STATISTICAL ANALYSIS IN THE GEOSCIENCES

Grunsky, E.C.
Geological Survey of Canada, Ottawa, Ontario, Canada

Keywords: Multivariate statistics, geochemistry, exploratory data analysis, principal components analysis, cluster analysis, canonical variate analysis, classification, robust estimation, regression, analysis of variance, compositional data, censored data, outliers

Contents

Summary

Geochemical survey data are commonly composed of thousands of observations and analyzed for 50 or more elements. This abundance of data provides an opportunity to discover a wide range of geochemical processes that may have occurred within a survey area. However, it is often a challenge to examine and interpret the significance of many of the elements, in terms of their exploration potential, and even more difficult to observe or understand the relationships between elements. The application of multivariate data analysis and statistical techniques help make the task of data interpretation and model building easier. Geochemical and other compositional data require special handling when measures of association are required. The application of log-ratios are required to eliminate the effects of closure on compositional data. Exploratory multivariate methods include: plots of all possible pairs of data, assessment of the marginal distributions, adjusting for censored and missing data, detecting atypical observations, computing robust means, correlations and covariances, principal components analysis, cluster analysis and knowledge based indices of association. Modeled methods of assessing multivariate data include: regression, analysis of variance, canonical variate analysis, and classification. These topics are covered with examples to demonstrate their application.

1. Introduction

The statistical analysis of geoscience data is a broad subject which cannot be comprehensively covered in this introductory contribution. What follows is a condensed summary of statistical methods that have been applied to geochemical data. Geochemical surveys are an important part of geoscience investigations in both mineral exploration and environmental monitoring. Geochemical specimens are commonly analyzed for as many as 50 elements. A soil or lake sediment survey can consist of collecting several thousand specimens. Analyzing and interpreting these large arrays of data can be a challenge. Data can be both categorical (discrete numeric or non-numeric) or continuous in nature. To extract the maximum amount of information from these large arrays of data there are a wide range of multivariate data analysis techniques available. In many cases, these techniques reduce these large arrays of data into a few simple diagrams that often outline the principal geochemical trends and assist with interpretation. Often, the trends that are identified include variation associated with underlying lithologies, zones of alteration, and in special cases, zones of potentially economic mineralization. From an environmental perspective, trends may also be observed that represent background and distinctive multi-element signatures associated with pollution. Areas of mineralization and some types of pollution are typically small in geographic extent. Thus, they can be considered as rare events relative to the regional geochemical signatures within an area of study and they will often be under-represented within a sample population. This means that they may often be observed as atypical or they can be masked by the main mass of the sample population.

The term “sample” usually refers to a selection of observations from a population in the statistical literature. In the lexicon of geoscientists, specimens of soil, rocks, stream sediments and other such media, are often called “samples”. This has been a source of confusion between the geoscience and the statistical communities. Within this chapter,

specimens that have been collected in the field are referred to as “observations” or as “specimens”. The terms variable and element are used interchangeably in this chapter. Elements are the geochemical entities that become variables in the application of statistics.

The amount of interpretation that can be inferred from geochemical survey data is dependent on the spatial density of the specimen sites of an area. In areas where the site collection density is low, only regional information is likely to be obtained. In areas where the specimen site density is high, much information can be obtained about the local variation in the surficial geology (soils) or bedrock lithologies as well as the extent and character of alteration and zones of mineralization.

The evaluation and interpretation of geochemical data relies on a sound understanding of the sample material. Different materials require different methods and techniques for the interpretation of the results. In the case of glacial till, lake and stream sediments, different size fractions of specimens may reflect different geological processes. The choice of size fraction in soils, streams, or lake sediments can result in very different geochemical responses and have a profound influence on the interpretation. In any geochemical survey the collected material should be carefully sampled and classified in order to provide any clues about the underlying geochemical processes

Quality control is an important part of assessing multivariate geochemical data. Initially, all data should be examined for analytical reliability and the identification of any suspect analyses. This is typically done using several exploratory data analysis methods. Issues of quality control, analytical accuracy and precision are beyond the scope of this chapter. Evaluating geochemical data without considering analytical quality control can be perilous.

With the availability of affordable desktop statistics packages and mapping systems (Geographical Information Systems and Image Analysis Systems), geochemical data and the results of multivariate analyses can now be interpreted together with other data (geology, geophysics, geomorphology) as a means to provide additional insight into spatial characteristics of data. The use of spatial information permits subdivision of the geochemical data into various distinct geographical regions that may be based on the underlying surficial processes or geology and enables more effective evaluation of the data for environmental monitoring or mineral exploration purposes. Multivariate datasets can be placed into a GIS and integrated with other geoscience information.

Three sets of data have been used in this chapter.

1) Lithogeochemical data from Ben Nevis township, Ontario, Canada (Figure 1).
Rock specimens (825) were collected as part of a study to examine the nature of alteration and associated mineralization in a sequence of volcanic rocks. The alteration consists of a large north-south trending zone of carbonate alteration with a central zone of silica enrichment and sulphide mineralization consisting of gold and copper. Small isolated zones of sulphide mineralization occur throughout the area. The specimens were not collected over a regular grid but were collected wherever rock outcrops could be located in the field. The geology of the area and the

specimen locations are shown in Figure 1. The lithogeochemical specimens were analyzed and the compositions of the specimens were reported as weight percent oxide values for the major elements and as parts per million (ppm) for the trace elements.

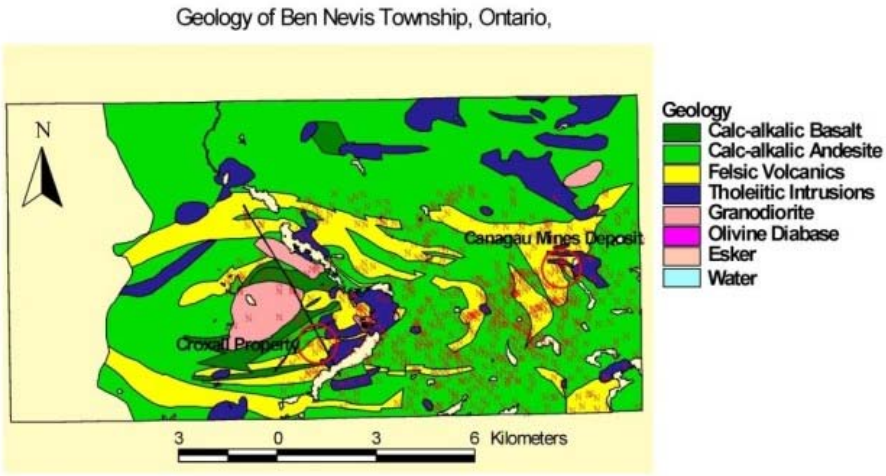

Figure 1. Geological map of the Ben Nevis township area, Ontario, Canada. The geology is dominated by mafic volcanic basalts, andesites and rhyolitic rocks A large zone of carbonate alteration occurs in the eastern part of the map. Two significant mineral occurrences are shown as red circles on the map.

2) Lake sediment survey data from the Batchawana district, Ontario, Canada (Figure 2)

This set of survey data was collected from a series of lakes that overlie a PreCambrian volcanic-sedimentary sequence that has been intruded by granitic rocks. Lake sediments were collected over an area as shown in Figure 2. The lake sediments in the area are derived from two primary sources; the underlying bedrock (shown in the legend) and the glacial overburden (not shown). Glacial till, outwash sand, lacustrine deposits and recent re-worked glacial deposits blanket the area in varying thickness. Bedrock exposure is less than 5% of the area with most of the glacial overburden being less than 3 m. The lake sediment survey was carried out in 1988 from which 683 lake sediments were collected.

3) Geochemical Database of Volcanic Rocks from Ontario, Canada

In the Canadian Shield area of Ontario, the volcanic rocks found in most areas belong to one of three major volcanic rock clans. These clans are known as Komatiitic, Tholeiitic and Calc-alkalic and are partly distinguished by their whole rock geochemistry. Their distinctive chemical compositions reflect specific volcanic environments. Many thousands of rock specimens have been collected over the volcanic terrains in Ontario, which were chemically analyzed and the results were placed in a database. From this database, a set of reference groups was established based on three volcanic rock clans mentioned above. These reference groups were

developed through a rigorous process of eliminating compositions that suggested alteration or were atypical for each clan. Most of the methods discussed in this chapter were used to establish and refine the reference groups.

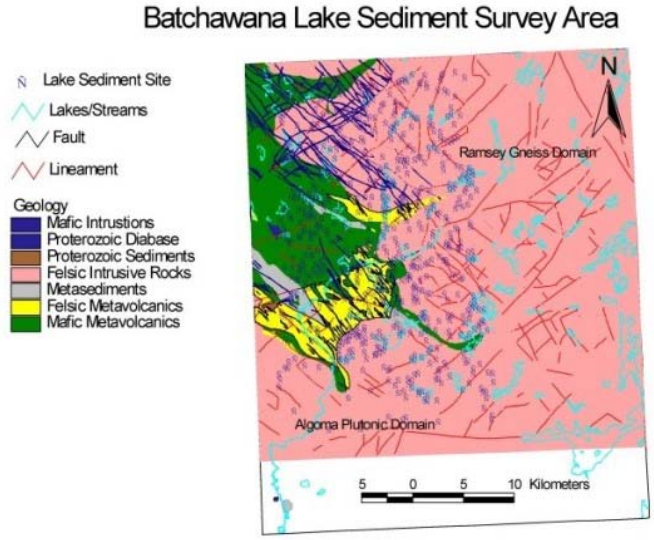

Figure 2. Geological map of the bedrock in the Hanes Lake area of the Batchawana district, Ontario, Canada. The area is underlain by Precambrian age metavolcanics, metasediments and grantitic rocks. The lake sediment survey sampling sites are shown as blue crosses. The area is covered by extensive deposits of glacial till and outwash sand.

1.1 Exploratory Data Analysis

Exploratory data analysis is concerned with analyzing geochemical data for the purpose of detecting trends or structures in the data. These methods can provide insight into the geochemical/geological processes from which geochemical process models can be constructed. Exploratory methods of data analysis include the evaluation of the marginal (individual) distributions of the data by numerical and graphical methods. Numerical methods include the use of summary tables (minimum, maximum, mean, median, standard deviation, minimum absolute deviation(MAD), coefficient of variation 1st and 3rd quartiles), measures of correlation, covariance and skew. Graphical methods include histograms, quantile-quantile plots, box plots, density plots and scatterplot matrices. The spatial presentation of data summaries can be incorporated into a GIS using features such as bubble, symbol plots and interpolated grids.

Multivariate methods include the use of principal components analysis, cluster analysis, χ^2 D^2plots, empirical indices and various measures of spatial association.

1.2 Target and Background Populations

Geochemical background represents a population of observations that reflect unmineralized (or unpolluted) ground. The background sample may be a mixture of

several populations (gravel, sand and clay). The separation of the background population into similar subsets that represent homogeneous multivariate normal populations is important and form the basis of the modeled approach of geochemical data analysis. This can be achieved using exploratory methods such as the methods described above, principal components analysis, methods of spatial analysis, χ^2 D^2plots and cluster analysis.

Observations that represent a group of elements being sought are termed "Target" populations. These populations are derived from specimens collected from orientation studies over known mineral deposits or areas of specific interest.

1.3 Modeled Data Analysis

Methods of modeled data analysis are based on the knowledge that certain geochemical patterns reflect particular geological processes (i.e., ore deposits, contaminated areas) as recognized through exploratory data analysis. Through orientation studies, geochemical characteristics can be obtained for specific geological processes. These geochemical characteristics are the models from which unknown specimens can be compared using classification methods. Modeled approaches to geochemical data involve the construction of reference and test datasets, as explained above. Methods such as discriminant analysis and analysis of variance (ANOVA) can be applied to test the uniqueness of the background and target groups. Unknown specimens can be tested for membership within none, one or more of the reference groups using classification procedures.

1.4 Special Problems

Problems that commonly occur in geochemical data include:

- Many elements have a "censored" distribution, meaning that values at less than the limit of detection and can only be reportedat that limit (i.e. < 5 ppm).
- The distribution of the data is not normal.
- The data have missing values. That is, not every specimen has been analyzed for the same number of elements.
- Combining groups of data in which there are distinctive differences for some or all of the elements. Leveling of the data is required.
- Not every element has been analyzed by the same method or the limits of detection of the method have changed over time.
- The constant sum problem for compositional data (i.e. the data sum to a constant value [100%]).

To overcome the problems of censored distributions, procedures have been developed to estimate replacement values for the purposes of statistical calculations. Non-normally distributed data can be transformed to normal or near-normal distributions. Procedures that assign replacement values have also been developed for data with missing values.

For geochemical data, a distribution that is not normal (positively or negatively skewed) may reflect mixtures of two or more distributions (i.e. different rock types). Rather than

apply a transformation, it might be better to split the data into separate distributions based on categorical criteria, such as rock type.

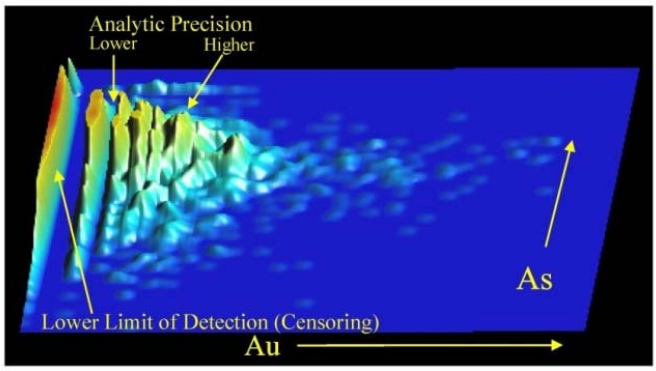

Figure 3. Bivariate distribution of As (y axis) and Au (x axis). The Au distribution is truncated by the lower limit of detection. The bivariate distribution also appears as discrete bands parallel to the y-axis indicating the limits of reporting resolution of Au at low detection levels.

The effect of censoring, departure from normality and low reporting accuracy can be expressed graphically (Figure 3). The image is a shaded relief map derived from the density of observations of As vs. Au. The "valleys" represent limits in data resolution near the lower limit of detection for Au. The actual limit of detection appears as a "wall" at the zero end of the Au axis. In contrast, As displays a continuous range of values without the same resolution or detection limit problems exhibited by Au.

1.4.1 Leveling Geochemical Data

In many studies, integration of several sets of data is necessary. The sets of data may represent sampling programs that used different methods of analysis for the same element. The detection limits may be different and there may be systematic shifts between the groups of data. In order to use the data effectively, one or more sets of data must be adjusted. This adjustment is known as leveling. One set of data is chosen against which all other sets of data will be levelled. The relationship of each element is compared and an adjustment is made through the application of a linear transformation. Given an observation x, with (i=1,…n) variables,

$$y_i = ax_i + b$$

x_i is the untransformed variable for observation x,
y_i is the transformed variable for observation x,
a represents the slope of the line in the transformation,
b represents the intercept or additive adjustment.

The adjustment transformation can be determined through regression methods. Non-linear transformations may also be applied if necessary. Figure 4 shows the types of leveling scenarios that can be encountered. The x and y axis of each figure shows the

values of the quantiles (values at 5, 10, 15, etc. percentiles) for the two variables. With the exception of Figure 4e, each scenario shows a possible relationship that will permit leveling. Figure 4e shows a random association between the two variables and in this case leveling is not possible.

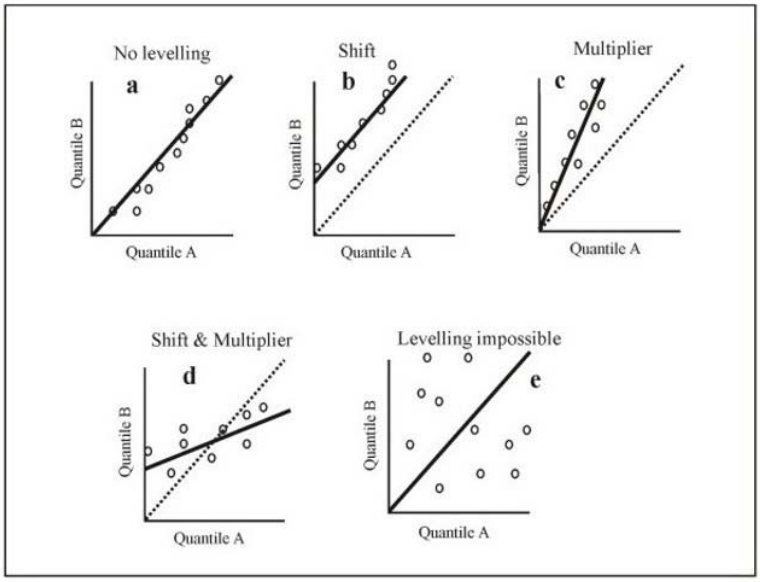

Figure 4. Various relationships between two sets of data for the same variable. Adjustments can be made to either sample A or sample B through shifts or multipliers of Figures 4a-d. Only the random association of Figure 4e cannot accommodate any adjustment

-
-
-

Bibliography

Agterberg, F.P. (1974). *Geomathematics*, Elsevier, Amsterdam, 596 pp. [This monograph is a comprehensive treatment of the theory and application of mathematics and statistics that existed at the time of publication. It is a unique contribution to the field of geomathematics and a useful reference for multivariate analysis and studies in spatial autocorrelation, harmonic analysis, Markov chains, multivariate stochastic process models and the spatial variability of multivariate systems.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*, New York: Methuen Inc., 416 pp. [This monograph is the standard reference for anyone interested in understanding the problems and solutions of compositional data. Continued research into the statistical analysis of compositional data is being carried out by Vera Pawlowsky and Cales Barcelo at the University of Girona, Spain].

Bailey, T.C. and Krzanowski, W.J. (2000) Extensions to Spatial Factor Methods with an Illustration in Geochemistry. *Mathematical Geology* 6, pp. 657 – 682. [This is a recent account on the state of multivariate spatial analysis. It provides a summary of research carried out].

Davis, J.C. (2002). *Statistics and Data Analysis in Geology*, New York: John Wiley & Sons Inc., third edition, 638 pp. [This is a standard reference for all students in the geosciences. It covers a wide range of topics including an extensive section on multivariate methods.

Garrett, R.G. (1989b). The chi-square plot. a tool for multivariate outlier detection. *Journal of Geochemical Exploration* **32**, 319-41. [This paper provides the foundation with practical demonstration on the use of χ^2 and D^2 plots].

Howarth, R.J. (1999). A History of mathematics in the geosciences ?

Howarth, R.J. (1983). *Statistics and Data Analysis in Geochemical Prospecting*, edited by R.J. Howarth, Vol. 2, in Handbook of Exploration Geochemistry, edited by G.J.S. Govett, New York: Elsevier, 437 pp. [This monograph is part of a series of books intended for exploration geochemistry. This volume offers a wealth of information on field specimen collection strategies, sample design, the statistics of laboratory duplicates and standards, univariate and multivariate statistics, map presentation and pattern classification. It is an essential reference for field based researchers who use geochemistry in their field sampling strategies. Chapters by R.G. Garret, R.J. Howarth, R.J. Howarth and R. Sinding-Larsen are oriented towards the processing of geochemical data].

Howarth, R.J. (2002). From graphical Ddisplay to Dynamic model: mathematical geology in the earth sciences in the \nineteenth and twentieth centuries, in *The Earth Inside and Out: Some Major Contributions to Geology in the Twentieth Century*, D.R. Oldroyd, editor, Geological Society of London, Special Publications, 192, p 59-97. [This article is a fascinating account of the development of statistics and mathematics in the earth sciences in the past two hundred years. It documents the advances made in the earth sciences as a result of developments in the fields of graphics, mathematics and statistics.

Jöreskog, K.G., Klovan, J.E. and Reyment, R.A. (1976). *Geological Factor Analysis*. New York: Elsevier, 178 pp. [This monograph is essential reading for geoscientists who want to understand the mathematics of factor analysis methods. It is somewhat data and has been superseded by the monograph by Reyment and Joreskog (see below)].

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis, A User's Perspective*, Oxford: Clarendon, Press, 563pp. [A monograph that discusses multivariate analysis on an applied level. It is well written and easy to read. Topics covered include principal components analysis, cluster methods, distribution theory and evaluating grouped and un-grouped data].

Reyment, R.A. and Jöreskog, K.G., (1993). *Applied Factor Analysis in the Natural Sciences*, Cambridge: Cambridge University Press, 371pp. [This book is an extension of the book *Geological Factor Analysis*, by Joreskog *et al*. (see above). It incorporates many developments since the previous publication and is one of the first texts for the natural sciences that emphasizes the need to evaluate compositional properly. A number of MATLAB procedures are provided in the text and are available from a website].

Reyment, R.A. and Savazzi, E. (1999). *Aspects of Multivariate Statistical Analysis in Geology*, Amsterdam: Elsevier, 285 pp. [This monograph offers many practical examples in the processing of compositional data and the use of discriminant and compositional data. It also provides a series of programs on a CD, which are awkward to use, but which provide a good introduction to the methods].

Rock, N.M.S. (1988). *Numerical Geology, A Source Guide, Glossary and Selective Bibliography to Geological Uses of Computers and Statistics*, Lecture Notes in Earth Sciences, Vol. 18, edited by Somdev Bhattacharji, Gerald M. Friedman, Horst J. Neugebauer and Adalf Seilacher, New York: Springer-Verlag, 427pp. [This monograph summarizes great deal of statistical definitions and methods in a geoscience context. Although it is outdated, it contains many useful references].

Sanford, R.F, Pierson, C.T., and Crovelli, R.A. (1993). An Objective Replacement Method for Censored Geochemical Data. *Mathematical Geology* **25**, 59-80. [This paper contains a detailed evaluation of the effectiveness of finding replacements for censored geochemical data].

Venables, W.N., and Ripley, B.D. (1999). *Modern Applied Statistics with S-Plus*, New York: Springer-Verlag, 501 pages, 3rd edition. [This book is an excellent reference for applying statistical methods in either S-Plus or R (two statistical languages and desktop environments). As the title indicates, the book is applied. Not much information on the theoretical details is provided. The authors have published a freely available library, MASS, which provides functions, examples and datasets].

Biographical Sketch

Eric Grunsky (B.Sc., M.Sc. University of Toronto; Ph.D., University of Ottawa) is a geologist with a background in field mapping, structural geology, remote sensing, geochemistry, statistical analysis and geoscience information management.. He is currently located, in Ottawa, Canada, at the Geological Survey of Canada.