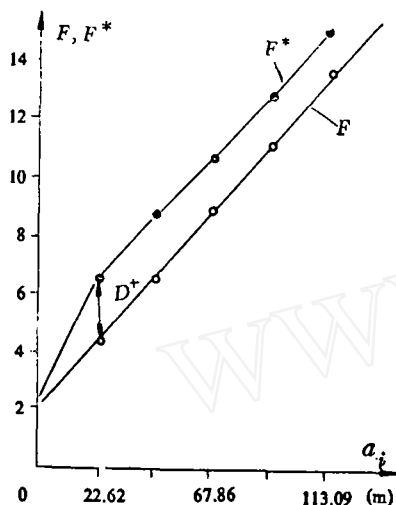


图 3 试验分布曲线与理论分布曲线

$$F = \frac{1}{N} + \frac{\Delta S}{S}$$

这样, 当 $\Delta S = 22.62 \text{ m}^2$ 时, $F = \frac{1}{463} + \frac{22.62}{126 \times 78}$
 $= 4.46$; 当 $\Delta S = 45.24 \text{ m}^2$ 时, $F = 6.7629$;

各信息样的距离权、丛集权及综合权

表 4

信息样号	1	2	3	4	5	6	7	8
距离权	1.00	0.671	0.579	0.575	0.484	0.483	0.417	0.342
丛集权	0.690	0.72	0.64	0.680	0.590	0.71	0.65	0.72
$W(j)$	0.690	0.483	0.371	0.391	0.286	0.343	0.271	0.246
$W^*(j)$	0.163	0.114	0.087	0.092	0.067	0.081	0.058	0.053
信息样号	9	10	11	12	13	14	15	16
距离权	0.307	0.247	0.240	0.199	0.168	0.139	0.133	0.110
丛集权	0.73	0.70	0.81	0.76	0.74	0.79	0.75	0.770
$W(j)$	0.224	0.173	0.194	0.151	0.124	0.110	0.100	0.089
$W^*(j)$	0.053	0.041	0.046	0.035	0.029	0.026	0.024	0.021

当 $\Delta S = 67.86 \text{ m}^2$ 时, $F = 9.0645$, 余类推 (表 3)。

在同一坐标系中, 画出对应于 ΔS 的理论累积频率曲线 F 及相应的试验累积频率曲线 F^* (图 3)。

(6) 计算 1 号信息样的丛集权 $WC(p)$

$$WC(p) = 4.46 \times 10^{-3} / 6.48 \times 10^{-3} \\ = 0.69$$

用同样的方法可求出其他信息样的丛集权。

5. 求每一信息样品的加权中位数权, 即综合权, 记为 $W(j)$

$$W(j) = WD(i, j) \times WC(j) \quad (22)$$

式中, $WD(i, j)$ 为信息样 i 对待估域 (或点) 的距离权;

$WC(j)$ 为信息样 j 的丛集权。

为保证估计的无偏性, 必须使

$$\sum_{j=1}^m W(j) = \sum_{j=1}^m WD(i, j) \times WC(j) \\ = 1 \quad (23)$$

所以, 综合权 $W^*(j)$ 为

$$W^*(j) = W(j) / \sum_{j=1}^m W(j)$$

应用图 1 及表 2 所列数据求出的各信息样的距离权、丛集权及综合权, 如表 4 所示。

6. 求待估域 (或点) 的加权中位数估值。

将估计邻域中所有信息样按观测值 (品

位) 从小至大排序, 并按原来的对应关系重新排列权函数, 在直角坐标系中, 以权函数为纵坐标, 以观测值 (品位) 为横坐标, 画出权的累积分布曲线, 求出权函数 $W=0.50$ 所对应的品位即为待估域 (或点) 的加权中位数估计值 (估计品位)。当应用图2及表1所列数据对待估块段A进行加权中位数估计时, 其加权中位数估值曲线如图4所示。从图3可知, 块段A的加权中位数估值为 $m=0.96$ 。

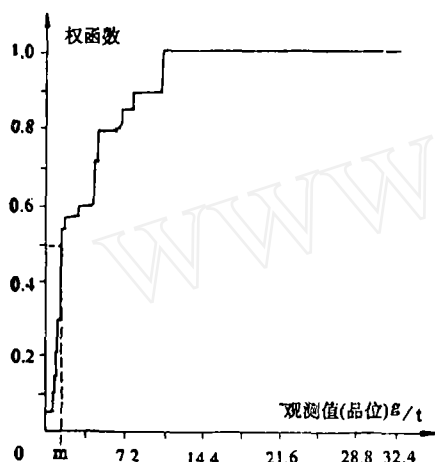


图4 加权中位数估值曲线

7. 若进行矿石储量计算时, 求出每一待估块段的加权中位数平均品位, 再计算其矿石量和金属量, 绘制采矿设计及矿山开采所需要的图件及表格。

加权中位数法在某金矿床储量计算中的应用

该矿床是含金石英复脉带型金矿, 围岩主要为变质岩。矿体是沿断裂裂隙带充填交代形成, 其走向330°, 倾向NE或SW, 倾角60~70°。矿体厚度变化大, 为1~60m。金主要呈自然金、金银矿和银金矿等产出, 所采90%样品中的Au均大于边界品位。

为应用加权中位数法计算矿床储量, 首先将全部样品换算成1m长的标准样, 经过

标准样的统计分布特征研究表明, 该矿床中Au品位呈对数正态分布。

金品位的变异函数可用球状模型拟合, 且在0°及90°方向各不相同。

$$0^\circ \text{方向上 } \gamma(h) = 115 + 44.5 \times \left[\frac{3}{2} \left(\frac{h}{6.2} \right) - \frac{1}{2} \left(\frac{h}{6.2} \right)^3 \right]$$

$$90^\circ \text{方向上 } \gamma(h) = 115 + 44.5 \times \left[\frac{3}{2} \left(\frac{h}{25} \right) - \frac{1}{2} \left(\frac{h}{25} \right)^3 \right]$$

呈现出几何异向性, 其各向异性比 $K=25/6.2=4.03$ 。

根据变异函数表明的金品位的变异性及信息样品的分布特征, 待估块段定为 $6 \times 6 \text{ m}^2$ 的正方形, 估计邻域为以待估块段中心点为圆心, 以6m为半径的圆形邻域, 应用上述的加权中位数法估计全部273个待估块

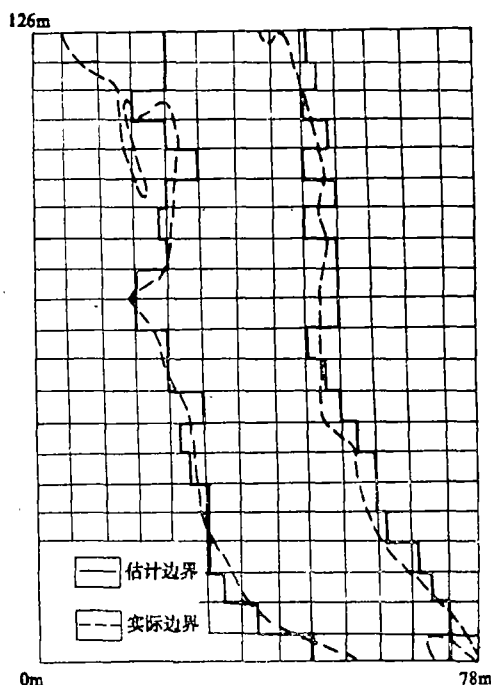


图5 加权中位数法估计矿体边界与矿体实际边界比较
(估计块段 $6 \times 6 \text{ m}^2$)

段的金平均品位,用加权中位数法计算储量圈定的矿体边界与实际开采的矿体边界相近(图5)。

结 论

1. 当处理地质、物化探数据或矿床品位值的分布律不服从正态分布呈现出长尾巴时,用经典系统统计方法估计样品点或估计域的平均值就会产生明显的偏差,影响估计精度,而加权中位数法则是一种良好的处理这种数据的非参数统计方法。

2. 加权中位数法的关键是正确的计算每一个信息样的距离权、丛集权和由此而形成的综合权系数。

3. 加权中位数法的优点在于该法无需假设数据来自某种特定分布的母体,也无需对原始数据进行变换,而且计算结果较为精确,计算方法也比指示克立格法简单。

参 考 文 献

- [1] 吴雨沛,有色金属(矿山部分),1990,第5期。
- [2] 侯景儒、张树泉、黄竞先,地质找矿论丛,1990,第3期,第93~102页。
- [3] 侯景儒,地质与勘探,1990,第3期,第28~36页。
- [4] 侯景儒、黄竞先等编译,《地质统计学的理论与方法》,地质出版社,1990年。
- [5] Henley Stephen, 1981, Nonparametric Geostatistics, Applied Science Publishers LTD.

Basic Theory, Methodology and Application of the Weighting Median Method

Hou Jingru Wu Yupei Xiang Yongsheng

Geological (including geophysical and geochemical) data usually have a non-normal distribution with their distribution curves characterized by a long tail. Such a distortion is frequently caused the presence of outliers in the data, a diversified mineralization types, or some other factors. In order to improve the quality of data processing and raise the precision of ore reserves calculation, a geostatistical method, the weighting median method, is presented. In this paper the basic theory and the computing methods of distance weighing, cluster weighing and sythetic weighing are introduced, and an example of ore-reserve calculation for a fold deposit is cited for illustration.

(上接第5页)

横断裂交叉可直接成为储矿构造。

3. 地层

中三叠统黄马青组下段富钙镁杂岩,是铁质交代最理想的对象。该岩性与下伏的周冲村组高硅铝灰岩的不整合面连通有矿液活动的断裂时,含矿岩浆可远源运液交代质地疏松的富钙镁杂岩成矿。

4. 围岩蚀变

结构致密的砂页岩呈角岩化;高硅铝灰岩呈弱大理岩化;质地疏松的富钙镁杂岩,则呈明显蚀变分带;相对远矿为浅色带,近矿为深色带。金云母、透辉石、磷灰石、绿帘石、绿泥石等,是典型蚀变矿物。

本文撰写中,参阅了我队原始地质资料,同时也参阅了华东地科所、地科院地矿所的有关专题研究资料,在此一并致谢。

The Ore Controlling Factor of the Zhong-Jiu Iron Deposit

Li Zhangang

The Zhongshan-Jiulianshan iron deposit is a major one of the Zhong-Jiu iron ore field. Proceeding from a study of the geological features, the ore controlling factor and the metallogenic mechanism of the deposit are discussed.