Paper 8

The Interpretation of Regional Geochemical Survey Data

Grunsky, E.C. [1]

[1] Geological Survey of Canada, Natural Resources Canada, Ottawa, Canada

ABSTRACT

Regional geochemical data are generally derived from government and industry geochemical surveys that cover areas at various spatial resolutions. These survey data are difficult to assemble and integrate due to their heterogeneous mixture of media, size fractions, and methods of digestion and analytical instrumentation. These assembled sets of data often contain thousands of observations with as many as 50 or more elements. Although the assembly of these data is a challenge, the resulting integrated datasets provide an opportunity to discover a wide range of geochemical processes that are associated with underlying geology, alteration, and mineralization. The use of data analysis and statistical methods combined with geographical information systems provides an effective environment for process identification and pattern discovery in these large sets of data, however it should be borne in mind that Areas of mineralization, because of their small areal extent, are generally underrepresented in regional geochemical data sets. Modern methods of evaluating data for associations, structures and patterns are grouped under the term "Data Mining". Mining data includes the application of multivariate data analysis and statistical techniques combined with geographical information systems can significantly assist the task of data interpretation and subsequent model building. Geochemical data require special handling when measures of association are required. Log-ratios are required to eliminate the effects of closure on compositional data. Exploratory multivariate methods include: plots of all possible pairs of data, 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. These topics are covered with examples to demonstrate their application.

INTRODUCTION

A review of contributions to the Exploration 1977, 1987 and 1997 conferences in the field of exploration geochemistry and the interpretation of regional geochemical survey data provides a perspective and appreciation of the very powerful tools that geoscientists now have at their disposal. Boyle (1979) described the first part of the twentieth century when rapid advancements were made in the recognition of primary and secondary dispersion haloes; development of accurate and rapid analytical methods; improvements in sampling technologies; the development of atomic absorption spectroscopy, flourimetry; chromatography; radiometric methods, neutron activation analysis, massspectrometry, airborne geochemical sampling methods; improvement in field techniques and access (helicopters); heavy minerals in glacial media; developments in statistical and computer techniques. At that time, Boyle also pointed out that further research was required to understand the trace and major element chemistry of rocks and their geochemical relationship to metallogenic belts. Boyle also noted that future research must focus on the identification of mineral deposits at depth, and for countries such as Canada, the evaluation of basal till geochemistry is an effective means of exploration for metallic mineral deposits. The role of government surveys in the collection of various geological media and subsequent geochemical analysis was considered paramount for a successful mineral exploration strategy for any country. Boyle discusses the term "vectors" as a means to identify mineral deposits through the evaluation of patterns and trends in geochemical data in both 2 and 3 dimensions.

At the time of Exploration 77, the use of geochemical data in glacial terrains, (Bølviken and Gleeson, 1979); non-glaciated terrains (Bradshaw and Thomson, 1979), lithogeochemistry (Govett and Nichol, 1979); biogeochemistry (Cannon, 1979; Brooks, 1979); stream sediment geochemistry (Meyer et al., 1979); lake sediments (Coker et al.., 1979) and hydrogeochemistry were well advanced. The fundamentals of these developments are still applicable today. There have been refinements in methods of extraction (digestion methods and selective leaches), improvements in detection limits and better understanding of the sedimentary environments of stream, lake, glacial and weathered environments. Howarth and Martin (1979) provided the basics of evaluating geochemical data, the principles of which are still in use today. The term "integration" was already in use in the 1970's when it was realized that several types of geoscience data could be merged using computer-based methods (Coope and Davidson, 1979).

The Exploration '87 meeting contained similar discussions along the lines of weathered terrains (Smith, 1989; Mazzucchelli,

In "Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration" edited by B. Milkereit, 2007, p. 139-182

1989; Butt, 1989), glaciated terrains (Shaw, 1989; Coker and DiLabio.,1989), stream sediments (Plant et al., 1989), lake sediments (Hornbrook, 1989); biogechemistry (Dunn, 1989) and bedrock geochemistry (Govett, 1989). In addition, the role of computers, databases and computer-based methods for use in mineral exploration were distinct contributions to the meeting (Garrett, 1989c; Holroyd, 1989; Harman et al., 1989) and expert systems were introduced as a means for decision making in exploration (Martin, 1989; Campbell, 1989). Exploration '87 also contained more results on the benefits of integrated exploration strategies.

Exploration '97 covered much of the same material of advances in geochemical exploration methods for the geochemistry of glaciated terrains (Klassen, 1997; McClenaghan et al., 1997), the geochemistry of deeply weathered terrains (Mazzucchelli, 1997; Smith et al., 1997), geochemistry of stream sediments (Fletcher, 1997), lake sediment geochemistry (Friske, 1997; Davenport et al., 1997), lithogeochemistry (Franklin, 1997; Harris et al., 1997), plus developments in extraction techniques for the enhancements of geochemical responses (Hall, 1997; Smee, 1997; Bloom, 1997). Closs (1997) emphasized careful sample design and objectives are the fundamental tenets of exploration geochemistry, which had not changed in the previous 30 years. Integrated exploration information management was a major focus at the Exploration '97 conference with significant contributions by Bonham-Carter (1997); de Kemp and Desnoyers (1997), Davenport et al. (1997) and Harris et al. (1997) along with the early developments on the use of the world wide web (internet) by Cox (1997).

Prior to the arrival of Geographic Information Systems and desktop statistical computing packages, exploration geochemistry was limited in scope in terms of extensive data analysis. Textbooks such as those by Hawkes and Webb (1962), Rose, Hawkes and Webb (1979) and Levinson (1980) provided the foundation for exploration geochemistry strategies and defined the principles for planning, executing and interpreting geochemical surveys. These texts were written before the development of geographical information systems or easily accessible statistical packages. As a result, they offered limited treatment for a statistical analysis of geochemical survey data. In the late 1980s Geographical Information Systems (GIS) began to play an increasingly important role in the display and management of spatially referenced data (e.g., geochemical data). These systems required large computers and specialists in the management and maintenance of the software. GIS's have evolved into "Desktop Mapping" systems that allow users of personal computers to display, query, manage, and to a limited extent analyze spatially referenced data.

Geochemical surveys are an important part of geoscience investigations in both mineral exploration and environmental monitoring. The International Geological Correlation Program (IGCP Project 259 (Darnley et. al, 1995) summarizes the value of geochemical surveys for both exploration and global change monitoring. This report contains recommendations for sampling strategies, data management, analytical methods and numerous other topics for the development of a global network of geochemical knowledge. A soil or lake sediment survey can consist of collecting several thousand specimens and be analyzed for 50 more elements. Analyzing and interpreting these large sets of data can be a challenge. Data can be categorical (discrete numeric or non-numeric) or continuous in nature. To extract the maximum amount of information from these data there are a wide range of multivariate data analysis techniques available. In many cases, these techniques reduce these large datasets 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. Areas of mineralization are typically small in geographic extent. Thus, they can be considered as rare events relative to the regional geochemical signatures within a study area and they will often be under-represented within a population. This means that they may often be observed as atypical or they can be masked by the main mass of the population.

The term "sample" in statistical literature, usually refers to a selection of observations from a population. 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 contribution, specimens (i.e. the geochemist's samples) that have been collected in the field are referred to as "specimens" and the data derived from them as "observations". Elements are the geochemical entities that become variables in the application of statistics. The terms variable and element are used interchangeably in this contribution. Specimen collection strategies are an important part of any geochemical survey program. Garrett (1983, Chapter 4) provides useful discussion on various approaches for sampling media for geochemical surveys.

The evaluation and interpretation of geochemical data relies on understanding the nature of the material that has been sampled. Different materials require a variety of methods and techniques for the interpretation of results. In the case of surficial sedimentary materials (glacial till, lake and stream sediments), different size fractions of specimens might reflect different geological processes. The choice of size fraction can have a profound influence on the interpretation of the geochemistry of an area. In any geochemical survey the material for study should be carefully collected and classified in order to provide any clues about the underlying geochemical processes.

Quality control is an essential part of assessing geochemical data. All data should be initially examined for analytical reliability and screened for the identification of suspect analyses. This is typically done using exploratory data analysis methods. Issues of quality control, analytical accuracy and precision are beyond the scope of this contribution, however it is briefly discussed in the section, "Special Problems".

Two sets of data have been used in this contribution.

1) Lithogeochemical data from Ben Nevis township, Ontario, Canada (Figure 1).

Rock specimens were collected as part of a study to examine the nature of alteration and associated mineralization in a sequence of volcanic rocks (Grunsky, 1986a, b). Two significant zinc-silver-copper-gold occurrences have been investigated in this area: the Canagau Mines deposit and the Croxall property (Grunsky, 1986a). The results of a detailed lithogeochemical sampling program outlined a zone of extensive carbonatization associated with the Canagau Mines deposit. A lesser zone of carbonatization is associated with the Croxall property. The alteration consists of a large north-south trending zone of carbonate alteration with a central zone of silica enrichment with gold and copper sulphide mineralization. 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.

Lithogeochemical sampling was carried out over the area in 1969, 1972 and 1979-1981. A total of 825 specimens were analyzed for SiO_2, Al_2O_3, Fe_2O_3, FeO, MgO, CaO, Na_2O, K_2O, TiO_2, P_2O_5, MnO, CO_2, S, H_2O^+, H_2O^-, Ag, As, Au, Ba ,Be, Bi, Cl, Co, Cr, Cu, F, Ga, Li, Ni, Pb, Zn, B, Mo, Sr, V, Y, U, Zr, Sc and Sn. Initially, the major element oxides were assessed using a multivariate procedure known as correspondence analysis and is documented in Grunsky (1986a). Details on the geology, sampling methodology and mineral occurrence descriptions can be found in Grunsky (1986b).

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

This set of survey data, consisting of 3047 lakes sediment specimens collected, from 1989-1995, from a series of lakes that overlie a PreCambrian volcanic-sedimentary sequence that has been intruded by granitic rocks (Grunsky, 1991). The lake sediments in the area are derived from the underlying bedrock (shown in the legend), glacial overburden and organic matter (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 meters.

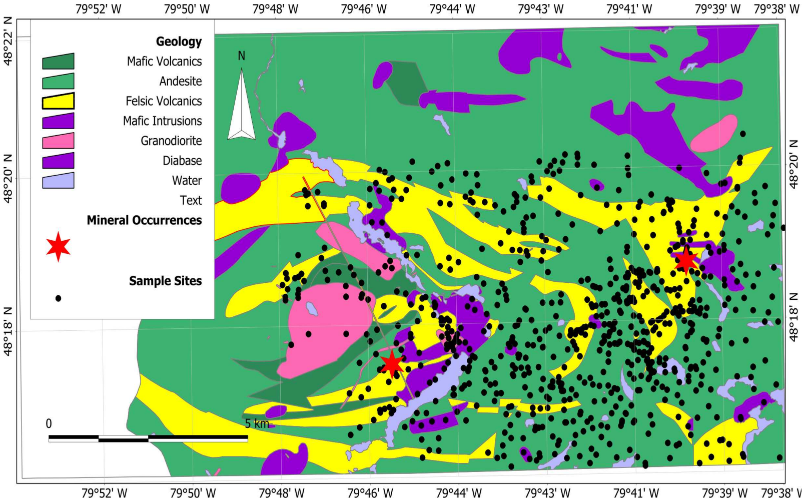


Figure 1: General Geology of the Ben Nevis Township area, Ontario, Canada.

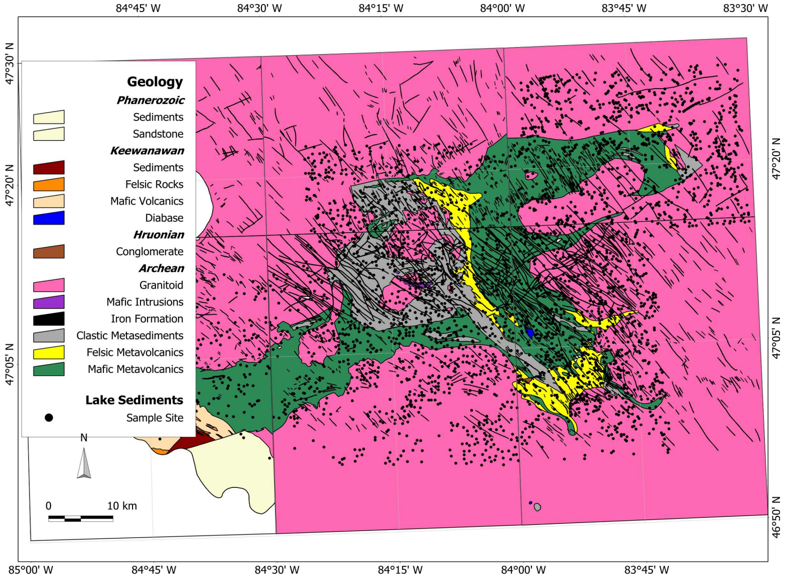


Figure 2: General Geology of the Batchawana area, Ontario, Canada.

GEOCHEMICAL DATA MINING

"Data mining (DMM), also called Knowledge-Discovery in Databases (KDD) or Knowledge-Discovery and Data Mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering, etc. Data mining is a complex topic and has links with multiple core fields such as computer science and adds value to rich seminal computational techniques from statistics, information retrieval, machine learning and pattern recognition." (Wikipedia, http://en.wikipedia.org/wiki/Data_mining, accessed 2007-Apr-19).

Common forms of data mining involve supervised and unsupervised pattern recognition. Unsupervised data mining includes techniques such as cluster analysis, principal components analysis, exploratory data analysis, multivariate ranking of data, neural networks and empirical indices. These methods vary from automatic, semi-automatic, to manual in the degree of pattern delineation. The use of a fully automatic method does not guarantee a result that necessarily represents the best view or meaningful structure in the data. Caution must be applied in using such techniques. Supervised methods include discriminant analysis, canonical variate analysis, model-based clustering, neural networks, support vector machines and cell automata. All require a priori assumptions and/or "target" and "background" definitions to which unknown data can be classified. Typically, target populations represent sets of geochemical data that define mineral exploration targets.

Visualization of Geochemical Data

Visualization is one the most effective ways of evaluating data. The human eye is very adept at recognizing patterns from pictures than with tables of numbers. Geochemists need to evaluate data comparatively in both the spatial domain (geographic location) and the variable (element/oxide) domain. When evaluating single elements data can be evaluated using simple plots such as probability plots (Sinclair, 1976; Stanley and Sinclair 1987, 1989; Stanley, 1987), histograms, or box plots. However, there are many other ways to evaluate data graphically. Many of these methods have been outlined by Cleveland (1993). Garrett (1988) developed a data analysis, statistics and visualization system, IDEAS, that provides a multitude of methods that are useful to the exploration geochemist.

Even the field of statistical evaluation of data has changed significantly in the past 10 years. This is exemplified by texts that combine extensive visualization techniques together with modern statistical methods (Venables and Ripley, 2002).

This contribution has made extensive use of the data analysis and statistical analysis software package, R (CRAN, 1999), which provides a number of powerful tools for manipulating and visualizing data. Most of the statistical graphics herein have been created using R. The application of this environment for geoscience applications is described by Grunsky (2002a). Recently a new library of statistical routines for the visualization of geochemical data (rgr) has been published on the Comprehensive R Archive Network (CRAN) (www.r-project.org) by Garrett (personal communication, 2007).

Geographical Information Systems

Geographic Information Systems represent digital visualization of spatially-based data on a map. Geographical Information Systems require a spatial definition of the data plus attribute tables that contain information relevant to the specified geographic locations and the representation of geochemical data. Examples of this have been presented by Bonham-Carter (1989a,b), Hausberger (1989), Gaál (1988), Kuosmanen (1988), Mellinger et al. (1984), Mellinger (1989), and George and Bonham-Carter (1989). In particular, a GIS facilitates the organized storage and management of spatially based data that are linked to a number of other features or other georeferenced data sets.

Bonham-Carter (1994) has written a monograph of geoscience applications using GIS and Harris (2006a) has edited a volume on GIS applications in the earth sciences covering a wide range of topics in which geochemistry is covered by (Grunsky, 2006; Cheng, 2006; Wilkinson et al., 2006; Harris, 2006b).

Depending on the nature of the geochemical data (stream sediment, soil, lake sediment, or lithogeochemical) various types of analysis can be performed that are dependent on the type of associated data present. Point, polygon (vector), and raster (regular array cells) features can be overlain, merged and analyzed through the associated map merging and database querying tools. Raster image grid cells can be considered as points provided there is an associated attribute record of data with each grid cell.

As geoscience information and data become available in ever-increasing volumes, exploration programs and government research programs involve significant amounts of data compilation. The compiled datasets are subsequently placed into a GIS and integrated with other geoscience information. Recent developments in the use of Geographical Information Systems together with data compilation programs have been discussed in Wilkinson et al. (1999); and Harris et al., (1997, 1999, 2000) and a book with a chapter on the evaluation of geochemical data using GIS's (Harris, ed., 2006a, Chapters 12-16).

Desktop mapping systems have evolved to the point that they are cheaper and less complex, are easier to use and offer an effective way for the geochemist to evaluate data. Thus, the goals of the geochemist can be achieved faster and at less cost. As digitally based map and attribute data are being continually created, there has been an increasing demand to view and assess these data without the use of complex GIS's. In its simplest form, a desktop mapping system has significant advantages in exploration geochemistry. Geochemical data can be loaded and visualized in both the geochemical space and the geographic space very quickly. Geochemical data can also be processed using a number of statistical or other data analysis techniques from which the results can also be loaded into a desktop mapping system. The permutations and combinations of data layer manipulation provide a wide variety of ways of examining and interpreting data.

Image Processing

When the sampling density of geochemical data is adequate, it is desirable to produce maps that represent smoothed gridded data and coloured/shaded surfaces. Smoothed, gridded data can be considered a raster image. Image analysis is primarily used for presentation purposes to enhance the results of an analysis or to show variation within data. Image analysis manipulates integer scaled raster data using a number of matrix based methods and after the use of additional integer scaling procedures represents the resulting transformed data on various graphical output devices using colour (e.g., intensity, hue, saturation, RGB, CMYK). Richards and Jia (1999) provides an introduction to image processing methods. Carr (1994) provides an introduction to image processing in geological applications and Gupta (1991) and Vincent (1997) provide comprehensive reviews of remote sensing applications in geology. Rencz (1999) contains a collection of papers covering the topic of remote sensing in the earth sciences and Pieters and Englert (1993) covers the topic of remote geochemical analysis through the evaluation of satellite spectroscopy.

Exploratory Data Analysis

Exploratory data analysis is concerned with analyzing geochemical data for the purpose of detecting trends or structures in the data. These features can provide insight into the geochemical/geological processes from which 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. These include the use of summary tables (minimum, maximum, mean, median, standard deviation, 1st and 3rd quartiles), measures of correlation, covariance and skewness. Graphical methods include histograms, probability (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 and symbol plots, and interpolated grids.

Multivariate methods include the use of principal components analysis, cluster analysis, Mahalanobis distance plots, empirical indices and various measures of spatial association.

Target and Background Populations

Geochemical background represents a population of observations that reflect unmineralized ground. Background may be a mixture of several populations (gravel, sand and clay or granitoid, volcanic and sedimentary lithologies). The separation of the background population into similar subsets that represent homogeneous multivariate normal populations is important and forms the basis of the modeled approach of geochemical data analysis. This can be achieved using exploratory methods such as principal components analysis, methods of spatial analysis, Mahalanobis distance plots and cluster analysis.

A group of specimens that represent an entity under investigation (features of geochemical alteration or mineralization) is termed the “sample” population, from which inferences will be made about the “target” population that cannot be sampled in its entirety. These populations are derived from specimens collected from orientation studies over known mineral deposits or areas of specific interest.

Sample populations, whether representing background or other populations, represent training sets with unique characteristics. These training sets are generally distinct from one another through their statistical properties although it is common for training sets to overlap. Unknown specimens can be tested against these populations to determine if they have similar characteristics. Probability based methods can determine if the unknown specimen belongs to none, one or more of the populations.

Developing training sets and testing unknown specimens is part of the modeled approach to evaluating geochemical data and will not be discussed here. This topic and associated references is discussed in Grunsky (2000).

Special Problems

Problems that commonly occur in geochemical data include:

- many elements have a "censored" distribution, meaning that values at less than the detection limit can only be reported as being less than that limit;
- 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. Often, missing values are reported as zero, which is not the same as a specimen having a zero amount of an element. This can create complications in statistical applications.
- combining groups of data that show distinctive differences between elements; where none is expected. This may be the result of different limits of detection, instrumentation or poor Quality Assurance / Quality Control procedures. Leveling of the groups is required;
- the constant sum problem for compositional data.

These problems create difficulties when applying mathematical or statistical procedures to the data. Statistical procedures have been devised to deal with all of these problems. In the case of varying detection limits, the data require separation into the original groups so that appropriate adjustments can be applied to the groups of data.

To overcome the problems of censored distributions, procedures have been developed to estimate replacement values for the purposes of statistical calculations. When data have missing values, several procedures can be applied to impute replacement values that have complete analyses. This will be discussed in more detail further on in the text.

Figure 3 summarizes the problems of censoring, non-normality and the discrete differences in the data due to analytical resolution. 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.

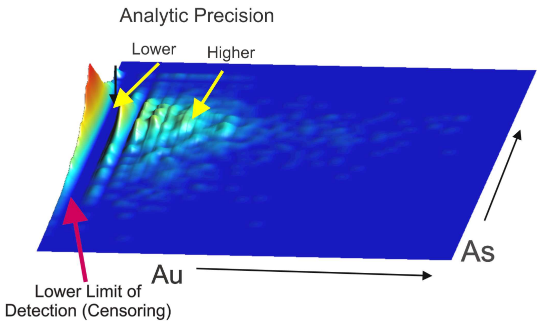


Figure 3: Density plot of Arsenic versus Gold displaying censoring and quantization of the analytical data.

Standard numerical and statistical methods have been developed for data analysis where the values being considered add to a constant sum (e.g. whole rock analyses summing to 100%). This is discussed in more detail below.

Quality assurance and quality control of geochemical data require that rigorous procedures be established prior to the collection and subsequent analysis of geochemical data. This includes the inclusion of certified reference standards, randomization of samples and the application of statistical methods for testing the analytical results. Historical accounts of Thompson and Howarth plots, for analytical precision studies can be found in Thompson and Howarth (1973, 1976a, 1976b, 1978). Additional discussion on the subject was most recently covered by (Stanley, 2003, 2006; Garrett and Grunsky, 2003)

Compositional Data

Geochemical data are reported as proportions (weight %, parts per million, etc.) For a given observation compositional proportions (i.e. weight %) always sum to a constant (100%). As a result, as some measures increase, others are "forced" to decrease to keep the sum constant. Because compositional data occur only in the real positive number space, the calculation of statistical measures such as correlation and covariance, can be misleading and result in incorrect assessment of correlation or other measures of association. However, the problem is seldom severe because the analysis of multi-element geochemistry using trace elements usually does not represent an entire composition, i.e. they only sum to a few percent of the total. The effect of closure, in most cases, may have little or no effect on the outcome of most statistical procedures. However it is dangerous to make the assumption that closure has no effect on the outcome of any statistical measure.

Aitchison (1986) developed a methodology for data analysis and statistical inference of compositional data using logratio transformations. These transformations project the compositional data into the entire (positive and negative) real number space, which allows standard statistical procedures to be applied. These methods are gaining popularity and examples of application to geochemical data are given by, Aitchison (1990), Grunsky et al. (1992) and Buccianti et al. (2006). The approach has also been extended into spatial data processing that is commonly used in ore reserve estimation (Pawlowsky, 1989). Recent work by von Eynatten et al. (2002, 2003), Pawlowsky-Glahn and Buccianti (2002), Martin-Fernandez et al. (1998, 2000) and Barcelo et al. (1995, 1996, 1997) document methods and issues around the treatment of compositional data. Aitchison (1997) provides a very readable account of compositional data issues. Appendix 1 provides a basic description of the use of logratios. Buccianti et al. (2006) provide the most recent developments in the field of compositional data analysis. A package for compositional data analysis (van den Boogaart and Tolosana-Delgado, in press), (compositions) provides a set of tools for evaluating compositional data using the R statistical package (www.r-project.org).

Most geochemical survey data are comprised of trace element measurements that are reported as parts per million (ppm). The reporting in parts per million constitutes the potential for closure, the trace element concentrations may interfere with each other particularly when one or more of the elements of interest are close to zero. The application of a centered logratio transformation (clr) will provide more reliable and statistically defensible results than the use of raw data and if balances can be constructed, an orthonormal basis of the variables will result for which statistical and vector calculations can be applied.

SUMMARIZING GEOCHEMICAL DATA

Univariate Data Summaries

The following description of data exploration is based on examining univariate populations. Exploratory Data Analysis (EDA) plots are shown in Figures 4a-d and 5a-d. These plots are often useful when grouped together as they provide different ways of summarizing data. Data summaries, in combined graphical and text form, provide a basis for context and comparison of different data types.

Histograms

The histogram is one of the most popular graphical means of displaying a distribution since it reflects the shape similar to theoretical frequency distributions. Figures 4a and 5a illustrate how the histogram can be used to display the distribution of Al and As in lake sediments. These two elements have been chosen to demonstrate two very different geochemical responses. Aluminum is ubiquitous in the lake sediments, mostly derived from alumino-silicates such as feldspars. Aluminum abundance is largely controlled by rock types such as granites and volcanic rocks. Figure 4a illustrates the range of Al values from sediments in lake catchments. The distribution appears polymodal, which could lead to the interpretation that the lake sediments have been derived from several different lithologies. In the Batchawana area of Ontario, these lithologies are granite gneiss, migmatite, granitoid intrusions, metasediments and metavolcanic rocks. However, on closer examination these "peaks" appear to be artifacts of analytical method (varying detection limits) and can create difficulties with the interpretation. Other graphical methods that are discussed below are better suited for interpreting these data.

Arsenic is much less abundant in the country rocks of the area. When it is present, it is usually associated with mineralization. Relative to Al, elevated amounts of As is a "rare event". This is reflected in the histogram of Figure 5a where most As values are below 10 ppm.

For constructing a histogram a number of objective procedures have been established as initial starting points for interval selection (see Venables and Ripley, 2002. page 112). If the nature of the distribution is normal or close to normal then Sturge's rule can be applied. Sturge's rule sets the number of intervals equal to log2n +1 where n is the number of observations. Sturge's rule does not work well if the distributions are not normal. If the number of intervals is too few, then the finer details of the distribution are smoothed over. If the number of intervals is too many, then the distribution appears discontinuous.

Histograms can be tuned by experimenting with starting points, cut off points and interval selections. This process is subjective and when the end points and intervals are well chosen, a meaningful interpretation is likely. Conversely, if the end points and intervals are poorly chosen, an incorrect interpretation, or no significant interpretation can be obtained.

Box Plots

The box plot is a method used to display order statistics in a graphical form (Tukey, 1977). The main advantage of the box plot is that its shape does not depend on a choice of interval as does the histogram. Providing the scale of presentation is reasonable, the box plot provides a fast visual estimate of the frequency distribution. A box plot for As in lake sediments is shown in Figure 5b.

Within a box plot, the box is made up of the median (50th percentile), left and right hinges (25th and 75th percentile, or first and third quartile). The "whiskers" are the lines that extend beyond the box. Several variations exist on the graphical presentation of box plots. The extreme ends (maximum and minimum values) of the data are marked by vertical bars at the end of the whiskers. Alternatively, the whiskers can extend to the "fences", which are defined as the last value before 1.5*midrange beyond the hinges of the data. Observations that plot beyond 3*midrange are plotted as bars or special symbols. The location of the median line within the box gives an indication of how symmetric the distribution is within the range of the upper to lower hinge (midrange). The lengths of the whiskers on each side of the box provide an estimate of the symmetry of the distribution. Notches can also be added to the diagram, which identify the width of the confidence bounds about the median. Notches are evident in the box plot of Figure 4b, where the distribution of Al is not highly skewed. The notches are less obvious in Figure 5b because of the skewed nature of the data and the scaling of the plot.

When using these plots to compare datasets representing different lithologies, etc., the notches provide an informal statistical analysis. If the notches do not overlap, it is evidence that the difference between the medians is significant.

Density Plot

The distribution of data can also be described graphically through the use of density plots. Density plots are smooth continuous curves that are derived from computing the probability density function of the data. The density plot is similar to the histogram however the curve actually represents an estimate of the probability density function. Density estimation involves the use of smoothing procedures to compute the curves and is described in Venables and Ripley (2002, p. 126-132). Density curves can be modified by specifying the range of the data from which the smoothing and estimation is calculated.

Figure 4c shows a density plot for Al in lake sediments. The polymodal nature of Al is shown more clearly than in Figures 5a and 5b. Figure 5c shows the density plot for As where the skewed nature of the distribution is illustrated by the sharp single peak followed by a long tail.

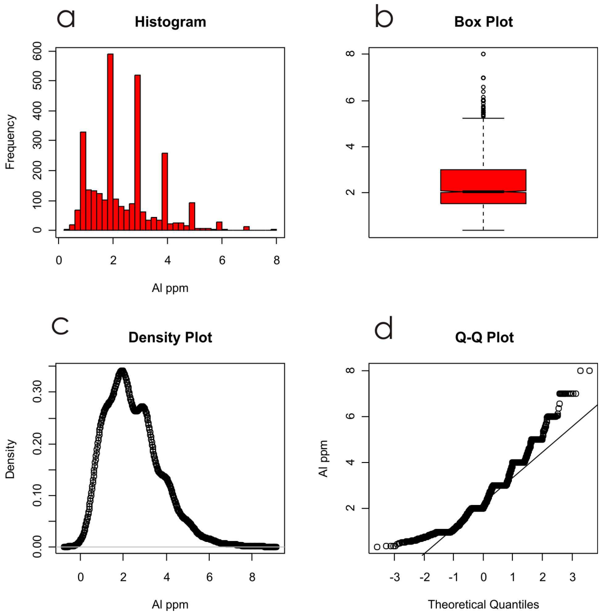


Figure 4: Exploratory Data Analysis (EDA) plot of Al in lake sediments, Batchawana area, Ontario. Note the distinct polymodal nature of the distribution. The Q-Q plot suggests that this polymodal appearance may be due to lack of precision in the chemical analysis.

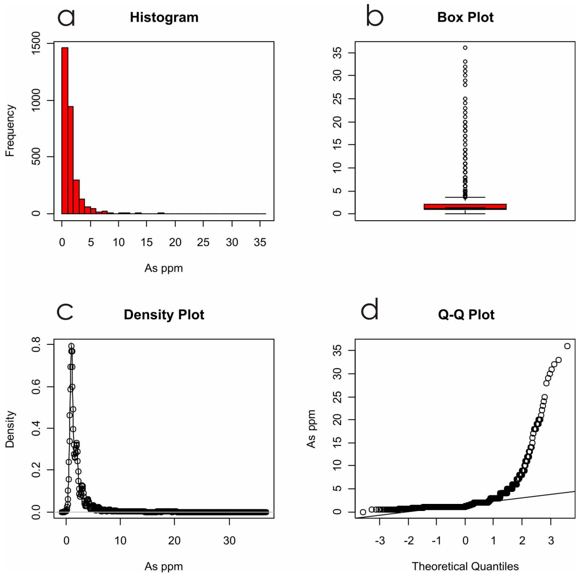


Figure 5: Exploratory Data Analysis (EDA) plot of As in lake sediments, Batchawana area, Ontario. Arsenic exhibits a log-normal type of distribution. Extreme values (outliers) influence the shape of the distributions in all four plots.

Quantile-quantile (Q-Q) Plots

Quantile-Quantile (Q-Q) plots are a graphical means of comparing a frequency distribution with respect to an expected frequency distribution, which is usually the normal distribution. Q-Q plots are equivalent to normal probability plots that have been extensively used by Sinclair (1976) for the analysis of geochemical data. Stanley and Sinclair (1987, 1989) and Stanley (1987) have written extensively on the use of probability plots for dissecting populations. A general description of Q-Q plots can be found in Venables and Ripley (2002, p. 108). These plots are generated by calculating quantile values for the normal frequency distribution (value of the normal frequency distribution over the range of probability, 0.0 to 1.0) and then plotting these against the ordered observed data. If a frequency distribution is normally distributed, when the quantile values are plotted against the ordered values of the population, the plot will be a straight line. If the frequency distribution of the population is skewed or the population is polymodal, the Q-Q plot will be curved or discontinuous. The advantage of the Q-Q plot is that each individual observation is plotted and thus the detailed characteristics of groups of observations can be observed.

Figure 4d shows a Q-Q plot for Al in lake sediments. The plot provides some insight into the nature of the data that is not shown by any of the other three plots (Figures 4a-c). The "stepped" nature of the plot suggests that the values of the data are not continuous but are reported as discrete values rounded off at the nearest part per million. The step-like pattern indicates that measurements were made in 1 ppm increments for some of the data and in 0.1 ppm increments for other data. In fact, the pattern that is observed is a mixture of four surveys, three of which have a resolution of 1 ppm for Al, and the fourth survey has a resolution of 0.1 ppm. The departure of the stepped plot from the straight line indicates that it is a slightly skewed distribution. Figure 5d shows the Q-Q plot for As which clearly reveals the non-normal nature of the distribution by its non-linearity. Q-Q plots are also useful for identifying extreme values at the tails of the distribution. The line that cuts through the data represents the intersection at the 25th and 75th percentiles of the data. In the case of the As data, it is clear that the distribution is very skewed.

Summary Statistical Tables

Summary statistical tables, are useful descriptions of data when quantitative measures are desired. Summary statistical tables commonly include listings of the minimum, maximum, mean, median, 1st quartile, and 3rd quartiles. Measures of dispersion include the standard deviation, median absolute deviation (MAD), and the coefficient of variation (CV). The coefficient of variation is useful because the dispersion is expressed as a percentage (the mean divided by the standard deviation), so it can be used as a relative measure to compare different elements. An example of a summary table for a selected group of elements from the lake sediment data is shown in Table 1. The table lists minimum, maximum, mean, median and selected percentile values for 35 elements and loss on ignition (LOI). Comparison of the mean and median values for each of the elements shows that many of them are significantly different from each other. This implies that the distributions for these elements are not normal and are likely skewed.

Summary tables are useful for the purpose of publishing actual values, however graphical methods, as previously described, provide visualization about the nature of distributions and the relationships between observations. The values of a summary table are best interpreted when used in combination with graphical summaries.

Spatial Presentation

It is particularly meaningful to display geochemical survey data in a geographical context. A GIS is a very useful tool for evaluating geochemical data during the exploratory analysis phase. Figure 6 shows a symbol plot of As from lake sediments in the Batchawana area of Ontario. Each symbol represents a collection site. The number of symbols and the symbol sizes were chosen based on an evaluation of the accompanying Exploratory Data Analysis (EDA) plot. An initial view of the EDA plot for As showed that the distribution was positively skewed and the plot was difficult to interpret. A log10 transform was then applied to the data values and the resulting EDA plot was much easier to interpret. The EDA plot of Figure 6 shows at least 4 distinct populations. The first population ranges in values from < 0.-02 to 0 log10 scale (.9 to 1 ppm) and is related to the many specimens with As values close to the detection limit. The second population ranges from 0 to 1.2 log10 scale (1 to 16 ppm) and reflects background As values associated with the geology. The third population ranges from 1.2 to 1.6 log10 scale (16 to 40 ppm) and occurs mainly in the south central part of the Batchawana greenstone belt in an area where there is known pervasive carbonate alteration associated with shear zones. The fourth population ranges from 1.6 to 2.0 log10 scale (40 to 100 ppm) and represents areas where there are known sulphides.

The choice of symbol size and colour can be used clearly illustrate patterns of similarity and difference between several elements in the data. If the goal is to illustrate atypical observations, then once a background range of values has been established, observations that exceed the limit of the background can be assigned unique colours or different sized symbols. If the distribution of the data is non-normal and the observations of interest are in the positive tail of the distribution, then a logarithmic scale can be used to assign symbol sizes.

Kürzl (1988) and Reimann et al. (2005) suggests a unique approach by creating symbols based on exploratory data analysis methods. Using the divisions within a box plot,
the median value (Q2) and the interquartile range Q1-Q3 (r),
the upper fence (Q3 + 1.5*(Q3-Q1),
the lower fence (Q1 – 1.5*(Q3-Q1),
lower outside values (Q1 – 3*(Q3-Q1)) and
upper outside values (Q3 + 3*(Q3-Q1)) can be used to define unique symbols which express the ranking of an observation. An example of a seven symbol set can be defined as:
1 < lower outside values
2 lower outside values to the lower fence
3 lower fence to Q1
4 Q1 to Median (Q2)
5 Median (Q2) to Q3
5 Q3 to upper fence
6 upper fence to upper outside values
7 > upper outside values 5 Q3 to Q3 + 1.5*r

Table 1: Summary Statistics for Lake Sediments, Batchawana Area, Ontario

	LLD	Num Obs	Min	1%	5%	10%	25%	50%	Median	Mean	75%	90%	95%	99%	Max	S.D.	MAD	C.V
LOI	2.96	3019	3	35.5	40.11	40.5	43.94	49.5	44	44	53.5	56.54	57.25	57.5	91.5	13.7	13.3	0.3
Ag	0.2	2900	0.2	0.2	0.2	0.2	0.2	0.3	0.5	0.7	1	1	1	1	72	1.5	0.4	2.3
Al	0.36	3047	0.4	1.385	1.955	2	2.19	2.22	2	2.5	3	3.5	3.5	4.17	8	1.2	1.4	0.5
As	0.5	3046	0.5	0.85	0.9	1	1.25	1.3	1.2	2.2	1.5	1.5	2	2	96	4	0.4	1.8
Au	1	3042	1	1	1	1	1	1.5	1	2.1	2	2	2.5	4	64	2.1	0	1
Ba	30	3047	30	132	156.5	160	160.5	175	148	167.8	178.5	195	235	295	680	85.2	71.2	0.5
Be	0.5	3047	0.5	0.5	0.5	0.5	0.5	0.75	0.5	0.8	1	1	1	1	54.1	1	0	1.3
Bi	2	3047	2	2	2	2	2	3	2	2.9	5	5	5	5	10	1.4	0	0.5
Br	1	3046	1	3.4	14.5	17.45	18.05	31	22	25.6	34.5	37.95	43	57.5	132	16.1	14.1	0.6
Ca	0.23	2685	0.2	0.71	0.805	0.87	0.915	1	1	1	1	1	1	1.08	9.1	0.4	0.1	0.4
Cd	0.2	3047	0.2	0.3	0.45	0.5	1	1	1	1	1	1.05	1.1	1.5	6	0.6	0.3	0.5
Co	1	3047	1	4	4.5	5	5.5	5.5	6	6.9	6.5	6.5	8	10.5	105	5	3	0.7
Cr	1	3047	1	18	25.5	26	31.5	32	27	31.3	32	41	41.5	47.5	328	18.2	13.3	0.6
Cu	2	3047	2	13	17	21	23.5	28	29	34.2	31.5	37.5	44	45.5	441	24.3	14.8	0.7
Fe	0.14	2649	0.1	0.4	0.45	0.965	1	1.5	1	1	1.5	1.5	1.505	1.745	15	0.7	0.3	0.7
Hf	1	3046	1	1	1.5	2	2	2.5	2	2.3	2.5	2.5	3.5	6	30	1.4	1.5	0.6
K	0.05	1809	0.1	0.19	0.265	0.33	0.37	0.43	0.3	0.5	0.56	1	1	1	2	0.3	0.3	0.7
La	1	3046	1	13	20.5	27	27	31	25	29	38	44.5	50	50.5	408	19.3	13.3	0.7
Lu	0.1	1605	0.1	0.15	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.25	1	2	0.2	0	0.7
Mg	0.04	1636	0	0.09	0.095	0.1	0.15	0.28	0.2	0.3	0.285	0.305	0.31	1	2	0.2	0.1	0.9
Mn	20	3047	20	76	89	97.5	101.5	127	114	159.8	134	142.5	160	309.5	3410	168	77.1	1.1
Mo	1	3047	1	1	1.5	1.5	1.5	2	2	2.3	2.5	2.5	2.5	3	84	3.2	1.5	1.4
Na	0.03	1999	0	0.17	0.355	0.44	0.52	0.94	0.5	0.7	1	1	1.055	2	4	0.5	0.5	0.8
Ni	3	3047	3	11.5	12	15	15.5	16.5	16	17.3	18	19.5	22.5	29	153	7.9	5.9	0.5
P	150	2197	150	300	515	650	825	830	820	941	970	1060	1105	2315	4700	508.6	474.4	0.5
Pb	2	3047	2	7	8	10	11	12	10	11.6	13.5	14	16	17	1340	27.3	5.9	2.4
Sb	0.1	1627	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.15	0.15	0.2	1	7	0.3	0	1.8
Sc	0.1	3046	0.1	2.6	4.5	5.35	5.4	6	5	5.2	6.35	6.5	7.5	7.5	19	2.2	1.5	0.4
Sr	12	3047	12	48	50	60.5	66	66	60	78.3	94	109.5	117	170	427	54.3	34.1	0.7
Ta	0.5	3046	0.5	0.5	0.5	0.5	0.5	2	2	1.4	2	2	2	2	3	0.7	0	0.5
Th	0.4	3044	0.4	1.9	2.4	2.5	3.25	3.3	3	3.3	3.65	4.5	5.5	8	26	1.7	1.5	0.5
Ti	0.009	1557	0	0.03	0.05	0.057	0.06	0.06	0.1	0.1	0.076	0.105	0.121	0.255	0.3	0	0	0.5
U	0.1	3009	0.1	1.9	2.3	2.5	2.65	2.95	2	4.2	4.1	4.5	5	18.5	195.5	7.5	1.5	1.8
V	5	3047	5	11	18.5	22.5	24	24.5	24	27.1	27.5	37	41	45.5	301	15.9	13.3	0.6
W	1	3046	1	1	1	1	1	1	1	1.7	1	1.5	1.5	2	46	1.7	0	1.1
Zn	13	3047	13	52	62.5	63.5	75.5	98	86	98.6	102.5	114	116.5	145	952	68.1	38.5	0.7

S.D. - Standard Deviation

C.V. - Coefficient of Variation

MAD - Median Absolute Deviation

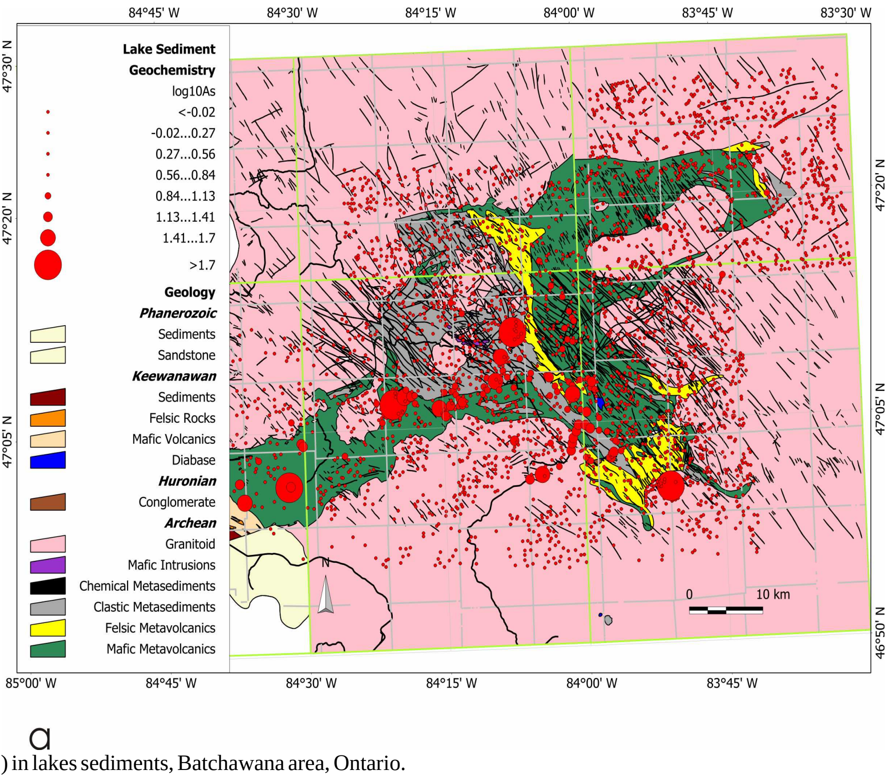


Figure 6a: Arsenic (log10) in lakes sediments, Batchawana area, Ontario.

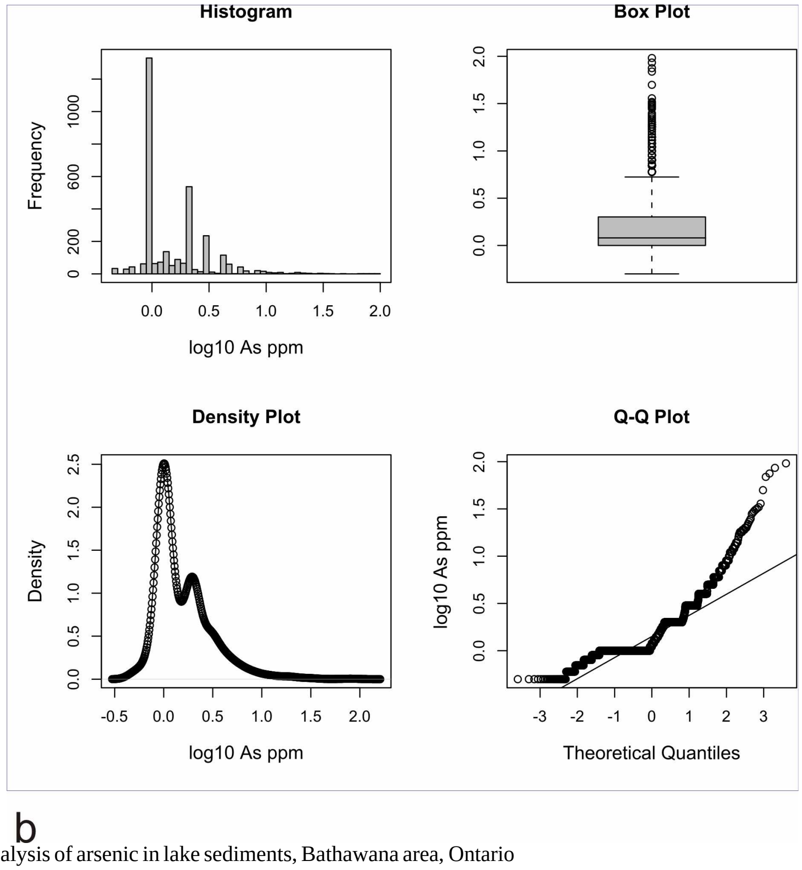


Figure 6b: Exploratory data analysis of arsenic in lake sediments, Bathawana area, Ontario

Application of Geostatistical Techniques for evaluating the spatial continuity of geochemical processes

Contouring or imaging techniques are most reliable when the sampling density is sufficient enough so that variation between sample sites is minimal for the purposes of the sampling survey. Subjective judgment is often employed for a decision to use contouring or imaging techniques. If the sampling density is high, but the geochemist/geologist believes that the geochemical response between sample sites is predictable, then contouring or imaging may be an appropriate means of visually describing the data. If the geochemical variability between sampling sites is unknown or large then it is better to use point or bubble plots as described previously. A quantitative way of assessing spatial variability can be carried out by the use of geostatistical procedures. The construction of a semi-variogram or correlogram can provide a measure of the spatial continuity / variability of a specific element. A semi-variogram measures the average variance between sample points at specific distances (lags). Generally, as the distance increases between any pair of points, the variance is expected to increase, the limit of which is the total variance of all of the data. In the correlogram, as the distance between any pair of points increase, the average correlation between the points decreases, eventually decaying to zero. Isaaks and Srivastava (1991, Chapter 4) describe a number of detailed methods for evaluating the spatial continuity of data. The effectiveness of employing geostatistical methods relies on an adequate sampling density in terms of representing the actual variation of the data as well as the spatial distribution of the points themselves.

A large number of freeware and commercial geostatistical software packages are now available for carrying out geostatistical analysis. The website www.ai-geostats.org provides a list of software that is currently available. A geostatistical package (gstat) has been written for the R programming environment (Pebesma, 2004), which is freely available from the Comprehensive R Archive Network (CRAN) (see: www.r-project.org). Deutsch and Journel (1997) provide a library of software routines in Fortran. A general introductory discussion on spatial statistics can be found in Venables and Ripley (2002, Chapter 15) and Davis (2002, Chapter 5).

If the spatial sampling density appears to be continuous then it may be possible to carry out spatial prediction techniques such as spatial regression modeling and kriging. A major difficulty with the application of spatial statistics to regional geochemical data is that the data seldom exhibit stationarity. Stationarity means that the data has some type of location invariance, that is, the relationship between any sets of points are the same regardless of geographic location. Thus, interpolation techniques such as kriging must be applied cautiously, particularly if the data cover several geochemical domains in which the same element has significantly different spatial characteristics.

Evaluation of the variogram or the autocorrelation plots can provide insight about the spatial continuity of an element. If the autocorrelation decays to zero over a specified range, then this represents the spatial domain of a particular geological process associated with the element. Similarly, for the variogram, the range represents the spatial domain of an element, which reaches its limit when the variance reaches the "sill" value, the regional variance of the element. Theoretically, at the origin, the variance should be zero at lag zero. However, typically, an element may have a significant degree of variability even at short distances from neighbouring points. This variance is termed the nugget effect.

Figure 7 displays 4 semi-variograms for Zn from the Batchawana lake geochemistry survey data covering an area of 95 km (east-west) and 62 km (north-south). Semi-variograms have been calculated for 4 preferred directions; East-West, (0 degrees), North-South (90 degrees), Northeast-Southwest, (45 degrees) and Southeast-Northwest (135 degrees). The y-axis of each figure is the correlation and the x-axis is the lag interval. The maximum lag distance was chosen as 20,000 meters and the lag interval was selected as 200 meters. The selection of a suitable lag distance can be made by visually examining the distribution of sample sites Geostatistical software packages can also determine optimum lag intervals. These figures were generated using the gstat package from R. Each figure has been fitted with an exponential model. The most regular semi-variograms appear for the 135 and 90 degree orientations. This is no surprise given that that there are two primary stratigraphic orientations in the area, one trending east-west and the other trending southeast to northwest. The orientations of 0 and 90 degrees display different nugget values, with the lowest nugget occurring with the east-west orientation, also suggesting better correlation between adjacent points in that direction. The 45 degree orientation displays a straight line indicating that for the range considered (20,000m), maximum variance is not achieved. It should also be noted that all four semi-variograms display periodicity, which indicates that there is heterogeneity in the spatial structure of the data, most likely reflecting changes in the underlying geology (granite versus greenstone).

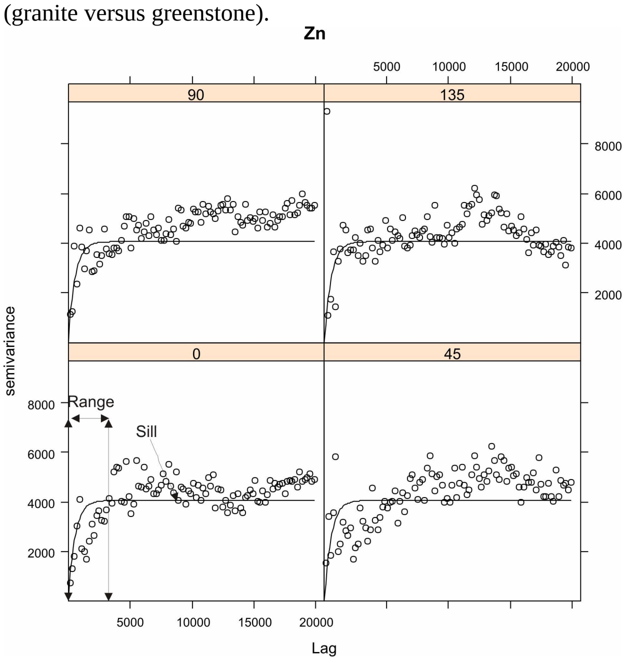


Figure 7: Semi-variogram of Zn from lake sediments, Batchawana area, Ontario. Semi-variograms are derived for four different orientations.

The use of kriging makes some assumptions about the spatial uniformity (stationarity) properties of the data. In many cases, particularly in regional sampling programs, there are several lithological domains in which elements have different spatial ranges. Kriging can account for various types of spatial drift in datasets, however the error in the kriged estimates tends to increase.

The use and application of geostatistical methods is a combination of art and science. Skill, knowledge and experience are required to effectively use geostatistical techniques. It requires considerable effort and time to effectively model and extract information from spatial data. The benefits of these efforts are a better understanding of the spatial properties of the data which permits better estimates of geochemical trends. However, they must be used and interpreted with the awareness about problems with techniques of interpolation and the spatial behavior of the data.

Fractal Methods

The use of fractal mathematics is playing an increasingly important role in the geosciences. Carr (1994) gave a good introduction into the use of fractal methods in the geosciences. Cheng and Agterberg (1994) have shown how fractal methods can be used to determine thresholds of geochemical distributions on the basis of the spatial relationship of abundance. They have shown that where the concentration of a particular component per unit area satisfies a fractal or multifractal model, then the area of the component follows a power law relationship with the concentration. This can be expressed mathematically as:

$$A(\rho\leq\nu) \propto \rho\text{-}\alpha1$$
$$A(\rho>\nu) \propto \rho\text{-}\alpha2$$

where $A(\rho)$ denotes an area with concentration values greater than a contour (abundance) value greater than ρ. This also implies that $A(\rho)$ is a decreasing function of ρ. If ν is considered the threshold value then the empirical model shown above can provide a reasonable fit for some of the elements.

In areas where the distribution of an element represents a continuous single process (i.e. background) then the value of α remains constant. In areas where more than one processes have resulted in a number of superimposed spatial distributions, there may be one or more values of α defining the different processes.

An example of the use of concentration versus area plots is shown for As derived from lake sediments collected over the Batchawana area. Figure 8 shows a colour contoured image of As values superimposed on the sample sites. As well, a plot of log10 As concentration versus log10 area occupied by each contour interval. Distinct changes in slope in the plot represent breaks based on the spatial distribution of the data and each break represents a threshold between populations of data possibly derived from different processes. There are three distinct trends shown on the concentration-area plot of Figure 8. The regional background is characterized by a straight line of points ranging from 0.7 (5ppm) to 1.3 (20ppm). Interpolated As values greater than 5ppm and less than 20 ppm are shown as red, blue and cyan. This represents the regional background of the area. The group of points that form a straight line from 1.3 (20 ppm) to 1.6 (40 ppm) represent the next population reflecting As associated with mineralization and anthropogenic effects. Anthropogenic effects are prevalent in the eastern part of the map area, whereas As values associated with potential mineralization are shown in the central and western part of the map area. Values above 1.6 (40ppm) represent a small population of observations that are greater than 40ppm (shown as orange and red on the map. These observations occur in the southeast portion of the map area and may represent areas of mineralization.

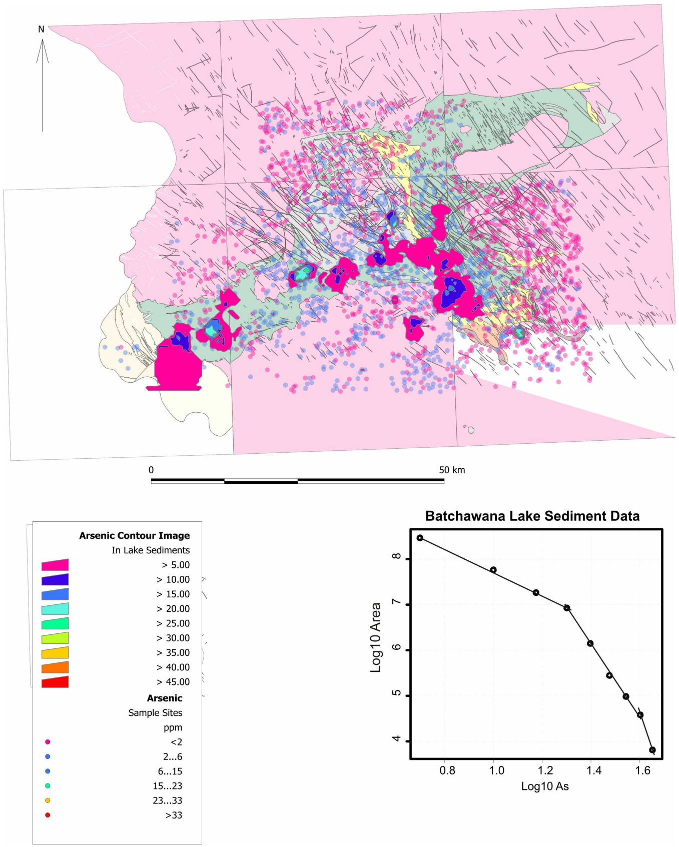


Figure 8: Arsenic from lake sediments, Batchawana area, Ontario. The contoured image reflects the area associated with each As contour level. The corresponding Concentration-Area plot display changes in slopes, which reflect changes in spatial patterns. These changes are associated in differences in geology, anthropogenic effects and mineralization.

Cheng et al. (2000) has also implemented the use of power-spectrum methods to evaluate concentration-area plots derived from geochemical data. By the application of filters, patterns can be detected related to background and noise, thus enabling the identification of areas that are potentially related to mineralization. More details on this methodology can be found in Cheng (2006).

Multivariate Data Summaries

Scatterplot Matrix

The scatterplot matrix a useful graphical multivariate methods for visually assessing the relationships between variables. When categorical information is available, colour can be used to show differences between the categories.

Two areas were chosen from the Ben Nevis mapsheet (Figure 9); one representing an area of carbonate alteration and the other, an area of metavolcanics without carbonate alteration. Figure 10 shows a scatterplot matrix of a selected number of elements from the two areas. The matrix highlights associations and patterns in the data. There is a clear distinction between the altered and unaltered observations for CO_2 with Co, Cu and Cr. CO_2 shows an overall increase for the altered specimens, whereas the abundances of Cu, Cr and Co vary widely in a suite of specimens from the carbonate alteration zone. The distribution patterns for these elements can be studied further using other graphical measures such as box plots.

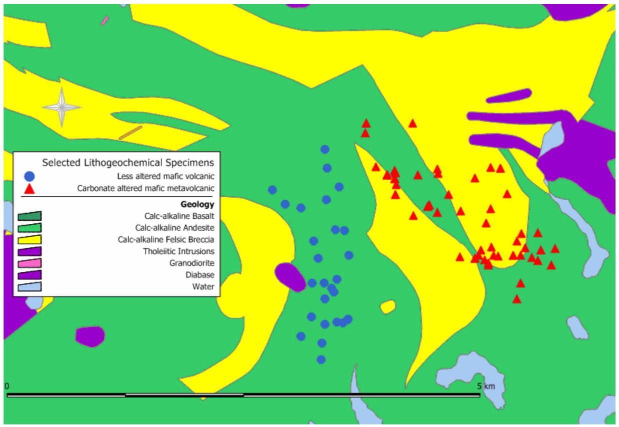

Selected Lithogeochemical Specimens
● Less altered mafic volcanic
▲ Carbonate altered mafic metavolcanic
Geology
Calc-alkaline Basalt
Calc-alkaline Andesite
Calc-alkaline Felsic Breccia
Tholeiitic Intrusions
Granodiorite
Diabase
Water
0
5 km

Figure 9: Map of altered/unaltered sampling sites in the Ben Nevis Township area.

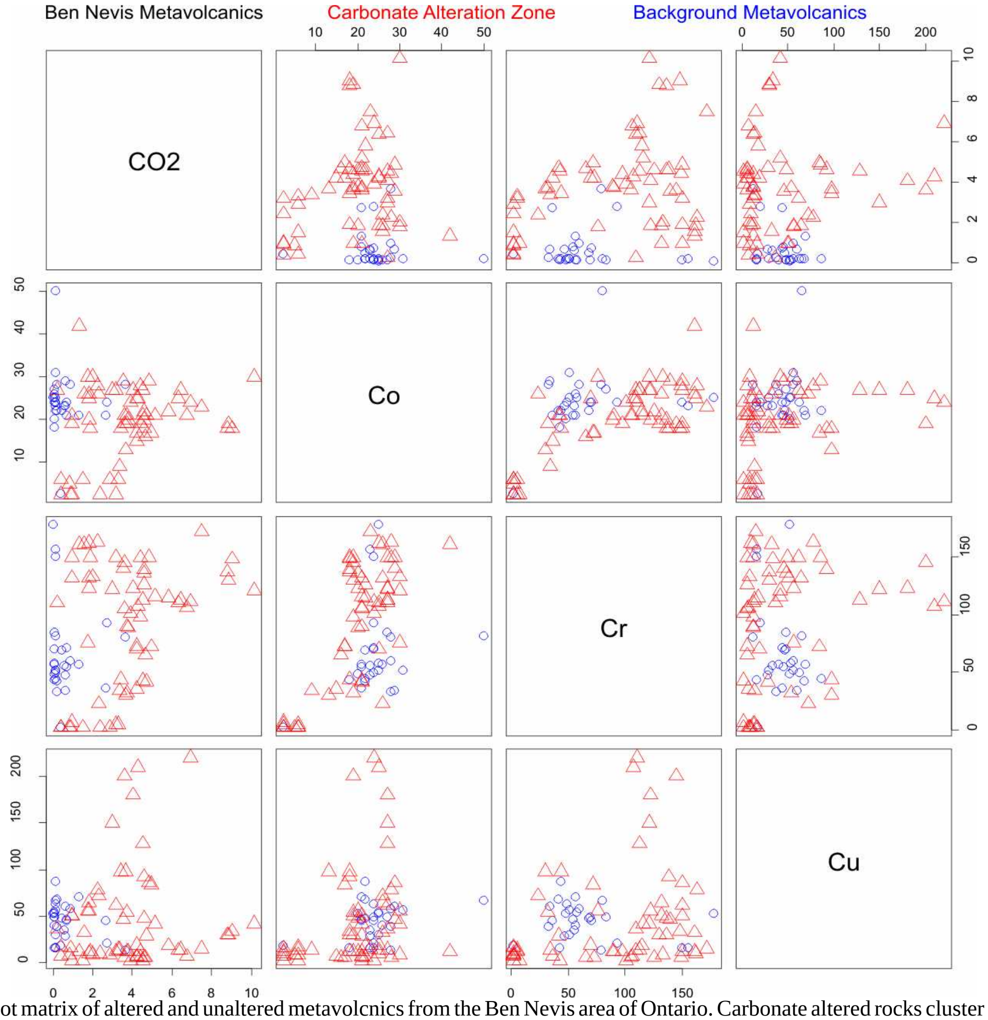


Figure 10: Scatterplot matrix of altered and unaltered metavolcnics from the Ben Nevis area of Ontario. Carbonate altered rocks cluster differently from the non-altered rocks.

Multiple Box Plots

In Figure 11 box plots for 9 elements from the Ben Nevis lithogeochemistry data show that there are clear differences in the geochemistry between the two areas. Box plots are a convenient way of summarizing the differences between groups of data. Note that there is a distinct shift in the median value data for CO_2, and Li (an increase) and a corresponding decrease in Ca and Sr for the specimens from the altered area. This is consistent with studies that indicate that there is overall loss of Ca and Sr in the zone of carbonate alteration, and an increase of Li and Na. Chromium, Na, Ni, Cu and Co show greater variability in the altered area. The greater variability is due to a breakdown of the original mineralogy accompanied with the addition of CO_2, Si, Li, Cu and several other elements that are associated with hydrothermal activity and mineralization.

DIFFERENTIATING GEOCHEMICAL BACKGROUND FROM ANOMALIES

The recognition of a geochemical anomaly requires that a geochemical background has been established, which in itself can be difficult to define. Geochemical values that depart from the background, that is, those values which are atypical, may be anomalous. Howarth and Sinding-Larsen (1983, p. 208) discuss the concept of anomaly and suggest that anomalous concentrations are those values that exceed a given threshold. Workshops held by the Association of Exploration Geochemists (AEG) in 1983 and 1985 (Garrett, 1984; Aucott, 1987) failed to give any formal definitions and concluded that an anomaly is a desired level of abundance in which the geologist has a particular interest and is different from the regional or background values. Joyce (1984, p. 15) discusses the definition of an anomaly in terms of an adequate definition of background.

Historically, values exceeding the 98th percentile were scrutinized for their potential to be identified as geochemical anomalies. As well, the threshold was defined as the mean ± 2 standard deviations (Hawkes and Webb, 1962; Howarth, 1983, p. 208). This definition was based on the assumption of normality of the data. However, with the introduction of computer-based methods for evaluating geochemical data, the ability to study sample populations and the nature of geochemical distributions has provided powerful tools for the identification of outliers and specimens that might be related to mineralization targets (anomalies). As a result, the use of choosing thresholds based on the calculation of the mean ± 2 standard deviations is no longer recommended (see Levinson, 1980; Rose, Hawkes, and Webb, 1979; and Garrett, 1989a). Filzmoser et al. (2005) describe an approach to outlier and anomaly detection using robust methods and adaptive techniques for recognizing outliers.

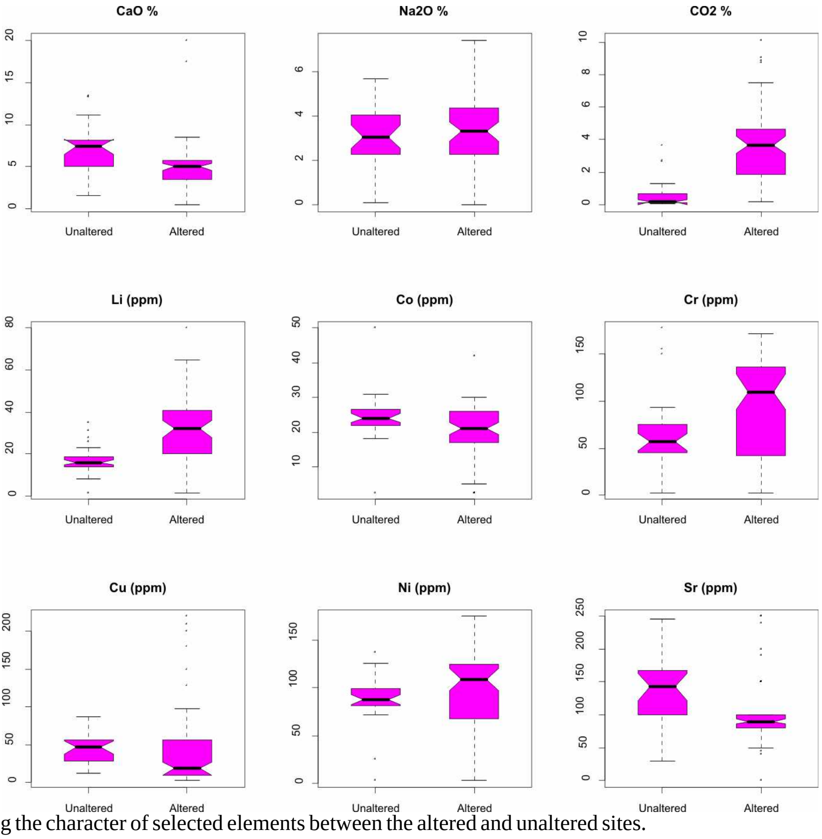


Figure 11: Box plots showing the character of selected elements between the altered and unaltered sites.

The Threshold and Pathfinder Elements

An important goal of the investigation of geochemical data is the detection of spatially continuous zones of elevated values of a strategic element that exceeds a specified threshold value. Observations that exceed the threshold are termed anomalies. Joyce (1984, p.9-13) provides a detailed description of indicator and pathfinder elements and minerals that can be used in exploration strategies. Garrett (1991) defined the threshold as the outer limit of background variation. The term "outer" is used instead of "upper". This allows the definition to include both "upper" and "lower" limits, as it is common in some geochemical environments for depletion haloes to be as important as enrichment haloes. Reimann et al. (2005) further refined the definition of threshold and background based on robust methods.

The concept of threshold can be extended from single element to multi-element data by the use of multivariate statistical methods such as the use of the Mahalnobis distance (Garrett, 1989b). In the multivariate case, the threshold can be selected on the basis of examination of Mahalanobis distance plots or some other more robust measure of background and departures from it.

Observations from distributions that represent processes of interest (mineralization or anthropogenic effects) usually overlap with observations from background distributions such that the threshold is more likely a range of values where the two distributions overlap. Rather than choose a specific threshold value, it may be better to assign a probability of the likelihood of an unknown specimen belonging to each population. In geochemical surveys, anomalies have a spatial association and are small and only occupy a fraction of the area that is covered by the regional population.

Figure 12 shows the threshold as determined by a visual inspection of the Q-Q plot. In this case, the threshold for K_2O is chosen at 2.5 %, which is considered above the usual range of values for volcanic rocks. The values that exceed the threshold can be identified on the map by choosing a symbol size or colour to identify them.

Mineral deposits are often characterized by a unique suite of elements whose values exceed the threshold of the surrounding background material. These elements are called pathfinder elements and often have a greater spatial extent relative to the target being sought. In the Ben Nevis metavolcanic sequence, K can be considered as a pathfinder element. Elevated values of K are typically associated with epithermal Au deposits. Examination of the distribution of K_2O in Figure 12b suggests that values above 2.5 wt% K_2O are atypical and that value defines the threshold. The map of K_2O values with in Figure 12a indicates that high K_2O values are associated with the two known mineral occurrences as well as several other sulphide bearing occurrences.

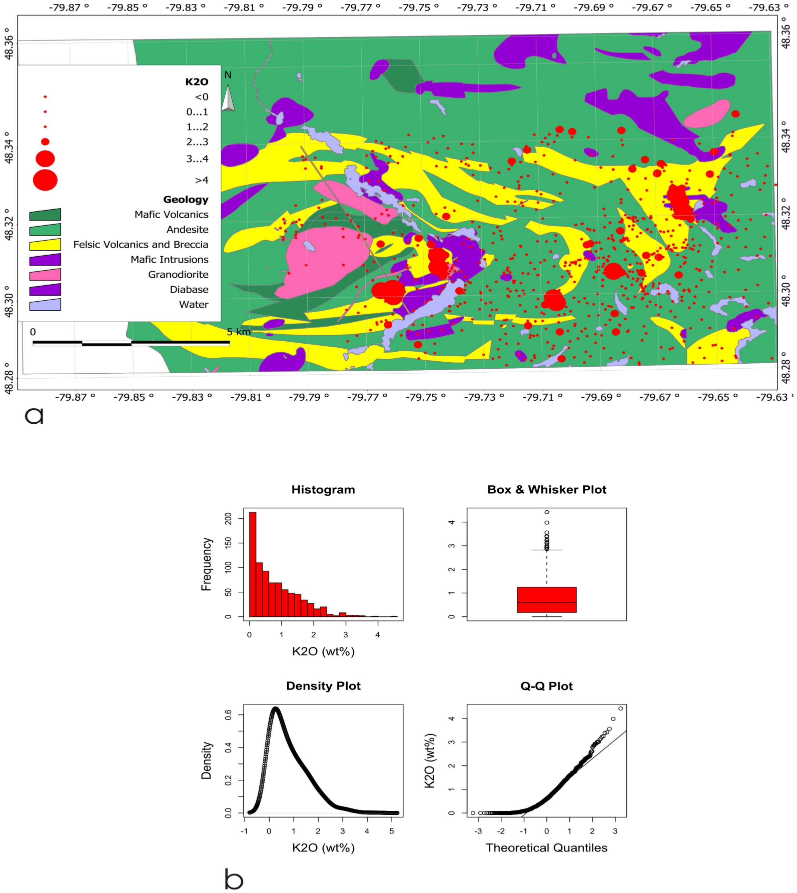


Figure 12: K_2O map across Ben Nevis Township. Separation of atypical K_2O values.

Outliers or Anomalies?

An outlier can be defined as an observation with a value that is distinctively different from observations with which it is intimately associated. If a threshold has been defined, then an outlier, by default, exceeds the threshold. Outliers may be of significance from an exploration or contamination point of view. An outlier may define a mineralized zone (anomaly) or a value that is above an accepted environmental background level. Outliers can also be artifacts of erroneous analytical results or data entries. An outlier can be identified as a geochemical anomaly if it exceeds the threshold, is not the result of an analytical problem, or assigned to an improper population. In other words, an anomaly is associated with a process of interest (alteration or mineralization), whereas an outlier is a value without an interpretation that requires further assessment.

Outliers should always be examined carefully to be certain that the observed values are not the result of an error. An observation that is an outlier in one group may be indistinguishable (masked) from other observations within another group. In practice, outliers are assessed by a graphical examination of the upper and lower rankings of the data and the identification of observations that occur as distinct breaks from the background population. The application of a transformation may be sufficient to separate the background from outliers.

Figure 13a shows a Q-Q plot of As from the lake sediment data. Arsenic, a pathfinder element, is commonly associated with gold deposits. An examination of the plot shows that “breaks” occur at the approximate values of 20, 25 and 35 ppm. In comparison with the fractal approach, the break at 20 ppm is equivalent the abrupt change in slope in Figure 8, where the concentration-area plot identifies a distinct change in the data population at a value of log10As=1.3 (19.95 ppm). These breaks most likely represent distinct populations that can be attributed to different source lithologies. The breaks are used as the basis for a change in symbol sizes on the map of Figure 13b. There are six extreme values, which occur above the level of 35 ppm, which is considered the threshold. These values can be considered as anomalies because of the break in the slope of the curve and the distance between these values and the bulk of the population. These outliers would be of interest in a mineral exploration program.

In the case of two or more (multi-modal) populations it will be necessary to decompose the populations into separate distinct populations through the analysis of Q-Q plots, probability plots or by computer-based means (Sinclair, 1976; Stanley, 1987; Bridges and McCammon, 1980). Garrett (1989b) and Filzmoser et al. (2005) have developed methods for outlier detection in multivariate data using a multivariate outlier plot, which identifies observations that appear to belong to a population different from the main population. This has obvious benefits in evaluating geochemical data for observations associated with alteration or mineralization.

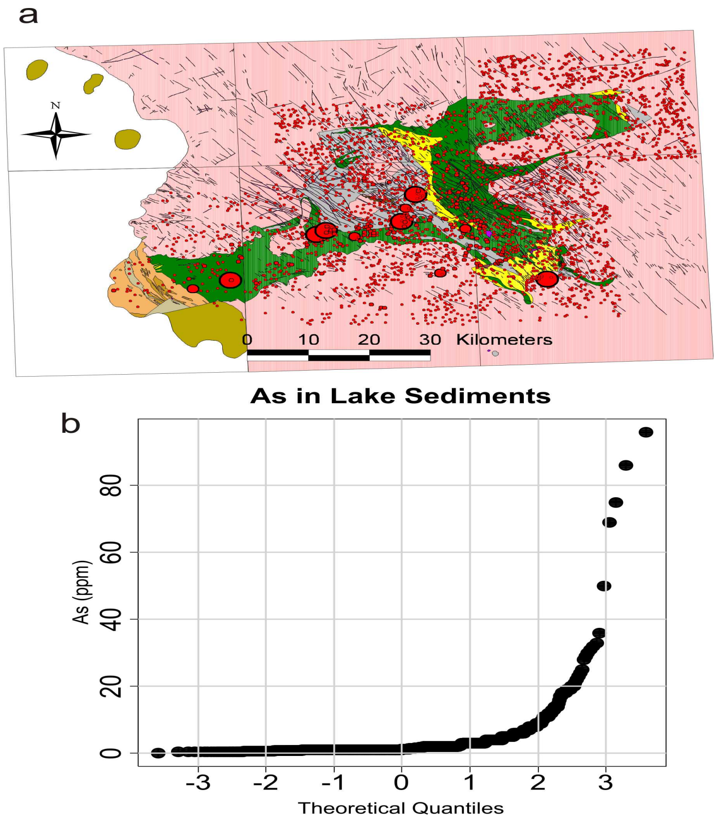


Figure 13: Map of atypical As (ppm) across the Batchawana area, Ontario.

Truncated and Censored Data

When an analytical procedure detects the presence of an element, but the value is too low to be accurately quantified, the value is reported as “less than the limit of detection” (lld). The same applies for values that exceed the upper limit of detection. The lower/upper limits of detection are the limits of reliable quantification by the analytical procedure. Typically, a laboratory will report the value prefixed with a “<” for a value less than the lld or “>” for a value that exceeds the upper limit of detection. When a group of values contains observations that exceed the detection limits, the effect is called censoring.

Figure 14 shows the distribution of Co in metavolcanics collected during lithogeochemical sampling program in the Ben Nevis township area of Ontario. The analytical procedure for Co has a lower limit of detection of 5.0 ppm and 85 out of the 824 observations fall below that limit. The histogram of Figure 14a shows a bar with a high frequency of observations at the lowest end of the scale. This bar represents the 85 values that are less than the detection limit. The Q-Q plot (figure 14b) shows these values as a flat part of the distribution at the left side of the figure. The density and box plots (Figures 14b,c) do not show the censored values as clearly. Historically, censored data were handled by applying a substitute value; somewhere between 1/3 to 1/2 of the actual detection limit. As the number of observations below the lld (censored) increases, then this estimate will produce inaccurate estimates of the mean and variance (see Sanford et al., 1993).

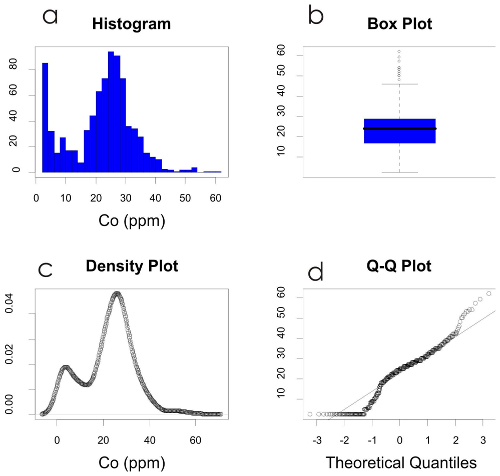


Figure 14: Cobalt (ppm) in metavolcanics, Ben Nevis Township, Ontario, Canada.

Several techniques have been developed to minimize the problem of censored data. The problem of censored data becomes more important when means of elements and covariances between elements are required. Using an arbitrary “replacement” value (i.e. ½ or 1/3 the lld) can introduce bias in the computation of the moments of the distribution. However, if the nature of the distribution can be assumed as normal, then the replacement value of the censored data and parameters of the distribution (mean, variance) can be estimated based on the portion of the distribution that is not censored. The process of finding suitable replacement values is known as “imputation” in the statistical literature. Estimates of the distribution parameters are obtained using the EM algorithm (Dempster, Laird, and Rubin, 1977), and is discussed by Campbell (1986) and Chung (1985, 1988, 1989). From these characteristics, an estimate can be made as to how the data is distributed below the (lld). The assumption of normality is essential for the EM algorithm to work. Campbell (1986) invokes an algorithm to transform the data to normality using Box-Cox. Sanford et al. (1993) have developed a method that allows for the calculation of a suitable replacement value based on a maximum likelihood approach. Helsel (1990) provides a detailed discussion on dealing with missing data in environmental studies. Chung (1985, 1989), Campbell (1986) and Lee and Helsel (2005, 2007) have published computer procedures that estimate the mean and variance of censored distributions by calculating a replacement value that is derived from the characteristics of the uncensored portion of the sample population. Dickson and Giblin (2007) have used self-organizing maps as a means of finding suitable replacement values.

Robust Estimation

The presence of extreme or atypical values in a sample population can have a dramatic effect on the estimation of the mean and variance, which in turn will affect the estimation of correlation and covariance with other variables. As these measures of association are used by many statistical techniques, it is useful to minimize the influence of atypical observations. Methods of robust estimation are primarily concerned with minimizing the influence of observations that are atypical. There are several methods for determining robust estimates of location (mean/median) and scale (variance). Robust estimation procedures can be applied to both single and multivariate populations. Good reviews on robust statistics can be found in Venables and Ripley (2002, Chapter 5.5) and Daszykowski et al., 2007).

Geochemical distributions are often positively skewed and log-normal in appearance. The skewed nature is commonly attributed to a mixture of different populations and/or the presence of outliers. For such distributions, a robust estimate of the mean will be less than the standard estimate of the mean because the influence of the long tail and outliers is reduced.

Methods for robust estimation of location and scale include Trimmed Means, Adaptive Trimmed Means, Dominant Cluster Mode, L-Estimates, M-Estimates and Huber W-Estimates (see Grunsky 2006).

Transformation of Data

Statistical testing and comparison between groups of data usually requires the estimation of means, variances and covariances. Most statistical procedures assume that the populations being tested are normal in nature. If there are outliers (extreme data values) or a mixture of populations (polymodal or skewed distributions) then the assumption of normality is violated. In right-skewed distributions (the most common effect observed with geochemical data) estimates of the mean exceed the median value. Similarly, the estimation of the variance is inflated for a skewed distribution. The skewed nature of the data can be overcome by applying a suitable transformation that shifts the values of the distribution such that it becomes normally distributed. It has been common in the geological literature to apply logarithmic transformations to data as a way to correct for a positive skew. The application of transformations to data should be carefully applied to avoid masking the presence of multiple populations and outliers (Link and Koch, 1975). If transformations are applied to data to minimize the effect of skewness, then quantile-quantile plots of the transformed data should be examined for changes in slope or breaks in the line, as these features might suggest the presence of two or more populations.

Transformations that can be applied are:

- Linear scaling,

  $y=kx$ or $y=(x_i - \bar{x})/s$

  where s is the standard deviation
- Exponential, $y=e^x$
- Box-Cox Generalized Power Transform,

  $y=(x^\lambda - 1)/\lambda$, $y=\ln(x)$ for $\lambda=0$

The linear scaling transformations do not change the shape of the distribution. However the degree of dispersion (variance) can change. The logarithmic, exponential, and Box-Cox generalized power transforms, or log10 modify both the shape and the dispersion characteristics of the distributions and are the transformations most commonly used. Howarth and Earle (1979) provided a computer program for estimating parameters for the generalized Box-Cox power transform based on the optimization of skew and kurtosis and the optimization of the maximum likelihood criterion of Box and Cox (1964). Lindqvist (1976) published a computer program (SELLO) for transforming skewed distributions based on minimizing skew.

In exploratory data analysis, transformations are useful in assessing whether outliers are the result of a non-normal frequency distribution or are truly atypical values. The distribution should be examined for outliers both before and after a transformation has been applied to the data. Once any outliers are eliminated, the data should be re-examined for outliers as above until all are identified and eliminated. Campbell (1986) prepared computer programs that account for atypical values in the estimation of transformations and robust estimates of means and variances. Stanley (2006) discusses the application of transformations to maximize geochemical contrast and improve data presentation.

Figure 15 shows the effect of applying four different transformations on Ni for lake sediments from the Batchawana area of Ontario. The data are represented on Q-Q plots. Figure 16a shows the untransformed data; Figure 15b shows the log10 transformation of the data; Figure 15c shows a square root transformation and Figure 15d shows a Box-Cox generalized transformation with a value of λ determined after the top 5% of the data were trimmed. The resulting value of λ=0.08 is close enough to zero that there is little difference between the log transform of Figure 15b and 15d.

Discussions on the application of transformations of geochemical data have traditionally been based on raw analytical values and the potential problems associated with closure have not been taken into account. Further research is required in this field.

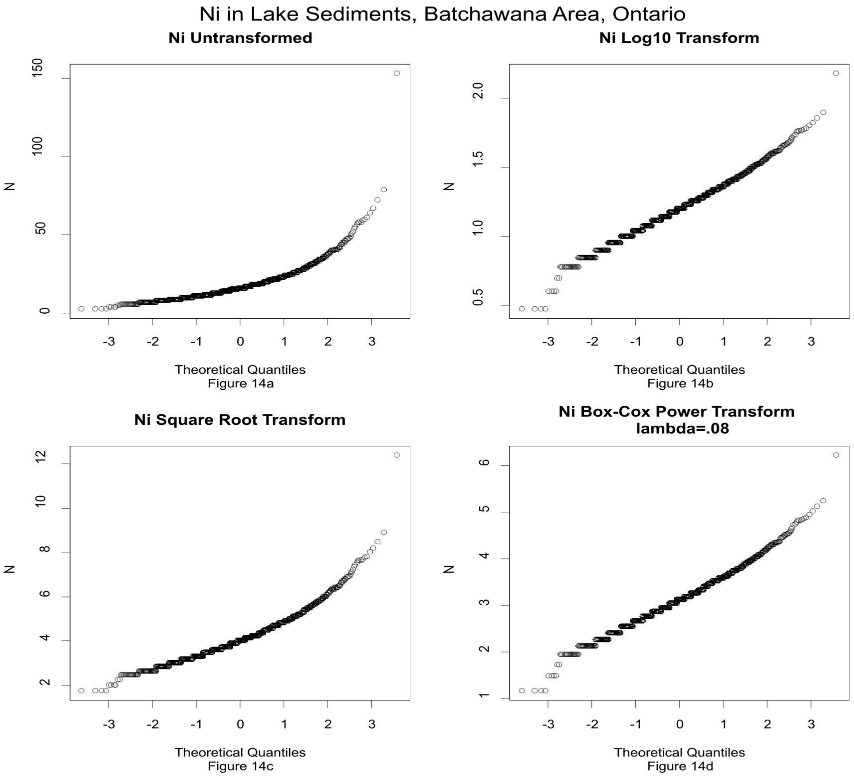


Figure 15: Ni in lake sediments, Batchawana area, Ontario.

LEVELLING GEOCHEMICAL SURVEY DATASETS

Regional exploration programs and integration projects often involve the assembly of diverse sets of data. A common problem associated with the assembly of geochemical survey datasets is known as levelling. Levelling involves the adjustment of values of an element from one survey to be similar to the values of another survey. This “similarity” implies that the means, medians and variations are similar, or in other words, have the same parametric characteristics. Levelling geochemical survey data involves many assumptions and is mitigated by many factors, which are discussed below.

In many geochemical studies, the integration of several sets of data is necessary. Geochemical surveys may have been carried out over an extended period of time during which field sampling methods, sample preparation, methods of digestion and analytical instrumentation may have changed. Thus, there is the potential for a large degree of heterogeneity in the data that is not based on the underlying geology. It is not advisable to level the results of geochemical data derived from different methods of collection (media), preparation (digestion) or analytical methods. The detection limits may be different and there may be systematic shifts between the groups of data. In order to use these data effectively, one or more sets of data must be adjusted. This is known as leveling. One set of data is chosen against which all other sets of data will 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,

$$yi = axi + b$$

xi is the unadjusted variable for observation x,
yi is the adjusted variable for observation x,
a represents the slope of the line in the transformation,
b represents the intercept or additive adjustment.

The adjustment can be determined through regression methods. Non-linear transformations may also be applied if necessary. Figure 16 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 exception of Figure 16e, each scenario shows a possible relationship that will permit leveling. Figure 16e shows a random association between the two variables and in this case leveling is not possible. A detailed example of leveling geochemical data is provided below.

There are several challenges in leveling data, first of which is the choice of data against which to level everything else. Considerable time should be spent on assessing the variability of each element across all of the surveys to be levelled. There may or may not be one set of survey data that can be used as the benchmark dataset, for all elements. Choosing when an element requires leveling must be carried out with caution. Comparing values on maps using bubble plots can be misleading, unless the data are evaluated using the same range and scaling.

Assembling a large number of geochemical surveys and evaluating the need for leveling can be a challenging problem. Trepanier (2006) developed an iterative and adaptive method for leveling a large number of surveys. The method assumes that for each element, one set of survey data represents the standard by which all other surveys will be levelled. All data is stored in a database and an automatic procedure is invoked to search through and adjust the data for each element. The method is computationally intensive and time-consuming.

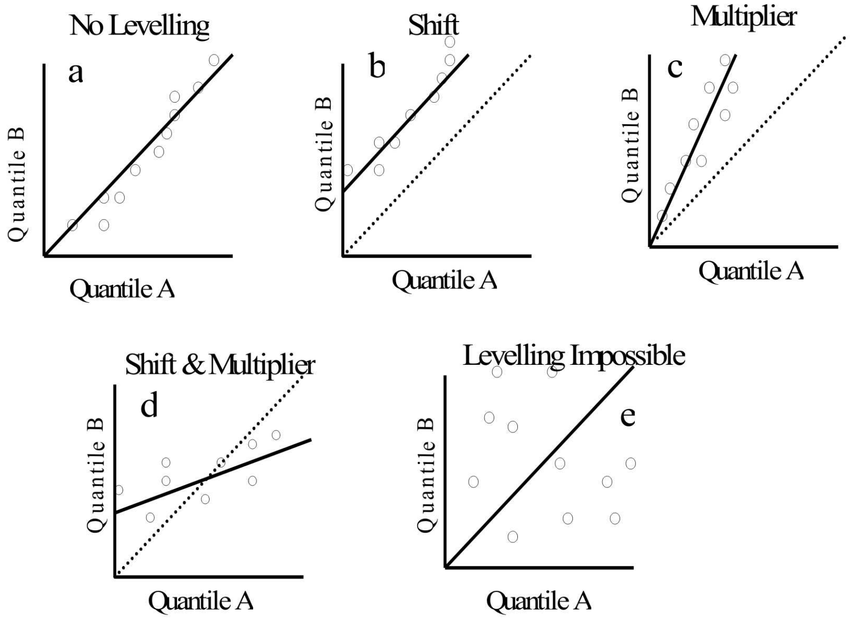


Figure 16: Levelling scenarios for geochemical data.

As shown previously in Figure 4, there are four typical scenarios for levelling between two datasets.

Note that in Figure 4, the values that are plotted are the values at specified quantiles of the data (i.e. 5, 10, 15,….90, 95th percentiles). The worst possible scenario is shown in Figure 4 e where no levelling is possible because no linear relationship exists between the two sets of data. It is also possible that a non-linear shift or multiplier will level two datasets. Graphical inspection of quantile plots between two sets of data should be carried out prior to assessing the type of leveling required.

Daneshfar and Cameron (1998) have demonstrated a method of levelling geochemical data described in Darnley et al. (1995) that accounts for the geology that underlies geochemical data survey sites. The method requires the use of a GIS and a statistical package that computes quantiles and linear regression.

A strategy for levelling several datasets involves the determination of which dataset should be chosen for all of the other databases to be leveled against. The choice of this dataset, the “standard dataset”, will depend on the following factors:

- Spatial proximity of the two datasets,
- Accuracy and precision of the standard dataset,
- The standard dataset contains enough specimens and enough elements so that the other datasets can be leveled to it.

The integration of geochemical survey datasets requires the identification of several key parameters so that the data can be accurately interpreted, i.e.

- Type of media
- Method of preparation
- Method of digestion
- Method of Analysis
- Lower and Upper limits of detection.

If levelling involves geochemical datasets where these characteristics are different then it may be unwise to attempt to level the data. An alternative approach is to map the departure from the median or some other measure that characterizes individual specimens against the distribution for a particular area. The following discussion describes some of the challenges associated with levelling geochemical survey datasets.

Non-spatial leveling is often required (i.e. adjusting location and scale) to remove boundary effects and the comparison of different analytical methods. This is a subject of research that is not well documented (D. Lawie, personal communication, 2007).

The lower and upper limits of detection are commonly different between geochemical survey reports. This is due to the nature of the method of analysis and the developments in the analytical procedures that have taken place over time. As the technology of geochemical analysis improves, the lower limits of detection also decrease. Thus, when merging geochemical survey datasets, the choice of a replacement value for the lower limit of detection (lld) may become an issue. A straight replacement method as indicated in Section 2.5 will not be sufficient because the replacement value is used only to ensure a better estimate of the mean and variance of the data. Varying detection limits within a large dataset assembled from many sources may create significant problems when choosing a replacement value. One approach is to set the lower limit of detection at the weighted median value for the range of lld’s in the dataset. A replacement value can then be determined based on the number of observations and associated lld’s.

Levelling Geochemical Survey Datasets – An Example from Lake Sediments in Northern Ontario

Figure 17 shows sites for 5 different lake sediment surveys in the Batchawana greenstone belt of Northern Ontario. These five surveys were collected during the 1980’s by Fortescue and Vida (1989, 1990, 1991a, 1991b). Hamilton (1995) describes the results of the survey conducted by Fortescue in the Cow River Area. The area is an Archean volcano-sedimentary terrane within the Abitibi-Wawa subprovince of the Superior Province. The geology of the area is described by Grunsky (1991).

Regional lake sediment surveys were carried out in five areas: Pancake Lake, Trout Lake, Hanes Lake, Montreal River and Cow River. The sampling program was carried out over several years and the methods of analysis were similar for all five datasets. However a levelling problem does exist between the survey areas. The greatest difference between geochemical data exists between the Cow River map sheet and the adjacent Montreal River and Hanes Lake survey areas.

Figure 18 shows the range of values for Zn over the five areas in the Batchawana area. The interquartile range, shown in the solid box is significantly higher for the Cow River data than of the other survey areas. However, the Cow River area also contains abundant mafic volcanics rocks of tholeiitic affinity that would naturally tend to have higher Zn values relative to the other survey areas that are composed of a mixture of tholeiitic, calc-alkcalic volcanics, sediments and granitoid rocks. Figure 19 shows a map of Zn values throughout the region. It is clear from the map that levels of Zn in the Cow River area (north east corner) are high relative to the other areas. There are a number of high Zn values within the centre of the volcanic sequence and these could be considered legitimate. However, the Cow River background Zn values appear to be 10 – 20 ppm higher than the background for the adjacent areas.

Using the approach outlined by Daneshfar and Cameron (1998) a quantile regression technique was applied. The procedure involves selecting “bands” of specific distances (5, 10, 15, 20, 25 km, or some suitable scale depending on the nature of the surveys) between adjacent map sheets from which quantile regression is carried out for each of the bands. The reasoning for choosing bands is that an optimum distance, which results in the selection of an optimal number of specimens, will result in a best-fit quantile regression formula for levelling.

Figure 20 shows the selection of bands that were made for levelling the Cow River survey area against the Hanes Lake survey area. Bands were selected at the 5, 10, 15, 20 and 25km ranges in a north-south direction.

For each of these bands, a linear regression was carried out based on the quantiles of the Zn distributions.

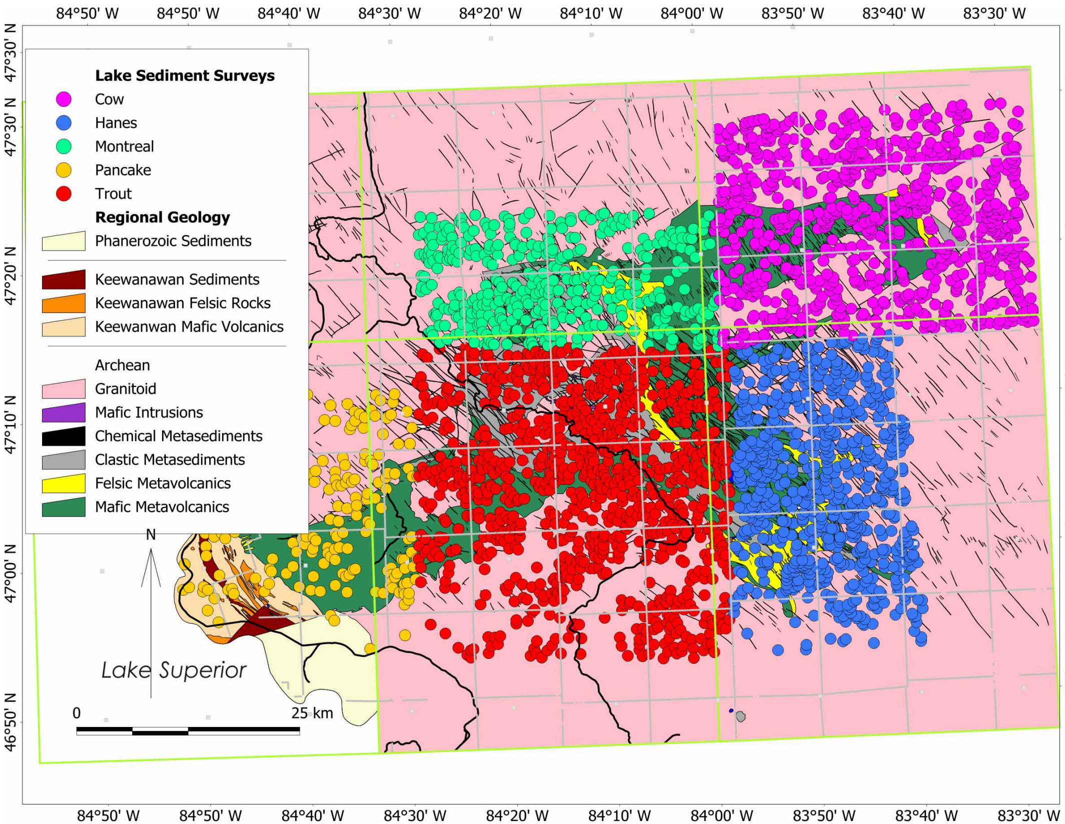


Figure 17: Lake Sediment Survey sites across the Batchawana area, Ontario.

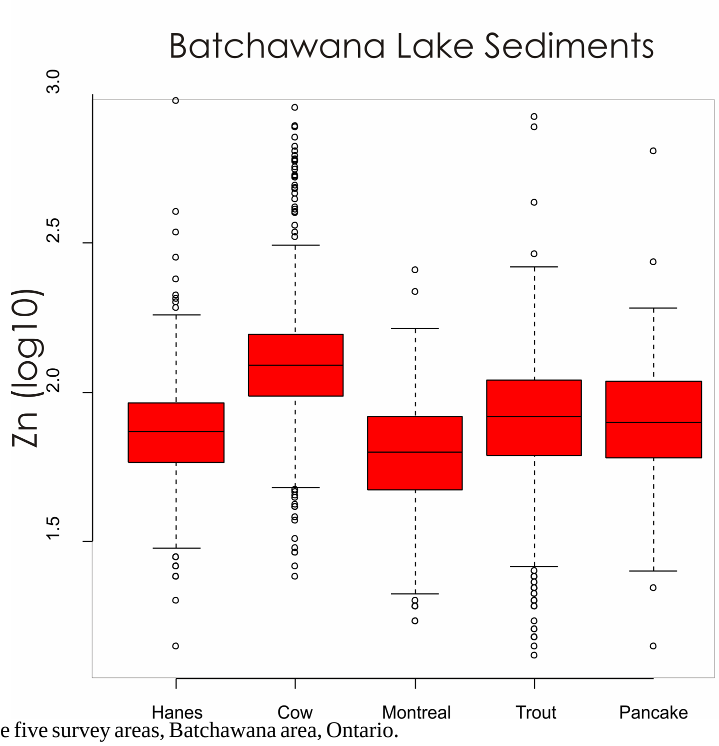


Figure 18: Boxplots of Zn from the five survey areas, Batchawana area, Ontario.

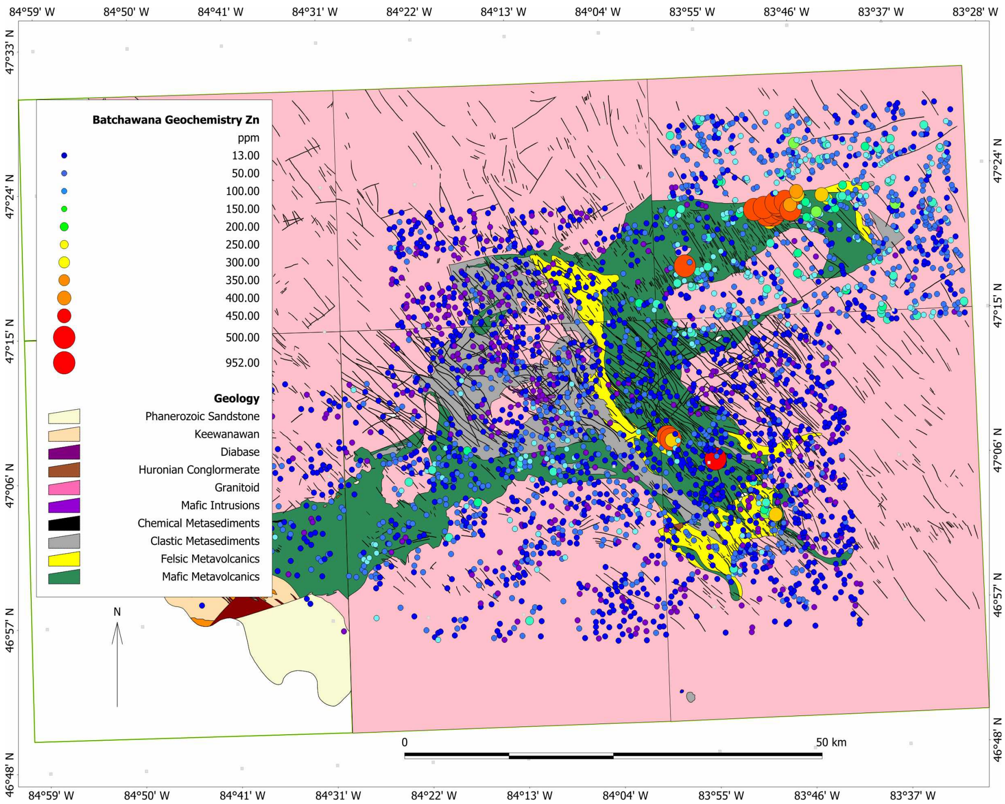


Figure 19: Unlevelled Zn values in lake sediments, Batchawana area, Ontario.

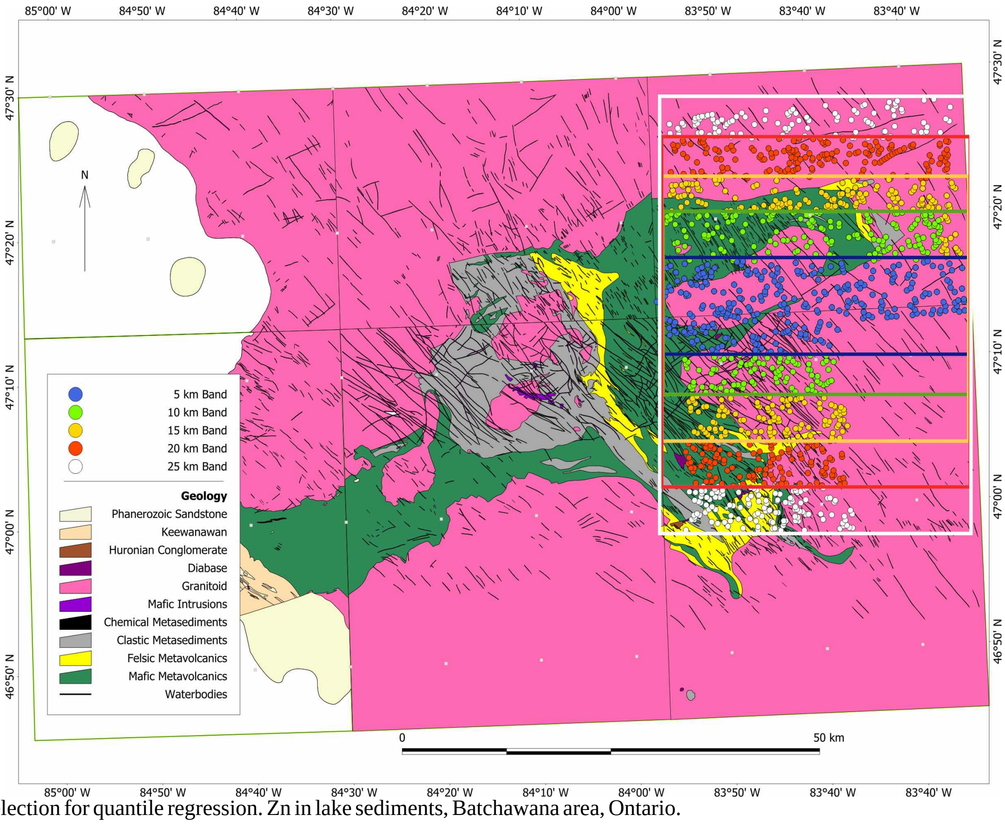


Figure 20: Band selection for quantile regression. Zn in lake sediments, Batchawana area, Ontario.

A measure, D is used to determine which band provides the best quantile regression. D is defined as:

$$D = \Sigma w_i[(q_i)_e - (q_i)_{e'}]^2 \text{ where}$$

w_i is the assigned weight to the ith quantile,
$(q_i)_e$ is the ith quantile in band of width e
$(q_i)_{e'}$ is the ith quantile in band of width e′ in the adjacent map sheet
e is the width of the band expressed as a measure of distance (i.e. meters or kilometers).

The weights favour quantile pairs at or near the median (50th percentile) of the distribution and are based on the ordinates of a normal distribution (weight for the median value = 0.399). These weights are listed in Table 2.

Table 2: Weights used for Quantile Regression in Levelling Geochemical Data

Regression Weights											
Quantile	5	10	20	30	40	50	60	70	80	90	95
Weight	0.1	0.18	0.3	0.35	0.39	0.4	0.39	0.35	0.3	0.18	0

The work by Daneshfar and Cameron (1998) was originally carried out in British Columbia where the adjoining map sheets show broad geological similarity. When the same approach was tried in the Batchawana area the selection of bands of appropriate size became problematic.

Because of the deformed nature of the rocks and the sub-vertical stratigraphy, there is a significant variation in geochemical character over short distances. Figure 21a shows the results of the values of D applied to the 5 band selections and it is clear that the 5km and 25km bands have the lowest D values.

The difference in D values for the different band selections is mostly due to the diversity of lithologies associated with each band. For the 5km band, the lithologies are similar on both sides of the survey boundary: mafic volcanics and granitoid rocks. However for the 10, 15 and 20 km bands, Figure 20 shows that there is a range of lithologies within the bands between the two surveys and the lithologies are most dissimilar for the 15km band. At the 25km band, it is not surprising that the D value is lowest for the similar range of lithologies between the two survey areas and was thus, the best band for the quantile regression methodology.

Quantile regressions were computed for both the 5 and 25km bands using the weights for each quantile, which are shown in Table 2.

In Daneshfar and Cameron (1998) the weight for the 95th percentile was chosen as 0.103. For this application, it was noted that many of the values for the Cow River Zn data were atypical and represented a group of specimens unique to Zn mineralization within the mafic volcanic sequence. There was no equivalent Zn response in the Hanes Lake survey area. Thus, the 95th percentile weight was changed from 0.103 to zero so that the effects of these large Zn values did not bias the leveling of the background.

The values of D, regression coefficients (intercept, slope) and plot of the quantiles for the 5km band selection are shown in Figure 21b and for the 25km band selection in Figure 21c. From the two plots, it can be seen that the 25km band is a better fit and the results from this regression were used to adjust the Zn values in the Cow River survey area. Note that the results of this regression are equivalent to the shift and multiplier effect as shown in Figure 16d.

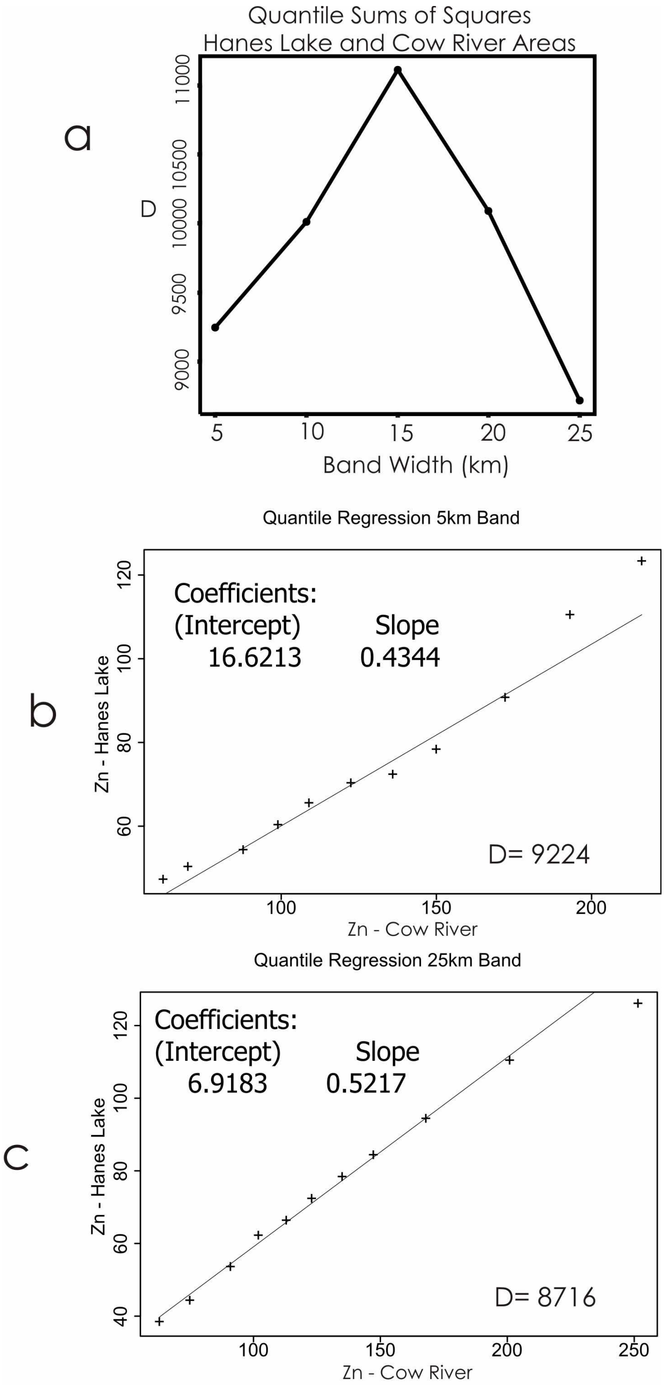


Figure 21: Selection of optimum band width and quantile regression for Zn in lake sediments, Batchwawana area, Ontario.

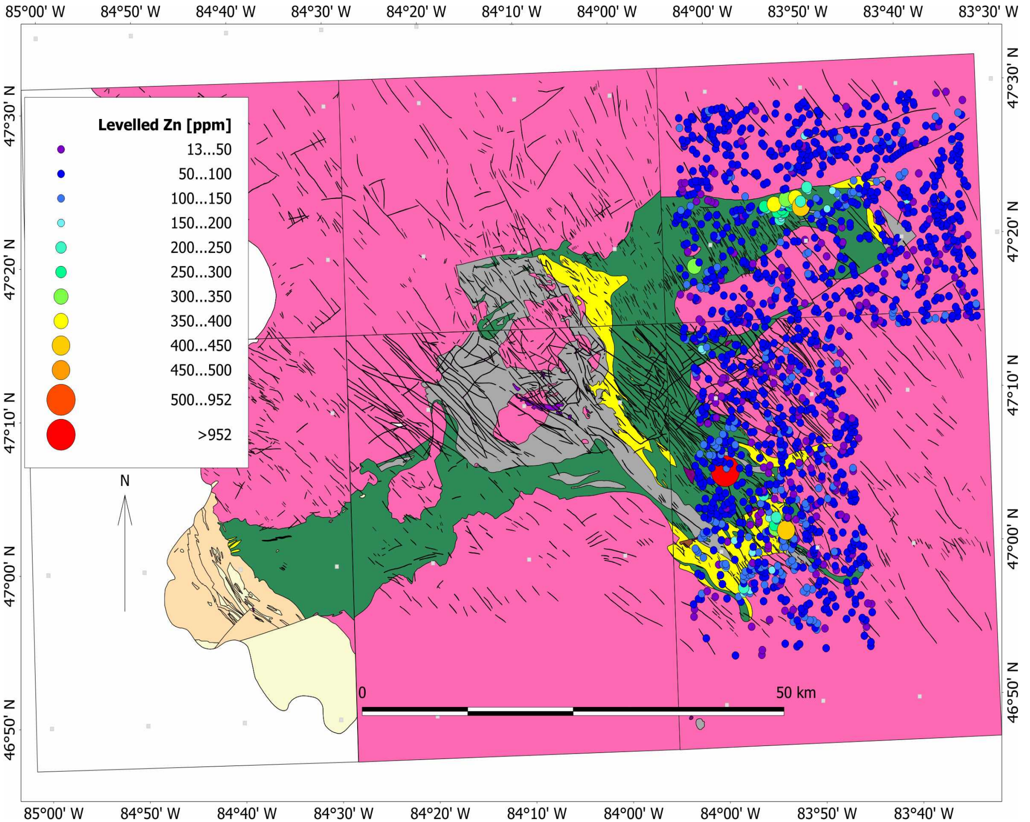


Figure 22: Levelled Zn values after applying quantile regression based on the 25km band selection. See text for a detailed explanation.

The results of applying the regression to the Cow River survey data for Zn are shown in Figure 22. The levelling procedure has had a significant effect on the lower values of Zn in the granitoid terrane but left the upper values, associated with the mafic volcanics and some Zn rich zones within the volcanic sequence relatively unaffected.

Levelling, using a GIS and statistical procedures can produce an optimal result and a combination of these tools is a recommended way to level geochemical survey data.

MULTIVARIATE DATA ANALYSIS TECHNIQUES

Multivariate data analysis techniques such as principal components analysis, cluster analysis, non-linear mapping and projection pursuit regression provide numerical and graphical means by which the relationships of a large number of elements and observations can be studied. These techniques typically simplify the variation and relationships of the data in a reduced number of dimensions, which can often be tied to specific geochemical/geological processes. The basics of multivariate data analysis techniques can be found in Jöreskog et al. (1976); Reyment and Jöreskog (1993); Davis, (2002); Krzanowski (1988) and Howarth and Sinding-Larsen, (1983). Mellinger (1987) provides a systematic approach to the application of multivariate methods in geological studies. Other methods include non-linear mapping (Sammon, 1969), projection pursuit (Friedman, 1987), multidimensional scaling (Kruskal, 1964) and self-organizing maps (Kohonen, 1995). A recent technique, independent components analysis (Comon, 1994), is similar to the method of projection pursuit.

Incorporation of the spatial association with multi-element geochemistry involves the computation of auto- and cross-correlograms or co-variograms. This field of study falls into the realm of geostatistics, which is not covered in this contribution. A number of texts are available that provide details on geostatistics (Isaaks and Srivastava, 1989; Journel and Huijbregts, 1978; David, 1977, 1988).

Grunsky (1986a) employed the use of principal components analysis and clustering methods to evaluate the lithogeochemistry of Archean volcanic terrains from which a number of geological processes were inferred, ranging from primary compositional variation to alteration and associated mineralization. This is discussed in greater detail below.

Multivariate techniques that have been developed specifically for geochemistry include various empirical techniques such as the chalcophile and pegmatophile indices developed by Smith and Perdrix (1983), which were used to outline areas of potential base and precious metal mineralization in the Yilgarn craton of Western Australia.

Robust Estimation of Mean and Covariance Matrices

Many multivariate methods require estimates of correlation or covariance so that interrelationships between the variables can be quantified. Estimates of correlation/covariance are sensitive to the presence of outliers in the data that can bias the results. The influence of outliers can be reduced by applying robust methods to the estimation of the means, correlations and covariances between variables. In multivariate analysis, the distance of an observation to a centroid is estimated by the Mahalanobis distance which depends on an estimate of the multivariate mean and covariance.

The Mahalanobis distance is defined as:

$$D^2 = [x - \bar{x}]' C^{-1} [x - \bar{x}]$$

where

x is a vector of variables for a given observations,
$\bar{x}$ is a vector of the group mean,
C^{-1} is the inverse of the covariance matrix.

There are many techniques for determining robust estimates of mean and variances for individual populations (Rock, 1987, 1988). Robust estimates can be determined for each individual variable or simultaneously for all variables. Multivariate estimates are affected by observations with missing values (no value) in any one of the individual variables. These must be discarded or have some suitable replacement value. Additionally observations that are censored (less than the detection limit) must have a proper replacement value as discussed previously. Campbell (1980) provided some early insight into the application of robust procedures in multivariate analysis. Venables and Ripley (2002, p. 336) provide a good discussion on robust estimation methods.

Two methods can be used to obtain robust multivariate estimates of means and covariance:

Minimum Volume Ellipsoid (MVE) - A multivariate method of determining means and correlations/covariances with minimal effect from outliers based on finding a hyperellipsoid that contains a subset of "good" observations that minimize the volume of the ellipsoid. A geochemical application of this method is given by Chork (1990).

Minimum Covariance Determinant (MCD) Estimation – This method works by minimizing the determinant (a measure of ellipsoid volume) of the covariance matrix based on a symmetric Gaussian hyperellipsoid. The method is faster than the minimum volume ellipsoid but has a lower breakdown point (Rousseeuw and van Driessen, 1999). The determinant is based on a minimum number of "good" observations. As the determinant decreases, the dispersion of the ellipsoid decreases with a corresponding drop in the estimates of central values, resulting in a "robust" estimate.

If there are many observations with values at the same detection limit, a condition of collinearity occurs, which has a direct effect on the covariance matrix. If there are too many identical observations, the method fails. However, by increasing the number of observations, the methods will generate less robust estimates. In the case of non-normal skewed distributions, the means and covariances will be affected. This type of problem is typically encountered when a percentage of the observations have elements with abundances below the detection limit (censored data) and increases the likelihood of collinearity problems.

An example of applying multivariate robust estimates is shown in Table 3 where estimates of the mean for 12 elements are given for 825 lithogeochemical observations from the Ben Nevis Township lithogeochemical data set. In this table, only estimates of the mean are shown. Classical estimates of the mean, based on univariate statistics, multivariate classical estimate, minimum volume ellipsoid and minimum covariance determinant methods are shown. Compared with classical methods of estimation, the robust estimate tends to minimize the effect of those distributions that are skewed.

For the minimum covariance determinant method, two estimates are shown based on two groups of "good" observations. The initial estimate for the MCD used 419 observations based on an initial starting formula of (825 observations + 12 variables + 1)/2. Because of the large number of observations with values at the detection limit, the initial MCD estimate was singular. The MCD was applied using 540 and 800 observations. Table 3 shows that as the number of "good" observations increase, the mean value tends towards the standard estimate where the effect of the long tailed skewed distribution increases the estimate of the mean for several elements.

PRINCIPAL COMPONENTS ANALYSIS

The objective of principal components analysis (PCA) is to reduce the number of variables necessary to describe the observed variation within a dataset. This is achieved by forming linear combinations of the variables (components) that describe the distribution of the data. These linear combinations are derived from some measure of association (i.e. correlation or covariance matrix). Davis (2002, Chapter 6) gives a very readable account on the mathematics of principal components analysis. More complete discussions on the theory and application of principal components analysis can be found in in Jöreskog et al. (1976), Jolliffe (2002) and Jackson (2003). Grunsky (2006, Appendix 1) provides a simple geometric description of principal components analysis.

Table 3: Robust and non-robust estimates of central values. Ben Nevis Township Lithogeochemistry.

Method	Ba	Co	Cr	Cu	Li	Ni	Pb	Zn	Sr	V	Y	Zr
Univariate Mean	208	23	83	56	17	78	17	89	135	132	24	132
Classical Robust Estimate	208	23	83	56	17	78	17	89	135	132	24	132
Univariate Median	170	24	68	42	14	85	5	74	120	150	21	130
Minimum Volume Ellipsoid	194	22	81	38	15	78	7	73	140	139	26	138
Minimum Covariance Determinant 800 observations	207	23	84	47	17	79	10	78	136	133	24	132
Minimum Covariance Determinant 540 observations	198	22	82	39	15	79	6	73	140	139	25	136
								Measures shown in parts per million (ppm)				

A method of principal components analysis known as simultaneous RQ-mode principal components analysis (Zhou et al., 1983) has the advantage of presenting the principal component scores of the observations and the variables (elements) at the same scale, which permits plots of the observations and variables on the same diagram. This method is similar to the biplot method of Gabriel (1971). The interpretation of the results of principal components is usually oriented on placing a geological/geochemical interpretation on the linear combinations of elements (loadings) that comprise the components. This method has been implemented in the S programming language (Grunsky, 2001).

Ideally, each principal component might be interpreted as describing a geological process such as differentiation (partial melting, crystal fractionation, etc.), alteration/mineralization (carbonatization, silicification, alkali depletion, metal associations and enrichments, etc.), and weathering processes (bedrock-saprolite-laterite). In lithogeochemical, weathered profile, lake sediment and stream sediment surveys, the first and second components commonly reveal relationships of observations and variables that reflect underlying lithologic variation. In areas of thick overburden such as glacial till, alluvium or colluvium the linear combinations of variables and the plots of the loadings may not be so easy to interpret as they may reflect a mixture of several surficial processes.

Maps of the principal component scores of the observations can be useful in understanding geochemical processes. If a component expresses underlying lithologies, then a map of that component will clearly outline the major lithological variation of the area Other components that outline other processes such as mineralization or alteration can also be clearly expressed on maps that display the component scores (e.g., Grunsky, 1986a).

The measure of association, or metric, can have a significant effect on the derivation of principal components. Covariance relationships between the elements reflect the magnitude of the elements and thus elements with large values tend to dominate the variance-covariance matrix. This has the effect of increasing the significance of these elements in the results of the principal components analysis. The correlation matrix represents the inter-element correlations, which is actually the standardized equivalent of the variance-covariance matrix. Other metrics of association can be used and this is discussed by Davis (2002) and Jöreskog et al. (1976). If the distributions of the elements are non-normal or there is a presence of outliers the estimates of correlation/covariance may be affected and it may be necessary to apply robust procedures (Zhou, 1985, 1989).

In situations where there are outliers or atypical observations, or where the marginal distributions are not normal a number of choices can be made:

If the marginal distribution is censored, find a suitable replacement value so that the mean and variance is a good estimate of the population mean and variance. This can be done:

- By assigning a replacement value that is around ½ to 1/3 the censored value.
- Use statistical procedures to estimate (impute) a replacement value based on the statistical characteristics of the un-censored portion of the data (i.e. the EM method) discussed previously.

If there are outliers present:

- Remove the outliers from the calculation for means and covariances,
- Apply robust procedures that minimize or eliminate the effect of these values.

Rare events, such as mineral occurrences or deposits, are usually under-represented in regional geochemical survey sampling schemes. A chemical signature that may be diagnostic of a unique geological event may show up as a linear combination of elements with a lesser principal component. Thus, it is important to scan all of the components to check for such features.

The following examples illustrate the use of PCA from the Ben Nevis metavolcanic data (see Figure 1). As it is a “compositional” set of data, it sums to a constant (100%). The data were transformed using the log-centred transformation method described previously. The distributions for these transformed variables are shown in Figure 23.

The results of the principal components analysis are shown in Table 4 where the eigenvalues, R-mode loadings, as well as the relative and actual contributions of the variables are presented. Results are shown for the first 7 components only, which accounts for more than 72% of the variation in the data. The accompanying screeplot displays the successive eigenvalues for all of the components.

The R-mode loadings are the eigenvectors are scaled by multiplying, in order, each of the eigenvectors by the square root of the eigenvalues. The first component accounts for 34% of the overall variation of the data as shown by the eigenvalues. The relative and actual contributions shown in Table 4 provide details on the relative significance of the variables. The relative contribution is the contribution that a variable makes over all of the components. The actual contribution is the contribution that a variable makes within a given component.

Examination of the relative contributions for the first component shows that elements such as Si, Al, Mg, K, Ba, Co, Cr, Ni, V and Zr are accounted for primarily by this component. The actual contribution shows that the variation is spread almost equally between Si, Mg, K, Ba, Co, Cr, Ni, V and Zr within the first component. A map of the first component (Figure 25) describes the compositional variation between the mafic and felsic metavolcanics. The relative contributions of the second component suggests alteration of the volcanic rocks with high loadings for CO_2, S, Li, Sr, Ti, Na, Ca, Fe^{+3} and Al. The relative contributions of the third component suggest alteration associated with more mafic rocks as indicated by Fe^{+2}, Mn, CO_2, S, H_2O^+, Cu and Li.

Biplots of PC1 vs.PC2 and PC1 vs. PC3 are shown in Figures 24 and 25. The scores of the observations are shown as crosses and the scores of the elements are shown as their name. Figure 24 (PC1 vs. PC2) shows that the mafic (Ni, Cr, Co, Mg, Fe) rocks plot on the positive side of PC1. Rocks reflecting felsic metavolcanics (Si, Zr, Ba, K, Y, Al) plot on the negative side of PC1. Observations with relative enrichment in CO_2, S, Li, Pb and Cu, plot along the positive side of the C2 axis. Figure 25 is a biplot of the first and third components where samples with relative enrichment in S and Cu plot along the negative side of the PC3 axis.

Figures 24 and 25 show that there is a distinct break between the rocks of mafic volcanic origin from those felsic volcanic origin. This is reflected by the break in the cloud of points along the C1 axis of both figures.

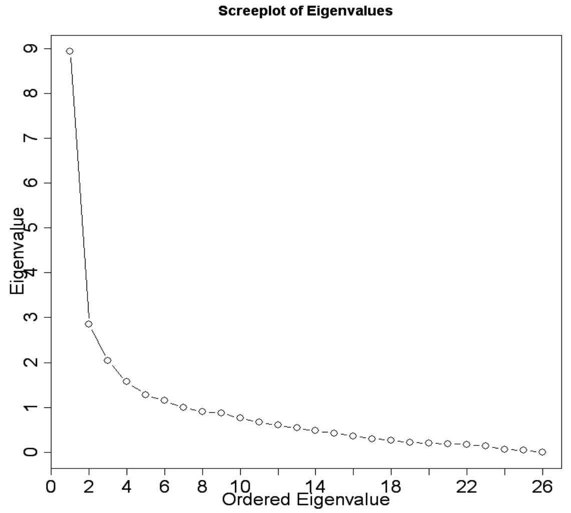
Table 4: Principal Components Analysis of Ben Nevis Lithogeochemistry. Analysis carried out on Log-centered.

Eigenvalue

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
λ	8.93	2.86	2.03	1.56	1.28	1.15	0.99
%	34.38	11.00	7.83	6.03	4.94	4.41	3.80
Σ%	34.38	45.38	53.22	59.24	64.18	68.59	72.39

R-Loadings Red values >0 Blue values <0

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SiO2	-0.87	-0.26	0.03	-0.06	0.04	-0.06	0.11
Al2O3	-0.72	-0.48	-0.01	-0.07	0.18	-0.15	0.18
Fe2O3	0.17	-0.48	-0.16	-0.55	-0.01	-0.06	0.18
FeO	0.63	-0.15	0.46	-0.25	-0.03	0.03	0.11
MgO	0.86	-0.03	0.16	-0.09	0.19	0.14	0.02
CaO	0.40	-0.47	0.01	0.40	-0.25	-0.28	0.10
Na2O	-0.36	-0.44	-0.06	0.40	0.15	0.15	0.04
K2O	-0.69	0.19	-0.08	-0.03	0.34	-0.16	-0.27
TiO2	0.43	-0.60	0.02	-0.12	0.02	-0.14	-0.08
P2O5	-0.12	-0.29	0.10	-0.01	-0.14	0.79	-0.24
MnO	0.20	-0.25	0.57	-0.01	-0.47	-0.31	-0.07
CO2	-0.35	0.42	0.37	0.51	-0.24	-0.16	0.02
S	-0.30	0.49	-0.41	-0.28	-0.37	0.07	0.07
H2Op	0.47	0.07	0.43	-0.38	0.27	-0.03	0.30
Ba	-0.76	0.00	-0.06	-0.03	0.39	-0.16	-0.20
Co	0.88	-0.15	-0.11	0.03	0.05	-0.04	-0.05
Cr	0.86	0.03	-0.03	0.12	-0.02	0.20	-0.11
Cu	0.31	0.29	-0.55	-0.24	-0.18	-0.15	0.04
Li	0.06	0.49	0.56	0.10	0.39	0.09	0.20
Ni	0.92	0.04	-0.08	0.10	0.07	0.08	-0.05
Pb	-0.47	0.33	0.04	-0.17	-0.14	0.12	0.44
Zn	-0.16	0.01	0.31	-0.40	0.02	-0.16	-0.55
Sr	-0.11	-0.53	-0.28	0.21	0.21	0.05	0.24
V	0.80	-0.07	-0.22	0.04	0.13	-0.13	-0.07
Y	-0.67	-0.37	0.21	-0.13	-0.27	0.13	-0.02
Zr	-0.80	-0.22	0.16	-0.13	-0.09	0.13	0.00



Relative Contributions Red Values>10 Blue values <10

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SiO2	76.63	6.93	0.11	0.42	0.16	0.38	1.13
Al2O3	51.37	23.50	0.01	0.46	3.28	2.32	3.21
Fe2O3	2.82	22.97	2.57	30.83	0.02	0.41	3.21
FeO	40.09	2.39	20.99	6.04	0.10	0.08	1.23
MgO	74.13	0.08	2.57	0.85	3.52	1.92	0.05
CaO	15.68	21.67	0.01	16.36	6.45	8.04	0.94
Na2O	13.08	18.99	0.42	16.27	2.17	2.15	0.19
K2O	48.18	3.78	0.66	0.09	11.60	2.44	7.38
TiO2	18.84	35.64	0.03	1.51	0.05	1.87	0.60
P2O5	1.53	8.51	1.03	0.01	1.89	62.58	5.68
MnO	4.19	6.41	32.38	0.01	22.15	9.47	0.43
CO2	12.34	17.52	13.93	25.77	5.67	2.47	0.05
S	8.95	24.51	16.44	7.59	13.50	0.43	0.54
H2Op	21.91	0.53	18.59	14.41	7.24	0.07	8.99
Ba	57.25	0.00	0.37	0.09	15.04	2.67	3.82
Co	78.41	2.27	1.10	0.06	0.24	0.14	0.24
Cr	74.12	0.08	0.12	1.35	0.04	4.06	1.18
Cu	9.44	8.52	30.45	6.00	3.17	2.19	0.15
Li	0.38	23.88	31.60	0.97	15.43	0.79	4.04
Ni	84.74	0.16	0.59	1.07	0.55	0.66	0.25
Pb	22.52	10.69	0.16	2.75	1.99	1.50	18.97
Zn	2.69	0.00	9.60	15.86	0.05	2.66	30.15
Sr	1.24	27.99	8.04	4.41	4.47	0.27	5.80
V	64.40	0.46	4.92	0.18	1.60	1.61	0.54
Y	45.31	13.64	4.50	1.62	7.16	1.74	0.04
Zr	63.61	4.98	2.45	1.69	0.85	1.80	0.00

Actual Contributions Red Values>10 Blue values <10

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SiO2	8.57	2.42	0.05	0.27	0.12	0.33	1.15
Al2O3	5.75	8.21	0.00	0.29	2.56	2.02	3.25
Fe2O3	0.32	8.03	1.26	19.68	0.02	0.36	3.24
FeO	4.48	0.84	10.31	3.86	0.08	0.07	1.25
MgO	8.29	0.03	1.26	0.54	2.74	1.67	0.05
CaO	1.75	7.57	0.01	10.44	5.02	7.01	0.95
Na2O	1.46	6.64	0.21	10.39	1.69	1.87	0.19
K2O	5.39	1.32	0.32	0.05	9.04	2.13	7.47
TiO2	2.11	12.46	0.01	0.96	0.04	1.63	0.61
P2O5	0.17	2.97	0.51	0.01	1.47	54.55	5.75
MnO	0.47	2.24	15.90	0.01	17.25	8.25	0.44
CO2	1.38	6.12	6.84	16.45	4.42	2.15	0.05
S	1.00	8.57	8.07	4.84	10.52	0.38	0.55
H2Op	2.45	0.19	9.13	9.20	5.64	0.06	9.10
Ba	6.41	0.00	0.18	0.06	11.71	2.33	3.86
Co	8.77	0.79	0.54	0.04	0.19	0.12	0.24
Cr	8.29	0.03	0.06	0.86	0.03	3.54	1.20
Cu	1.06	2.98	14.95	3.83	2.47	1.91	0.15
Li	0.04	8.35	15.52	0.62	12.02	0.69	4.08
Ni	9.48	0.05	0.29	0.68	0.43	0.58	0.26
Pb	2.52	3.74	0.08	1.75	1.55	1.31	19.20
Zn	0.30	0.00	4.72	10.13	0.04	2.32	30.52
Sr	0.14	9.78	3.95	2.81	3.48	0.24	5.87
V	7.20	0.16	2.42	0.12	1.24	1.40	0.55
Y	5.07	4.77	2.21	1.03	5.58	1.52	0.04
Zr	7.12	1.74	1.20	1.08	0.66	1.57	0.00

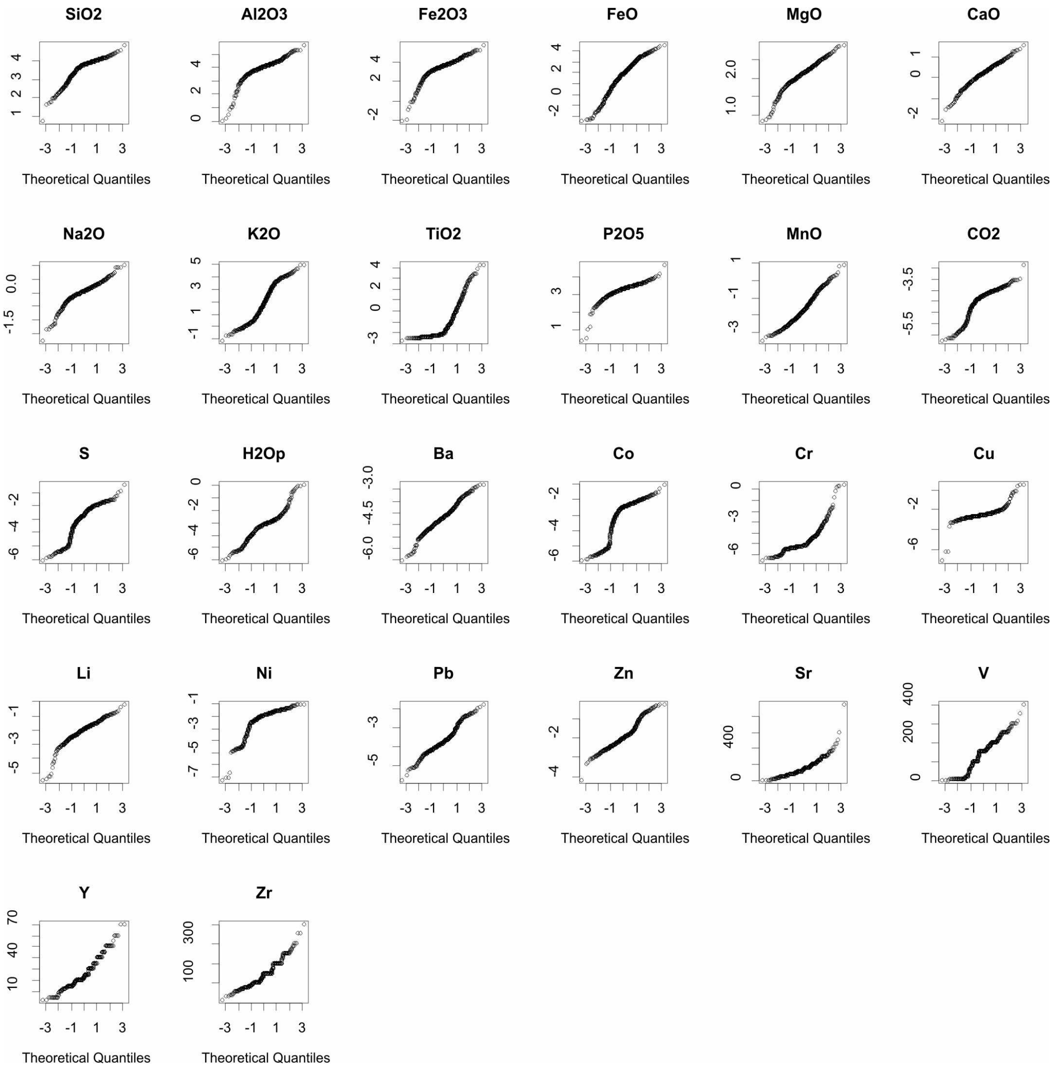


Figure 23: Quantile-Quantile plots of log-centred major and trace elements for the Ben Nevis lithogeochemical data.

The Q-mode scores were interpolated to a 100 meter resolution grid by kriging.. Figure 26 shows an interpolated image of the first principal component. The distinction between the mafic and felsic volcanic rocks is evident by the colour map of the image. Green and blue areas are associated with felsic rocks and red to yellow areas are associated with mafic rocks as shown in the relationships of the observations and elements in Figure 24.

Figure 27 shows an image of the second principal component, which accounts for 11% of the variation in the data. The plot of PC1 vs. PC2 in Figure 24 shows that the second component has Cu, Li, S, Pb and CO_2 are associated with positive values of PC2. The image of Figure 27 shows that areas in red-yellow correspond to the zones of carbonate alteration and mineralization that are present around the Canagau Mines deposit and the Croxall property.

Figure 28 is an image of the third principal component (7.8% of the variation in the data). Areas associated with sulfur and Cu enrichment are evident, most notably around the Canagau Mines Cu-Au deposit in the eastern part of the image. These areas are also adjacent to areas of CO_2, Li, and Zn enrichment, which represent altered and mineralized country rocks that surround the S-Cu zones of relative enrichment.

Much more information can be obtained by examining all of the principal components. Other components exhibit zoning of Ca around the main zone of carbonate alteration and K has an association with S at the mineral occurrences. The fourth component highlights the relationship between Zn and S at both the Canagau and Croxall properties. However, the illustration of the first three components shows that PCA is an effective method for exploring the structure of the geochemical data and assisting in deriving models of geochemical processes by the use of graphics and geographic representation.

Principal components analysis has many different uses in evaluating geochemical data, including the development of empirical indices for specific element targeting (see sections on Empirical Indices and Weighted Sums).

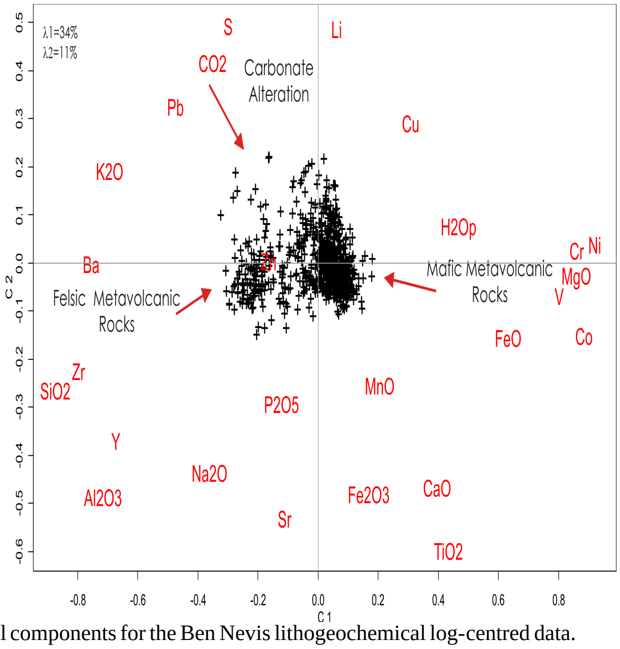


Figure 24: Biplot of the first two principal components for the Ben Nevis lithogeochemical log-centred data.

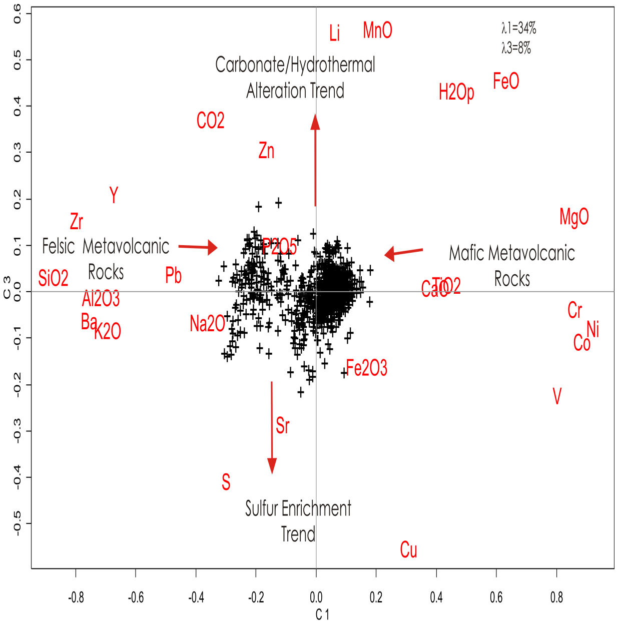


Figure 25: Biplot of the first and third principal components for the Ben Nevis lithogeochemical log-centred data.

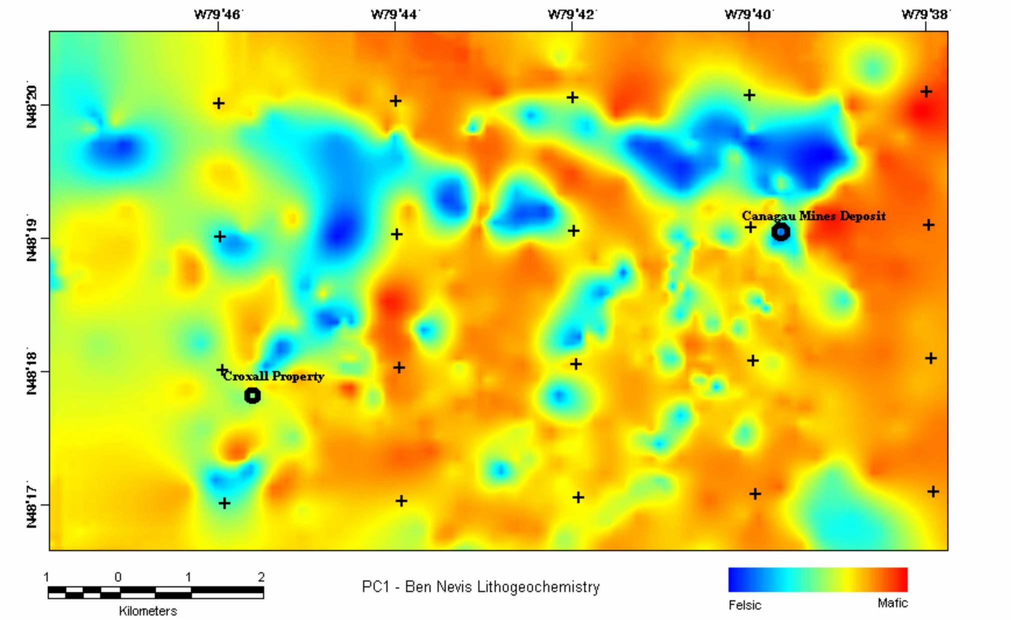


Figure 26: Image of the first principal component derived from the log-centred lithogeochemical data, Ben Nevis Township, Ontario. This image outlines the lithological variation.

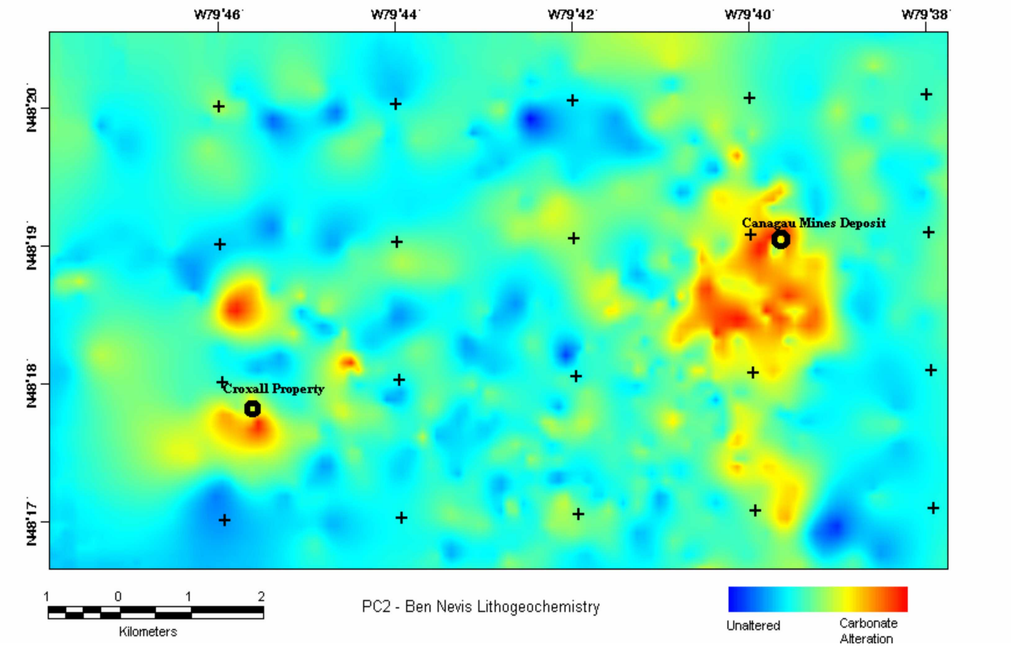


Figure 27: Image of the second principal component derived from the log-centred lithogeochemical data, Ben Nevis Township, Ontario. This image outlines the zones of carbonatization.

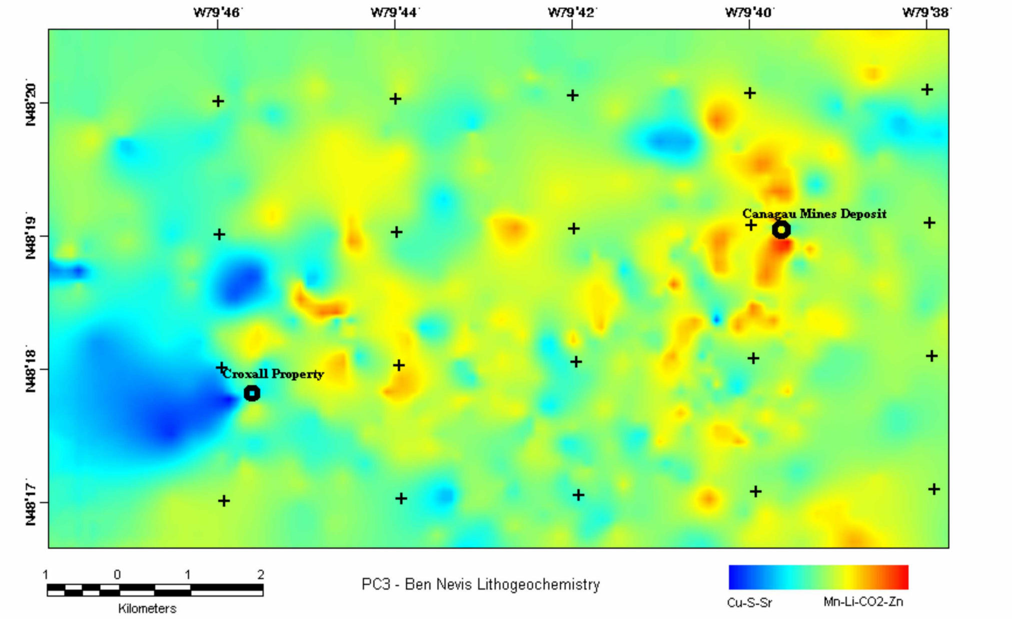


Figure 28: Image of the third principal component derived from the log-centred lithogeochemical data, Ben Nevis Township, Ontario. This image outlines the sulphide and mineralized occurrences.

CLUSTER ANALYSIS METHODS

Cluster analysis methods are useful as an exploratory tool for detecting groups of multi-element data that may not be readily observable in simple scatter plots or through the use of methods such as principal components analysis. The main objective of clustering algorithms is to identify distinct natural groupings within multidimensional data. Clustering methods can be broadly divided into hierarchical and non-hierarchical methods. The following example shows the use of k-means clustering as a method for partitioning multivariate geochemical data. Davis (2002) provides a good introductory review of clustering methods. Sinding-Larsen (1975) used clustering methods for the initial subdivision of a heterogeneous geochemical area. Jaquet et al. (1975) provides a detailed analysis of lake sediment geochemistry using clustering procedures. Howarth and Sinding -Larsen (1983) provide a general discussion of clustering methods applied to geochemical exploration. Grunsky (1986a) has shown how dynamic cluster analysis was used to detect different types of mineralization based on distinct geochemical differences between the mineral occurrences. The use of fuzzy clustering methods in geochemistry has been introduced (Bochang and Xuejing, 1985).

Hierarchical clustering is based on the linking of variables (R-mode) or observations (Q-mode) through measures of similarity. The relationships between the variables or observations can be graphically expressed using a dendrogram. Individual clusters can be discriminated by choosing an appropriate value of linkage, which separates internally similar groups of objects into dissimilar groups. Hierarchical clustering assumes that all variables are linked at some level, which may not be a reasonable assumption in many instances.

The correlation coefficient (R-mode) is the most common measure of similarity for clustering. For Q-mode analysis (similarities between the observations), the Euclidean distance can be used as a measure of proximity by which observations can be clustered. In the case of Q-mode analysis, the size of the similarity matrix that contains the measure of the distance metric between points can become so large that computation becomes intractable.

Arbitrary Origin Methods are non-hierarchical and may offer some advantage over hierarchical methods since the clusters are formed based on multivariate similarities (proximities) rather than individual correlation coefficients. These methods start with an initial number of cluster centres that can be specified or randomly chosen. Each observation is allocated to one of the groups based on proximity to the group centres. The process is iterative and group centres change until a stable solution results. Methods such as K-means (McQueen, 1967; Everitt, 1974, Hartigan, 1975) or dynamic cluster analysis (Diday, 1973) are examples of these techniques. Kaufman and Rousseeuw (1990) also describe a number of clustering methods.

K-Means Clustering

K-means cluster analysis is a method that starts with an initial “guess” of the cluster centers. The distance of each observation from each cluster center is measured and then provisionally assigned to the closest cluster center. A new cluster center is calculated based on the designated observations for each previous center. The process is iterative until it converges on stable centers. The method requires an initial choice of the number of cluster centers. If the number is too great, there will be small clusters that have few points. If the number of centers is too few, then the structure of the data may not be realized. A disadvantage of the procedure is that a less than optimal clustering may result if the initial cluster centres do not fall in distinct clusters (Davis, 2002, p 500). Venables and Ripley (2002) provide a method by which a suitable number of starting clusters may be determined by using a combination of hierarchical clustering and principal components analysis.

It is common to apply non-hierarchical clustering methods to principal component scores. If one or more principal components can be inferred to represent specific geological/geochemical processes, then the application of cluster analysis can provide further insight in how those processes may be related. Additionally, the component plots provide a reduced set of dimensions for viewing the multi-element associations of the data and thus provide additional visual assistance in examining grouped associations.

K-means clustering was applied to the log-centred transformed Ben Nevis township metavolcanic data. The number of clusters was set at 10, based on the perceived variation in the rock types (felsic metavolcanics, mafic volcanics, mafic intrusions, granite) as well as the two known mineralization zones that have surrounding alteration. The results of the clustering are shown in Figure 29. Each observation is labeled with the group number to which it was assigned. Several clusters (Groups 1, 2, 5, 6, 8 and 10) are associated with the distinctions between mafic and felsic metavolcanics. Groups 3 and 9 are directly associated with mineralization. Observations that belong to these groups occur where there is known mineralization. There are also two clusters associated with carbonate alteration (Groups 4 and 7), which occur in the eastern part of the map-area. It is apparent that the observations assigned to each group not only share similar geochemical characteristics but also have close spatial associations, which is clearly shown in Figure 29.

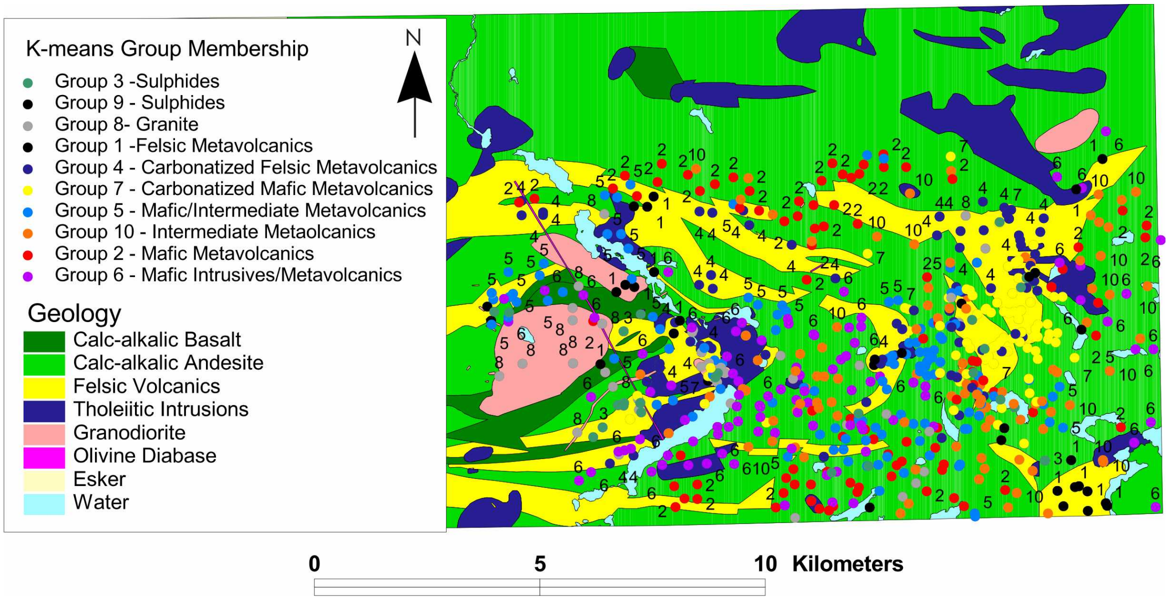

K-means Group Membership
N
● Group 3 -Sulphides
● Group 9 - Sulphides
● Group 8- Granite
● Group 1 -Felsic Metavolcanics
● Group 4 - Carbonatized Felsic Metavolcanics
● Group 7 - Carbonatized Mafic Metavolcanics
● Group 5 - Mafic/Intermediate Metavolcanics
● Group 10 - Intermediate Metaolcanics
● Group 2 - Mafic Metavolcanics
● Group 6 - Mafic Intrusives/Metavolcanics
Geology
Calc-alkalic Basalt
Calc-alkalic Andesite
Felsic Volcanics
Tholeiitic Intrusions
Granodiorite
Olivine Diabase
Esker
Water
0
5
10 Kilometers

Figure 29: K-mean clustering of the log-centred lithogeochemical data, Ben Nevis Township, Ontario. Specific groups are associated with distinctive lithologies and zones of alteration and mineralization.

Multivariate Ranking using the Mahalanobis Distance: A multivariate extension of Q-Q Plots

The use of the covariance matrix as a tool for distinguishing background from anomalous populations is well established in geochemical research (Garrett, 1989b, 1990; Chork, 1990). Filzmoser et al. (2005) have written a library of routines (mvoutlier) that is available as part of the R environment (www.r-project.org/cran/). The covariance matrix contains information on the variability of the elements as well as their inter-relationships. The multi-element data constitute a hyper-ellipsoid in multidimensional space. The mean value of each element defines the centroid of this hyper-ellipsoid and the distance from each observation point to the centroid is the Mahalanobis distance. In a multivariate normal population, most observations lie within an expected radius of the centroid and is, by definition, the background group of observations. However, if outliers are included in the data, the shape of the hyper-ellipsoid will change. This resulting distortion affects the location of the centroid and thus affects the Mahalanobis distance for all of the observations. In such cases, the application of robust procedures is recommended.

Outliers can be distinguished from the main background population by determining the Mahalanobis distance of each observation from the group centroid. The distances can be compared to the "expected" distances of a multivariate normal population (cumulative probability with the number of degrees of freedom defined as the number of variables) by the use of x^2 values as defined by Garrett (1989b). If the population is multivariate normal, then the plotted pairs form a straight line. If the population contains outliers, then the observed Mahalanobis distances are greater than the expected x2 quantiles and the plot becomes non-linear. However, the x2 distribution is long tailed near the extreme ends of the distribution and this property may mask outliers with large Mahalanobis distances. An alternative to the use of the x2 values is the cubed root of a normal distribution, which does not have the long tail property of the x2 distribution and thus less likely to mask outliers.

The lake sediment survey data from the Batchawana area of Ontario was evaluated for the potential to host copper, zinc and precious metal deposits. A suite of elements (Cu, Zn, As, Sb and W) was chosen to test the possibility that these elements could identify potential mineral deposits. For these data, censored values were replaced with estimates from the EM method for determining replacement values for censored distributions. Because these data are compositional, they were normalized to a constant sum and then transformed using log-ratios.

Figure 30 shows a series of ranked Mahalanobis distance plots versus the cubed root of a normal distribution for different degrees of trimming. The first figure shows a plot of all of the observations. The plot displays a curved line with several outliers at the positive end of the curve, suggesting that there are observations which are not part of a multivariate normal population. Each successive plot is the data with the outliers from the previous plot removed. For each plot, a new centroid and corresponding Mahalanobis distances were re-computed. Trimming of the data in the 7% to 10% range yields a reasonably straight curve which suggests that the trimmed observations could be considered atypical and warrant further investigation.

The 10% of data that were trimmed data were then re-inserted into the data matrix from which the D^2 values were computed based on the covariance from the other 905 of the data The ranked multivariate distance values are plotted on the map and graph in Figure 31.. Observations with high D^2 values are locales of interest and warrant further investigation. Note that observations, which are atypical, are not necessarily geochemically "anomalous". No multivariate equivalent of a threshold was established, although the 10% trim could be used as an initial starting point in establishing the threshold.

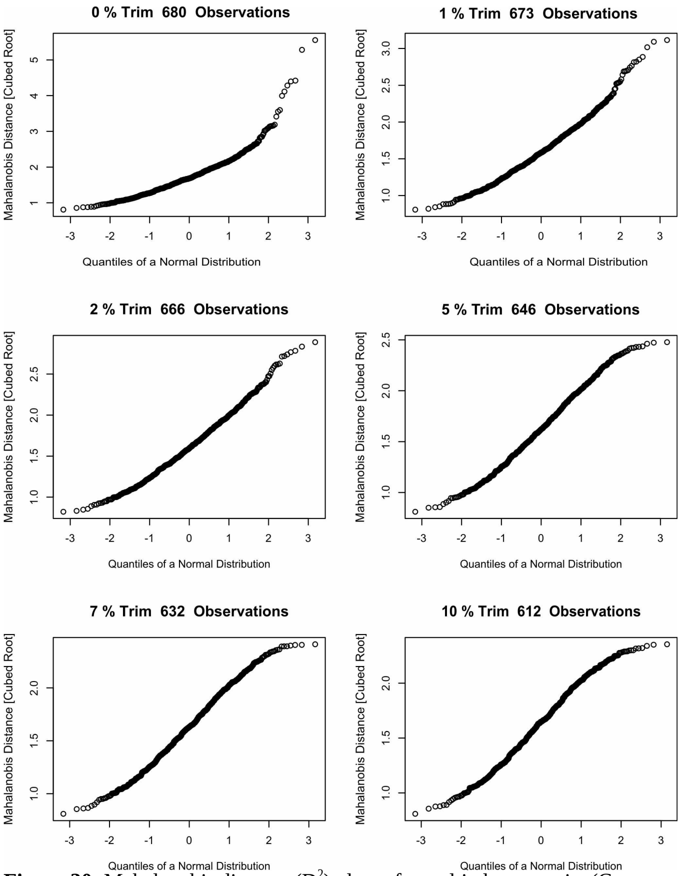


Figure 30: Mahalanobis distance (D^2) plots of a multi-element suite (Cu, Zn, As, Sb, W) of lake sediment data. Successive trimming of the outliers defines a homogeneous background population. The deleted outliers are then follow-up for their potential as sites of mineralization.

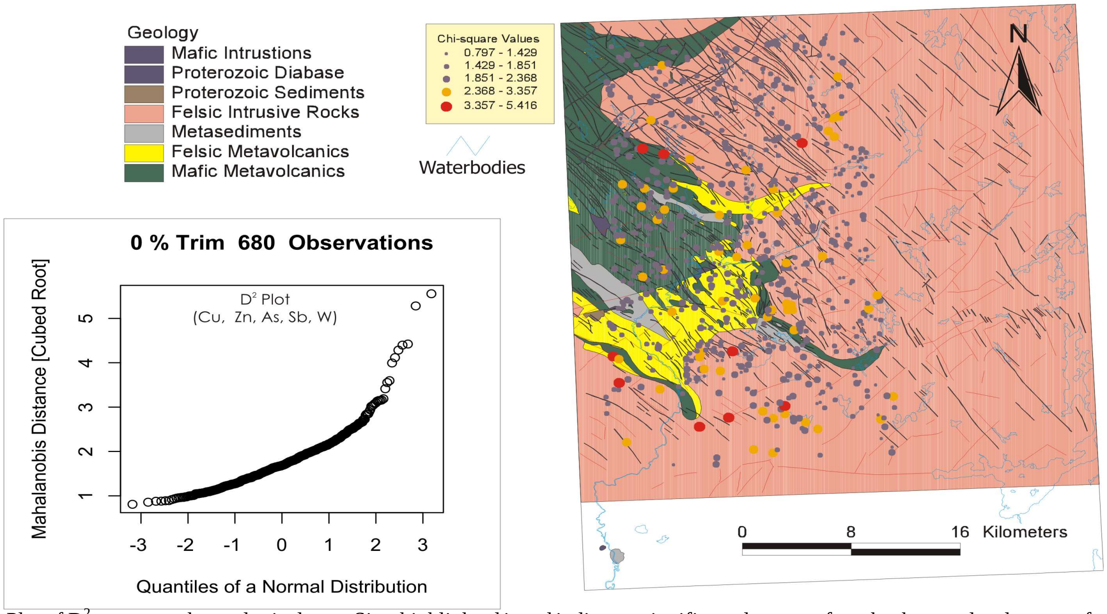


Figure 31: Plot of D^2 scores on the geological map. Sites highlighted in red indicate a significant departure from background and warrant further evaluation.

The Use of Empirical Indices

The existence of pathfinder elements has prompted the use of several numerical procedures through which selected elements can be used in an exploration program by creating mineralization potential indices based on the weighted sum scores of the pathfinder elements. Empirical indices can be determined from selected elements that are associated with specified geochemical processes. The techniques used in this approach are described by Garrett(1991), Garrett et al. (1980), Smith and Perdrix (1983), Smith et al. (1987), and Chaffee (1983). Garrett and Grunsky (2001) have reviewed objective comparisons of various weighting schemes used to highlight observations defined by pathfinder elements.

In many geochemical studies, several pathfinder elements may be identified for defining target areas (mineralization, anthropogenic sources). These pathfinder elements may be chosen based on geological/geochemical knowledge of the processes of interest. Combining these pathfinder elements together through a multivariate ranking scheme is a potentially useful tool for defining multi-element anomalies. Defining the pathfinder elements can be based on geological knowledge or through the use of data analysis/discovery procedures discussed previously, such as principal components and cluster analysis. These methods can reveal relationships in the data that may be directly related to underlying lithologies or processes of interest (mineralization, anthropogenic effects) from which pathfinder elements can be determined. Methods such as principal components analysis can help determine which elements are positively /negatively associated with an element of interest and can be a starting point for developing an empirical index.

Chaffee (1983) developed a method of scoring observations for anomaly potential. Each element is evaluated such that the range of values are subdivided into 4 groups, by thresholds, with corresponding scores that represent background (0), weakly anomalous (1), moderately anomalous (2), and strongly anomalous (3). These ranges are derived from orientation studies over areas where the range of values and underlying geochemical distributions are reasonably well understood. Each is then assessed with respect to each element. Observations with the highest scores are considered anomalous and are targeted for further follow-up.

Smith et al. (1987), Smith and Perdrix (1983) and Smith et al. (1989) made use of three indices derived from geochemical trends that were noted in the laterite geochemistry of the Yilgarn Block of Western Australia.. A group of pathfinder elements, As, Sb, Bi, Mo, Ag, Sn, and W form the basis of these empirical indices known as, CHI-6*X, NUMCHI, and PEG-4. These indices show elevated values of these pathfinder elements in lateritic materials associated with greenstone belts, shear zones, base metal and precious metal deposits (CHI-6*X and PEG-4). These indices are based on simple equations as follows:

The coefficients provide weighting to the elements such that observations with elevated chalcophile values have high CHI-6*X or PEG-4 indices. These coefficients were derived for lateritic materials only. The coefficients must be altered for other materials. The CHI-6*X index is suited more to isolating observations with elements associated with precious metal deposits, while the PEG-4 index is suited for isolating observations with element associated with pegmatophile environments, such as Sn deposits within granitoid terrains.

The NUMCHI index is a score of the number of elements that exceed the threshold for each element. Thus for a given specimen, if nine elements exceed their respective thresholds, then the NUMCHI index will have a value of 9. As discussed previously, threshold values are chosen from visual inspection of summary tables, order statistics, Q-Q plots etc.

Weighted Sum Index

Garrett et al. (1980, p.144) have suggested the use of a linear combination of a group of indicator elements that give a weighted sum. In a multi-element survey, those elements, which are considered pathfinders are given more weight than elements which may be more diagnostic of background. The choice of weights may be based on the knowledge of the investigator. Alternatively, principal component loadings may be used as a starting point. Examples of the use of this index are given by Garrett et al. (1980), and Garrett and Grunsky (2001).

INTEGRATION OF MULTI-ELEMENT GEOCHEMISTRY AND DIGITAL TOPOGRAPHY

Modern methods of data management including the use of desktop database management systems (DBMS) combined with Geographical Information Systems (GIS) that can produce images of multiple datasets simultaneously provide significant assistance in the management and presentation of geochemical data. In many areas of the world digital base maps can be acquired from local governments that typically include lakes, rivers, streams, road networks and other topographic information that is useful in the orientation and interpretation of geochemical data. In addition, digital topography is often available that provides a topographic relief backdrop for the interpretation of geochemical data. Digital geological maps are now routinely provided by many Geological Surveys, together with mineral occurrence inventory databases that have been accumulated from Geological Survey and private company data.

Digital topography offers a unique view of data in that it provides a "real world view" of the data over the terrain. When digital air photos or satellite imagery are integrated with digital topography and viewed using image processing systems with three dimensional rendering ability, the viewer gets a sense of looking at the terrain from an aircraft. Interpolated geochemical images can often be interpreted more effectively when merged with digital topography and viewed in a similar manner. Grunsky and Smee (1999) demonstrated the usefulness of integrating digital elevation data with multi-element geochemistry. Geochemical patterns, that were otherwise obscure, became meaningful when multi-element patterns were draped over the digital elevation model for the study area.

As an example, the Campo Morado mining camp in the Guerrero state of Mexico hosts seven precious metal bearing volcanogenic massive sulphide deposits in the complexly folded and faulted Guerrero terrain (Oliver, et al., 1996, Rebagliati, 1999). Approximately 29, 221 samples were collected over a soil grid comprised of 25 meter sample interval along lines and spaced 100m apart. The field samples were analyzed for: Al, Fe, Ca, K, Mg, Na, Ti, Au, Ag, As, Ba, Cd, Co, Cr, Cu, Hg, Mn, Mo, Ni, P, Pb, Sc, Sr, V, W and Zn using aqua regia digestion and ICP-ES finish. A DEM was created at 25 metre resolution. Principal components analysis was carried out on the data and revealed several significant patterns related to lithological variation and mineralization. Because of the high topographic relief in the area, the problem of transported material from weathering has the potential to result in "false anomalies" that are often due to hydromorphic dispersion and down-slope creep. When the results of the principal components analysis are draped over the topography, there is an increased ability to distinguish anomalies associated with hydromorphic dispersion from those associated with a bedrock source.

Figure 32 shows a planimetric image of the second principal component over a shaded relief image of the DEM. Felsic volcanic rocks (red and yellow) are distinguished from mafic volcanic rocks (blue). Felsic rocks show relative enrichment in K, and Na while the mafic rocks show relative enrichment in Fe, Co, Ti, Mg, Cr, Al, Sc, and V. The areas highlighted in green represent lithologies of intermediate compositions and are mostly mudstones, argillites and sandstones. These are the host rocks for several of the mineral deposits in the Campo Morado area. The first principal component highlights areas of relative enrichment of Ag, Zn, Au, As, Pb, Hg, Sb and Cu. These areas, shown in red and yellow, are potential sites of mineralization (Figure 33). This image is a three dimensional rendering over the DEM. Examination of these areas in conjunction with the DEM assists in setting priorities for follow-up. Anomalies that lie along riverbeds or show significant dispersion must be treated with caution due to the effects of hydromorphic and downslope creep dispersion effects.

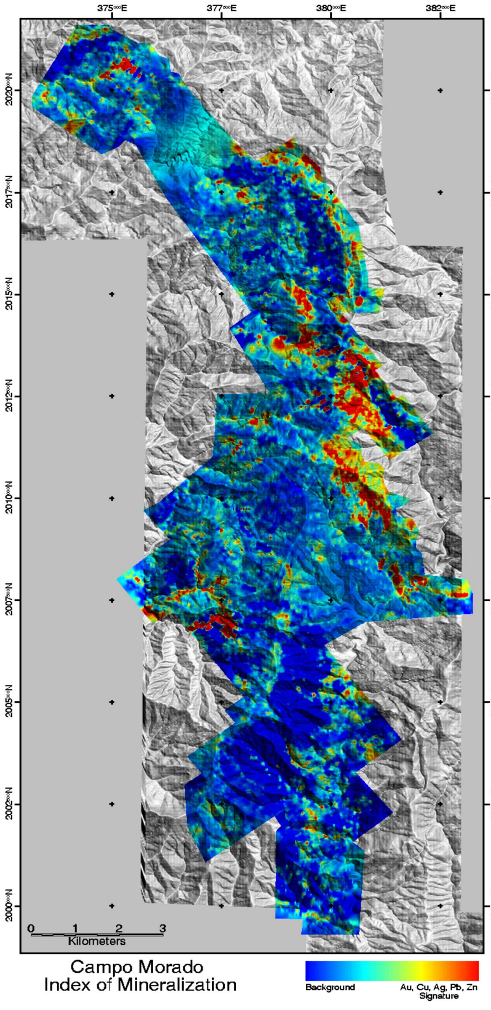


Figure 32: Plot of the interpolated PC1 scores over the digital terrain model in the Campo Morado area, Mexico. Areas highlighted in red are elevated in Au,Cu, Ag, Pb and Zn values. The image is termed as an "index of mineralization".

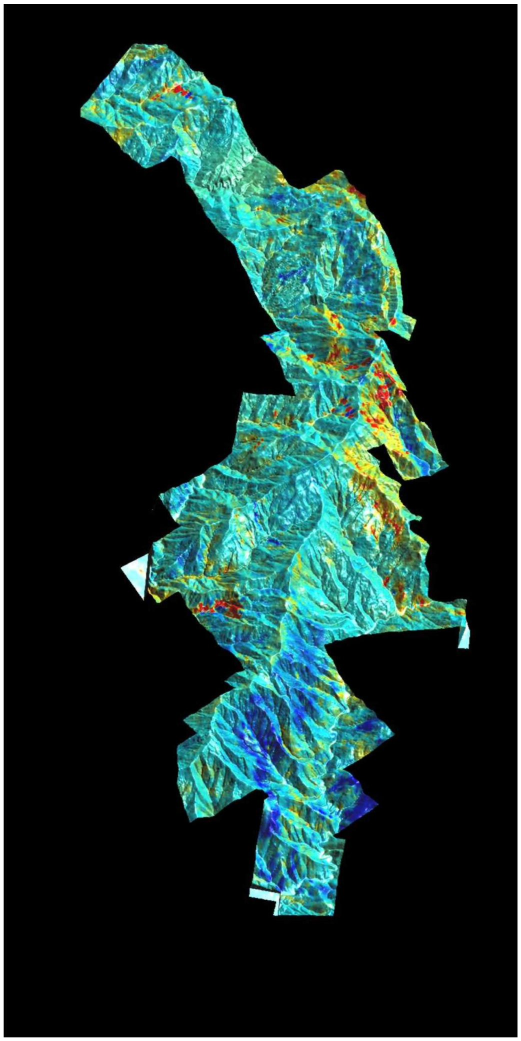


Figure 33: The same image as Figure 32, however the index of mineralization is draped over the digital terrain model and rendered in 2.5D. This enhances the interpretation of mineralization with respect to the terrain variation.

SUGGESTED SEQUENCE OF DATA ANALYSIS

The following list of suggested ways to evaluate data should be considered in any investigation. Of course, not all steps are necessary or appropriate, but should serve as a guideline for a thorough investigation of geochemical data.

Preliminary Data Analysis

- Know your data! There is no substitute for spending time by evaluating the data using a wide variety of procedures so that associations and structures in the data can be identified.
- Examine each element with histograms, boxplots, Q-Q plots, scatter plot matrix and summary tables.
- Use bubble or symbol maps to show the range and spatial variability of the elements of interest. Interpolated images can be used where appropriate.
- Trim the distribution of each element of gross outliers.
- Investigate outliers for each element; analytical error, or atypical value?
- Adjust data for censored values if required.
- Consider the application of log-ratio transformations (logcentred, isometric logratio) so that compositional data can be evaluated without the effect of “closure”. This is necessary if measures of association are required (correlation, covariance).
- Apply measures of association using standard measures as well as robust procedures. Examine the differences and scrutinize the outliers.
- Test the data to see if the identification of patterns and outliers are improved by the use of transformations. Apply Box-Cox power transformations using observations below the 95th-98th percentile to determine the optimal transformation. The choice of transform parameters can be chosen visually (q-q plots, boxplots, histograms) or by semi-automatic means.
- Examine scatter plots, and quantile-quantile plots for the presence of multiple populations.
- If assembling datasets, examine the requirement for leveling.

Exploratory Multivariate Data Analysis

A summary of exploratory multivariate techniques follows:

- Create a scatter plot matrix of the raw data and transformed (logcentred ratios, isometric logratios) data. Look for trends/associations.
- Use robust estimates to compute means and covariances to enhance the detection of outliers.
- Apply dimension reducing techniques such as principal components analysis to identify patterns and trends in the data. Other methods such as non-linear mapping, multi-dimensional scaling and self-organizing maps may help discover structure in the data.
- Use geographic maps of the component scores to assist in identifying spatially based geochemical processes.
- Apply methods such as cluster analysis to isolate groups of observations with similar characteristics and atypical observations. Specific groups of interest can often be isolated using these methods. Maps of the locations of the groups can help to examine the spatial continuity of the groups.
- Use robust Mahalanobis distance plots (D^2) applied to transformed data to assist in isolating outliers based on a selected number of elements of interest. Maps of large distances (>95th percentile) can assist in identifying observations or groups of observations of interest.
- Calculate specifically tailored empirical indices in areas where multi-element associations are well understood. The indices are based on a linear combination of pathfinder elements with coefficients that are selected for each area and commodity being sought. Observations with high indices can be investigated for mineralization potential.

CONCLUDING COMMENTS AND FUTURE DIRECTIONS

Garrett (1989c) stated that the power of computers and capability of software would continue to grow along with a corresponding decrease in price. Almost 20 years later, that prediction still holds. Computers are not only more powerful, but they are more portable, which permits the most sophisticated processing even in the most remote parts of the planet. Developments in software, in terms of the amount of data capacity, developments in visualization and statistical methods have made enormous contributions to the way that exploration geochemists can evaluate and integrate all types of geoscience data. The rapid expansion of the internet has allowed new statistical communities to grow, such as the R project (www.r-project.org) in which thousands of statisticians and users throughout the world develop and contribute to an open source statistical software environment. Recent developments in freely available software (Grunsky, 2002b) will make it easier to integrate geochemical data with geospatial data In the R community, new statistical developments can be available to users within weeks and to anyone who has internet access. There is no doubt that this type of cooperative approach to the sharing of knowledge will increase the ability of geoscientists to extract as much information from their data as possible.

Another factor that has contributed to very significant advancements in evaluating regional geochemical data is the ubiquitous development of internet resources for geochemical data availability. In addition, internet resources have contributed significantly to information on how to evaluate geochemical data. The internet itself is one of the first places one starts to “mine” for data.

Discussions on the application of transformations of geochemical data have traditionally been based on raw analytical values and the potential problems associated with closure have not been taken into account. Further research is required in this field. There is ongoing research at the University of Girona, Spain, where the issues of evaluating compositional data is being addressed. Emphasis is being placed on research and on the development of tools for the user.

Surprisingly, the scientific literature on leveling geochemical data is sparse. Levelling is routinely carried out in geophysical and geochemical programs, however a formal review of procedures has not yet been published. A full review of leveling methods applied to geochemical survey data is due.

Integrating spatially referenced data together with multivariate observations is an area that is undergoing many interesting developments. The use of fractals has been shown to highlight different spatial patterns that are attached to multivariate patterns and trends (e.g. Cheng and Agterberg, 1994). Similarly the integration of multivariate statistics with geostatistical analysis is developing and will lead to new methods for extracting spatially-dependent multivariate patterns and trends.

Current implementations of statistics with geographical information systems are not fully integrated and spatial statistics that are employed by geographic information systems or image analysis systems offer limited analytical and developmental capability. Increased integration of multivariate methods together with spatial analysis will provide a comprehensive approach to assessing all spatially reference multivariate data. Multivariate geostatistics, which incorporates both the spatial and inter-element relationships, has been studied by only a few. Grunsky and Agterberg (1988, 1992), Grunsky (1990) and Wackernagel and Butenuth (1989) discuss two approaches to multivariate geostatistics. Bailey and Krzanowski (2000), Christensen and Amemiya, (2003) and Krzanowski and Bailey (2007) discuss approaches to “spatial factor” methods. Spatial factor methods will permit the simultaneous evaluation of geochemical processes within the geochemical and geospatial domain. The long term benefit of this will be to identify geochemical processes as a function of spatial scale (sampling density) and will permit further discrimination between geochemical background and mineralization.

ACKNOWLEDGMENTS

The author wishes to acknowledge helpful discussions with colleagues at CSIRO, Australia and the Geological Survey of Canada, most notably, Frits Agterberg, Graeme Bonham-Carter, Bob Garrett, Norm Campbell, Harri Kiiveri, Jeremy Wallace and Ray Smith. This manuscript has also benefited from reviews by Robert Jackson, David Lawie and Graham Closs.

The author wishes to acknowledge thanks to the following for permission to use their data:

- Ontario Geological Survey and the Ontario Ministry of Natural Resources for the provision of the digital elevation data for the Ben Nevis area of Ontario.
- Mark Rebagliati of Hunter Dickinson Inc., Vancouver, for permission to present the results of the Campo Morado geochemical study.

REFERENCES

Aitchison, J., 1986, The Statistical Analysis of Compositional Data, Methuen Inc.

Aitchison, J., 1990, Relative Variation Diagrams for Describing Patterns of Compositional Variability, Mathematical Geology, 22, 487-511.

Aitchison J., 1997, The one-hour course in compositional data analysis or compositional data analysis is simple. V. Pawlowsky-Glahn, ed., in: Proceedings of IAMG '97, the Third annual conference of the International Association for Mathematical Geology, 3-35.

Aucott, J.W., 1987, Workshop 5. Geochemical Anomaly Recognition, Journal of Geochemical Exploration, 29, 375-376.

Bailey, T.C. and Krzanowski, W.J., 2000, Extensions to Spatial Factor Methods with an Illustration in Geochemistry, Mathematical Geology, 32, 657-682.

Barcelo C, Pawlowsky V., Grunsky E., 1995, Classification problems of samples of finite mixtures of compositions. Mathematical Geology, 27, 129-148.

Barcelo C., Pawlowsky V., Grunsky E., 1996, Some aspects of transformations of compositional data and the identification of outliers. in R. A. Olea, ed., Geostatistics. Mathematical Geology. 28, 501-518.

Barcelo-Vidal C, Pawlowsky-Glahn V., Grunsky E.C., 1997, A critical approach to the Jensen diagram for the classification of a volcanic sequence. In, V. Pawlowsky-Glahn, ed., Proceedings of IAMG '97, the Third annual conference of the International Association for Mathematical Geology, 117-122.

Bloom, L., 1997, The Critical Importance of Monitoring Chemical Analyses in Frontier Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 295-300.

Bochang, Y. and Xuejing, X., 1985, Fuzzy cluster analysis in geochemical exploration, Journal of Geochemical Exploration, 23, 281-292.

Bølviken, B. and Gleeson, C.F., 1979, Focus on the Use of Soils for Geochemical Exploration in Glaciated Terrane, in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, .295-326.

Bonham-Carter, G.F., 1989a, Integrating Global Databases with a Raster-Based Geographic Information System, in J.N. Van Driel and J.C. Davis, eds., Digital Geologic and Geographic Information Systems, American Geophysical Union Short Course in Geology, 10, 1-13.

Bonham-Carter, G.F., 1989b, Comparison of Image Analysis and Geographic Information Systems for Integrating Geoscientific Maps, G.F. Bonham-Carter and F.P. Agterberg, eds., in: Statistical Applications in the Earth Sciences, Geological Survey of Canada Paper 89-9, 141-155.

Bonham-Carter, G.F., 1994, Geographic Information Systems for Geoscientists, Modelling with GIS, Volume 13, Computer Methods in the Geosciences, Pergammon Press.

Bonham-Carter, G. F., 1997, GIS Methods for Integrating Exploration Data Sets, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 59-64.

Boyle, R.W., 1979, Geochemistry Overview in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 25-31.

Box, G.E.P., and Cox, D.R., 1964, An Analysis of Transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252.

Bradshaw, P.M.D. and Thomson, I., 1979, The Application of Soil Sampling to Geochemical Exploration in Nonglaciated Regions of the World, in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 327-338.

Bridges, N.J., and McCammon, R.B., 1980, Discrim. A computer program using an interactive approach to dissect a mixture of normal or lognormal distributions, Computers & Geosciences, 6, 361-396.

Brooks, R.R., 1979, Advances in Botanical Methods of Prospecting for Minerals Part1 – Advances in Biogeochemical Methods of Prospecting in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 397-410.

Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (eds), 2006, Compositional Data Analysis in the Geosciences: From Theory to Practice, Geological Society, London, Special Publications, 264, 212p.

Butt, C.R.M., 1989, Geomorphology and Climatic History – Keys to Understanding Geochemical Dispersion in Deeply Weathered Terrains, Exemplified by Gold, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 323-334.

Campbell, A.N., 1989, Putting Expert System Technology to Work, p. 825, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 825.

Campbell, N.A., 1980, Robust procedures in multivariate analysis. I Robust covariance estimation. Applied Statistics, 29, 231-237.

Campbell, N.A., 1986, A General Introduction to a Suite of Multivariate Programs, CSIRO Division of Mathematics and Statistics, unpaginated unpublished report.

Cannon, H., 1979, Advances in Botanical Methods of Prospecting for Minerals Part1 – Advances in Geobotanical Methods in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 385-396.

Carr, J.R. 1994, Numerical Analysis for the Geological Sciences, Prentice Hall.

Chaffee, M. A., 1983, Scoresum- A Technique for Displaying and Evaluating Multi-Element Geochemical Information, With Examples of its use in Regional Mineral Assessment Programs, Journal of Geochemical Exploration, 19, 361-381.

Cheng, Q., 2006. GIS-based multifractal anomaly analysis for prediction of mineralization and mineral deposits, in J. Harris, ed., GIS for the Earth Sciences, Geological Association of Canada Special Publication 44, 285-297.

Cheng, Q., and Agterberg, F.P., 1994, The separation of geochemical anomalies from background by fractal methods, Journal of Geochemical Exploration, 51, 109-130.

Cheng, Q., Xu, Y. and Grunsky, E.C., 2000, Integrated Spatial and Spectrum Analysis for Geochemical Anomaly Separation, Natural Resources Research, 9, 43-51.

Chork, C.Y., 1990, Unmasking multivariate anomalous observations in exploration geochemical data from sheeted-vein tin mineralization near Emmaville, N.S.W., Journal of Geochemical Exploration, 37, 205-223.

Christensen, W.F., Amemiya, Y., 2003, Modeling and prediction for multivariate spatial factor analysis, Journal of Statistical Planning and Inference, 115, 543-564.

Chung, C.F., 1985, Statistical treatment of geochemical data with observations below the detection limit; in Current Research, Part B, Geological Survey of Canada, Paper 85-1B, 141-150.

Chung, C.F., 1988, Statistical analysis of truncated data in geosciences, in Sciences. de la Terre, Series. Inf., Nancy, 27, 157-180.

Chung, C.F., 1989, FORTRAN 77 program for constructing and plotting confidence bands for the distribution and quantile functions for truncated data, Computers & Geosciences, 15, 625-643.

Cleveland, W.S., 1993, Visualizing Data, Hobart Press.

Coker, W.B. and DiLabio, R.N.W., 1989, Geochemical Exploration in Glaciated Terrain: Geochemical Responses, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 336-383.

Coker, W.B., Hornbrook, E.H.W. and Cameron, E.H., 1979, Lake Sediment Geochemistry in Geophysics and Geochemistry, in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 385-396.

Coope, J.A. and Davidson, M.J., 1979, Same Aspects of Integrated Exploration, in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 575-592.

Comon, P., 1994. Independent component analysis. A new concept?, Signal Processing, 36, 287-314.

Cox, S., 1997, Delivering Exploration Information On-line Using the WWW: Challenges, and an Australian Experience, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 135-143.

Closs, L.G., 1997, Exploration Geochemistry: Expanding contributions to Mineral Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 3–8.

CRAN, 1999, The Comprehensive R Network, http://cran.r–project.org.

Daneshfar, B. and Cameron, E., 1998, Levelling Geochemical Data Between Map Sheets, Journal of Geochemical Exploration, 63, 189-201.

Darnley, A.G., Bjorklund, A., Bolviken, B., Gustavsson, N., Koval, P.V., Plant, J.A., Steenfelt, A., Tauchid, M. and Xie Xuejing, 1995, A Global Geochemical Database for Envrionmental and Resource Management, Recommendations for International

Geochemical Mapping, Final Report of IGCP 259, with contributions by R.G. Garrett and G.E.M. Hall, Earth Sciences Report 19, UNESCO Publishing.

Daszykowski,M., Kaczmarek, K., Vander Heyden, Y., and Walczak, B., 2007, Robust statistics in data analysis - A review: Basic concepts, Chemometrics and Intelligent Laboratory Systems, 85, 203-219.

Davenport, P.H., Kilfoil, G.J., Colman-Sadd, S.P. and Nolan, L.W., 1997, Towards Comprehensive Digital Geoscience Data Coverages for Newfoundland and Labrador, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 161-164.

David, M. 1977, Geostatistical Ore Reserve Estimation, Elsevier Scientific Publishing Company.

David, M. 1988, Handbook of Applied Advanced Geostatistical Ore Reserve Estimation, Elsevier.

Davis, J.C., 2002, Statistics and Data Analysis in Geology, John Wiley & Sons Inc., third edition.

de Kemp, E.A. and Desnoyers, D.W., 1997, 3-D Visualization of Structural Field Data and Regional Sub-Surface Modelling for Mineral Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 157-160.

Dempster, A.P., Laird, N.M., and Rubin, D.B., 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B, 39, 1-38.

Deutsch, C.V. and Journel, A.G., 1997, GSLIB: Geostatistical Software Library and Users Guide, Oxford University Press, second edition.

Diday, E., 1973, The dynamic clusters method in non-hierarchical clustering, International Journal of Computer Informatics, 2, 61-88.

Dickson, B.L. and Giblin, A.M., 2007, An evaluation of methods for imputation of missing trace element data in groundwaters, Geochemistry: Exploration, Environment, Analysis, 7, .173-178.

Dunn, C.E., 1989, Developments in Biogeochmical Exploration, in G.D. Garland, ed., Proceedings of Exploration ’87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 417-438.

Davenport, P.H., Friske, P.W.B., and Beaumier, M., 1997, The Application of Lake Sediment Geochemistry to Mineral Exploration: Recent Advances and Examples From Canada, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 261-270.

Everitt, B., 1974, Cluster Analysis, Heinemann, London, 122, 2nd Edition, 1980.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras G and Barceló-Vidal, C., 2003: Isometric logratio transformations for compositional data analysis. Mathematical Geology 35, 279-300.

Filzmoser, P., Garrett, R.G., Reimann, C., 2505, Multivariate outlier detection in exploration geochemistry, Computers & Geosciences, 31, 579-587.

Fletcher, W.K., 1997, Stream Sediment Geochemistry in Today’s Exploration World, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 249-260.

Franklin, J.M., 1997, Lithogeochemical and Mineralogical Methods for Base Metal and Gold Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 191-208.

Friske, P.W.B., 1997, Putting It All Together— Surficial Geochemistry Maps for Large Areas of Canada, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 363.

Fortescue, J.A.C. and Vida, E.A., 1989, Geochemical Survey of the Trout Lake Area; Ontario Geological Survey, Map 80803.

Fortescue, J.A.C. and Vida, E.A., 1990, Geochemical Survey, Hanes Lake Area; Ontario Geological Survey, Map 80806.

Fortescue, J.A.C. and Vida, E.A., 1991a, Geochemical Survey, Montreal River Area; Ontario Geological Survey, Map 80808.

Fortescue, J.A.C. and Vida, E.A., 1991b, Geochemical Survey, Pancake Lake Area; Ontario Geological Survey, Map 80807.

Fortescue, J.A.C., 1992, Landscape geochemistry: retrospect and prospect - 1990. Applied Geochemistry, 7, 1-53.

Friedman, J.H., 1987, Exploratory Projection Pursuit, Journal of the American Statistical Association, 82, 249-266.

Gaál, G. (Editor), 1988, Exploration target selection by integration of geodata using statistical and image processing techniques: an example from Central Finland. Geological Survey of Finland, Report of Investigation 80, Part 1, Text, 156 pages, 109 figures, 18 tables.

Gabriel, K.R., 1971, The biplot graphical display of matrics with application to principal component analysis, Biometrika 58, 453-467.

Garrett R.G., 1983, Sampling Methodology, Chapter 4, Statistics and Data Analysis in Geochemical Prospecting, edited by R.J. Howarth, 2, in Handbook of Exploration Geochemistry, edited by G.J.S. Govett.

Garrett, R.G., 1984, Workshop 5. Thresholds and Anomaly Interpretation, Journal of Geochemical Exploration, 21, 137-142.

Garrett, R.G., 1988, IDEAS: an interactive computer graphics tool to assist the exploration geochemist, in Current Research, Part F, Geological Survey of Canada, Paper 88-1F, 1-13.

Garrett R.G., 1989a, A Cry from the Heart, Explore, Newsletter of the Association of Exploration Geochemists, 66, 18-20.

Garrett, R.G., 1989b, The chi-square plot. a tool for multivariate outlier detection, Journal of Geochemical Exploration, 32, 319-41.

Garertt, R.G., 1989c. The Role of Computers in Exploration Geochemistry, in G.D. Garland, ed., Proceedings of Exploration ’87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 586-608.

Garrett, R.G., 1990, A Robust Multivariate Procedure with Applications to Geochemical Data. in F.P. Agterberg and G.F. Bonham-Carter, eds., Statistical Applications in the Earth Sciences, Geological Survey of Canada Paper 89-9, 309-318.

Garrett, R.G., 1991, The management, analysis and display of exploration geochemical data. in Exploration Geochemistry

Workshop, Geological Survey of Canada, Open File 2390, 9.1-9.41.

Garrett, R.G., and Grunsky, E.C., 2001, Weighted sums – knowledge based empirical indices for use in exploration geochemistry. Geochemistry, Exploration, Environment, Analysis, 1, 135-141.

Garrett. R.G., and Grunsky, E.C., 2003, S and R functions for the display of Thompson-Howarth plots, Computers & Geosciences, 29, 239-242.

Garrett, R.G., Kane, V.E., Zeigler, R.K., 1980, The Management and Analysis of Regional Geochemical Data, Journal of Geochemical Exploration, 13, 113-152.

George, H. and Bonham-Carter, G.F., 1989, An example of spatial modelling of geological data for gold exploration Star Lake area, in F.P. Agterberg and G.F. Bonham-Carter, eds., Statistical Applications in the Earth Sciences. Geological Survey of Canada, Paper 89-9, 171-183.

Govett, G.J.S. and Nichol, I., 1979, Lithogeochemistry in Mineral Exploration in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 339-362.

Govett, G.J.S., 1989, Bedrock Geochemistry in Mineral Exploration, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 273-200.

Grunsky, E.C., 1986a. Recognition of Alteration in Volcanic Rocks Using Statistical Analysis of Lithogeochemical Data, Journal of Geochemical Exploration, 25, 157-183.

Grunsky, E.C., 1986b, Recognition of Alteration and Compositional Variation Patterns in Volcanic Rocks Using Statistical Analysis of Lithogeochemical Data, Ben Nevis Township Area, District of Cochrane, Ontario; Ontario Geological Survey, Open File Report 5628.

Grunsky, E.C., 1990, Spatial Factor Analysis: A Technique to Assess the Spatial Relationships of Multivariate Data, in F.P. Agterberg and G.F. Bonham-Carter, eds., Statistical Applications in the Earth Sciences, , Geological Survey of Canada Paper 89-9, 329-347.

Grunsky, E.C., 1991, Geology of the Batchawana Area, District of Algoma; Ontario Geological Survey, Open File Report 5791.

Grunsky, E.C., 2000, Strategies and Methods for the Interpretation of Geochemical Data in Exploration Geochemistry in Today's World, Queen's University, Kingston, 11-17 March, 2000.

Grunsky, E.C., 2001, A Program for Computing RQ-Mode Principal Components Analysis for S-Plus and R, Computers & Geosciences, 27, 229-235.

Grunsky, E.C., 2002a, R: a data analysis and statistical programming environment – an emerging tool for the geosciences, Computers & Geosciences, 28, 1219-1222.

Grunsky, E.C., 2002b, Shareware and freeware in the Geosciences II. A special issue in honour of John Butler, E.C. Grunsky, ed., Computers & Geosciences, 28.

Grunsky, E.C., 2006, The evaluation of geochemical survey data: Data analysis and statistical methods using Geographic Information Systems, in J. Harris, ed., GIS for the Earth Sciences, Geological Association of Canada Special Publication 44, 229-283

Grunsky, E.C. and Agterberg, F.P., 1988, Spatial and multivariate analysis of geochemical data from metavolcanic rocks in the Ben Nevis area, Ontario. Mathematical Geology, 20, 825-861.

Grunsky, E.C. and Agterberg, F.P., 1992, Spatial Relationships of Multivariate Data, Mathematical Geology, 24, 731-758.

Grunsky, E.C., Easton, R.M., Thurston, P.C., and Jensen, L.S., 1992, A Statistical Approach to the Characterization and Classification of Archean Volcanics Rocks of the Superior Province, Geology of Ontario, Ontario Geological Survey Special 4, Part 2, 1397-1438.

Grunsky, E.C. and Smee, B.W., 1999, The differentiation of soil types and mineralization from multi-element geochemistry using multivariate methods and digital topography, Journal of Geochemical Exploration, 67, 287-299.

Gupta, R.P., 1991, Remote Sensing Geology, Springer-Verlag,

Hall, G.E.M., 1997, Recent Advances in Geoanalysis and Their Implications, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 293-294.

Hamilton, S., 1995, Lake Sediment Geochemistry of the Cow River Area, Ontario Geological Survey, Open File Report 5917.

Harman, P.G., Bye, S.M., and Munro, A.G., 1989, Image Processing of Geophysical and Geochemical Exploration Data Sets, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 822.

Harris, J.R. Grunsky, E.C., Wilkinson, L., 1997, Developments in the Effective Use of Lithogeochemistry in Regional Exploration Programs: Application of GIS Technology, in A.G. Gubins, ed., Proceedings of Exploration '97: Fourth Decennial International Conference on Mineral Exploration, 285-292.

Harris J.R., Wilkinson, L., Grunsky, E.C., Heather, K. and Ayer, J., 1999, Techniques for analysis and visualization of lithogeochemical data with applications to the Swayze greenstone belt, Ontario. Journal of Geochemical Exploration, 67, 301-334.

Harris J.R., Grunsky, G., Wilkinson, L., 2000, Effective use and interpretation of lithogeochemical data in regional mineral exploration programs: Application of Geographic Information System (GIS) technology, Ore Geology Reviews. 16, 107-143.

Harris, J.R., 2006a. Statistical, mathematical and geostatistical methods for dealing with glacial dispersal: Application of GIS technology to till data from the Swayze greenstone belt and Cape Breton Island, in J. Harris, ed., GIS for the Earth Sciences, Geological Association of Canada Special Publication 44, 317-368.

Harris, J.R., 2006b. Integration of geoscience data for mapping potassic alteration, Swayze greenstone belt, Ontario, Canada, in J. Harris, ed., GIS for the Earth Sciences, Geological Association of Canada Special Publication 44, 369-396.

Hartigan, J.A., 1975, Clustering Algorithms, Wiley.

Hausberger, G., 1989, GIS and Computer-Mapping Aspects of the Austrian Stream-Sediment Geochemical Sampling Project. in J.N. Van Driel, and J.C. Davis, eds., Digital Geologic and Geographic Information Systems, American Geophysical Union Short Course in Geology, 10, 25-45.

Hawkes, H.E. and Webb, J.S., 1962, Geochemistry in Mineral Exploration, First Edition, Harper and Row.

Helsel, D.R., 1990, Less than obvious: Statistical treatment of data below the detection limit, Environmental Science and Technology, 24, 1766-1774.

Holroyd, M.T., 1989, The Relevance of Data Base Technology to Resource Exploration Data, in G.D. Garland, ed., Proceedings of Exploration ’87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 811-821.

Hornbrook, E.H., 1989, Lake Sediment Geochemistry: Canadian Applications in the Eighties, in G.D. Garland, ed., Proceedings of Exploration ’87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 405-416.

Howarth, R.J., 1983, Mapping, Chapter 5, Statistics and Data Analysis in R.J. Howarth, ed., Geochemical Prospecting, 2, in G.J.S. Govett, ed., Handbook of Exploration Geochemistry, Elsevier, 111-205.

Howarth, R.J. and Earle, S.A.M., 1979, Application of a Generalized Power Transformation to Geochemical Data, Mathematical Geology, 11, 45-62.

Howarth, R.J. and Martin, L., 1979, Computer-based Techniques in the Compilation, Mapping and Interpretation of Exploration Geochemical Data, Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 544-574.

Howarth, R.J. and Sinding-Larsen, R., 1983, Multivariate Analysis, Chapter 6, Statistics and Data Analysis in R.J. Howarth, ed., Geochemical Prospecting, 2, in G.J.S. Govett, ed., Handbook of Exploration Geochemistry, Elsevier. 207-289.

Isaaks, E,H., and Srivastava, R.M., 1989, An Introduction to Applied Geostatistics, Oxford University Press.

Jaquet, J.-M., Froidevaux, F., Bernet, J.-P., 1975, Comparison of Automatic Classification Methods Applied to Lake Geochemical s, Mathematical Geology, 7, 237-266.

Jackson, J.E., 2003, A User’s Guide to Principal Components, Wiley-Interscience.

Jolliffe, I.T., 2002, Principal Components Analysis, 2nd edition, Springer.

Jöreskog, K.G., Klovan, J.E. and Reyment, R.A., 1976, Geological Factor Analysis. Elsevier Scientific Publishing Company.

Journel, A.G. and Huijbregts, C.J., 1978,: Mining Geostatistics, Academic Press.

Joyce, A.S., 1984, Geochemical Exploration, The Australian Mineral Foundation Inc.

Kaufman, L, Rousseeuw, P.J., 1990, Finding Groups in Data, An Introduction to Cluster Analysis. John Wiley.

Klassen, R.A., 1997, Glacial History and Ice Flow Dynamics Applied to Drift Prospecting and Geochemical Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 221-231.

Kohonen, T., 1995, Self-Organizing Maps, Springer-Verlag.

Kuosmanen, V. (Editor), 1988, Exploration target selection by integration of geodata using statistical and image processing techniques: an example from Central Finland. Geological Survey of Finland, Report of Investigation 84, Part 2, Atlas, 47 pages, 5 figures, 1 table and 40 plates.

Kürzl, H, 1988, Exploratory data analysis: recent advances for the interpretation of geochemical data, Journal of Geochemical Exploration, 20, 309-322.

Kruskal, J.B., 1964, Multidimensional scaling by optimising goodness of fit to non-metric hypothesis, Psychometrika, 29, 1-27.

Krzanowski, W.J., 1988, Principles of Multivariate Analysis, A User's Perspective, Clarendon, Press.

Krzanowski, W.J. and Bailey, T.C., 2007, Extraction of Spatial Features Using Factor Methodss Illustrated on Stream Sediment Data, Mathematical Geology, 39, 69-85.

Lee, L., and Helsel, D., 2005, Statistical analysis of water-quality data containing multiple detection limits: S-language software for regression on order statistics, Computers & Geosciences, 31, 1241-1248.

Lee, L., and Helsel, D., 2007, Statistical analysis of water-quality data containing multiple detection limits II: S-language software for nonparametric distribution modeling and hypothesis testing, Computers & Geosciences, 33, 696-704.

Levinson, A.A., 1980, Introduction of Exploration Geochemistry, Second Edition, Applied Publishing.

Lindqvist, L., 1976, SELLO, A Fortran IV program fo the transformation of skewed distributions to normality, Computers & Geosciences, 1, 129-145.

Link, R.F., and Koch, G.S., 1975, Some consequences of applying lognormal theory to pseudolognormal distributions, Mathematical Geology, 7, .117-128.

Martin, L., 1989, Expert Systems and Their Use as Exploration Assistants, in G.D. Garland, ed., Proceedings of Exploration ’87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 826-834.

Martin-Fernandez, J.A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V., 1998, A critical approach to non-parametric classification of compositional data, in A. Rizzi, M. Vichi, and H.H. Bock, eds., Advances in data science and classification, Springer, 49-56.

Martin-Fernandez, J.A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V., 2000, Zero replacement in compositional datasets, in H. Kiers, J. Rasson, P. Groenen, and M. Shader, eds., Studies in classification, data analysis, and knowledge organization: Spriner, Berlin(D), 155-160.

Mazzucchelli, R.H., 1989, Exploration Geochemistry in Areas of Deeply Weathered Terrain: Weathered Bedrock Geochemistry, in G.D. Garland, ed., Proceedings of Exploration ’87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 300-311.

Mazzucchelli, R.H., 1997, Geochemical Exploration in Areas Affected by Tropical Weathering—An Industry Perspective, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 315-322.

McClenaghan, M.B., Thorleifson, L.H., and DiLabio, R.N.W., 1997, Till Geochemical and Indicator Mineral Methods in Mineral

Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 233-247.

McQueen, J., 1967, Some methods for classification and analysis of multivariate observations, 5th Berkeley Symposium on Mathematics, Statistics, and Probability, 1, 281-298.

Mellinger, M. 1987, Multivariate Data Analysis. Its Methods, Chemometrics and Intelligent Laboratory Systems, 2, 29-36.

Mellinger, M., 1989, Computer tools for the integrative interpretation of geoscience spatial data in mineral exploration. In Statistical Applications in the Earth Sciences. Bonham-Carter, G.F. and Agterberg, F.P. (Editors). Geological Survey of Canada Paper 89-9, 135-139.

Mellinger, M., Chork, S.C.Y., Dijkstra, S., Esbensen, K.H., Kürzl, H., Lindqvist, L., Saheurs, J.-P., Schermann, O., Siewers, U., and Westerberg, K., 1984, The Multivariate Chemical Space, and the Integration of the Chemical, Geographical, and Geophysical Spaces, Journal of Geochemical Exploration, 21, 143-148.

Meyer, W.T., Tehobald, Jr., P.K., and Bloom, H., 1979, Stream Sediment Geochemistry in Geophysics and Geochemistry in the Search for Metallic Ores, Proceedings of Exploration 77 – an international symposium held in Ottawa, Canada in October 1977, Geological Survey of Canada Economic Geology Report 31, 411-434.

Oliver, J., Payne, J, and Regabliati, M., 1996: Precious-metal-bearing Volcanogenic Massive Sulfide Deposits, Campo Morado, Guerrero, Mexico, Exploration Mining Geology, 6, 119-128.

Pawlowsky, V., 1989, Cokriging of Regionalized Compositions, Mathematical Geology, 21, 513-521.

Pawlowsky-Glahn V. and Buccianti A., 2002, Visualization and modeling of sub-populations of compositional data; statistical methods illustrated by means of geochemical data from fumarolic fluids. International Journal of Earth Sciences. 91, 357-368.

Pawlowsky-Glahn, V. and Egozcue, J.J, 2006, Compositional data and their analysis, in A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, eds., Compositional Data Analysis in the Geosciences: From Theory to Practice, Geological Society, London, Special Publications, 264, 1-10.

Pebesma, E.J., 2004, Multivarilabe geostatistics in S: the gstat package, Computers & Geosicences, 30, 683-691.

Pieters, C.M. and Englert, P.A.J., 1993, Remote Geochemical Analysis: Elemental and Mineralogical Composition, Cambridge University Press.

Plant, J.A., Hales, M. and Ridgway, J. 1989, Regional Geochemistry Based on Stream Sediment Sampling, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 384-404.

Rebagliati, M., 1999: Applied Exploration Geochemistry: Campo Morado Precious-Metal-Bearing Volcanogenic Massive Sulphide District, Guerrero, Mexico, 19th International Geochemical Exploration Symposium, Vancouver, British Columbia, Canada, April 10-16, 1999, Abstract.

Reimann, C., Filzmoser, P., Garrett, R.G., 2005. Background and threshold: Critical comparison of methods of determination, Science of the Total Environment, 346, 1-16.

Rencz, A.N., 1999, Remote Sensing for the Earth Sciences, in A.N. Rencz, ed., Volume 3 in R.A. Ryerson, ed., Manual of Remote Sensing, Third Edition, John Wiley & Sons.

Reyment, R.A. and Jöreskog, K.G., 1993, Applied Factor Analysis in the Natural Sciences, Cambridge University Press.

Richards, J.A. and Jia, X., 1999, Remote sensing digital image analysis, An Introduction. Third, Revised and Enlarged Edition, Springer-Verlag.

Rock, N.M.S., 1987, Robust, An Interactive Fortran-77 Package for Exploratory Data Analysis using Parametric, Robust and Nonparametric Location and Scale Estimates, Data Transformations, Normality Tests, and Outlier Assessment, Computers & Geosciences, 13, 463-494.

Rock, N.M.S., 1988, Numerical Geology, A Source Guide, Glossary and Selective Bibliography to Geological Uses of Computers and Statistics, Lecture Notes in, S. Bhattacharji, G. Friedman, H.J.. Neugebauer and A. Seilacher, Earth Sciences, 18, Springer-Verlag.

Rousseeuw, P. J. and van Driessen, K., 1999, A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212-223.

Rose, A.W., Hawkes, H.E., and Webb, J.S. 1979, Geochemistry in Mineral Exploration, Second Edition, Academic Press.

Sammon, J.W., 1969, A non-linear mapping for data structure analysis. IEEE Transactions in Computing, C18, 401-409.

Sanford, R.F, Pierson, C.T., and Crovelli, R.A., 1993, An Objective Replacement Method for Censored Geochemical Data, Mathematical Geology, 25, 59-80.

Shaw, J., 1989, Geochemical Exploration in Areas of Glaciated Terrain: Geological Processes, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3, 335.

Sinding-Larsen, R., 1975, A computer method for dividing a regional geochemical survey area into homogeneous subareas prior to statistical interpretation. in I.L, Elliott and W.K. Fletcher eds., Geochemical Exploration 1974, Elsevier, 191-217.

Sinclair, A.J., 1976, Application of Probability Plots in Mineral Exploration, Association of Exploration Geochemists Special Publication 4.

Smee, B.W., 1997, The Formation of Surficial Geochemical Patterns Over Buried Epithermal Gold Deposits in Desert Environments: Results of a Test of Partial Extraction Techniques, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, 301-314.

Smith, R.E., 1989, Using Lateritic Surfaces to Advantage in Mineral Exploration, in G.D. Garland, ed., Proceedings of Exploration '87: Third Decennial International Conference on Geophysical and Geochemical Exploration for Minerals and Groundwater, Ontario Geological Survey, Special Volume 3,. 312-322.

Smith, R.E., Anand, R.R., and Alley, N.F., 1997, Use and Implications of Paleoweathering Surfaces in Mineral Exploration, in A.G. Gubins, ed., Proceedings of Exploration 97: Fourth Decennial International Conference on Mineral Exploration, p. 335-346.

Smith, R.E., Birrell, R.D., and Brigden, J.F., 1989, The implications to exploration of chalcophile corridors in the Archaean Yilgarn

Block, Western Australia, as revealed by laterite geochemistry, Journal of Geochemical Exploration, 32, 169-184.

Smith, R.E. and Perdrix, J.L. 1983, Pisolitic laterite geochemistry in the Golden Grove massive sulphide district, Western Australia, Journal of Geochemical Exploration, 18, 131-164.

Smith, R.E., Perdrix, J.L., Davis, J.M., 1987, Dispersion into Pisolitic Laterite from the Greenbushes Mineralized Sn-Ta Pegmatite System, Western Australia, Journal of Geochemical Exploration, 28, 251-265.

Stanley, C.R. 1987, PROBPLOT, An Interactive Computer Program to Fit Mixture of Normal (or Log normal) Distribution with Maximum Likelihood Optimization Procedures, Association of Exploration Geochemists Special Volume 14, 1 diskette.

Stanley, C.R., 2003, THPLOT.M: a MATLAB function to implement generalized Thompson–Howarth error analysis using replicate data. Computers & Geosciences, 29, 225-237.

Stanley, C.R. 2006, On the special application of Thompson-Howarth error analysis to geochemical variables exhibiting a nugget effect Geochemistry: Exploration, Environment, Analysis, 6, 357-368.

Stanley, C.R. and Sinclair, A.J., 1987, Anomaly recognition for multi-element geochemical data- A background characterization approach. Journal of Geochemical Exploration, 29, 333-53.

Stanley, C.R. and Sinclair, A.J., 1989, Comparison of probability plots and the gap statistic in the selection of thresholds for exploration geochemistry data. Journal of Geochemical Exploration, 32, 355-357.

Thompson, M. and Howarth, R.J., 1973, The rapid estimation and control of precision by duplicate determinations. The Analyst 98, pp. 153–160.

Thompson, M. and Howarth, R.J., 1976a, Duplicate analysis in practice––Part 1. Theoretical approach and estimation of analytical reproducibility. The Analyst 101, 690–698.

Thompson, M. and Howarth, R.J., 1976b, Duplicate analysis in practice––Part 2. Examination of proposed methods and examples of its use. The Analyst 101, 699–709.

Thompson, M. and Howarth, R.J., 1978, A New Approach to the Estimation of Analytical Precision, Journal of Geochemical Exploration, 9, 23-30.

Trépanier, S., 2006, Identifcation de domains géochemiques à partier des levés régionaux de sediments de fond de lacs, Projt 2004-09, Consortium de recherche en exploration minérale, Presentation.

Tukey, J.W., 1977, Exploratory Data Analysis, Addison-Wesley.

Venables, W.N., and Ripley, B.D., 2002, Modern Applied Statistics with S, fourth Edition Springer-Verlag.

Vincent, R.K., 1997, Fundamentals of Geological and Environmental Remote Sensing, Prentice Hall.

van den Boogaart, K.G. and R. Tolosana-Delgado, R., in press.: "compositions": a unified R package to analyze compositional data, Computers & Geosciences. doi:10.1016/j.cageo.2006.11.017

von Eynatten, H., Pawlowsky-Glahn, and Egozcue, J.J., 2002, Understanding perturbation on the simplex: A simple method to better visualize and interpret compositional data in ternary diagrams, Mathematical Geology, 34, 249-258.

Von Eynatten H., Barcelo-Vidal C., Pawlowsky-Glahn V., 2003, Composition and discrimination of sandstones; a statistical evaluation of different analytical methods. Journal of Sedimentary Research, 73, 47-57.

Wackernagel, H, and Butennuth, C., 1989, Caractérisation d'anomialies géochemiques par la géostatistique multivariable, Jouranl of Geochemical Exploration, 32, 437-444.

Wilkinson, L., Harris, J.R. and Grunsky, E.C., 1999, Building a Lithogeochemical Database for GIS Analysis; Methodology, Problems and Solutions, Geological Survey of Canada Open File 3788.

Wilkinson, L., Harris, J.F., Kjarsgaard, B.K., and McClenaghan, 2006, Till geochemistry for kimberlite exploration: Using GIS to visualize, analyze and decide, 297-316, in J. Harris, ed., GIS for the Earth Sciences, Geological Association of Canada Special Publication 44, 297-316.

Zhou, D., 1985, Adjustment of geochemical background by robust multivariate methods, Journal of Geochemical Exploration, 24, 207-222.

Zhou, D., ROPCA, 1989, A Fortran Program for Robust Principal Components Analysis, Computers & Geosciences, 15, 59-78.

Zhou, D., Chang, T. and Davis, J.C., 1983, Dual Extraction of R-Mode and Q-Mode Factor Solutions, Mathematical Geology, 15, 581-606.

APPENDIX 1

Logratios and Compositional Data

Compositional data should be adjusted by the use of log-ratios. A compositional vector x defined by D component variables (elements). By definition, this vector will sum to a constant (100%) and as a result, the composition can be described by D-1 of the variables. A composition x can be transformed by

$$y_i = \log(x_i/x_D) \quad (i = 1, \ldots, D\text{-}1)$$

There is no loss of information by choosing one of the variables as a divisor. This transformation is known as the additive log-ratio (alr). The resulting logratio coordinates cannot be projected onto orthogonal axes because the axes are at 60º (Pawlowsky-Glahn and Egozcue, 2006) and creates difficulties when comparing compositions using different denominators. In particular, measures of distances between alr-transformed observations are not equal when using different denominators and the angles between vectors cannot be computed using a standard Euclidean inner product.

An alternative way of transforming a compositional vector is by applying the log-centered ratio, namely:

$$z_i = \log(x_i/g(x_D)) \quad (i = 1, \ldots, D),$$

where $g(x_D)$ is the geometric mean of the composition The log-centered ratio (clr) is useful because it preserves all of the variables in the composition. However, the inverse of the covariance matrix for this transform is singular, which requires a special generalized inverse procedure for computation.

An important aspect of assessing compositions is the calculation of an adequate measure of variability. This is done by the creation of a variation matrix, *T* defined by:

$$\tau_{ij} = \mathrm{var}\{\log(x_i/x_j)\} \ (i=1,\ldots,d;\ j=i+1,\ldots,D)$$

and the mean, E, is expressed as:

$$\xi_{ij} = E\{\log(x_i/x_j)\} \ (i=1,\ldots,d;\ j=i+1,\ldots,D)$$

The variability matrix *T* summarizes the contribution that any pair of variables makes in a sub-compositional analysis. For example, consider a major element oxide composition consisting of SiO_2, Al_2O_3, MgO, FeO, CaO, Na_2O, K2O, TiO_2, and MnO. A sub-composition may be interested in examining the relationships of MgO, FeO and Na_2O. The amount of compositional variability that these elements will account for can be expressed by the sum of ($\tau_{MgO,FeO,}$ $\tau_{MgO,Na2O,}$ $\tau_{FeO,Na2O}$). This is an important concept in understanding the significance of sub-compositional data which will never fully explain the overall variation of the data.

More recent developments by Egozcue et al. (2003) have identified the isometric logratio (ilr), which is a transformation that defines compositional vectors in an orthonormal basis. A very simple explanation of this transformation is described in Pawlowsky-Glahn and Egozcue (2006). The application of the ilr transform requires the construction of “balances”, which are ratios of selected variables into groups (i.e. elements associated with a fractionation process versus elements associated with alteration). These balances are used to construct new variables that exist in an orthonormal base from which standard Euclidean measures can be calculated (mean, variance, etc.).