ROBUST ANALYSIS OF MULTIDIMENSIONAL GEOCHEMICAL DATA

Pavel Machek, Department of Geology and Mineralogy, Mining University, Ostrava

Analytical data are of prime interest in geochemical and petrological data studies. These data represent contents of various chemical elements in rocks and minerals. They are typically multidimensional: calssical sillicate analysis estimates contents of at least 13 elements (mostly expressed as percentages of corresponding oxides), i.e. SiO_2, TiO_2, Al_2O_3, Fe_2O_3, FeO, MnO, MgO, CaO, Na_2O, K_2O, P_2O_5, H_2O+, H_2O-, CO_2; often also contents of the so-called trace elements are estimated (all remaining elements of the periodical table). Processing of these data is connected with a lot of problems which follow from the nature of geochemical data. First of all, they are notoriously nonnormal. Geochemical (especially trace element) data sets are repeatedly positively skewed and leptokurtic, have large outliers (the distributions have heavy tails). Trace element data very often show analytical truncation, i.e. there are samples in which the estimated elements' contents lie below apparatus detection limits. Another important feature is inhomogeneity of the data resulting in bi- or polymodal distributions. A very difficult problem to solve is also the constant-sum closure of silicate analyses (the sum of the oxides is a constant, equal to 100%, common to all samples) which destroys the usability of elementary independence testing.

The complexity of real geochemical data amazingly has been ignored in their processing. Today, the the increasingly advanced analytical technology allows huge data sets to be produced. But simultaneously, the advancement of data processing methods lags behind the analytics. All textbooks of geochemistry and petrology more or less quietly suppose normal distribution for the analytical data. This is possibly one of the reasons why classical statistical methods are mulishly used for "nonclassical" data. This very often results in very bizarre interpretations.

The use of robust statistics is a very promising alternative in real geochemical data processing. A basic outline of the use and performance of several robust estimates of location (simple median, various combined, trimmed, adaptive, skipped, L, M, and W estimates) and nonparametric estimates of scale using data sets with various type of analytical and/or geological error are illustrated by ROCK (1988). Robust estimates of location and scale can be used for highly effective and objective summary statistics as representations of large bodies of data. As a consequence of the positive skewness and the presence of outliers, the arithmetic mean and standard deviation may even several times (especially for trace elements) overestimate absolute values for both location and scale. Robust estimates generally give smaller but consistent values. They yield "good" results also for "bad" data which contain large outliers and/or are truncated. A confrontation of different robust estimates helps reliably to reveal the influence of outliers or below-detection-limit values. The inconsistency between avrious estimates increases with the number of the "bad" values and the estimates become statistically less efficient. On the other hand, if the values of various estimates derived on quite different fundaments are consistent, then these estimates must approximate some geologically objective value.

Robust estimates therefore should not be computed and used separately, because an important information of the data quality would be lost.

By this time, more than 70 robust location estimates have been derived. However, from this great number not more than one fourth can be recommended for practical purposes in geochemistry with respect to the unstability and undue computing complexity of the majority of the known estimates. Trimmed means, Gastwirth median, trimean, Andrew's estimate, and Hampel's series seem the most preferable for extensive applications in many geochemical situations. On the other hand, the frequently recommended dominant cluster mode (ELLIS at al., 1977), Johns' (1974) estimate, and shorth (shortest half) estimate (ROCK, 1987) do not give stable results (instability of the algorithm?) with geochemical data.

As for robust scale estimates, the following can be used in geochemical applications: mean deviation from the mean, mean deviation from the median, median deviation from the median, semi-interquartile range (or half H-spread). These estimates are listed in the absolute order of their values which is common in geochemistry (the standard deviation providing the highest values).

All methods of hierarchical and nonhierarchical agglomerative cluster analysis were derived for data with normal distributions. Their most frequently used strategies take arithmetic mean in computations of

distances between two clusters. E.g. the group average strategy defines the distance between two clusters as the arithmetic mean of all intercluster distances. Provided the data are nonnormally distributed, esp. because of the presence of outliers, the results of clustering may be entirelly misleading. A typical outcome of nonnormality is the instability of the results of clustering. In other words, the classical clustering methods are nonrobust.

Employing some of the robust estimates of location can make robust also the clustering process. The most accessible from hierarchical methods is the group average strategy. Substituting even the simple median for the arithmetic mean in computing the distances between clusters the clustering is no more sensitive to outliers, skewed distributions, etc. Performances with trimean and Gastwirth median are even more safe. Unfortunately, in such a procedure the intercluter distances cannot be computed recursively as in classical combinatorial strategies and the complete dissimilarity matrix between individuals is required for all fusions. So, this modification results in large computing memory requirements and needs skilful programming.

In an analogical way robust varieties of nonhierarchical strategies can be obtained. Most of these methods look for the so-called typical points whichare computed as arithmetic means of all the variables fo the individuals placed into clusters variously defined in various methods. Instead of arithmetic means robust location tsimates can be employed. There is not sufficient experience as for making a statement of the quality of clustering.

The robust estimates of location and scale are included in the program package SHLU designed at the Department of Geology and Mineralogy, Mining University, Ostrava by the author. The SHLU package serves for multidimensional analysis of geological objects, but may be used also for other types of data. It contains procedures for the principal components and extended Q-mode factor analysis (including very elegant and robust fuzzy clustering) and supposrting procedures for testing the quality of clustering. It can use data prepared with dBase or other database packages. The SHLU package produces outputs which can be used for graphical presentation of the results of clustering using SURFER, STATGRAPHICS or BOEING packages.

Literature:

1. ELLIS, P.J. et al. (1977): Estimation of the mean by the dominant cluster method. Geostat.Newslet. (International Working Group. Association Nationale de la Récherche Téchique, Paris). 1,3,123-130.

2. JOHNS, M.V. (1974): Nonparametric estimation of location. Jour.Am.Stat.Assoc. 69,346,453-460.

3. ROCK, N.M.S. (1987): ROBUST: An interactive FORTRAN77 package for exploratory data analysis using parametric, robust and nonparametric location and scale estimates, data transformations, normality tests and outlier assessment. Comput.Geosci. 13,463-494.

4. ROCK, N.M.S. (1988): Summary statistics in geochemistry: A study of the performance of robust estimates. Math.Geol. 20,3,243-274.