第9卷 第4期 岩 矿 测 试 Vol. 9, No. 4
1990年12月 ROCK AND MINERAL ANALYSIS Dec. 1990

多参数统计定值模式及其在标样定值中的应用

罗 代 洪
地矿部岩矿测试技术研究所，北京，100037

摘要 本文用多参数统计定值的新模式进行标样定值。对离群值的处理提出了新方 案。采用11种统计参数对原始数据进行处理，并将这11种参数的算术平均值、中位值、几何 平均值、选择平均值、Hampel *M*估计值和主族众数的平均值作为最佳定值。在NCR TOWER-1632多用户高档微机*UNIX*操作系统下开发了多参数模式定值软件MPDPS，对 6 个新研制的标样进行了定值。

关键词：数据处理，标样定值，多参数模式

标准物质的研制要求多个实验室用多种分析方法对尽可能多的项目定值，数据处理的工作量很大。尽管目前对标样主成分的测定方法已较成熟,但作者在考察了GSR 7-12主成分的分析数据后发现约有 $^1/_3$ 的 主成分测试数据偏离正态分布（表1）。痕量元素的测定手续复杂，样品的分解及分离富集等环节都可能引入误差。再加上仪器检测中的各种影响因素，分析数据较离散，这就给标样的数据处理和定值带来了困难。

表 1 GSR 7-12 主成分测试数据的分布❶

成 分	GSR系列					
	7	8	9	10	11	12
Al_2O_3	+	+	+	+	+	−
CaO	+	+	−	+	+	+
CO_2	−	+	+	+	+	+
FeO	+	−	+	+	+	+
T Fe_2O_3	−	+	+	+	+	+
K_2O	+	+	+	−	+	−
MgO	−	−	+	+	−	+
MnO	+	+	+	−	+	−
Na_2O	+	+	+	+	+	−
P_2O_5	+	+	+	−	−	−
SiO_2	+	−	−	+	+	−
TiO_2	+	+	+	+	−	+

❶ +为正态，−为非正态

本文采用多参数统计定值模式以寻求适合各种数据分布模型的数据处理方法，利用计算机快速有效地求出较为准确的推荐值，以提高标样定值的效率及定值的准确性。

一、标样定值模式

在较早的标样研制中，国外有的研制者选定权威人士或某一可靠方法的测定数据直接定值。这种处理方法由于测定数据少，难以克服测定过程中的偶然误差，带有一定的片面性。

目前国内外普遍采用多个实验室用多种分析方法协同分析定值的模式。这种处理模式可克服上述定值方法中较为片面的缺点，但由于参加测试的实验室在仪器设备及技术水平上的差异等原因，使测试数据较为复杂；数据量也较大，增加了数据处理的工作量和定值的难度。较早采用多个实验室协同定值研制的标样一般用算术平均值定值，但对于非正态分布的测试数据，定值效果欠佳。O. H. Christie[1], P. J. Ellis[2,3], B. Lister[4] 等人先后提出用中位值，选择平均值，众数及几种强估计值作为标样定值，但究竟哪种参数更优越，一直存在争议[5]。S. Abbey[6]提出的选择实验室方法，认为实验室间的因素是协同分析中数据离散的主要原因，通过剔除不好的实验室数据来克服实验室间系统误差对定值的影响。该法需要主观判断的因素较多，不便于使用计算机处理

数据，其使用受到了限制。

地矿部研制的GSD 1-12，GSS 1-8及GSR 1-6按分析方法将数据集分为若干个子集，剔除差异较大的方法子集的数据后计算各子集的算术平均值、几何平均值、选择平均值及中位值，将其平均值作为分析方法的估计值，然后将各方法估计值平均作为定值[7]。这种处理方法涉及大量分析数据的剔除，没有充分利用可利用的数据，对定值的准确性会产生不利影响[8]。

二、多参数定值模式

本文在考察了上述的标样定值模式后针对标样分析数据量大，数据离散，分布复杂的特点，提出用多种参数统计定值的模式，以适应不同数据分布类型的定值要求。

多参数模式的计算步骤如下：

1．整理各实验室报出的数据，将数据按从小到大排序。

2．剔除离群值，采用Dixon和Grubbs法则判断离群值，保留只有一种方法判为离群的数据，剔除两种方法都判为离群的数据。

3．计算剔除离群值后数据集的以下11种统计参数：算术平均值、几何平均值、选择平均值、中位值、三均值、1/4修正值、Gastwirth中位值、众值、修正中位值、主族众数、Hampel M估计值。

几种统计参数的计算方法如下[9]：

选择平均值：原始数据集算术平均值一倍标准偏差内数据的算术平均值。

三均值$=0.5M+0.25(UQ+LQ)$

其中M为中位值，UQ、LQ分别为较大及较小数据端的四分位点数据值。

1/4修正值：剔除原始数据两端各1/4数据后剩余数据的算术平均值。

Gastwirth中位值$=0.4M+0.3(UT+LT)$

其中M为中位值，UT、LT分别为较大及较小数据端三分位点数据值。

$$\text{众值}=G+\frac{c(f_n-f_{n-1})}{(f_n-f_{n-1})+(f_n-f_{n+1})}$$

式中G为众值所在组的下限，f_n、f_{n-1}及f_{n+1}分别为众值所在组及其左右邻组的频数，c为组距（$n\geqslant 10$时计算众值）。

修正中位值：剔除与中位值相差较大的一个端值（若两端值与中位值之差绝对值相等，同时剔除），求剩余数据中位值，重复上述运算至只剩下一个数据，若剩两个数据，平均之。

主族众数：①剔除$x>|\bar{x}+ks|$及$x<|\bar{x}-ks|$之x值；②计算剩余数据之算术平均值$\bar{x}_1$及标准偏差s_1；③用$\bar{x}_1$代替$\bar{x}$，s_1代替s取新的k值重复上述计算并循环。当a）剩余数据为5个；b）剩余数据相等；c）循环20次时停止，剩余数据之算术平均值为主族众数。k的取值依次为4.0，3.0，2.6，2.3，2.0，1.9，1.8，1.7，1.6，1.5，1.4，1.3，1.2，1.1，1.05，1.02，1.01，1.005，1.002，1.001。

Hampel M估计值：

$$\varphi(x,a,b,c)=\operatorname{sgn}x\cdot\begin{cases}|x| & 0\leqslant|x|<a\\ a & a\leqslant|x|<b\\ \dfrac{c-|x|}{c-b}-a & b\leqslant|x|<c\\ 0 & |x|\geqslant c\end{cases}$$

求解下列方程求出T即为Hampel M估计值。

$$\sum_{i=1}^{n}\left[\varphi\left(\frac{x_i-T}{s_1},a,b,c\right)\right]=0$$

式中s_1为$|x_i-M|$的中位值，M为原始数据中位值；常数a，b，c的三种取值模式为：12A：a=1.2，b=3.5，c=8.0；17A：a=1.7，b=3.4，c=8.5；25A：a=2.5，b=4.5，c=9.5。

4．从11种参数中定出最佳值：得到的

11种参数值之间还有较大差异，尤其对测试数据分散，非正态分布的数据差异更大。为此应对这11个参数再进行统计处理，计算它们的算术平均值、中位值、几何平均值、选择平均值、Hampel M 估计值和主族众数，将这 6 个参数的算术平均值作为最佳估计值。

经上述处理得出的最佳估计值可克服由于数据分散，数据量少及分布复杂等不利因素给定值带来的偏差，具有较好的应用效果。

三、多参数模式的应用

作者在NCR TOWER-1632多用户高档微型计算机UNIX操作系统下开发了MPDPS数据处理及定值软件，为 6 个新研制的地球化学标准参考样 GSR7—12 进行了定值。参加测试定值的实验室共20个，报出测定数据23394个，定值项目438个，最后建议定为推荐值412个，定为参考值26个。这6个标样已通过有关部门审查，将作为国家一级标准物质。

多参数统计定值模式的定值数据与其它定值方法的数据对比列于表 2。表 3 为这几个元素的测定数据。从表 2 的数据可以看出，对于服从正态分布的测试数据，本法的定值数据与证书的定值数据十分吻合，而对于非正态分布的测试数据，两种定值方法的定值数据有所差异。对原始数据端值的剔除方法不同，也是导致这些差异的原因之一。如表3，GSR-3 Sb 的定值数据，证书定值中将0.27作为离群值剔除，而本法则没有剔除这一数据。

表 2　两种定值方法的比较

样品	元素	分布	证书定值			本法定值		
			BV (ppm)	n	s	BV (ppm)	n	s
GSR-3	Sb	−	0.083	23	0.049	0.079	24	0.060
GSR-3	Tl	+	0.12	16	0.08	0.12	16	0.08
GSR-4	Hg	+	8.4	19	4.1	8.4	19	4.1
GSR-6	Tl	−	0.36	16	0.19	0.34	16	0.19
GSS-1	Cl	−	78	12	41	74	12	41

注：+为正态，−为非正态，BV为推荐值，n为数据数，s为标准偏差

表 3　原始数据表[4]（ppm）

GSR-3		GSR-4	GSR-6	GSS-1
Sb	Tl	Hg	Tl	Cl
0.03	0.014	1.3	0.17	39.6
0.04	0.026	3.8	0.22	50.5
0.049	0.033	6.0	0.25	53.0
0.05	0.060	6.7	0.25	60.0
0.05	0.070	7.0	0.28	71.3
0.052	0.080	7.1	0.28	72.2
0.053	0.11	7.2	0.29	75.0
0.060	0.12	8.18	0.35	80.0
0.07	0.13	8.4	0.35	84.6
0.07	0.14	8.4	0.35	123
0.07	0.18	8.8	0.37	159
0.08	0.18	9.0	0.44	164
0.08	0.20	9.0	0.46	
0.081	0.21	9.2	0.70	
0.098	0.22	10.0	0.74	
0.1	0.32	11.0	0.80	
0.1	0.70	11.5		
0.11		16.0		
0.13		19.0		
0.14				
0.18				
0.2				
0.2				
0.27				

四、讨论

1．关于离群值　在样品均匀的前提下，数据分散程度体现了现有分析方法的技术水平。过多地剔除离群值（端值）以追求剩余数据的一致性，不仅未充分利用可利用的原始数据，而且还掩盖了原始数据的离散性，对标样定值无任何益处。作者认为应尽可能少剔除端值数据，以保证定值数据的准确。

2．关于不确定度　对于正态分布的总

体，期望值的置信区间（不确定度）的计算已有定论。但对于非正态分布的总体，由于不知其分布模式，不确定度无法计算，如果套用正态总体不确定度的方法计算，反而显得牵强附会。建议对标样定值时只给出推荐值、数据个数、标准偏差。对于正态总体用算术平均值作为推荐值时才给出不确定度。

参 考 文 献

〔1〕 O. H. Christie et al. Geostandards Newsletter Vol.1, 47, 1977.

〔2〕 P. J. Ellis, Geostandards Newsletter Vol. 6,207, 1982.

〔3〕 P. J. Ellis et al, Geostandards Newsletter Vol.1, 123, 1977.

〔4〕 B. Lister, Geostandards Newsletter Vol 8, 171, 1984.

〔5〕 S. Abbey, Geostandards Newsletter Vol 5, 13, 1981.

〔6〕 S. Abbey, R. M. Rousseau, Geostandards Newsletter Vol.9, 1, 1985.

〔7〕 地球化学标准参考样研究组，《地球化学标准参考样的研制与分析方法》，地质出版社，1987。

〔8〕 R. Dybczynski, Analytica Chimica Acta, Vol.117, 53, 1980.

〔9〕 B. Lister, Geostandards Newsletter Vol. 6,175,1982.

（收稿日期：1990年4月23日）

A Multi-parameter Data Processing Model for Derivation of Composition Values of RMs

Luo Daihong

(Institute of Rock and Mineral Analysis, Ministry of Geology and Mineral Resources, Beijing, 100037)

A software system MPDPS, dedicated to NCR TOWER-1632 computer, was devised which allows the data from collaborative laboratories to be processed to obtain arithmetic mean, geometric mean, preferance mean, median, trimean, 25% trimmed mean, Gastwirth median, mode, modian, dominant cluster mode and Hampel M-estimate. All the estimators are again processed to obtain arithmetic mean, geometric mean, preferance mean, median, dominant cluster mode and Hampel M-estimate, and their arithmetic mean is assigned for the certificate value for an element. Composition values of six newly-prepared geochemical RMs (GSR 7-12) were derived by this multiparameter data processing model.

Key words: data processing, derivation of composition values of RMs, multiparameter model