勘查数据分析（EDA）技术的应用

史长义

（地矿部物化探研究所·河北廊坊）

勘查数据分析（EDA）技术是一种新的处理数据的非常规统计学方法。本文利用 Boxplot 图示，通过对山西和新疆某地区的区域地球化学数据的分析，以 Boxplot 对比为基础确定成矿有利地质体、研究异常元素组合和进行异常评序。结果表明，该方法简单、实用，并能抵抗特高值的影响。

关键词 勘查地球化学　异常解释推断　数据处理　EDA 技术

引　言

目前，化探数据的处理与解释主要依据经典统计学方法，如各种常规单变量和多变量统计方法。这些经典统计学方法都具有很严格的假设条件，即假设数据服从正态分布。而化探数据的复杂性使得它很难满足这种假设，极端偏离数据主体的少数测量值或小额量总体可明显地影响估计量，至少部分掩盖数据的固有信息（Kürzl,1988）。为此，Tukey(1977) 提出了一种新的数据处理方法——勘查数据分析技术，即 EDA(Exploratory Data Analysis) 技术。

EDA 技术是一种处理数据的非常规统计学方法，它利用稳健统计学，并引入各种简单而有效的图示技术，从中可迅速看出数据的结构和特点。它不需要任何假设条件，而是根据数据本身所固有的模型来识别异点（outlier），由此来确定背景总体和异常总体。利用这些特征我们可以直接解释原始数据（史长义，1990）。业已证明，在单元素地球化学数据的描述和分析中，EDA 技术非常有效。EDA 技术包括：① 5 参数综合法；② 框图；③ 密度轨迹；④ 一维散点图；⑤ 分位数图（Kürzl,1988）。

异常解释推断是当前化探工作者的重要课题，探索和引入新的方法，是解决这一难题的重要一环。本文引入 EDA 技术，以山西某老变质岩区和新疆某干旱景观的中生代火山活动区为例，主要讨论 EDA 技术中的 Boxplot 图示法在区域化探数据分析和解释推断中的应用。

5 参数综合法和 Boxplot（框图）简介

这是一组选择顺序统计量的方法，可用来描述单一数据集的最重要特性。为了找出这 5 个参数，首先将原始数据按由大到小的顺序排队，然后从中找出中位数和上、下 4 分点，即上、下节点（史长义，1990）。这样，按顺序排列的一批数据的中值、两个节点和两个端点值就构成了 5 参数综合法，其图解就是 Boxplot(图 1)。从 Boxplot 上我们可以迅速发现一批数据的下列特征：① 位置；② 散度；③ 偏度；④ 尾长；⑤ 外围数据点数。

框图的画法很简单。在上、下节点之间划一框，这个框包含了 50% 的数据点。其中的中位数以一横线来代表，为更明显起见也可在横线上加一个“×”，构成一个星号。这个框描述了经验分布的内散度（或 $h-$ 散度）。框内中值的具体位置表示分布

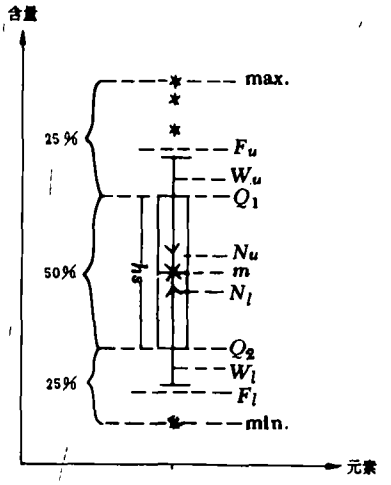

图 1　Boxplot 示意图

m－中位数；N_u－上切口；N_l－下切口；Q_1－上节点；Q_2－下节点；W_u－上支杆；W_l－下枝杆；F_u－异点下限；F_l－异点上限；hs－内散度；max－最大值；min－最小值；*－表示异点

的中心（主体）部分的对称性和偏度。此外，还可以用切口（N_u,N_l）代替“×”表示中位数，在 0.05 置信水平上，切口可提供一种不同中值之间差异性的大致估计值。切口在图上可标可不标。外围数据点的特性用支杆（W_u,W_l）来说明，每一支杆代表节点和端点之间 25% 的数据。两个支杆一直分别扩展到 F_u 和 F_l 之内的两个极值点。F_u、F_l 叫做异点临界值，$F_u=Q_1+S$，$F_l=Q_2-S$，其中 $S=1.5\times hs$。位于 F_u、F_l 之外的所有数据点即为异点，图 1 中以星号表示。其中 F_u 以上叫做上异点（相当于正异常），F_l 以下为下异点（相当于负异常）。这种框图在相当程度上强化了对单一数据集上述特性的直观评价，而且它的一个非常重要的特性是能够抵抗干扰和“野”数据（当然它们也可能是异常）的影响。就是说，25% 的数据点可能是“野”的，这些“野”值对中位数和节点不产生明显影响。而异点临界值仅由内散度确定，所以，它就不会受异点的影响。与稳健统计量相反，只要有一个极值点就会给算术平均值和标准离差带来很大的影响。

研究微量元素区域分布规律

元素的集中与分散，与地质背景密切相关。利用各地质单元中元素的含量变化特征，可以分析元素的时空分布规律，探讨元素分布与地质背景的关系。

用 Boxplot 对比不同数据子集，是一种简单而非常有效的方法。从中既可鉴别出相似性，又可识别出差异性。较之常规的单一参数（如均值 $\bar{x}$、离差 s、变异系数 C_V 等）方法更直观、客观，更具优势。图 2 示出了 4 个不同数据子集间 Ni 的框图对比结果。从中可清楚地看出 2、3 组元素含量的数据结构明显地与 1、4 组不同。

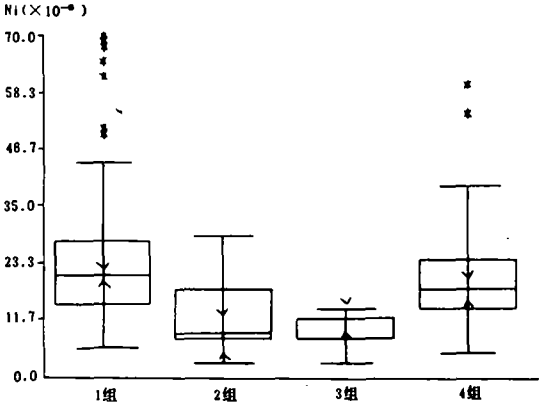

图 2　一个元素的 4 个不同数据子集间的框图对比

（据 Kürzl, 1988）

利用框图对比研究不同地质单元中元素的分散与集中趋势，主要依据 Boxplot 的结构特性，包括：① 异点数多少；② 中位数高低；③ 内散度大小。异点数多，中位数高，内散度大，表示集中趋势；反之，则表示分散趋势。当然，这种分散与集中，只是一种相对趋势，并没有给定任何一个临界值。上述特征从 Boxplot 上均可直观看出。图 3 为山西某地区各地质单元内，水系沉积

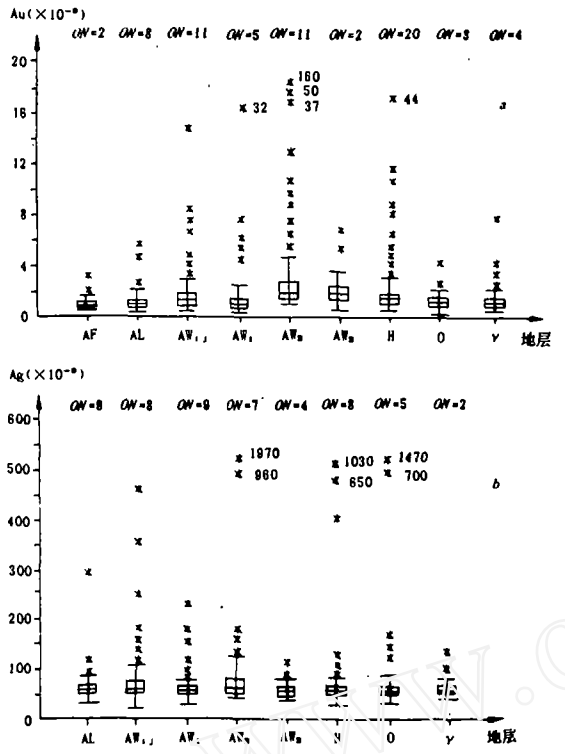

图 3　山西某地区各地质单元中 Au、Ag 的 Boxplot 对比图

(*ON*= 异点数)

物中 Au、Ag 含量的 Boxplot 对比图。由图 3*a* 可知（图中 *ON* 数与星号数不符是因为有相等量值的异点重叠造成的，下同），在 AW_2、AW_{1j} 和 H 单元中，Au 的异点数、中位数和内散度均较高，呈相对集中趋势。在 AF 和 O 单元中，Au 的异点数、中位数和内散度均很低，呈相对分散趋势。在 AL 中，Au 的异点数虽多，但其中位数和内散度却很低，说明 AL 地质单元中，Au 含量普遍较低，数据主体离散性不大，虽有个别高值点，但在总体上还是呈分散趋势。同样，从图 3*b* 中也可得出，Ag 主要在 AW_2、AW_{1j}、AL 和 H 地质单元中集中，而在 AW_3、γ 和 O 单元中可能呈分散趋势。

因此，本区 Au 的成矿有利地质体为 AW_2、AW_{1j} 和 H;Ag 的成矿有利地质体为 AW_2、AW_{1j}、AL 和 H 地质单元。

但是，若以传统的 $\bar{x}$、s、C_V 等参数作为衡量标准，可能会因个别高值点的存在而得出不恰当的结论。表 1 为 Ag 在各地质单元中的 $\bar{x}$、s、C_V 参数表，与图 3*b* 相比，AW_3、H、AW_2 和 O 地质单元中因有特高值而使其 $\bar{x}$、s、C_V 加大。这样，如果以其中任何一个参数作为标准来观测各地质单元中 Ag 的富集与分散状态，就会笼统地把 $\bar{x}$、s、C_V 较大的几个地质单元当作成矿有利地质体，从而产生不客观的估计。

表 1　山西某地区各地质单元中 Ag 的特征参数

地质单元	$\bar{x}$	s	C_V
AL	67.29	28.21	0.42
AW_{1j}	69.37	46.79	0.67
AW_1	61.78	27.65	0.45
AW_2	90.79	187.57	2.07
AW_3	72.50	97.39	1.34
H	72.65	98.44	1.35
O	108.37	243.53	2.25
γ	64.49	15.87	0.25

确定异常元素组合

找矿实践证明，仅仅依靠单元素异常特征筛选异常已显不足，必须充分研究多个异常在空间上的相互关系及异常元素组合和区域异常与区域地质的关系等，才能更有效地筛选出有意义的异常，特别是不显著异常。

Overstreet(1981) 曾指出，解释地球化学数据的目的之一，就是要区分出异常元素组合，并把它们与一定的矿床类型相联系。区域异常元素组合的研究是一个复杂问题，它既受内生成矿作用的影响，又受表生地球化学作用的控制；既有元素垂直分带与水平分带的问题，又有空间上不同成矿期的叠合以及不同地质过程的叠加问题。多年来，多元素异常元素组合的研究主要是靠肉眼观察单

元素异常在空间上的关系。其后由于多变量统计学与电子计算机技术的发展，因子分析等多变量统计学方法才逐渐成为化探工作者研究异常元素组合的重要手段。但是，这些统计学方法所取得地结果都是所谓“全盘性”的，难于在不同图幅之间进行对比，不利于今后对全国或全省的异常进行综合研究（谢学锦，1990）。

Kürzl(1988)指出，单元素数据的直观对比可强化解释推断能力。根据几个出现异点的元素的含量特征和Boxplot形状不同，可以很容易地确认与矿化作用有关的指示元素组合。

表2是根据山西和新疆两个地区5个Au异常的框图对比所确定的异常元素组合。其

表2 Au异常元素组合

地区	异常号	异常元素组合
山西	9	Au－W－Bi－Ag－Pb－Hg－Cu－Zn
	75	Au－Bi－Cu－Ag－Hg－W－Mo－Pb
	45	Au－Bi－W－Hg－Ag－As
新疆	70	Au－Hg－Sb－As－W－Li－Cd－Be－Ag－Mo－Sn－Cu
	44	Au－As－Mo－B－Sb－W

中9号和70号Au异常有关元素的框图分别示于图4和图5。可见每个元素，即使是与同一异常有关的指示元素的框图也明显不同，反映出各元素的数据结构和特点不同。单就异点数（*ON*）来分析，图4中W最多，为12个；其次是Bi(11个)和Ag(9个)；

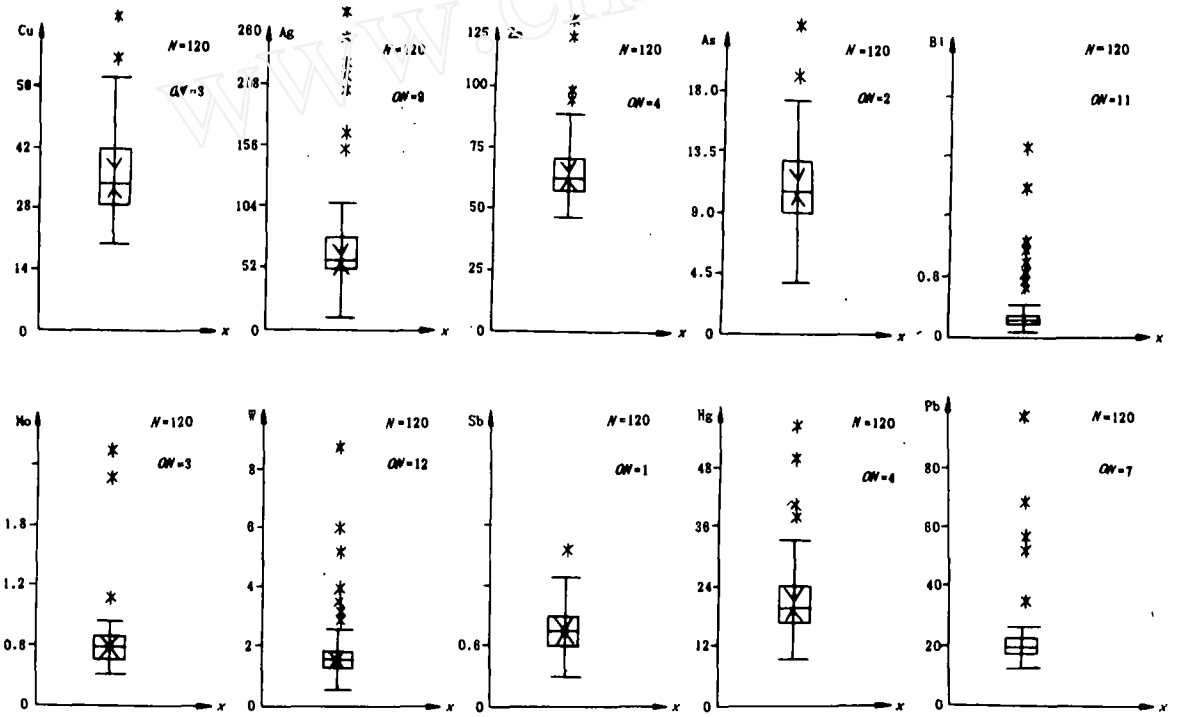

图4 山西某地区9号Au异常各元素Boxplot对比图

（图中*N*为样品数；*ON*为异点数；Ag、Hg×10^{-9}；其余×10^{-6}；后图同）

Sb最少，只出现1个异点。异点虽是一条重要标志，但并不是唯一标准，内散度和中位数大小也是评价异常的重要参数。因此，以是否出现异点为基本条件，综合考察、对比各元素Boxplot的结构特征来确定异常元素组合。如图4中，Cu、Mo异点数均为3，

但 Cu 的内散度较 Mo 大，所以 Cu 列入了异常元素组合而 Mo 没有。图 5 中，As 的异点数为 6，排在 W ($ON=14$)、Ag ($ON=10$)、Li($ON=9$)之后，但由于 As 的内散度较 W、Ag、Li 大，所以在异常元素组合中 As 排在 W、Ag、Li 之前。相反，B 的内散度虽然很大，但却只有 1 个异点，故其没有列入元素组合。

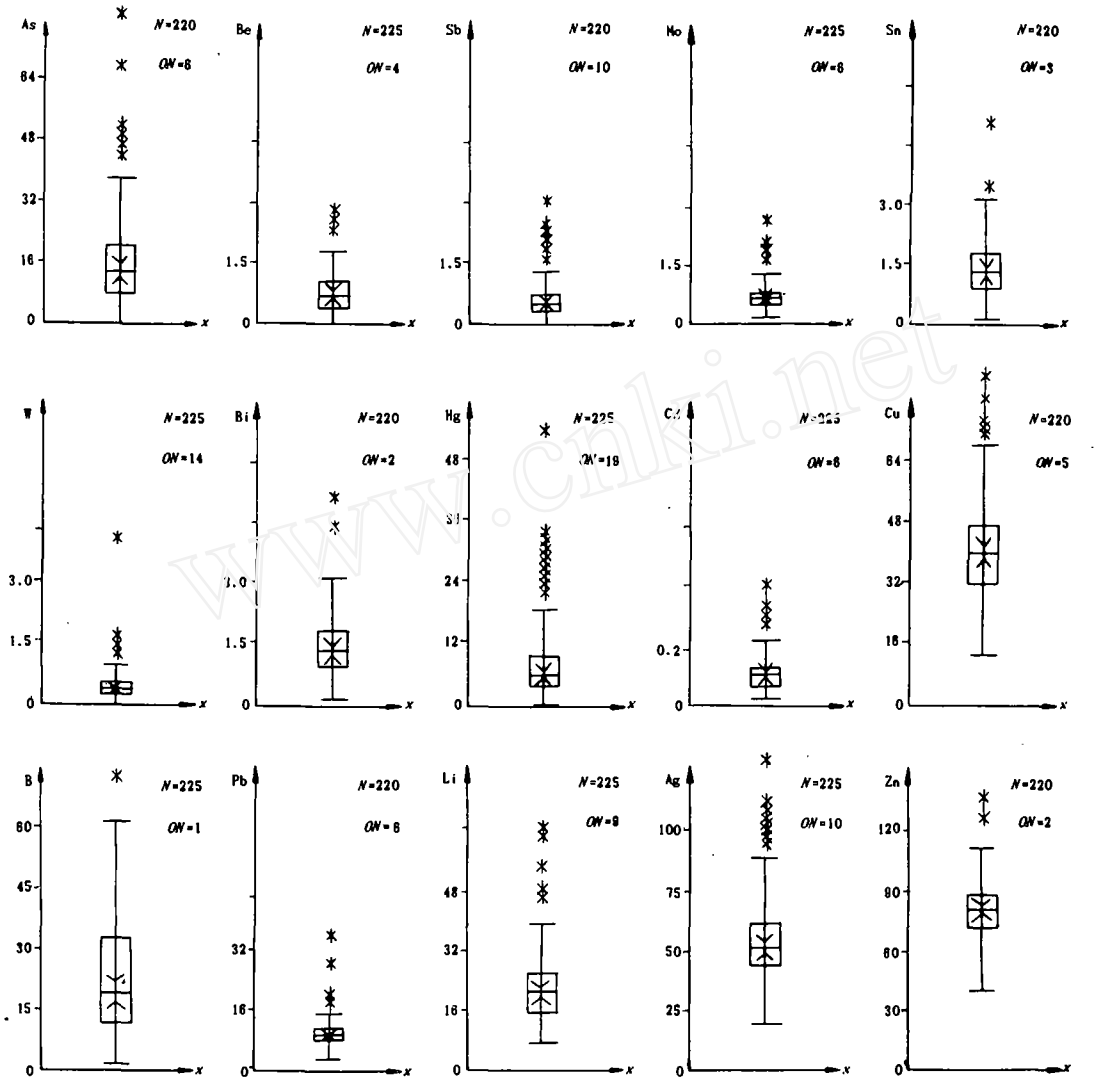

图 5　新疆某地区 70 号 Au 异常各元素 Boxplot 对比图

从表 2 还可看出，山西某地区的 Au 异常元素组合中均有 Bi 出现，而且排位亦较靠前；新疆某地区的 2 个 Au 异常中均没有 Bi。这种现象在上述两地区其他异常中也普遍存在。因此，在利用已知矿的异常元素组合进行类比或评价未知异常时，要充分考虑地质背景和景观条件不同所带来的同类或同种矿床异常元素组合之差异性。

异常评序

区域化探异常的评序，一般是以异常面积、异常强度和异常规模为依据，综合元素组合特征和地质背景特征进行的。但异常面积、强度和规模受异点的影响很大，一个特

高值就可以使元素含量平均值增大很多。本文以 Boxplot 对比为基础，根据其特征参数，采用异常参数排序总得分法进行异常评序的尝试。方法是将各异常的有关参数分别排序，赋以得分，然后计算各异常的总得分，按总得分多少排定次序。

图 6 为山西某地区 6 个 Au 异常的 Boxplot 对比图，它清晰地反映出各异常的数据分布特征及其差异性。按框图的结构，6 个异常可划为 3 组，9、32、75 号异常可归为第一组，54、45 号异常可归为第二组，8 号异常为第三组。以此为基础的评序结果示

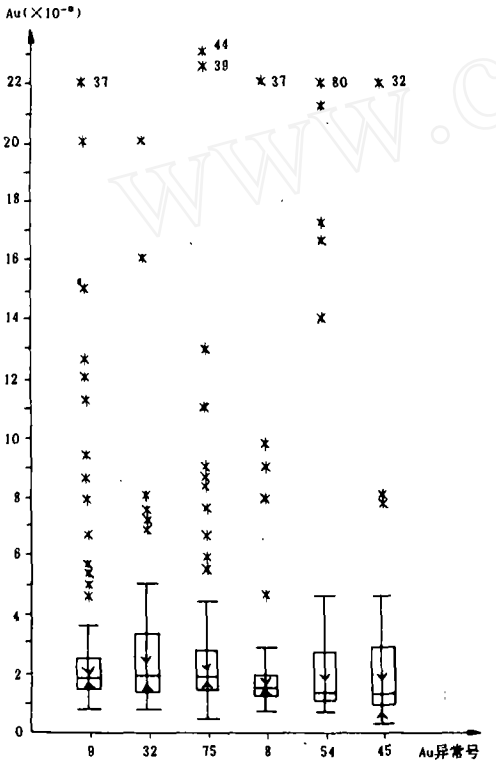

图 6　山西某地区 Au 异常 Boxplot 对比图

于表 3。表中同时还给出了各异常的 NAP 值（即规格化面金属量）。很明显，Boxplot 对比排序总得分法的评序结果与 NAP 值大小顺序不尽一致。9 号异常的 NAP 值比 75 号异常大，但却排在 75 号异常之后，这是因为 75 号异常的异点临界值和内散度较 9 号异常大，表示 75 号异常的异常下限和离差均较 9 号异常高。同样，45 号异常（NAP＝237）排在 54号异常（NAP＝531）之前也缘于此故。

表 3　山西某地区 Au 异常评序结果

异常号	总得分	排序	NAP 值
32	7	1	651
75	14	2	551
9	15	3	618
45	19	4	237
54	19	5	531
8	23	6	231

图 7 示出了新疆某地区 8 个 Au 异常的框图。从中可清楚地看出各异常的特征及其相对的“优劣”程度。70、44、16 和 27 号异常与其他 4 个异常的特征明显不同。其中 70 号异常的中位数、内散度和异点数均较高，从数据结构分析，它富集成矿的可能性

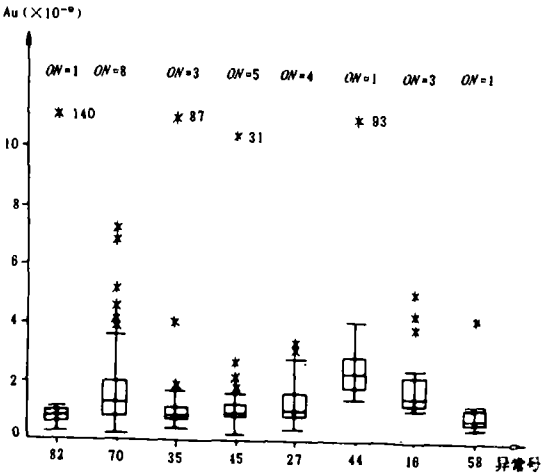

图 7　新疆某地区 Au 异常 Boxplot 对比图

最大。其次是 44、16 和 27 号。但 44 号异常的内散度和中位数虽高，却只有 1 个异点。一般来说，单点异常较之多点异常见矿的概率要小得多。所以，44 号异常的排序位置应往后挪。综合考虑各异常 Boxplot 的结构，排序结果见表 4。从表中亦可看出

Boxplot 评序与 NAP 值大小排列不尽一致。35、44 和 82 号异常 NAP 值偏大可能是因其分别有 87、93 和 140×10^{-9} 的特高点所致。

表 4　新疆某地区 Au 异常评序结果

异常号	总得分	排序	NAP 值
70	7	1	451
16	12	2	2252
27	15	3	180
45	20	4	181
35	21	5	434
44	9	6	191
82	24	7	671
58	25	8	88

讨　论

EDA 技术为我们提供了一种在数据性质未知的情况下，直接处理原始数据的简单方法。与经典统计学不同，Boxplot 可以在没有任何假设模式的条件下直接描绘出经验分布的各种特性，并具有很强的抵抗“野”数据干扰的能力，可用于确定异常总体和背景总体。

用 Boxplot 对比研究元素时空分布规律，选择成矿有利地质体，较之常规方法更简单、直观而客观，可以避免个别特高值的影响。依据 Boxplot 的结构特征可以很直观地确定与主成矿元素有关的异常元素组合。以 Boxplot 为基础，采用异常参数排序总得分法进行异常评序，可以排除个别特高值对异常强度、异常规模的影响。本文所讨论的异常评序没有考虑异常所处的地质背景等因素。然而，地质背景是异常评序和异常评价的重要条件，如何将这种定性标志转化成异常评价中的定量参数，是异常解释推断中急待解决的一个问题。所以，Boxplot 作为异常解释推断的一种新方法需做进一步开发研究。

这里所提供的只是初步研究成果，因资料和水平所限，不妥之处敬请批评指正。

本文承蒙谢学锦导师的精心指导，并有幸得到林存山高级工程师的指教，特此致谢。

参考文献

[1] 谢学锦等，地矿部地球物理地球化学勘查研究所所刊，1990，第 4 号，第 181～222 页。

[2] 史长义，国外地质勘探技术，1990，No.1，p.38～40。

[3] Kürzl, H., Journal of Geochemical Exploration, 30 (1988), 309～322.

[4] Overstreet, W. C., Economic Geology 75th Anniversary Volume, 1981, p.775～805.

Application of the Exploratory Data Analysis Technique

Shi Changyi

Exploratory Data Analysis (EDA) technique is an unconvential statistical method of data processing. Boxplot graphical technique is used as the method to distinguish the element associations of multi-element anomalies in the paper according to the regional geochemical data of two regions, one in Shanxi, one in Xinjiang. It is also used to try to do the ranking of significant anomalies and study the distribution of trace elements in time and space so as to identificate favourable geological units. The results proved that Boxplot is a new simple and practical method in the analysis and interpretation of geochemical exploration data.