文章编号: 1001—1749(2006)03—0272—05

如何利用 Excel处理化探数据

春乃芽
(辽宁有色葫芦岛地质勘查院, 辽宁 葫芦岛 125000)

摘 要: 利用 Excel可以实现对化探数据多种项目的处理,如正态性检验;背景值及异常下限的确定;一元及多元回归分析;相关系数(R型聚类分析)以及一次趋势分析等,操作直观,目前在没有流行地质行业软件的情况下,不失为一种较为可取的选择。通过这种方法,可以较为简洁地实现对一定数量化探数据的处理,为地质找矿工作提供了化探方面的可靠依据。

关键词: 化探数据处理; Excel; 数据分析
中图分类号: TP 274 **文献标识码**: A

0 前 言

在化探找矿工作中,大量数据的处理是一项重要的前期基础工作,数据处理的及时、准确与否,将直接涉及到下一步勘查工作的部署。作者在本文中以 Microsoft Office XP Excel工具~数据分析功能为工具,简述了这些功能在化探数据处理当中的一些应用。

1 准备工作

缺省的 Windows Professional XP一般不安装用来数据处理所需要的数理统计功能,故需重新加载,步骤如下:工具 -加载宏 -分析工具库 -确定。之后,还需检查在工具菜单下有无数据分析选项。在大量数据的录入过程中,可以设定工具 -语音选项。在某个数据输入完毕之后,按 Enter键即可语音朗读(需要音响或耳麦),这样可以实现数据录入的同步检查,确保录入数据的准确性。

2 数据的处理

2.1 正态性检验

数据是否服从正态分布(或对数正态分布)是化探数据处理的一个重要前提条件,因此必须先进行数据正态性检验,以文献[1]中 Cu对数数据为例,参照文献[1]的方法,用偏度和峰度二个指标来加以检验。步骤如下:工具 -数据分析 -描述统计 -确定。在"描述统计"对话框选项中,输入相关的项目,输入区域的数据要注意"分组方式"的选择,如果表格(即 sheet)的字段名是放在列上,即同一列为同一个元素数据,则选择"逐列",否则选"逐行",不能选错(见图 1)。

平均数置信度一般默认为 95%,单击确定即可获得相关的统计量(见下页表 1)。

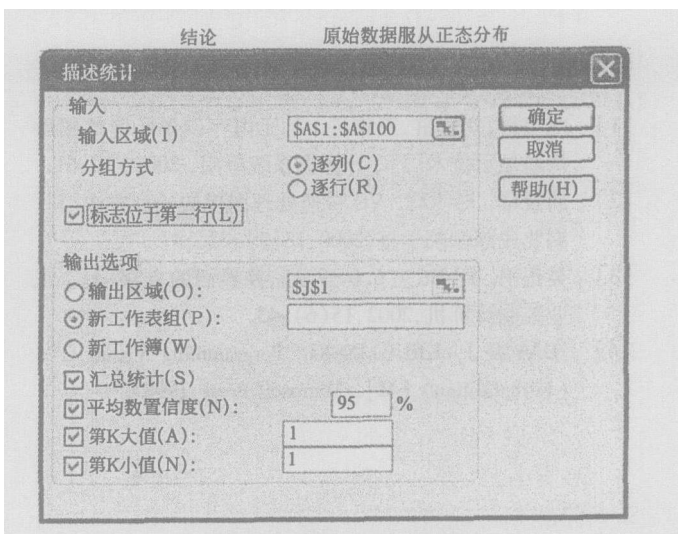

图 1 描述统计对话框

Fig 1 The dialog box of the the descriptive statistics

收稿日期: 2005 - 01 - 05

表 1　描述统计量

Tab. 1　The descriptive statistics

平均	0.906	区域	0.8
标准误差	0.020 588 317	最小值	0.5
中位数	0.9	最大值	1.3
众数	0.9	求和	90.6
标准差	0.205 883 168	观测数	100
方差	0.042 387 879	最大(1)	1.3
峰度	-0.360 475 696	最小(1)	0.5
偏度	-0.117 583 703	置信度(95.0%)	0.040 851 686

值得注意的是,一般认为微量元素在地质体当中是服从对数正态分布的,因此在检验之前,应将原始数据转换为对数(实例当中提供的数据为对数数据),方法是:首先选择某一个空单元格如 F1(假设第一个数据单元格为 D1),在公式栏中输入 =LOG10(D1),然后按"确定"键,再次选择该单元格 F1,用充填托柄将整个对数单元格填满即可,详细方法见文献[2]。

在大子样(样品个数 $n > 100$)和置信度 = 95%的二个条件下,如果偏度 $R1$ 和峰度 $R2$ 的绝对值同时满足如下条件,则数据服从对数正态分布,反之不然:

$$|R_1|^2 < 24/n$$

$$|R_2|^2 < 2 * 24/n。$$

在本例当中,计算之后可以得到:

$|R_1|^2 = (-0.117\,583\,703)^2 = 0.001\,38 < 0.24$; $|R_2|^2 = (-0.360\,475\,696)^2 = 0.01299 < 0.48$,二项要求均满足,故对数的数据服从对数正态分布。本例的数据为分组之后的数据,利用 Excel无需分组,故在计算中用组中值和组中值个数来代替未分组的原始数据(比如组中值 = 0.9,频数 = 40,则用 40个 0.9代替原始数据),所以描述统计量与文献的结果有差距。

2.2　背景值及异常下限的确定

背景值及异常下限的数学公式为

背景值 C_0 =数据的平均值

在背景区,异常下限 $C_A = C_0 + 2 \sim 3 * Sx$;

Sx =标准差

在异常区,背景值 C_0 =众值,异常下限 $C_A = C_0 + 2 \sim 3 *$"Sx",由于描述统计量是由对数数据求得的,因此还要反算回正常的数据值。

在本例当中,对数背景值 C_0 =数据的平均值 = 0.906,背景值 = $10^{0.906}$ = 8.05 ppm;对数异常下限 $C_A = C_0 + 2 \sim 3 *$ 标准差 = 0.906 + 2 * 0.205 883 168 = 1.318,异常下限 = $10^{1.318}$ = 20.8 ppm;

在异常地区,"Sx"应由众数频数的一半个数 d 和其之前的所有数据来求得。在本例中,众数 = 0.9,其频数 = 40,即用 40/2 + 20 + 8 = 48之前的数据来求"Sx",方法是重复描述统计或在公式栏中输入 = Sqrt(Var(数据区域))来求得"Sx" = 0.145 865(图 2)。

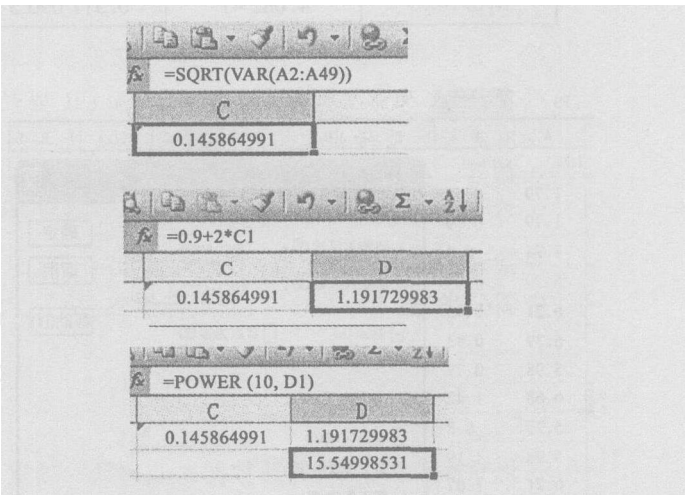

图 2　背景值及异常值计算公式输入栏

Fig. 2　Input box for the steps of calculating background value & anomaly threshold

异常区异常下限 $C_A = C_0 + 2 \sim 3 *$"Sx" = 0.906 + 2 * 0.145 865 = 1.19 173,异常下限 = 101.191 73 = 15.55 ppm。

2.3　一元及多元回归分析

回归分析多用来求解二个以上变量之间的数理统计上的关系,重要的是要检验所得到的回归方程是否显著。在输入数据后(文献[1]最好按列输入,Excel缺省的回归变量即列 = 16),操作步骤如下:工具 - 数据分析 - 回归 - 确定。在回归对话框选项当中,选择相应的选项,然后按"确定"键。在选择"输入区域"时,注意要包含字段名即 Y 和 X 的变量名 LnCu和 K_2O/Na_2O 并选择"标志"(见下页图 3),获得的统计变量如下页表 2所示。

在表 2中,有二个重要指标: Significance F和 P - value。如果 Significance F的值 < 0.05(0.05 = 1 - 置信度 95%),则回归方程线性关系显著,若否则不显著;截距 Intercept的 P - value < 0.05,则截距 ≠0,即截距存在,否则截距无法确定; K_2O/Na_2O 的 P - value < 0.05表明 LnCu与 K_2O/Na_2O 存在线性关系,但其显著性要由 Significance F来判定。

表 2 回归统计量表

Tab. 2 The regression analysis statistics

方差分析					
Significance F	df		SS	MS	F
回归分析	1	54.495 862	54.495 9	62.33	1.733 92 E-08
残差	27	23.606 262	0.874 31		
总计	28	78.102 124			
	Coefficients	标准误差	t Stat	P-value	Lower 95%
Intercept	1.549 11	0.530 275 8	2.921 34	0.007	0.461 077 278
K_2O/Na_2O	4.082 41	0.517 090 5	7.894 96	2E-08	3.021 427 666

图 3 “回归”对话框

Fig. 3 The dialog box of the regression analysis

在本例中的方程回归显著,变量系数 =4.082 41,截距 =1.549 11。

最终的回归方程: $LnCu = 4.082\ 41\ K_2O/Na_2O + 1.549\ 11$,与原文献的回归方程完全一致,实际应用中可以用 K_2O/Na_2O 的含量去预测 Cu的含量,或反过来也可以。

事实上 Excel的回归功能 并不局限在一元回归上,它可以求解变量 (列)小于 16的多元回归方程,显著性判断的方法与上述方法相同。另外,置信度也是依据实际情况来确定的,在多数情况下取 95%或 90%。具体操作方法详见参考文献[2]。

2.4 相关系数(R型聚类分析)

R型聚类分析主要是求解各个元素之间的相关矩阵,然后依照 F聚类原则进行聚类。现将文献[1]中的原始数据录入到表格 3的 C4: H9单元格当中,并进行聚类分析。

2.4.1 数据的转换

一般认为微量元素多为对数正态分布 (这是地质数理统计的一个重要前提条件),所以要将其转换为以 10为底的 Log对数形式。选择单元格 C10,在公式栏中输入 =Log10(C4)之后,按 Enter键,重新选择单元格 C10,拖动‘充填托柄’至 H15,即可实现全部数据的转换,如表 3 C10: C15单元格。

2.4.2 对数数据的标准化

将对数数据标准化的目的是为了将差距较大的数据转换在同一度量的水平上,标准化数据的公式 =(某个元素的某个对数数据 - 该元素的全部数据的平均值)/该元素的全部数据的标准离差,操作步骤如下。

(1) 平均值。选择 C16,在公式栏中输入 =AVERAHE(C10: C15),按 Enter键。再次选择单元格 C16,拖动‘充填托柄’至 H16,即可得到全部数据的平均值,如下页表 3中 C16: H16单元格。

(2) 标准离差。选择 C17,在公式栏中输入 =VAR(C10: C15),按 Enter键。再次选择单元格 C17,拖动‘充填托柄’至 H17,即可得到全部数据的方差,如表 3中 C17: H17单元格。同样,选择 C18,在公式栏中输入 =SQRT(C17),按 Enter键。再次选择单元格 C18,拖动‘充填托柄’至 H18,即可得到全部数据的标准离差,如表 3的 C18: H18单元格。

(3) 标准化数据。选择 C22,在公式栏中输入 =(C10 - C16)/C18,按 Enter键;选择单元格 C23,在公式栏中输入 =(C11 - C16)/C18,按 Enter键。选择 C24在公式栏中输入 =(C12 - C16)/C17,按 Enter键。依此类推,将 C25: C27单元格填满,即将公式当中的 C列单元格依次改为 C25～C27后按 Enter键即可。选择单元格 C22: C27,拖动‘充填托柄’至 H27即可实现全部数据的标准化,如下

页表 3中 C22: H27单元格。

2.4.3　求解相关矩阵

选择工具栏 -工具 -数据分析,在其选项当中选择"相关系数",按"确定"键,如下页图 4所示。

在"输入区域"中输入 C21: H27单元格,分组方式选择"逐列",选择"标志位于第一行","输出区域"输入 C29: H35(也可选择"新工作表组",这种选择较为方便简洁,可以通过"粘贴"把关系矩阵与原始数据放在同一 Sheet内,便于数据的对比),按"确定"键即可得到关系矩阵,如表 3中的

表 3　聚类分析原始数据与计算结果

Tab. 3　The raw data & the result of the data processing

1	B	C	D	E	F	G	H
2		元素平均含量(ppm)					
3		Ni	Co	Cu	Cr	S	As
4		1 903	273	160	1 178	8 163	4
5		2 328	79	6	3 175	586	14
6	原	744	26	1	841	425	3
7	始	2 782	273	150	2 400	8 234	37
8		1 775	94	13	3 140	54	1
9		1 046	44	6	2 093	104	4
10		3.279	2.436	2.204	3.0711	3.912	0.602
11		3.367	1.898	0.778	3.501 7	2.768	1.146
12	对	2.872	1.415	0	2.924 8	2.628	0.477
13	数	3.444	2.436	2.176	3.380 2	3.916	1.568
14		3.249	1.973	1.114	3.496 9	1.732	0
15		3.02	1.643	0.778	3.320 8	2.017	0.602
16	平均	3.205	1.967	1.175	3.282 6	2.829	0.733
17	方差	0.047	0.171	0.752	0.055 5	0.852	0.301
18	离差	0.218	0.413	0.867	0.235 7	0.923	0.549
19							
20							
21		Ni	Co	Cu	Cr	S	As
22		0.341	1.135	1.187	-0.897	1.173	-0.24
23	标	0.744	-0.17	-0.46	0.9299	-0.07	0.754
24		-1.53	-1.34	-1.35	-1.518	-0.22	-0.47
25	准	1.1	1.135	1.154	0.414 2	1.177	1.523
26		0.202	0.015	-0.07	0.909 5	-1.19	-1.34
27		-0.85	-0.78	-0.46	0.162	-0.88	-0.24
	相关系数矩阵						
29		Ni	Co	Cu	Cr	S	As
30	Ni	1					
31	Co	0.846	1				
32	Cu	0.758	0.98	1			
33	Cr	0.643	0.242	0.181	1		
34	S	0.498	0.728	0.712	-0.304	1	
35	As	0.56	0.424	0.393	0.199 8	0.672	1

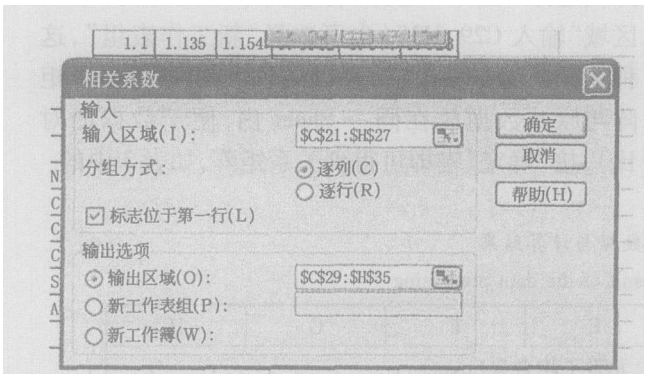

图 4 相关系数对话框

Fig 4 The dialog box of the correlation analysis

C29: H35单元格,所得到的相关矩阵与参考文献[1]完全相同。

2.4.4 聚类分析

依照 F模糊聚类原则,在无需证明关系矩阵的自反性、对称性和传递性的前提下,可以依据直接聚类法或编网法,通过取不同的 λ来得到不同水平下的分类。

例如,利用编网法,λ = 0.7,将单元格中小于0.7的数值删掉,将含有数值的单元格视为行和列的"结点"。通过结点可以相互连接起来的元素为同一类别,并可分为三类: Cr - As - S、Cu、Co、Ni,见表 4。依此类推,λ值分别取 0.8、0.6,可以得到不同的类, Co、Cu - Ni - As - S - Cr和 Co、Cu、Ni、As、S - Cr,分类结果与原参考文献的结论基本相同,详细分类方法见参考文献[3]。

表 4 λ = 0.7水平下的相关性分类

Tab.4 The result of cluster analysis as λ = 0.7

Ni	Co	Cu	Cr	S	As	Ni
Co	0.846	1				
Cu	0.758	0.98	1			
Cr				1		
S	0.728	0.712			1	
As					1	

3 结束语

Micro Office XP Excel具有强大的数理统计功能,多数的地质数理统计都直接或间接地可以用 Excel来实现。这些操作虽然略显繁琐,但在没有大众化地质行业软件或程序的情况下,这种选择仍不失为一个较为理想的处理问题的方法,特别是适合于生产单位和广大一线地质人员,容易了解和掌握。但是,这些数理统计功能属 Excel的高级用法(推荐参考文献[2]及其系列书籍),要了解相关多元统计学的基础知识和 Windows的熟练操作,这样才能运用自如。

参考文献:

[1] 王崇云,陶正章,马超洁,等. 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]. 中国青年电子出版社. 2005.

[3] 杨纶标. 高英仪. 模糊数学原理及应用(第三版)[M]. 广州五山: 华南理工大学出版社, 2002.

[4] 黄文清. 周子勇. 利用 Mathcad数学软件进行物化探数据处理[J]. 物探化探计算技术. 2004, 26(1): 66.

[5] 时艳香. 纪宏金. 利用水系沉积物地球化学数据判别浅覆盖区岩性与构造——欧氏距离法[J]. 物探化探计算技术. 2004, 26(3): 243.

[6] 傅水兴. 祝新友. 走向地学新世纪——首届有色系统青年地质工作者学术讨论会论文集[C]. 北京: 冶金工业出版社. 1995.

[7] 汪荣鑫. 数量统计[M]. 西安:西安交通大学出版社. 1986.

[8] 赵伦山,张本仁. 地球化学[M]. 北京:地质出版社. 1988.

作者简介: 春乃芽 (1969 -),男,内蒙古科左后旗人,地质高级工程师,长期在辽宁省西部从事野外地质找矿工作。

VC^{++}

SUN Sheng, NIU Bin-hua (School of Geophysics and Information Technology, China University of Geosciences, Beijing 100083, China). *COMPUTING TECHNIQUES FOR GEOPHYSICAL AND GEOCHEMICAL EXPLORATION*, 2006, 28(3): 0268

The propagation of seismic wave in a specific medium is a dynamic process with respect to time. We can study this process under the help of snapshots obtained from the arithmetical solution of wave equations. Projection of snapshots can resume the relationship between discrete data and time, so that it will be a great help to the extensive study of the propagation of seismic wave. A process of programming realization of the 2-D snapshot projection and methods of data loading, memory device imaging, and timer message processing are presented in this paper. After these presentations, it also discusses a possible format of snapshot data and another data loading and imaging method. The theory in this paper proves to be feasible after testing of an example program developed by MFC of Visual C^{++}.

Key words: snapshot; projection; memory device; timer events

THE METHOD PROCESSING THE GEOCHEMICAL EXPLORATION DATA WITH THE MICROSOFT EXCEL

CHUN Nai-ya (Liaoning HuLuDao nonferrous geological institute, Liao ning HuluDao 125000, China). *COMPUTING TECHNIQUES FOR GEOPHYSICAL AND GEOCHEMICAL EXPLORATION*, 2006, 28(3): 0272

A variety of geochemical exploration data can be processed in the Microsoft Excel, such as: test of normality; ascertainment of the background value & anomaly threshold; unity & multianalysis of regression; the correlation analysis or cluster analysis and trend analysis etc.. And the operation is easy. It can be one advisable choice in the case of unavailable of the geological professional software at the present time.

Key words: geochemical exploration; the microsoft excel; data processing

DESIGN AND REALIZATION OF BORE HOLE HISTOGRAM AUTOMATIC DRAWING SYSTEM BASED ON ARCGIS ENGINE

HUANG Hai-feng[1,2], XIA Bin[1], BAO Shi-tai[1,2], et al (1. Guangzhou Institute of Geochemistry Chinese Academy of Sciences, Guangzhou 510640, China; 2. Graduate School of Chinese Academy of Sciences, Beijing 100039, China). *COMPUTING TECHNIQUES FOR GEOPHYSICAL AND GEOCHEMICAL EXPLORATION*, 2006, 28(3): 0277

Engineering investigation information system based on GIS is now prevalent with the increasing application of GIS technology in engineering investigation. This paper discusses bore hole histogram automatic drawing system, which is a vital part of engineering investigation information system. Based on the ArcGIS Engine, which is a new GIS software development platform in ArcGIS 9.0, the paper first discusses the design of "template" and system, and then expound the realization of bore hole histogram drawing system, and last summarize some techniques of ArcGIS Engine.

Key words: GIS; ArcGIS engine; bore hole histogram; template; engineering investigation

IBM RATIONAL SOFTWARE CONFIGURATION MANAGEMENT (SCM)

SHAN Yan-ming, WU Jun (Exploration and Development Research Institute of Daqing Oilfield Limited Company, Daqing, Heilongjiang 163712, China). *COMPUTING TECHNIQUES FOR GEOPHYSICAL AND GEOCHEMICAL EXPLORATION*, 2006, 28(3): 0282

This paper explains the functions, running environment and management strategies of IBM Rational SCM. With integrated version control, automated workspace management, parallel development support, process control, and build and release management, distributed development support, change request management, it provides the capabilities needed to create, update, build, deliver, reuse and maintain business-critical software assets. It is a powerful tool of the software project management.

Key words: SCM; version control; workspace management; parallel development support