利用Excel实现R型聚类分析

春 乃 芽
(辽宁有色葫芦岛地质勘查院,辽宁 葫芦岛 125000)

摘 要:R型聚类分析是对若干个元素进行数量化相似程度分类的一种数理统计方法,主要步骤包括:原始数据转换;求解相关系数;对结果聚类。利用Excel的数据分析工具实现R型聚类分析的方法和步骤,对野外一线地质人员的工作相当适用。

关键词:Excel; 数据分析;相关系数; 显著性检验
中图分类号: P632 **文献标识码**: A **文章编号**: 1000-8918(2007)04-0374-03

化探工作中,聚类分析可提供数量化的衡量元素或样品相似程度的指标,利用这些指标可以将元素或样品划分为不同的类别,从而揭示元素或样品之间本质上的联系,分析元素的共生组合和对岩体异常等的分类评价。聚类分析一般分为2种:R型聚类分析(对元素分类)和Q型聚类分析(对样品分类)。笔者介绍运用Excel数据分析工具实现R型聚类分析的方法和步骤。

利用Excel数据分析工具实现R型聚类分析的操作流程为:①加载分析工具库;②原始数据输入;③数据转换;④求解相关矩阵; ⑤聚类分类;⑥分类结果解释。

以参考文献[1]的原始数据为例,介绍上述流程。

1 加载数据分析工具库

缺省的Windows并不安装数据分析功能,需要重新加载,步骤如下:工具栏→工具→加载宏→分析工具库→确定。

2 原始数据输入

为了快速而准确地输入原始数据,除了按照正常的Excel方法输入数据以外,可以设定"工具-语音"选项,1个数据输入完毕之后,按Enter键即可语音朗读,实现数据输入的同步检查,确保其准确性。例子的原始数据如表1所示。输入数据应按行输入字段名(元素符号),相同一列按行输入同一元素的数值,所有数据输入完毕之后,所有字段名之下的其他行单元格内不再输入任何内容,以保证在5.2步用Count()函数求样本数 n 时不发生错误。在第4步求解相关矩阵时应选择逐列,表明不同的列代表不同元素的数据。

表1 几种元素的原始数据 10^{-6}

Ni	Co	Cu	Cr	S	As
1903	273	160	1178	8163	4
2328	79	6	3175	586	14
744	26	1	841	425	3
2782	273	150	2400	8234	37
1775	94	13	3140	54	1
1046	44	6	2093	104	4

3 数据转换

一般认为岩石当中常量元素服从正态分布,而其他微量元素多为对数正态分布,而且数据过于离散[1](这是地质数理统计的一个重要前提条件,利用"数据分析-描述统计"当中的偏度/峰度,依据文献[1]所介绍的方法进行检验,笔者直接使用文献[1]的数据,未做检验),所以要将其转换为常用对数。

选择单元格"J3",在公式栏中输入"=Log10(J3)"之后按Enter键,重新选择单元格"J9",将鼠标放在该单元格的右下角的复制控点上,鼠标变成黑色实心"+"形状,按住鼠标左键将其拖拽至"J8",完成Ni列数据的转换;重新选择"J3",以同样方法拖拽至"O3",完成第1行数据的转换,选择"J4"将其拖拽至"O4",完成第2行数据的转换,依此类推完成所有数据转换(表2)。

4 求解相关系数

选择工具栏→工具→数据分析→相关系数→确

收稿日期:2005-11;修回日期:2006-03

表 2　元素含量的对数数据

Ni	Co	Cu	Cr	S	As
3.27944	2.43616	2.20412	3.0711	3.91185	0.60206
3.36698	1.89763	0.77815	3.5017	2.7679	1.14613
2.87157	1.41497	0	2.9248	2.62839	0.47712
3.44436	2.43616	2.17609	3.3802	3.91561	1.5682
3.2492	1.97313	1.11394	3.4969	1.73239	0
3.01953	1.64345	0.77815	3.3208	2.01703	0.60206

定,显示相关系数对话框(图 1),在输入区域中输入"J2:O8",分组方式选择逐列,选择标准位于第一行,在输出区域输入"Q2:W9"(可选择新工作表组,较为简洁,可以通过粘贴把相关系数矩阵与原始数据放在同一 SHeet 内,便于数据的对比),按确定即可得到关系矩阵(表 3),结果与参考文献[1]完全相同。

图 1　相关系数对话框

表 3　相关系数

	Ni	Co	Cu	Cr	S	As
Ni	1					
Co	0.84585	1				
Cu	0.75761	0.97997	1			
Cr	0.64302	0.24197	0.18136	1		
S	0.4979	0.72799	0.71246	-0.3039	1	
As	0.56015	0.42401	0.39289	0.19984	0.67218	1

5　聚类分析

R 型聚类分析是以相关系数为基础进行的元素分类,必须对相关系数进行显著性检验。在样本数一定的情况下,是否显著相关与显著水平 α 的大小有关。一般情况下,α 值越大,相关元素个数会越多。笔者参照参考文献[2-8],依据显著水平 α 的大小进行显著性检验,将在某一置信度 α 之下显著相关的元素归为一类,逐渐增大显著水平 α 值,将显著相关的元素逐一归类并画出谱系图。

5.1　显著性检验——r_α 检验

求得相关系数 r 后,按如下方法判断显著性:如果 $|r| > r_\alpha$ 则表明元素之间相关,可以归类,否则,元素之间没有关联。$r_\alpha = t_\alpha / \sqrt{t_\alpha^2 + (n-2)}$,其中,$t_\alpha$ 为利用自由度 $n-2$ 的 t 分布求得的(双尾);n 为样本数。

5.2　利用 Excel 进行 r_α 检验和元素归类

在 Y1 单元格内输入"=COUNT(B3:B100)",求得样本数 $n=6$(某一个元素的样品个数),在 Y2 单元格内输入"=Y1-2",求得自由度 $n-2=4$,在 X4 至 X8 单元格内分别输入显著水平 α 不同的值,在 Y4 单元格中输入"=TINV(X4,Y2)/SQRT((TINV(X4,Y2))^2+Y2)",求得 $\alpha=0.025$ 条件下的 $r_\alpha=0.867\,96$,选中 Y4 单元格内容,拖拽其右下角的复制控点,将其内容复制到 Y5:Y8 单元格,将 Y5 单元格公式中的 Y3 改为 Y2,Enter,求得显著水平 $\alpha=0.05$ 条件下的 $r_\alpha=0.811\,401$,依次类推分别求得显著水平 $\alpha=0.075$、0.10 和 0.15 条件下的 r_α 值(表 4)。

表 4　不同 α 条件下的 r_α

α	r_α
0.025	0.867962
0.05	0.811401
0.075	0.767176
0.1	0.729299
0.15	0.66445

在显著水平 $\alpha=0.025$ 即 $r_\alpha=0.867\,96$ 条件下,相关系数满足显著性检验公式 $|r| > r_\alpha$ 的只有 Co 和 Cu,即两者相关性极其显著,可归为一类;在显著水平 $\alpha=0.05$,即 $r_\alpha=0.811\,401$ 条件下,Ni 与 Co、Cu 显著相关且与 Co 的关系更近一些,归为一类;在显著水平 $\alpha=0.15$ 条件下,As、S 相关性显著且 S 与 Co、Cu 的关系更近些,可以归为一类。在显著水平 $\alpha \leqslant 0.15$ 条件下,Cr 与上述元素相关性不显著,单独归为一类。上述分类按先后顺序绘制成变量分群谱系图(图 2)。值得一提的是,虽然 Cr 与 S 的相关系数 $-0.303\,9 < 0$,但由于 $|-0.303\,9| < 0.664\,5$,从显著性检验公式来判断,在显著水平 $\alpha \leqslant 0.15$ 条件下并不能就此可以确定两者是负相关的。

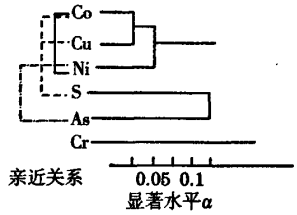

图 2　变量分群谱系

在绘制变量分群谱系图时,参考文献[1]是以相关系数为横轴来归类的,笔者以置信水平进行分

类,似乎更易理解。例如在显著水平 $\alpha=0.05$ 条件下,Ni 与 Co、Cu 显著相关,就可以理解为:在测试条件下,有95%的把握认为 Ni 与 Co、Cu 关系极为密切,类似地,可以说有85%把握判定 As、S 与 Ni,Co、Cu 关系密切。

6 统计结果的解释

任何一个数理统计结果必须得到合理的解释才能对实践有指导意义。上述实例聚类分析之后可以判定:6个元素在置信水平较小的情况下($\alpha=0.05$),可以划分为3组 Co、Cu、Ni,As、S 和 Cr,其中 Cu、Co 高度相关,相对而言 Ni 与 Co 的关系更密切些,而 S 与 Cu、Co 的关系较之 Ni 更为亲密,亲氧元素 Cr 与其他5个元素关联并不显著,而这些现象符合一般意义上的地质和化探规律,所以 R 型聚类分析所获得的结果是可取的。

7 结语

Micro Office XP Excel 具有强大的数理统计功能,在许多方面非 Access 等常用数据库可比,多数的地质数理统计都可以直接或间接地得以实现,特别是对小数据量的情况尤为合适,例如化探数据的正态或对数正态检验,背景值的确定,异常走向的判断(F 检验),样品化验数据误差的检验(F 检验),一次趋势面分析等等都可以用 Excel 来实现,这些操作虽然略显繁琐,但在大众化地质行业软件欠缺的情况下,依然不失为较为理想的选择,对一线地质人员较为适用,值得了解和掌握。数据分析属 Excel 的高级用法,具备相关的多元统计学知识和 Excel 的熟练操作,运用起来更会得心应手。

参考文献:

[1] 王崇云. 地球化学找矿基础[M]. 北京:地质出版社. 1986.
[2] 杨世莹. Excel 数据统计与分析范例[M]. 北京:中国青年电子出版社. 2005.
[3] 章晔,张文斌,范正国. 航空伽玛能谱测量数据分类图的自动编图系统[J]. 铀矿地质,1992,8(5):297.
[4] 赵荣军. 河南卢氏县杜关地区地球化学异常及找矿效果[J]. 物探与化探,2001,25(6):447.
[5] 王硕儒,葛宗侠. 鄂东南中酸性小岩体含矿性评价的模糊聚类法[J]. 物探与化探,1990,14(1):63.
[6] 杜光伟,徐开锋. 藏东"三江"地区地球化学特征及其找矿意义[J]. 物探与化探,2001,25(6):425.
[7] 胡远来. 大样本模糊聚类的快速计算法及应用[J]. 成都地质学院,1998,15(1):
[8] 赵玉琛. 多功能聚类分析程序[J]. 物探化探计算技术,1991,13(1):81.

THE UTILIZATION OF EXCEL TO THE PERFORMANCE OF R-MODE CLUSTER ANALYSIS

CHUN Nei-ya

(*Huludao Institute of Geological Exploration, Liaoning Nonferrous Metals Company, Huludao* 125000, *China*)

Abstract: The R-mode cluster analysis is a mathematic statistical method for obtaining the quantitative similarity of several elements. Its procedure includes: the conversion of the original data; the solution of the relevant coefficient ; the clustering of the result. The above operation can be realized by using the data analysis tool of Excel. This method is quite suitable for field utilization.

Key words: Excel; R-mode cluster analysis; data analysis; relevant coefficient; significance test

作者简介:春乃芽(1969-),男,高级地质工程师,长期在辽宁省西部地区从事野外地质找矿工作。