1 INTRODUCTION

1.1 The meaning and relevance of data analysis

Data analysis is a matter of condensing large volumes of information into summary information that is more concise, but no less accurate and truthful, from which conclusions and/or decisions can be more easily and efficiently made. They may not be "snap" decisions but equally, they are not decisions that allow time for the decision-maker to study the data in their raw form. Depending on the audience for whom they are intended, these summaries may range between two extremes:

- Slightly abridged versions of the original data set; for example, the conversion of a file of co-ordinates and analyses into a map or series of maps where the numerical values of the analyses are "posted" against the locations of the samples.

- Single-sentence conclusions such as "Follow-up work is justified."

The first example is appropriate to a detailed exposition of a project, prepared for an interested potential joint-venture partner, who would nevertheless be wise to request a file containing the raw data, and have it examined by an impartial expert, before committing to a major investment. The second example is appropriate to a press release or annual report summarizing the results of several different projects, though more detail might be expected if only one project was being described. Between these two extremes lie such inappropriately-chosen summaries as:

- Oral presentations incorporating tables of figures that are rarely, if ever, displayed long enough for the information they contain to be digested by their audience, and

- Press releases incorporating so-called "anomaly maps" in which no actual assays, analyses, or even contour labels, are incorporated

All of the methods covered in this presentation, whether statistical, graphical or spatial, are concerned with the extraction of important information from data in which it is obscured by what is perceived, perhaps temporarily, to be less important information. It is important that this judgement (as to what is important, and what is less important) be as well-informed as possible, with regard to possibly changing needs and circumstances.

Facilities for computer-assisted data processing are now available to most explorationists, and even with standard computer software, and certainly with packages like Statistica and Systat, sophisticated methods of data analysis have become very easy to apply, and equally easy to misapply. They do not issue warnings that the chosen method may be inappropriate for the problem at hand, when the results they present are based on inadequate or unsuitable data, or when the necessary conditions for their proper application have not been satisfied; nor do they have a requirement that their attractively-packaged conclusions should make geological sense.

The methods of data analysis that will be presented are not intended as a means to avoid careful geological thought, and any conclusions that arise from them must not fly in the face of geological and other observations. On the other hand, they need not necessarily confirm previous inferences (however dearly cherished) from geological and other lines of investigation. If we limit the conclusions we are prepared to draw to those that confirm our current theories, we eliminate our chances of discovering anything new – and this does not only apply to statistical data processing, of course.

The cost of a program of data analysis on a data set of 2,000 samples, assuming that hardware and software costs have already been amortized, is currently of the order of $4,000-$6,000 (amounts in Canadian currency); in other words, $2 - $3 per sample. Analytical costs vary between $5 and $15 per sample (or more, if specialized techniques are required), and collection costs between $10 and $30 (depending on local salary scales; more, if the work is helicopter-supported or involves other complex logistics). Therefore, a comprehensive data review will add less than 10% to the costs of a geochemical program.

1.2 Material Covered in These Notes

The notes for this presentation will begin with a description of basic methods of checking the integrity of a data set and its suitability for data analysis. The importance of incorporating all available information, prior to beginning a geochemical interpretation, will also be stressed. The remainder of the material covered will be divided into univariate and multivariate methods, and two others (the calculation of multi-element indices, and the calculation of correlations) that straddle the boundary between the two methods.

Univariate methods will be described from the point of view of concisely summarizing larger data sets by **statistical** and **graphical** methods, and the identification and validation of geochemical samples, or groups of samples, whose analyses justify follow-up work: or "**anomalies**", in geochemical shorthand. The usefulness and importance of **percentiles** will be stressed in this section.

The use of the **correlation coefficient** is important in a number of multivariate techniques and in quantifying the relationship between different elements in the geochemical environment. The calculation and representation of correlation coefficients and matrices will be covered briefly. The combination of elements known to be of application in the search for certain mineral-deposit types, or certain key lithologies, into **multi-element indices** has a special appeal since it requires no advanced statistical knowledge, it is readily accomplished in a spreadsheet, and can be tailored to suit the individual needs of the company, project, or geologist.

The scope of multivariate methods in geochemistry is greater since they can be used to detect "anomalies" *per se*, with appropriate data, but are more commonly used to identify large-scale processes and features, including lithological features as an aid to mapping in areas where exposure is poor or published geological information is lacking. The methods described in detail comprise **regression, factor** and **discriminant analysis.** C**luster analysis** will be also covered briefly although it is less widely used and generally-available software packages do not greatly facilitate or encourage its use in geochemical exploration.

Appendix A describes the scope of a geochemical orientation survey.

Step-by-step instructions for carrying out some of the methods of data analysis that will be covered, using MS-Excel, Systat and Surfer, will be distributed at the time of the practical session and can also be found at the website www.dirtbagger.com/methods.html. It is hoped that this will encourage participants to attempt analysis of their own data subsequently.

1.3 An Underutilized Resource

Reliable multielement analyses are available in North America for as little as $6 per sample and have been requested increasingly for routine geochemical exploration programs, but many geologists who use geochemical methods in exploration are unaware of the potential uses to which these data may be put. Consequently, such data sets tend to be underutilized.

Whether or not they are subjected to sophisticated statistical treatment, multielement analyses constitute a valuable resource of information. A typical package, whether derived from induction-coupled plasma spectrometry (ICP) or instrumental neutron-activation analysis (INAA), or both, will probably contain analyses of some ore elements (e.g. Au, Cu, Pb, Zn) and pathfinder elements (e.g As, Sb, Mo, W). It may also contain elements that if interpreted with care, can assist in geological mapping in areas of poor exposure or where geological maps are inadequate or absent (e.g. Ni, Ca, Cr, K, Sr, rare-earth elements), and elements that act as monitors of surficial or environmental processes (Br, partial/total ratios of Fe, Co etc.). A more detailed summary of the uses of commonly-analyzed elements, indexed alphabetically and by the periodic table, may be found at the website www.dirtbagger.com/applabs/elements.html.

2 CHECKING DATA INTEGRITY

2.1 Bad Data

The truth of the expression "garbage in, garbage out" is nowhere more vividly demonstrated than in data analysis, including when applied to geochemical exploration. A significant proportion of the time expended on any interpretation exercise is (or should be) taken up with checking the integrity of the data set. Typically, a file of geochemical data may contain shortcomings such as the following:

- missing analyses recorded as zeros;
- variable detection limits
- "undetectable" values expressed as zeros, inequalities or negative numbers;
- displaced or mixed columns of analyses and coordinates, and
- quality-assurance data that have not been removed for separate evaluation.

The more well-meaning hands a digital geochemical database has passed through, the greater the likelihood that these will be present. Although not feasible with very large data sets, it is often beneficial to spend a couple of hours visually scanning through the data in a spreadsheet or text editor. Potential problems like mixed detection limits, text entries in numeric fields, and missing (or unwanted) data can be spotted and allowance made for them before data processing begins. Some of

these problems would not be picked up by the processing software and would lead to incorrect conclusions, which would probably not be recognized as such.

There are, however, certain procedures for proofing the integrity of data sets more rapidly, once they have been loaded into a spreadsheet or database:

- The data can be sorted column by column, which will put "suspect" values at one or other end of the sequence where they can be examined more readily. It is advisable to insert a temporary sequential index number before this exercise is carried out, so that the original order of the samples can be restored

- The MIN or MAX function can be applied to each column, which will draw attention to atypical entries

- Preliminary plots (to be clearly labelled as such) will draw attention to such features as poor calibration between laboratory batches, or mixed coordinates.

- Even if duplicate samples have not been labelled as such, sorting the data simultaneously by east-west and north-south coordinate will cause samples collected from the same site to be brought together in the sequence. Calculating the geographic distance between each sample point and the one following it, and searching for zeros, will enable duplicates to be rapidly identified and removed.

2.2 Data Transformations

Many statistical techniques are **parametric**; that is, they require the fulfilment of certain assumptions about the data for any conclusions derived from them to be valid. The most important of these is that the frequency distribution of every variable conform to a Normal distribution; that is, their histogram is bell-shaped, their cumulative frequency distribution is shaped like a stylized "f" and their probability plot (see below) forms a straight line. Furthermore, the distribution should be free of outliers and most values (at least 70%) should exceed the analytical detection limit. Many geochemical variables do not, in their "raw" state, display such distributions.

If the application of parametric statistical methods is envisaged, it is therefore necessary to inspect the frequency distributions of the elements in a geochemical data set, and apply appropriate transformations prior to data analysis, or in some cases to omit certain elements altogether.

The most rapid and convenient means of characterizing the nature of a frequency distribution is in the form of a **probability plot**. This is described in more detail in Section 4.1.2.

Some authorities maintain that the time spent "parameterizing" the data is not justified and that perfectly valid conclusions, that make geological sense, can be drawn from data that have not been subjected to such time-consuming processing. This apparently tedious work can, however, be expedited with good organization, and practice; an example of such an approach is shown at www.dirtbagger.com/methods.html.

2.3 Detection Limits

When some of the analyses, for a particular element in a data set, are reported as "less than the detection limit" or "undetectable" (some labs even report them as "zero", though this is not very good practice) the problem is known as "censoring". Although such entries can be readily stored in databases and spreadsheets, most data-processing utilities will "crash" on encountering the "less than" sign (<). Therefore, it is necessary to convert the inequality to a numerical value. A common procedure has been to replace the "less than" sign with a minus sign (-). This has the effect of making the data accessible to utilities that require numerical input, but it can result in the creation of misleading results if the negative values are not converted, prior to data processing, to something realistic. The presence of negative values representing inequalities is one of the most common flaws that are encountered in "inherited" databases that geochemists are asked to interpret, and a close eye should be kept out for it.

Censoring only becomes a problem, in the application of parametric statistics, if more than about 30% of the analytical values fall below this limit. For percentages of "undetectables" less than 30%, a common and acceptable practice is for them to be converted to the value of half of the detection limit (though more complex procedures have also been suggested). Meaningful results can then be obtained from the data set, using parametric statistical procedures. If there are more "undetectables" than this (and regrettably, this limitation often applies to gold analyses), the carrying out of parametric statistical procedures may lead to unreliable results. However, the calculation of percentiles above the detection

limit is still possible. Therefore (for example) a “threshold” value equivalent to the 97.5-percentile can in theory be calculated (see Sections 4.1.3, 5.3.2) even if only 1 analysis in 40 is reported as detectable, although analytical precision tends to be poor for analyses that only exceed the detection limit by a small margin.

A more serious problem occurs when detection limits vary from sample to sample. This is a particular problem in neutron-activation (INAA) analysis, whose sensitivity, when other irradiation parameters are constant, is determined by the quantity of sample material, which may vary considerably. More recently, samples have been sorted by weight prior to INAA analysis, so that irradiation can be increased to compensate for the smaller samples in a particular batch. However, the problem crops up in some older data sets, including those relating to government-sponsored reconnaissance programs in both Canada and the U.S. It is unrealistic to assign a value of half the detection limit under these circumstances because detection limits can sometimes be so high that halving them creates an “anomalous” value from what is essentially “unknown”. The only option, which is admittedly not very satisfactory, is to assign to all “undetectables” the value of half of the lowest detection limit in the data set for that variable.

3 APPLES AND ORANGES

One of the more frustrating experiences for a geochemist charged with extracting useful new information from a geochemical database is to announce the identification of certain geochemical inhomogeneities in the database, and their possible link to mineralization, bedrock geology, or perhaps hitherto-unidentified variations in regolith type, and to be told by the on-site exploration staff that “We knew that already”. This stems from a rather common misconception among non-geochemists that while the confirmation of other observations by analysis of geochemical data is reassuring and therefore desirable, the real power of the method lies in adding additional information that cannot be gained by other methods. This is dependent on the incorporation of all available information into the database before the interpretation exercise begins, so that the existing information can be used as a platform from which to proceed. In other words: incorporate as much information as you can about the inhomogeneity of the data and you are likely to identify more. Ignore the known inhomogeneities and all you are likely to do is rediscover them.

"Apples-and-oranges" situations are rather common in geochemical exploration. Just as it is not meaningful to generalize about the acidity of a basket of fruit containing apples and oranges collectively (since the latter are, in general, considerably more acid than the former) it is not meaningful to generalize about the Cr content of a suite of volcanic rock samples if it is known that the volcanic rocks range in composition from basalt through rhyolite.

The response to mineralization is in many cases very subtle, compared to much stronger responses due to differences in the lithology from which the sample material is derived, or the position in the soil profile from which the sample was collected, to name but two. The latter problem is particularly acute in the thick lateritic profiles in which so much exploration has been carried out over the last few years.

In certain cases (for example, where there is a thin, evenly-distributed soil profile) it is possible to reduce the introduction of unwanted variance by encouraging samplers to restrict their sampling as much as possible to one sample type. In areas of dissected lateritic terrain this is not possible; nor is it possible (or even desirable), in most regional or semi-regional geochemical studies, to select only certain lithologies from which to collect samples. In such cases one must collect whatever material can be reached and sampled without expending inordinate amounts of time, effort and money to homogenize the data set.

In these circumstances interpretation is aided greatly if the collectors of the samples make and record certain key characteristics of the material they are sampling and the site from which it is collected. Contrary to widely-held supposition, this need not be a difficult or time-consuming exercise, and standard software packages facilitate the customized design of suitable sample sheets. Setting up a system of coded descriptors facilitates data entry and subsequent computerized data-processing; it is also more adaptable to situations where field crews do not speak or write in a common language, provided that the coding instructions themselves are translated into the vernacular.

If information regarding mapped geology is available, the principal source lithology associated with each sample can be incorporated into the database at a later stage, prior to interpretation. When interpretation takes place, the data set can be broken down into major lithological, pedological and other subsets, and the values re-expressed in terms of the properties of each subset, prior to recombination. The conversion of the analyses into percentiles of each subset (so that all values range between 0 and 100) is recommended, although other methods such as subtracting the subset mean, and

dividing by its standard deviation ("normalizing") have been used. After adjustments have been made for gross compositional variations between the subsets, the data can be recombined and more subtle features, which may be related to mineralization, investigated more readily. An example of the application of this method can be seen in Figures 3.1a and 3.1b.

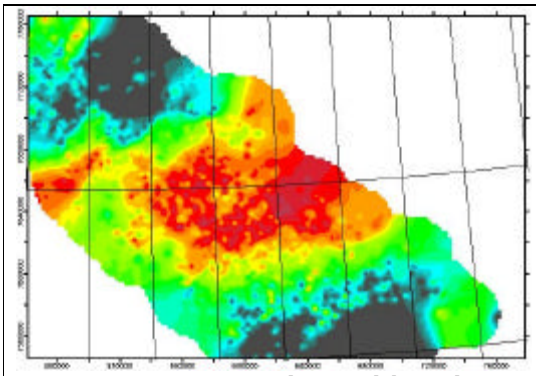

Figure 3.1a: Raw As values in lake sediment, Central Baffin Island. Plotted as percentiles: from black (lowest) through blue, green, yellow to red (highest).

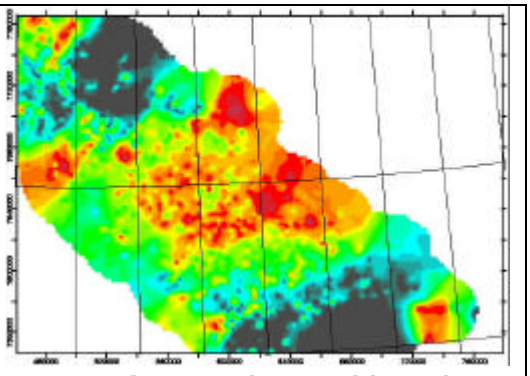

Figure 3.1b: As values in lake sediment, compensated for source lithology, same area. Note the appearance of zones of elevated values in the southeast and northwest.

4 UNIVARIATE METHODS

Univariate methods deal with data for which only one variable is considered at a time. These methods are fundamental to virtually all statistically-oriented geochemical studies, including those concerned principally with multivariate analysis, since a thorough understanding of individual variables is essential to the interpretation of results of multivariate methods. In fact, in many cases, the results of multivariate studies can be predicted by a detailed univariate approach, particularly if combined with a simple correlation (bivariate) study.

In this section, coverage will extend to both graphical techniques and those that involve making calculations to derive summary statistics from the data.

4.1 Methods of Characterizing Distributions

4.1.1 Histograms

Histograms consist of a series of contiguous upright rectangles, whose base width corresponds to a constant interval width, within the overall range of values encountered for that particular variable, and whose height expresses the frequency with which observations occur within that range. The latter can be expressed either in absolute terms (number of observations) or as a percentage of the total number of observations. Artificially-created histograms for a Normal (bell-shaped) distribution, and a positively skewed distribution rather typical of geochemical data, are shown in Figures 4.1a and 4.1b.

Histograms are a familiar method of displaying numerical information. Some obvious advantages of histograms as a means of visual representation of data are as follows:

- The total range of the data in the sample, and nature of the distribution, are apparent.
- Modes (most commonly-encountered values) can be recognised easily.
- The range of the values can be estimated rapidly, and
- Where the distinction is clear-cut, histograms can be used to distinguish between "background" and "anomalous" populations (although under most circumstances, probability plots are superior in this respect)

As well as being a fundamental component of most statistical packages, histograms may be readily prepared in spreadsheets such as Excel and QuattroPro. The generation of a histogram is demonstrated at www.dirtbagger.com/methods.html.

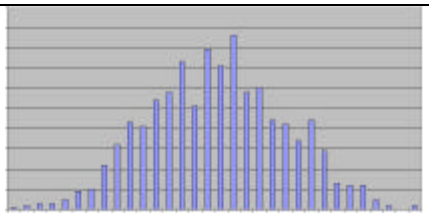

Figure 4.1a: Histogram of Normally-distributed data

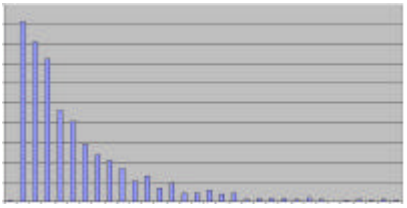

Figure 4.1b: Histogram of positively-skewed data

4.1.2 Cumulative Frequency Plots and Probability Plots

Whereas each bar in a histogram indicates the number of samples that fall within the upper and lower limits of an interval, plotted against that interval, the cumulative frequency diagram shows the number of entries (or, much more commonly, the percentage of the total) whose values fall below a certain value, plotted against that value. The cumulative frequency curve for a Normal distribution is shaped like a rather stylized "S" (Figure 4.2).

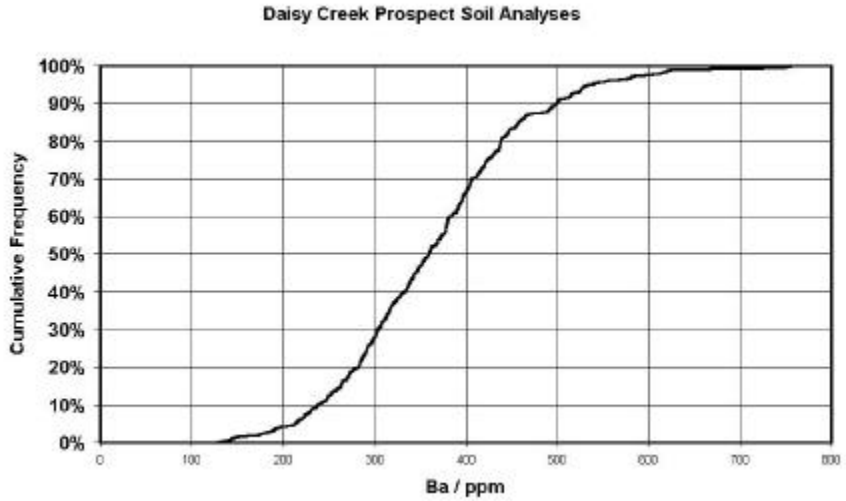

Figure 4.2: Cumulative frequency plot of Ba soil analyses, Daisy Creek Prospect, Montana

A special adaptation of the cumulative frequency curve is known as a probability plot, in which the y-axis is scaled in such a way that the "S" shaped cumulative frequency curve, representing a Normal distribution, plots as a straight line. Probability plots were already mentioned as a useful method of characterizing the nature of a distribution (Section 2). Any deviations from Normality can be quickly identified and another transformation experimented with, until a reasonable straight-line plot is arrived at (Figure 4.3).

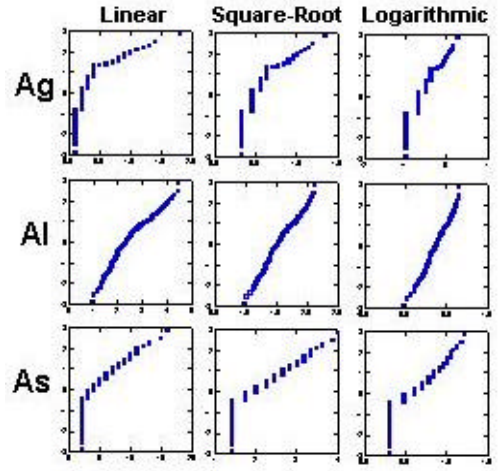

Figure 4.3: Linear, square-root and logarithmic probability plots for Ag, Al and As in soils, Daisy Creek Prospect, Montana. On the basis of these plots, Ag values would be subjected to a log-transformation and Al values to a square-root transformation, while As values would probably be omitted from the data set because too many of them (nearly 50%) fall below the analytical detection limit (represented by the straight vertical line at lower left). These plots are sized merely to characterize the shape of the distributions and it is not necessary that the axes be legible in this case.

Although it can be achieved using various iterative procedures, not available in most statistical packages, creation of a perfect straight line (representing a perfect Normal distribution) is unnecessary and indeed often undesirable, since deviations from rectilinearity may also be due to the presence of more than one population, which can be separated by methods that will be described later. Generally, either a square-root transformation, a logarithmic transformation or, very rarely a double-logarithmic (log-log) transformation will suffice. In some cases no transformation is necessary. It should never be assumed, however, that the nature of an element's distribution is a fundamental property of that element; it is advisable to check the nature of every element's distribution in every new data set.

Probability plots have also found application in the splitting of univariate, polymodal geochemical populations into unimodal subpopulations to facilitate the identification of anomalies; this is described in Section 5.6.

For the purpose of displaying data, cumulative frequency diagrams and probability plots are superior to histograms in a number of respects. Cumulative frequency diagrams are more suitable than histograms for the comparison of multiple populations, both visually and by certain statistical tests, although Box Plots may also be used (see Section 4.2.2). As shown above, probability plots can be used to select an appropriate data transformation for the creation of a Normal distribution and in certain cases, to partition the data set into subsets which may have geological or economic significance (see Section 5.6).

Cumulative frequencies may be readily calculated in MS-Excel using the "Data Analysis" option in the tools menu, as demonstrated at www.dirtbagger.com/methods.html.

4.1.3 Descriptive Statistics

Typical descriptive statistical parameters commonly used in geochemistry include the following:

- Minimum: Lowest value
- Maximum: Highest value
- Range: Absolute value of the difference between the lowest and highest values
- Mean: "Average" of the total data set; sum of all values, divided by the number of entries
- Median: The value that equally divides the data set, representing the middle case in the sequence
- Mode: Most frequent value in a data set
- Standard Deviation: A measurement of dispersion either side of the mean; the root-mean-square deviation from the mean
- Percentile: A one-hundredth division of the total size of a data set

Percentiles are a powerful and useful method of handling univariate geochemical data, and some elaboration of their meaning and application is called for. The first percentile in a data set is the value, of the variable in question, below which one percent of the entries lie; the fiftieth percentile (or 50-percentile) divides the data set into two equal parts, and is better known as the median. Some other familar percentiles include the 25- and 75-percentile, otherwise known as quartiles, which are used to calculate the interquartile range (IQR). Percentiles are robust and non-parametric; that is, their calculation is not seriously affected either by outliers, or by frequency distributions that are not Normal (see Section 2.2). Also, in a population of analytical values that is censored by the analytical detection limit (see Section 2.3), percentiles in the range above the detection limit are unaffected by the analyses that fall below it, although the more numerous the latter are, the fewer meaningful percentiles can be calculated. They can be used as direct input in the compilation of cumulative frequency and probability plots and even in their tabulated, unplotted form can facilitate the comparison of populations. They can also be rapidly determined, even for large data sets, using spreadsheet applications (see www.dirtbagger.com/methods.html for a demonstration).

4.2 Methods of Comparing Distributions

4.2.1 Multiple Cumulative Frequency Curves

A non-parametric method called the Kolmogorov-Smirnov test (Massey, 1951) exists to determine whether two populations of values of the same variable are significantly different. The maximum amount, in percentage terms, by which two cumulative-frequency curves differ is measured and compared to a critical value dependent on the sizes of the two populations and the desired confidence level.

4.2.2 Box-and-Whisker Plots

Because of their intuitive impact, Box-and-whisker plots (or simply, "box plots") are preferable for comparing populations, by means of the frequency distributions of single variables. Essentially, the box plot is a graphical display of the summary characteristics of a data set. The plot is based mainly on the median and the lower and upper quartiles, or 25- and 75-percentiles (see Section 4.1.3) of the population, and is, therefore, very resistant both to the effects of non-Normal distributions and to outliers.

Box plots provide a rapid method of categorizing and comparing subsets of data; for example, to illustrate differences of elemental concentration between fresh rocks and their altered or weathered equivalents, or between different regolith or protolith types. In this application, the box plots for a series of previously-defined subsets are plotted side by side with a common concentration scale, and are referred to as "Side-by-Side" or "Categorical" box plots.

In geochemical exploration, interest is often focussed on values at the upper extremity of the range, as these could reflect possible mineralization. On the other hand, if it is necessary to decide whether to partition a data set into previously-defined subsets as a solution to the "apples and oranges" problem (see Section 3), comparison of the overall concentration levels is more important and the central accumulations become the subject of study.

Although box-plots are non-parametric, extremely high values force the main part of the plot to be compressed in order to accommodate them. In such a situation, logarithms of the data values may be used in order to make the data range more manageable; or they may be converted to percentiles (see Section 4.1.3).

Box plots are constructed around a rectangular box that encompasses the central 50% of the data set (Wilkinson, 1999). The lower and upper bounds of the box (referred to as "hinges") are, respectively, the 25th and 75th percentile values (i.e. the box is drawn between the first and third quartile, Q1-Q3). This range of the data is therefore often called the **inter-quartile range** or IQR. The median (that is, the 50th percentile value or second quartile) is represented by a line across the box at the appropriate value. The plotted positions of parameters reflect the symmetry and skewness of the data, being more central, and closer to each other, if the data are symmetrically distributed.

The data values falling outside the hinges are displayed differently, according to how far removed they are. "Whiskers" (continuous lines) extend for a distance representing 1.5 times the IQR, below and above the hinges (Q1 and Q3). The ends of the whiskers are referred to as "inner fences". Any value more extreme than these whisker extents is considered an outlier and is marked with a small asterisk, unless it lies more than 3 IQRs from the box, in either direction (referred to as "outer fences") when it is marked with an circle. There are other conventions in use for the whisker lengths but that described above has proved to be applicable in most situations.

An example of a box-and-whisker plot is shown in Figure 4.4. The procedure whereby the plot was generated (in Systat 9) is demonstrated at www.dirtbagger.com/methods.html. The data plotted are real, and relate to As analyses on lake-sediment samples from Central Baffin Island that were shown earlier in map form in Figure 3.1a. There are no tests of significance for similarity or dissimilarity between boxes designed in this way (although more advanced designs have been proposed which do not have this limitation) but it can be intuitively grasped, in this case, that the As content of lake sediments derived from the Aphebian "Sulphide Schist", mafic volcanics, dolomites and greywackes is conspicuously higher than for those derived from Archean and Aphebian granites, gneisses and migmatites. This assessment led to the data being divided into lithology-based subsets for the purpose of percentile calculation, before being recombined for plotting in Figure 3.1b.

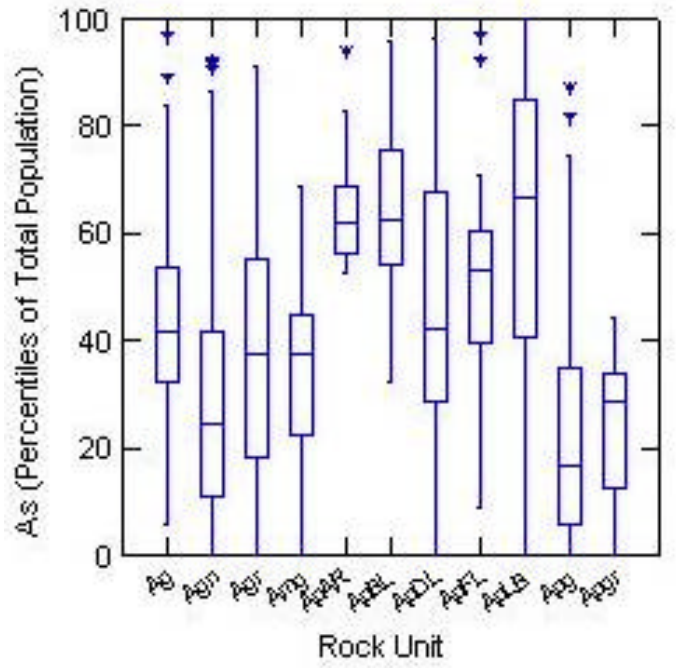

Figure 4.4: Box-and-whisker plot of As analyses in lake sediments derived from various bedrock lithologies, central Baffin Island. Lithologies are as follows: Prefix "**A**" = Archean; **Ag** = granite; **Agn** = quartz monzonite-granodiorite gneiss; **Agr** = weakly-foliated quartz-monzonite-granodiorite; **Amg** = migmatite; prefix "**Ap**" = Aphebian; **ApAR** = "Sulphide Schist"; **ApBL** = mafic volcanics; **ApDL** = quartzite; **ApFL** = dolomite, marble & calcsilicate gneiss; **ApLB** = greywacke; **Apg** = granite; **Apgr** = granodiorite. Data from Friske *et al*. (1999)

5 THE IDENTIFICATION OF GEOCHEMICAL ANOMALIES

5.1 Definition

An anomaly is defined as a "Deviation from law; irregularity" (Penguin English Dictionary). Except in advanced follow-up programs, mineralization, and the geochemical responses associated with it, tend to be statistical rarities. This means that the presence of mineralization exerts little influence on the statistics that are extracted from a large data set. These are more likely to reflect stronger and more widespread influences such as lithological or pedological differences, or co-precipitation phenomena (though as shown above, it is sometimes possible to compensate for some of these by good sample design, or treating subsets of the data separately).

Consequently, the search for anomalies by statistical means has been largely, but not exclusively, concerned with screening out a small proportion of very high (or, under certain circumstances, very low) values. A statistical rarity in geochemical data is not, of course, necessarily related to mineralization.

5.2 Orientation Survey

The principal aim of geochemical exploration remains the detection not just of "high" values of potentially economic elements, but more generally of variations in the chemical composition of naturally-occurring materials related to the presence of a potentially economic commodity. Such responses are best recognized during routine exploration if they can be measured beforehand in material, of the type that will be sampled in the main survey, in the vicinity of a known mineralized occurrence of the type sought, and preferably known to be derived, by natural processes, from it. This is the essential function of an **Orientation Survey** which has no equal as a means of establishing cutoffs or **thresholds** by which the economic significance of geochemical data can be assessed. Geochemists define the threshold as "the upper limit of background variation"; this is, of course, specific to a particular analytical parameter, sample medium and environment. The term "environment" may refer in this context to a region, single property, or homogeneous area within a property. Anything with values in excess of this threshold represents something other than background, and is therefore worthy of follow-up. Further details of the scope and organization of orientation surveys may be found in Appendix A.

5.3 Statistical Methods

In the common, often unavoidable absence of orientation information, or reliable data from the literature, it is necessary to examine the data themselves for unusual behaviour and this leads to the concept of the **Statistical Anomaly**. Various ways of identifying such anomalies, for which the more cumbersome term “areas meriting follow-up” is nevertheless more precise, will be described below.

It cannot be overemphasized that while geologists and geochemists should retain the flexibility to modify their interpretational procedures to suit a particular situation, the procedures, and assumptions underlying them, should be fully documented. It is not sufficient to describe a particular gold value as "anomalous" in a report or recommendation, or on a map; the definition of the term "anomalous", as applied by the geologist or geochemist responsible for the interpretation, must be close at hand.

5.3.1 An Inappropriate Method (Mean plus Two Standard Deviations)

A method of selecting threshold values that is still much used, and much abused, involves calculating the mean (**m**) and standard deviation (**s)** of the data set and applying the classification “anomalous” to those values that exceed the value of (**m + 2s**). The application of this method is no longer justified in most circumstances but it has become so entrenched that its merits and demerits will be discussed in some detail.

The use of the value of the mean plus two standard deviations is not based on some fundamental, mystic relationship to mineralization; it is merely equivalent, in a Normal, unskewed distribution, to the 97.5-percentile (see Section 5.3.2). If the equivalency is satisfied, then the use of this parameter as a threshold is equivalent to selecting the uppermost 2.5 percent of any sample population for follow-up work (that no more, and no less, than this proportion of any data set, however acquired, will justify follow up is in itself an assumption that requires further examination and this will be dealt with in Section 5.4). This indirect method of deriving the 97.5-percentile was developed before computers were widely available and summary statistics for large populations of analytical values were often derived, element by element, by grouping the total data range into equal intervals and marking a tick in each interval for every sample whose value fell within that range. Summary statistics were then calculated by multiplying the midpoint of each interval by the number of samples that fell within that interval and creating weighted averages.

5.3.2 A Better Method (The Use of Percentiles)

With the wide availability of computers, many of which have MS-Excel installed with its useful “Rank and Percentile” utility, percentiles can be readily calculated and the 97.5-percentile can be estimated reliably without resorting either to time-consuming or imprecise methods like the one described above. Furthermore, the relationship between the mean plus 2 standard deviations and the corresponding percentile varies depending on the skewness of the data distribution; this has been demonstrated by a random simulation whose results are shown in Table 5.1. The definition of the term “skewness” will not be given in mathematical terms; however, the greater the positive magnitude of the skewness, the longer the tail of the histogram extends to the right, and the more of the data are concentrated in the first few histogram intervals (see Figure 4.1b).

Table 5.1: Relation between skewness and percentile equivalent of the mean plus two standard deviations

Skewness	**Percentile Corresponding to (m + 2s)**
0	97.5%
0.5	96.3%
1	95.5%
2	95.0%
3	95.1%
4	95.5%
5	95.9%
6	96.2%
7 (typical lognormal)	96.6%
9	97.5%
10	97.9%
20 (typical log-lognormal)	98.8%

For a skewness of zero (the familiar bell-shaped distribution), the corresponding percentile is indeed 97.5; as skewness increases it decreases rather sharply to a minimum of slightly less than 95% for a skewness of about 2; in other words, in moderately skewed distributions, use of this method of defining

anomalous samples will result in twice as many samples being identified thus, as for unskewed distributions. Then it rises again, and for highly skewed distributions (skewness of 8-9, greater than for a typical lognormal disribution) it is once again 97.5%. The percentile continues to rise but levels off at about 99%.

5.4 Appropriateness of Using A Single Threshold

The foregoing was intended to demonstrate the unpredictable nature of statistics based on the mean and standard deviation of typical geochemical populations, and the preferability of using percentiles to derive cutoffs for identifying follow-up targets. However, even the selection of those "magic" values that exceed the 97.5-percentile, more rigorously estimated, is merely a way of screening out most of the values in the data set, on the originally well-founded assumption that since mineralization is a statistical rarity, its geochemical manifestations will be equally rare.

However, all distributions have a 97.5 percentile. The total scores of five dice, thrown more than 60,000 times in a computer simulation, approximate a Normal distribution. The 97.5-percentile of this distribution is 25 (and, incidentally, the mean plus two standard deviations is 25.1). However, there is no reason to believe that the laws of chance governing the 5-dice score were any different for instances when scores exceeded this cutoff, from instances when scores were lower. The same is true for most geochemical data: every distribution of analytical values will have a 97.5-percentile, even though the proportion of values in a population that are truly worthy of follow-up can be zero, or much greater than 2.5%. The first instance would arise if the tract of land chosen for the survey were barren; the second, if the area covered by the survey constituted a "target" selected for follow-up after a regional survey, or if the company's area selection was particularly well-informed.

To extend the dice analogy a little further, however, the incidence of two scores exceeding the 97.5 percentile occurring next to one another is much more unusual; on only 16 occasions (0.024% of the total) did two successive throws both return scores in excess of 25. In geochemical terms, this means that for two spatially-adjacent samples to exceed the 97.5-percentile is so unusual purely by chance that when that happens, it is always worthy of further investigation. In practice, of course, the assessment as to whether or not to follow up a group of samples that exceed the "threshold" should always be made after examining their relative spatial dispositions in map form.

Even when the selected threshold is based on more rigorous data, for example that derived from a *bona fide* orientation survey, its use should be implemented with caution. With knowledge of the behaviour of an element in the natural environment that is usually extremely limited, it is reasonable to question the wisdom of expending a large sum of money to follow up samples returning a certain analytical value, and consigning permanently to "outer darkness" samples returning another value, when the compositions of the two values may differ by only a few parts per million but are separated by the threshold. A group of spatially-associated samples whose values fall just below the threshold are more worthy of follow-up than a spatially-isolated sample whose value falls just above it (see Figure 5.1).

5.5 A Suggested Approach

- For each element in turn, percentiles are calculated and **two** cut-points are selected. The first corresponds to the 97.5-percentile and is equivalent to the "threshold" discussed above. The second corresponds to the 90-percentile. Values that exceed the 97.5%-percentile can, if necessary, be referred to as "probably anomalous", while those that fall between the 90-percentile and 97.5-percentile are "possibly anomalous" or perhaps "elevated". Significant spatial groupings of samples in this second category are almost as important as isolated samples in the higher category.

- The data are plotted in the form of coloured circles; in this case, values exceeding the 97.5-percentile are coloured red; those that fall between the 90-percentile and 97.5%-percentile are coloured green; and the remainder are coloured grey. The choice of colours is arbitrary, but it should be applied consistently.

- The resulting plot is examined for spatially-grouped red circles (First Priority Anomalies), and isolated red circles or grouped green circles (Second Priority Anomalies). Figure 5.1 shows the results of applying this procedure to reconnaissance data from central Yukon Territory, and there are a number of concentrations of green circles that might have been overlooked if the "threshold" were applied too rigorously. (The procedure is demonstrated on the same data in step-by-step form, using Surfer, at www.dirtbagger.com/methods.html).

- The anomalies are assessed in the light of known geology, if any, other available information such as ground or airborne geophysics, remotely sensed data etc., and co-association with other geochemical anomalies (e.g. Au and As together might indicate epithermal gold mineralization; Zn and Pb might indicate Sedex or Mississippi-Valley type mineralization, and Cu and Zn might indicate VHMS mineralization)

- The sites of the anomalies are revisited for follow-up.

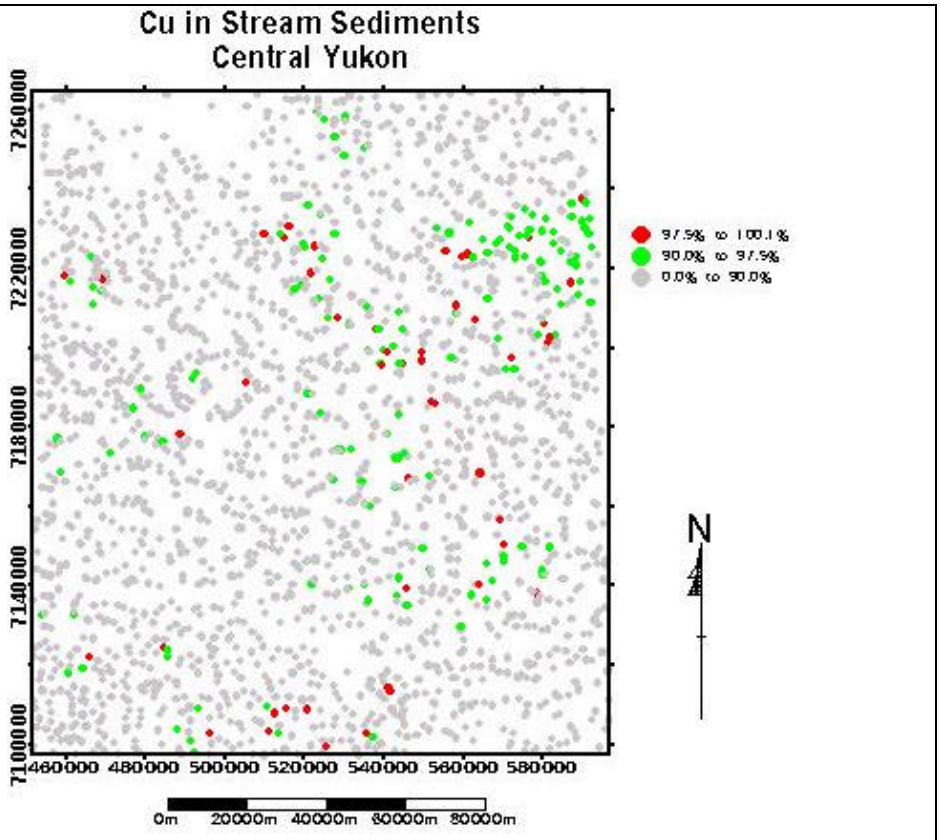

Figure 5.1: Cu Stream Sediment Anomaly Map, Central Yukon

Red circles indicate values exceeding 97.5-percentile (140 ppm Cu); green circles indicate values exceeding 90-percentile (84 ppm Cu). Data from GSC Open File 2175

5.6 The Use of Probability Plots to Partition Data and Clarify Anomaly Identification

The derivation of thresholds for the identification of anomalies involves the calculation of percentiles on the entire population. This serves a useful purpose as a first-pass screening of the data but even if account has been taken of known subdivisions in the data like variable source lithology (see Section 3), it may be an oversimplification, inasmuch as hitherto-unknown subpopulations of the data may be present. A method described by Sinclair (1974) involves the use of log-probability plots to identify and characterize subpopulations of polymodal univariate distributions. As was described in Section 4.1.2, a unimodal log-normal distribution plots as a straight line on such plots. The program PROBPLOT (Stanley, 1987) allows this task to be implemented interactively.

The frequency distribution for Cr data from a soil survey in Jamaica is presented in histogram form in Figure 5.2. It is clear that the distribution is polymodal, and that separate criteria for establishing degree of anomaly (as described in Section 5.4) need to be applied to each subpopulation – or, one subpopulation may possibly be related to mineralization in its entirety. It is difficult, however, to discern from the histogram where one population ends and the next begins. What are the summary statistics of the constituent subpopulations?

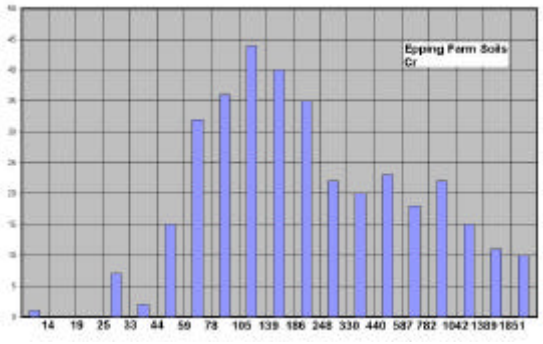

Figure 5.2: Histogram of Cr values in soils, Epping Farm, Jamaica

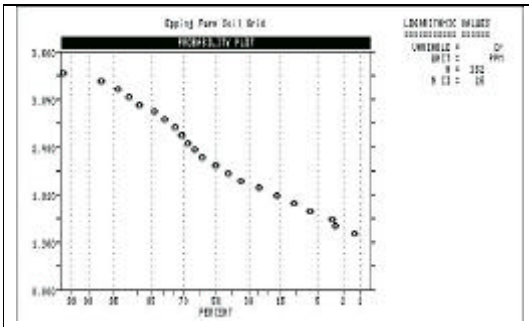

Figure 5.3a: Probability plot of Cr values in soil, Epping Farm, Jamaica. Cr (y-) axis is scaled logarithmically. Plot created in PROBPLOT (Stanley, 1987).

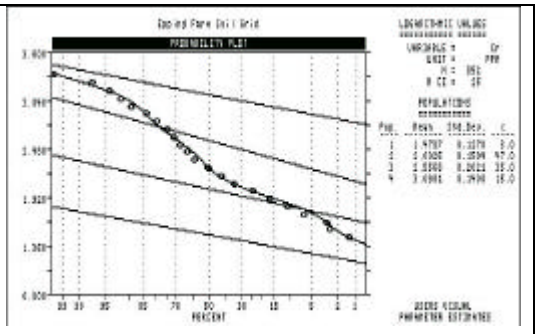

Figure 5.3b: First four-subpopulation estimate. The curved line represents the recombination of the four theoretical populations represented by the straight lines; its correspondence to the observed distribution (represented by the individual points) can be observed and improved interactively.

The first plot created in this program, which is a simple log-probability plot of the data as a whole, is shown in Figure 5.3a.

The percentage of the total entries falling into each of the sub-populations is estimated visually at the outset from the inflection points of the transition curves. These are estimated to be at the 3%, 50% and 85% marks -- in other words, out of the 352 entries in the data set, the subpopulations comprise 10, 165, 123 and 54 samples. The cumulative percentages of the total population are now recalculated for each of the sub-populations, with the assumption being made that the influence of one population on the other ones is negligible. Probability plots, represented by the four straight lines in Figure 5.4b., are generated for each subpopulation. For each of the populations represented by these lines, the medians can be read off against the 50% line on the probability axis. To estimate the standard deviation, use is made of the properties of the Normal distribution that the median and mean are the same, and that 68% of the population (34%, on each side of the mean) is enclosed by the compositional range delineated by the mean plus one standard deviation, on one hand, and the mean minus one standard deviation, on the other.

In this case, the four population medians are estimated as 1253, 360, 101 and 30 ppm Cr. The recombination of these four modelled populations results in the curve that approximates the line joining the original points, but the correspondence is not perfect. At this point a better fit can be sought by choosing new percentiles for splitting, or PROBPLOT's own optimization procedure can be used to seek a better fit. This approach has been taken here with the result shown in Figure 5.3c. The medians are now revised to 1004, 382, 115 and 35 ppm Cr. As defined by adding two standard deviations to the median (which may or may not be appropriate in the identification of samples deserving follow-up), the "thresholds" for the same sub-populations are 2417, 1480, 312 and 80 ppm Cr. The classifications of the samples are plotted in map form in Figure 5.3d.

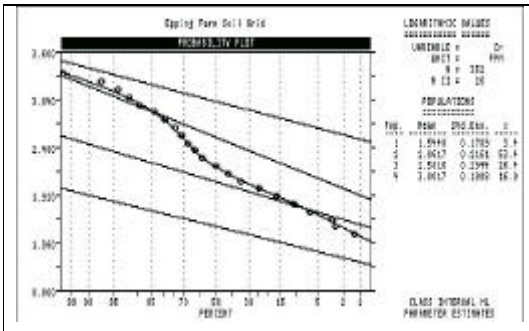

Figure 5.3c: Results of revision of the first estimate by four optimisation steps to maximize correspondence between points and curve.

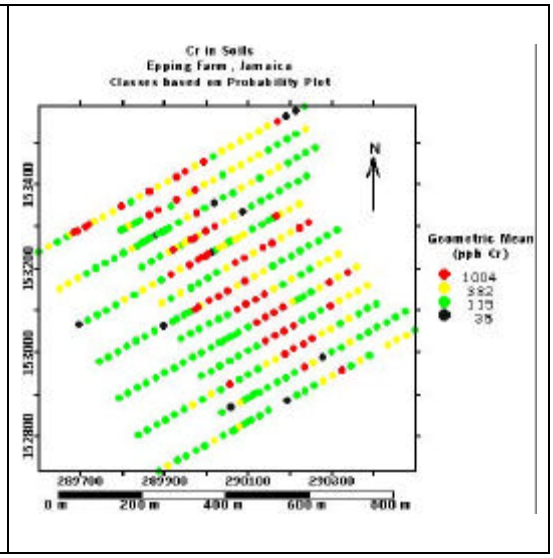

Figure 5.3d (opposite): Areal distribution of populations defined by population-splitting exercise. Geometric (logarithmic) means of the populations are 1004, 382, 115 and 35 ppm Cr.

This method is often successful in identifying sub-populations that display elevated values with respect to the data set as a whole, but it should not be assumed that these sub-populations, or those identified by any other statistical method, are necessarily associated with economic mineralization. In the example shown here, the subpopulation displaying the highest Cr values is probably associated with an ultramafic intrusion or flow. However, selection of the uppermost 2.5% of this subpopulation might draw attention to accumulations of chromite mineralization.

6 BIVARIATE CORRELATIONS BETWEEN ELEMENTS

6.1 X-Y Plots

In an X-Y plot, or scatterplot, the values of one variable are plotted against those of another, for the same group of samples. Although the assumption of a Normal distribution is not necessary for an X-Y plot to be effective, log-transformation may be of assistance in scaling the data so that the accommodation within the plot of extremely high values does not result in the compressing of most samples to the point where they are too crowded to interpret.

Such plots serve as a means of visually estimating the relationship between two variables, and may highlight clusters within the data, and multivariate outliers. The latter purpose is best achieved where the data are displayed in the form of a scatterplot matrix or “draughtsman’s plot” where X-Y plots for every variable against every other variable are displayed simultaneously, arranged in triangular matrix form like a correlation matrix (see next section).

6.2 Correlation Coefficients

The correlation coefficient, if correctly applied, is a useful quantification of the degree of rectilinear interdependence between two variables -- in other words, how closely the X-Y plot of the two variables can be fitted by a straight line. Although various forms of the correlation coefficient have been devised, the most commonly used, termed the Pearson Correlation Coefficient and abbreviated to **r** (the uppercase version is also used), is defined as the covariance of the two variables divided by the product of their standard deviations. R-mode factor analysis (see Section 8.3) is so called because it is based on the correlation coefficients between a large number of variables.

Correlation coefficients, which are dimensionless, range between +1.0, indicating a perfect positive linear relationship, to -1.0, indicating a perfect negative relationship. Figure 6.1 shows a series of X-Y plots of data with different correlation coefficients. In the last plot, the perfect circular relationship of the points nevertheless has a correlation coefficient of zero, because this parameter measures the degree of **rectilinear** interdependence of two variables.

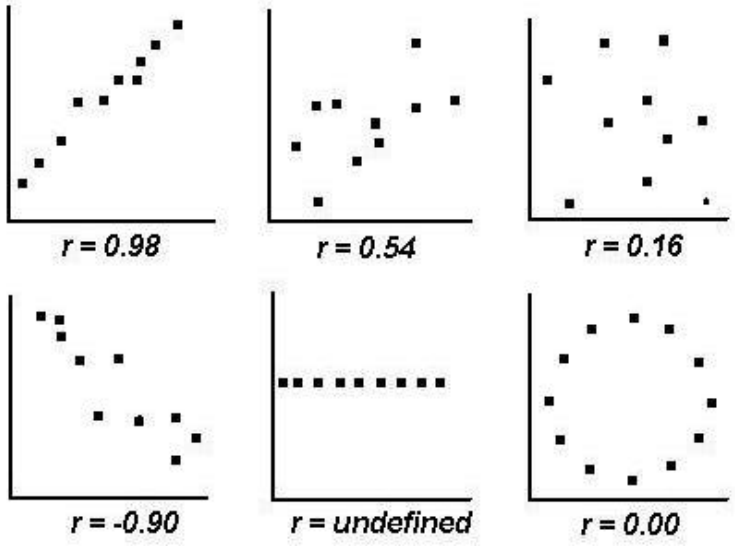

Figure 6.1: Scatterplots of two variables, showing different correlations (from Davis, 1973)

Perfect correlation, whether positive or negative, is rarely achieved from real data; indeed, it suggests a calculation error. It is, therefore, necessary to quantify the correlation coefficient's significance, which is dependent on the size of the population from which it was calculated. Table 6.1 shows significance levels at 90%, 99% and 99.9% confidence levels for the correlation coefficient in populations of various sizes. It is noteworthy that in sample populations of more than 100 cases, correlation coefficients of considerably less than 0.5 are significant at the highest level.

Table 6.1: Critical values of the correlation coefficient at various confidence levels and population sizes

		Significance Level		
		90%	**99%**	**99.9%**
Degrees of Freedom (N-2)	**5**	0.649	0.889	0.959
	10	0.436	0.699	0.826
	20	0.298	0.512	0.642
	40	0.207	0.366	0.473
	60	0.168	0.299	0.391
	120	0.118	0.212	0.279

Like other summary statistics, the correlation coefficient is susceptible to abnormalities in the nature of the distribution which must be rectified before any important conclusions are drawn from the data, or if the correlation coefficient is used as input to other statistical methods like factor analysis (see Section 8.3). Examples of this are shown in Figures 6.2a and 6.2b.

In the first plot, the correlation coefficient of a group of essentially random bivariate data is only 0.015, but the presence of a single outlier causes the correlation coefficient to rise to a misleadingly high (and misleadingly significant) value of 0.85 in the second plot..

The intercorrelations in a data set with k variables are generally summarized in the form of a triangular **correlation matrix**, which consists of $k*(k-1)/2$ entries. The understanding of such a matrix is improved by converting the magnitude of the correlation coefficients to a series of coloured or sized symbols. An example is given in the section dealing with factor analysis (Section 8.3).

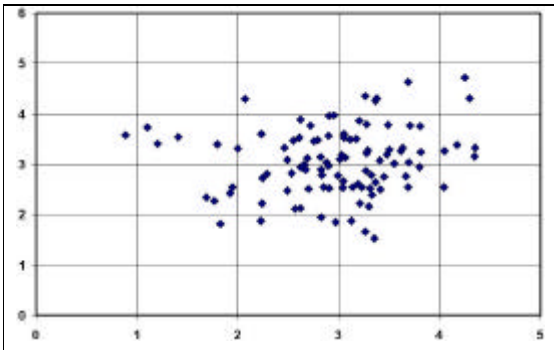

Figure 6.2a: X-Y plot of essentially random data with correlation coefficient of 0.15

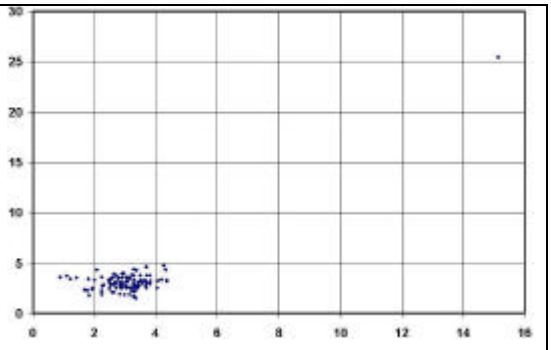

Figure 6.2b: X-Y plot of data from Figure 6.2b, with a single outlier. Correlation coefficient increases to 0.85.

7 MULTIELEMENT INDICES

7.1 Advantages

Methods exist for dealing with multielement data that do not involve multivariate statistics in the strict sense of the term. The calculation of multielement indices is an example of how element associations, perceived by virtue of previous work or experience to be diagnostic of specific types of mineralization or key lithologies, can be applied to optimize the response to such features, merely by summing their values together in an unweighted fashion. Under certain circumstances it may be felt that certain elements are deserving of greater weighting in such an index because of their greater importance as pathfinders for the deposit type sought. Gold and arsenic, for example, might be accorded greater weight than copper or lead in an index designed to detect manifestations of lode-gold mineralization.

A spreadsheet application enables these calculations to be readily performed on large data sets. A special case of the "apples and oranges" problem, described above, must be dealt with because the raw values of the input variables vary greatly in range; even if the weightings of the two elements are equal, copper values ranging between 10 and 300 ppm will exert greater influence on the value of an index than silver values ranging between 0.2 and 3 ppm. In order to compensate for this effect the values of each variable can be normalized or converted to percentiles of the overall population, or of subpopulations based on observable criteria such as mapped geology or regolith type.

The use of indices, along with some of their drawbacks, is demonstrated with two examples; one with regional stream-sediment data from Jamaica (Bondar Clegg & Co., 1988; Siriunas, 1992) and one with deposit-scale lithogeochemical data from the Canadian Shield (McConnell, 1976; Amor, 1983).

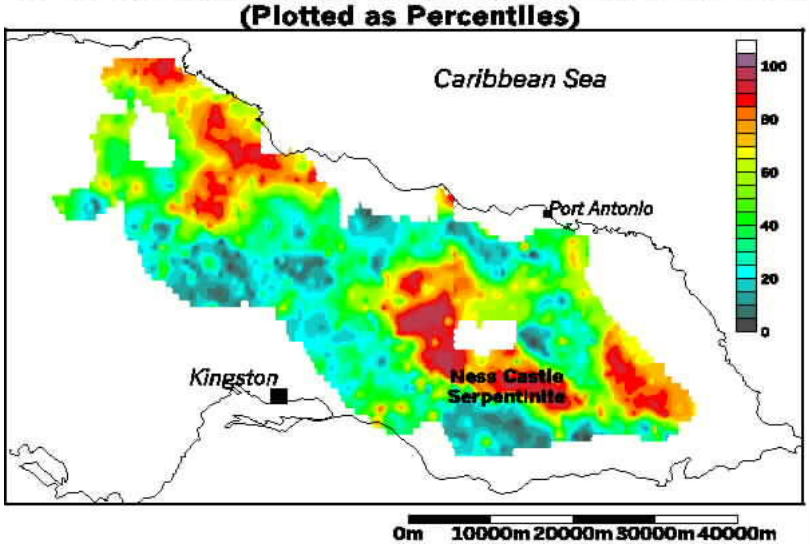

Figure 7.1a: Distribution of "mafic-ultramafic index" in stream sediments, Eastern Jamaica

7.2 Application to Regional Stream Geochemical Data

For the Jamaican example, Figures 7.1a and 7.1b are colour-contour image plots of two indices calculated from multielement stream sediment data over the Blue Mountains and adjoining areas of eastern Jamaica. The first is a "mafic-ultramafic index" comprising summed, unweighted percentiles of Cr, Ni and Co, and the second is a "porphyry-copper" index comprising Cu, Mo, Te and Au, combined in a similar fashion. For the purpose of generating colour-contour image plots at the same scale, the indices were converted to percentiles a second time.

Mapped ultramafic rocks are rather rare in Jamaica but the mafic-ultramafic index clearly highlights a complex known as the Ness Castle Serpentinite, and there are suggestions of other ultramafic occurrences, perhaps hitherto-undiscovered, in the southeast and northwest. The porphyry copper index is particularly high around the margins of the Above Rocks Granodiorite in the southwest. In the northeast, where extensive alteration zones (and minor Cu-Au mineralization) are known and the presence of felsic intrusive rocks at shallow depth is suspected (Amor and Elliston, 1993), the index indicates a local maximum only.

Other indices calculated similarly for the Geochemical Atlas of Jamaica (and not shown here) were combinations of Pb+Zn+Ba+Cu (Base Metal Index), Au+As+Sb+Hg+Ag (Epithermal Precious Metal Index), Pb+Zn+Ba+Cu+Cd+Ag (Base Metal Index II), U+Th+Hf ("UHT" Index), Eu+Tb+Yb (Rare Earth Index), As+Sb+Hg+Te+Se (Epithermal Sulphide Index) and Ta+W (Ta-W Index).

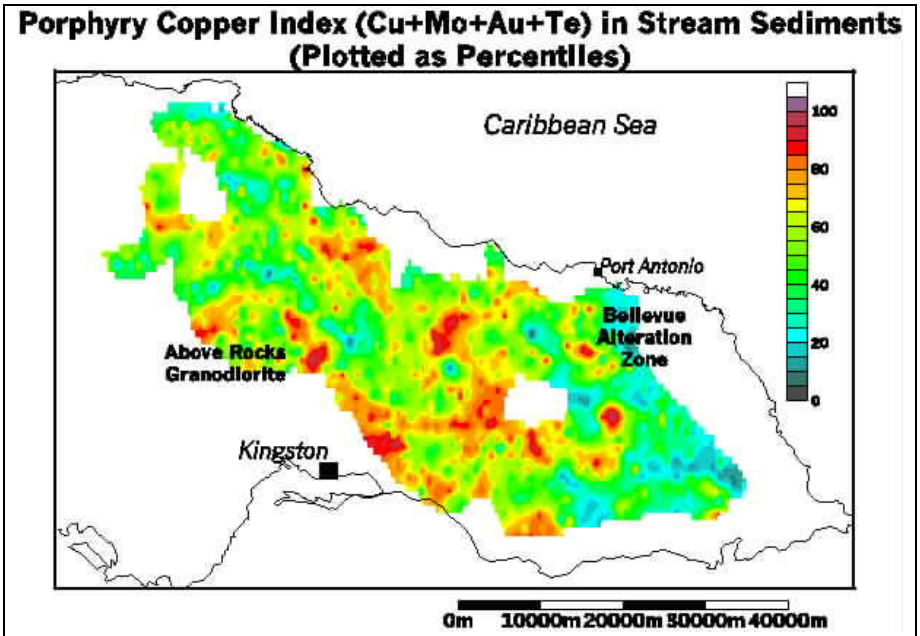

Figure 7.1b: Distribution of "porphyry copper index" in stream sediments, Eastern Jamaica

7.3 Application to Deposit-Scale Rock Geochemical Data

In an investigation of the chemical changes taking place in the wallrocks of Archean volcanic-exhalative massive sulphide deposits, McConnell (1976) concluded that in general, alteration involved the addition of Fe_2O_3 and MgO, and removal of CaO and Na_2O, and that these changes were highlighted if the effects of igneous differentiation were compensated by multiple regression (see Section 8.2.2). Furthermore, the individual propensities of each of these variables to indicate the presence of alteration could be compounded if they were combined in a single index, termed the Standardized Net Residual (SNR), defined as $Fe_2O_{3R} + MgO_R - CaO_R - Na_2O_R$ (the subscript $_R$ denoting the regression residual of the oxide in question). This was proposed as a useful parameter for exploration for similar mineralization. However, in a subsequent re-examination of the same data Amor and Nichol (1983) showed that at the seven base-metal deposits studied, two distinct alteration styles could be observed, which shared depletion in Na_2O but in which Fe_2O_3 and MgO behaved oppositely; both were enriched in one alteration type, and both were depleted in the other. At a deposit where the latter predominated, such as South Bay, incorporation of these two oxides into the index actually reduced the effectiveness of Na_2O alone in delineating the alteration zone. The comparison of the alteration zone defined by the SNR, and by a discriminant score specific to this type of alteration, is shown later, in Figures 8.10a and 8.10b.

This illustrates one of the drawbacks of user-defined indices: their effectiveness is decreased when the constituent elements behave unusually. Furthermore, while the element associations are well enough documented in particular mineral-deposit types, the associations in commonly-sampled media may be modified by surface processes, especially in areas of deep weathering, that are not so well understood.

In such situations it may be preferable to derive the indices empirically based on how the data respond to the presence of known mineralization, and interpret their significance *a posteriori*; this is the primary role of multivariate methods in exploration geochemistry, as will be described below.

Furthermore, while indices are useful in many situations, they are an example of an alluring method that can lead to meaningless or misleading results, if attention is not paid to geological and geochemical realities. In the case of the "Mafic-Ultramafic Index" derived for the Jamaican stream sediment analyses, there is no "threshold" value corresponding to the definite presence of mafic or ultramafic rocks, and local maxima would be expected to be present even where the source rocks consisted of arenaceous sediments with no igneous rocks in the study area.

8 MULTIVARIATE STATISTICAL METHODS

8.1 General Statement

Computer processing is mandatory in the application of most multivariate statistical methods as it would be prohibitively difficult and time-consuming to undertake it manually. These techniques can be usefully applied without detailed understanding of the underlying mathematics, although some understanding of how they work (and the circumstances under which they do not) is advantageous. The key to the understanding of these apparently arcane methods lies in the graphical demonstration of a 2D (bivariate) situation and its intuitive extension into higher dimensions ("hyperspace").

When deriving any kind of summary statistics, multivariate or otherwise, from a large data set it is important to decide whether the feature sought constitutes a statistical rarity in the sampled population. In a regional drainage survey, for example, the major controls on each sample's composition are likely to consist of gross lithological and surficial or environmental agencies. The predominant element associations (factors), sample associations (clusters) or inter-element and intersample relationships of other kinds, like regression equations, are likely to reflect these controls, and unlikely to reveal much about mineralization, if its presence is manifested in only a few samples. On the other hand, in the follow-up survey of a previously-defined anomaly, the presence or proximity of mineralization is more likely to exert a discernible influence on the data as a whole.

This does not negate the usefulness of "global" statistics in geochemical exploration, since it is in the (often subtle) deviations from these large-scale features that indications of mineralization are most likely to be found. Furthermore, where the geology of a concession is not well known, the composition of the sample medium (especially if multielement data are available) may provide an important aid to geological mapping and the identification of favourable geological environments for mineralization.

8.2 Regression Analysis

Whereas the correlation coefficient is a measure of the strength of the relationship between two variables, regression analysis provides a means of expressing its nature in quantitative terms. In the case of simple linear regression (not a multivariate method), a set of bivariate data, expressed graphically as an X-Y plot, is fitted with a straight line, that may or may not pass through the origin. This line represents the best estimate of the relationship between what is termed the **dependent** variable (which is normally plotted on the y-axis) and the **independent** variable (x-axis) though no cause-and-effect relationship need be implied. Regression modules are a feature of all statistical software packages and many spreadsheets.

8.2.1 Simple Regression

An example of the application of simple regression to regional geochemical data comes from the Bathurst district of New Brunswick (I. Nichol, personal communication, 1999; see also Section 8.3.7.1) where there is potential for various types of zinc mineralization but where at least two other agencies are believed to contribute zinc to the stream sediment: co-precipitation with iron and manganese oxides, and mafic rocks in bedrock. Since nickel is also strongly influenced by these two agencies, but is not enriched in mineralization of the type sought, regression can be used in an attempt to separate the component of the zinc that is attributable to mineralization, from that which is not. Figures 8.1a and 8.1b show the raw Zn values, and regression **residuals** (that is, the difference between predicted and measured values) for Zn against Ni. In general the patterns indicated by the two parameters do not differ greatly, indicating that in this case at least, that the dispersion of zinc is more complex than that postulated by this model.

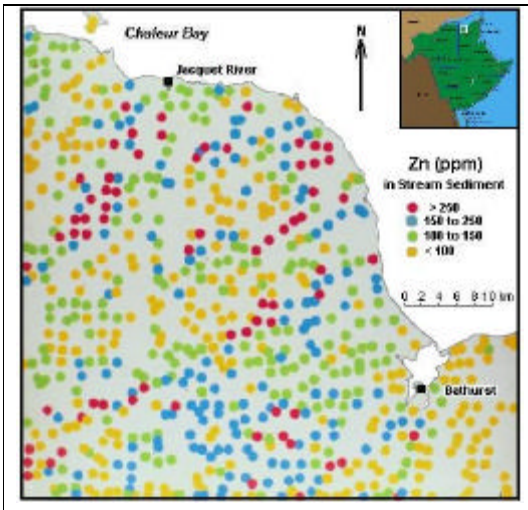

Figure 8.1a: Zinc in stream sediments, Bathurst District, New Brunswick

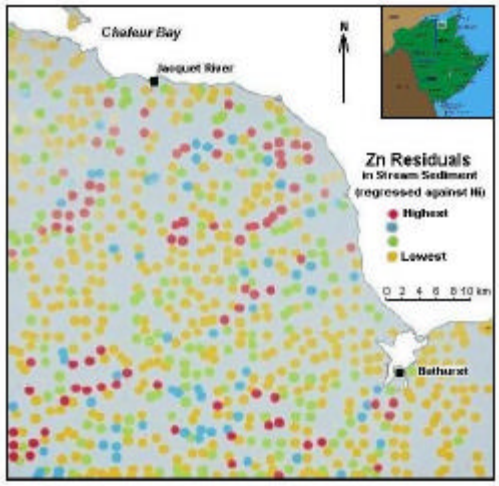

Figure 8.1b: Zinc residuals (regressed against nickel) from same samples

8.2.2 Polynomial and Multiple Regression

Two developments of simple regression as described above are **polynomial regression**, which also operates in two dimensions only and involves the fitting of a curve, rather than a straight line, to the scatterplot of dependent vs. independent variable, and **multiple regression** which involves the admission of more than one independent variable and is analogous to fitting a plane or curved surface, rather than a line, to a set of points in three (or more) dimensions.

Polynomial regression is illustrated in Figure 8.2, where a curve is fitted to a plot of Fe_2O_3 values (dependent variable) and SiO_2 values (independent variable) for a suite of samples of Archean metavolcanic rock. The equation of the curve is a 4^{th}-order polynomial of SiO_2.

In the most commonly-applied form of regression, the "fitting" procedure seeks a line or surface that minimizes the total deviation of the predicted values from the observed values. As a negative deviation is of equal importance to a positive one, the deviations are squared to remove the effects of sign (plus or minus) and the method is known as "least-squares".

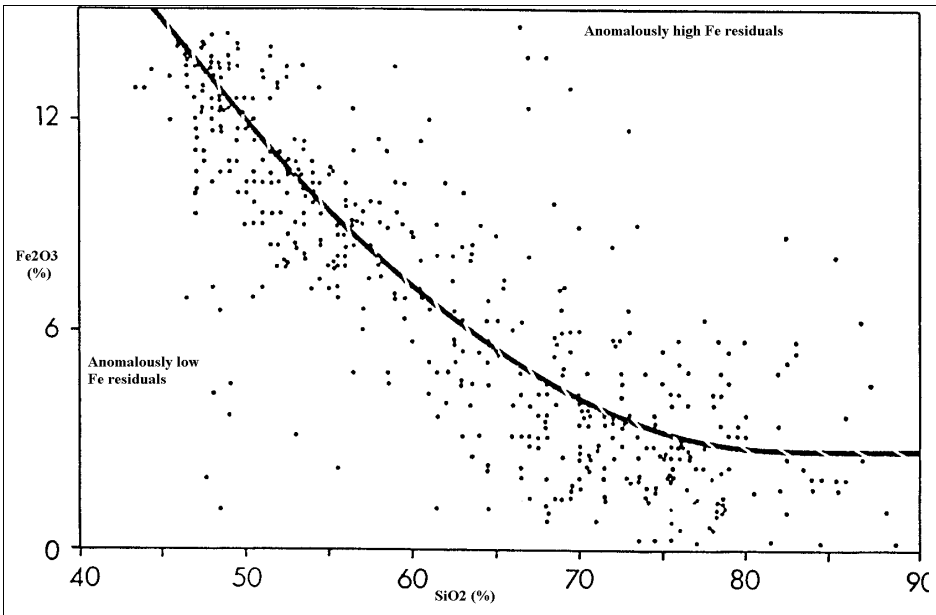

Figure 8.2: Graphical demonstration of curve-fitting, and residuals, in regression analysis. Iron oxide is regressed against silica. Data are for ostensibly unaltered Archean volcanic and volcaniclastic rocks, N.W. Ontario. From Lavin (1976).

The use of regression analysis to remove the gross effects of igneous differentiation, so that more subtle features attributable to alteration or mineralization may be highlighted in terms of regression residuals (as demonstrated in Figure 8.2), has been applied with encouraging results in studies of wallrock alteration in volcanic-exhalative massive sulphide deposits (Lavin, 1976; McConnell, 1976; Sopuck *et al*., 1980; Amor and Nichol, 1983). It is, however, dependent on the assumption that while the dependent variables (e.g. Fe_2O_3, MgO, CaO and Na_2O) are mobile during the mineralization and alteration processes, the independent variables (e.g. SiO_2, TiO_2) are not: an assumption that has been challenged.

In reconnaissance stream-sediment geochemistry in Pennsylvania, Rose *et al*. (1970) successfully used regression analysis to filter out geological effects so that the zinc response to mineralization was highlighted (Figures 8.3a and 8.3b). The anomalous values of this element in their unadjusted form tend to highlight inliers of unmineralized Precambrian gneiss, granite, gabbro and serpentinite, rather than potentially economic mineralization. However, regression residuals highlight at least one area, (the Bamford District, northwest of Lancaster), that hosts several lead-zinc deposits that were completely missed by the simpler procedure. Significantly, this is not an area with the strongest response, and many other targets are apparent.

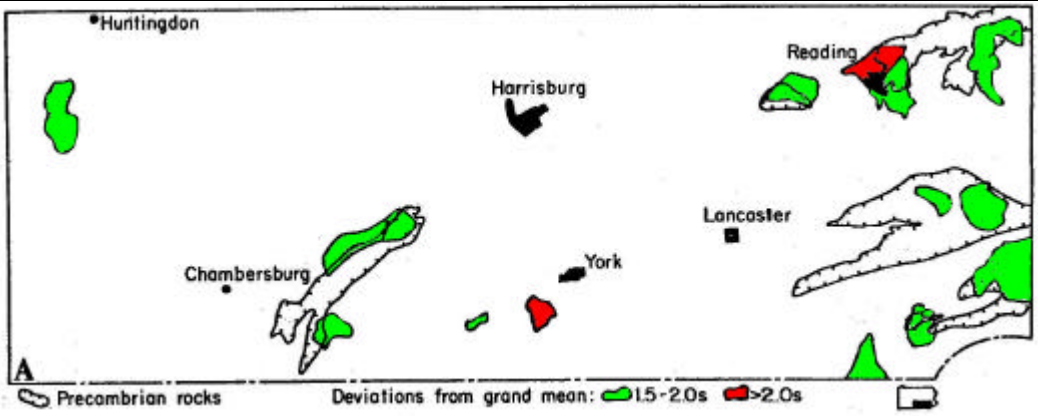

Figure 8.3a: Zinc stream anomalies in a region of Pennsylvania. Data are unadjusted for lithological variations and other factors, and zinc highs generally indicate only Precambrian inliers.

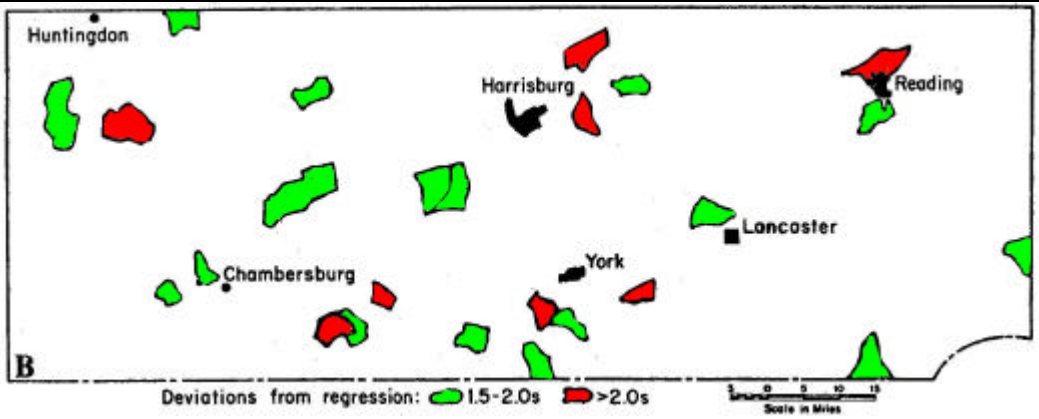

Figure 8.3b: Zinc stream anomalies after adjusting of data for lithological variations and other factors, highlighting areas of mineralization. From Rose *et al.* (1970).

8.3 R-Mode Factor Analysis

8.3.1 Process and Response

The analytical data in a typical multielement set of soil analyses, for example, consist of analyses for up to 30 different chemical variables, but it is reasonable to assume that these elements are not emplaced in the soils by 30 different processes, each one unique to a single element response. Furthermore, the amount of a particular element in a sample is unlikely to be the result of only one process acting on the sample material. The strength of the correlations between certain elements in

most naturally-occurring media bears witness to this. Typically, for example, one might expect the iron, vanadium, nickel, chromium and zinc content of a lateritic soil to be the result the interaction of at least two, but less than five agencies: the composition of the soil's protolith, and the leaching and reconcentration of the elements by weathering processes, are likely hypotheses.

"R-mode" factor analysis (because it is based on **r**, the correlation coefficient and deals with relationships between variables; "Q-mode" factor analysis, which is little used in geochemistry, deals with relationships between samples) is a general term given to a variety of related techniques which seek to identify a limited number of controls on a much greater number of observational variables. These controls are modelled in the form of linear combinations of those variables which are known as "factors". In geochemistry, it is reasonable to suppose that such factors will be more closely related to the processes that have acted on the naturally-occurring medium in question, than are the individual variables themselves, although this does not necessarily follow.

A useful bi-product of R-mode factor analysis is that it often provides a means of concisely describing and summarizing the behaviour of a large number of elements in a geochemical data set.

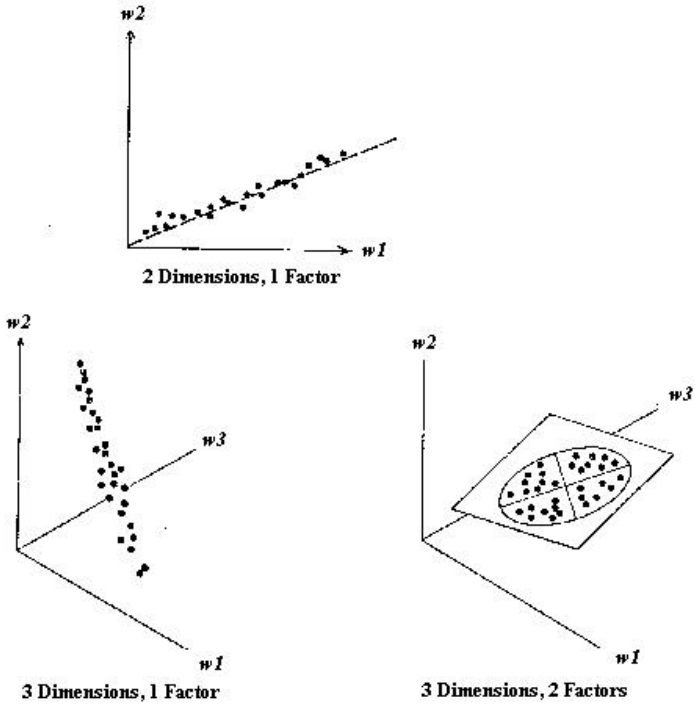

Figure 8.4: Graphical demonstration of factor analysis. In the first plot, points plotted in 2 dimensions fall on a single line; in the second, points plotted in 3 dimensions fall on a single line; in the third, points plotted in three dimensions fall within a single plane. Adapted from Koch and Link (1970).

8.3.2 Graphical Depiction

In mathematical terms, factor analysis involves the extraction of the eigen-vectors of the correlation matrix, although for non-mathematicians the process is more readily demonstrable graphically (Figure 8.4).

In the simplest, two-dimensional case, a "straight-line" graphical relationship between two elements in a naturally-occurring medium implies that the compositions of both elements are essentially controlled by the same process (and not that the content of one of the two elements controls that of the other). Of course, the use of the term "straight-line" normally means not literally that, but at best that the data points concentrate in the form of a rather elongate, cigar-shaped ellipse whose long axis represents the straight line, and whose short axis represents some minor component of random variation.

If this analogy is extended into space of greater dimension than 2, the multielement analyses of a group of samples can be conceptualized as a hyperellipsoidal cluster of points in multidimensional space. Factor analysis and the closely-related technique of principal-components analysis (PCA) can be seen as involving the extraction of the principal axes of this hyperellipsoid. These are expressed as linear

combinations of the input variables; they are mutually at right angles to one another and are, therefore, uncorrelated and mutually independent.

The dimensionality of the modelled hyperellipsoid is limited by the number of dimensions in which the data are “plotted”, and in principal-components analysis, the number of eigen-vectors (or principal axes) extracted is always the same as the number of input variables. If a certain amount of variability is ascribed to random agencies, most of the variance and covariance can often be expressed by a much smaller number of such axes and this is the role of factor analysis; the somewhat limited number of axes of the hyperellipsoid (compared to the dimensionality of the data) are known as **factors**.

The strength of an individual input variable in a factor is referred to as its **loading**. It is important to note that the loadings are determined internally by the factor-analysis process and are not set by the user, although the user determines which input variables will be used, and how many factors will be extracted.

8.3.3 Factor Rotation

Some form of factor rotation is generally applied in factor analysis. This has the effect of increasing the loadings on the more important input variables in a factor, and reducing the loadings on the less important variables. The interpretation of the factors, in terms of familiar natural geochemical procedures, is thus facilitated. The most commonly applied method of factor rotation is referred to as Varimax, which retains the mutual orthogonality of the factors and appears to be suitable for most geochemical applications.

8.3.4 Number of Factors to Extract

The number of factors to extract must be determined before the commencement of the computations. The "correct" number of factors is not determined automatically (although some software packages may have default settings that may make it appear that this is what is taking place) and there are no universally-agreed criteria for determining it. Amongst the methods that can be applied are the following:

1. Continuing to extract factors until the eigen-value of a new factor drops below 1.0 (meaning that the new factor explains less of the total variance than a single input variable)

2. Observing the plot of Factor Number vs. Eigen-value and looking for a flattening out of the line. This indicates a sudden reduction in the capacity of the factor-extraction process to account for the variance of the data.

3. Observing the loadings on the factors and only extracting those factors that can be recognized as representing familiar geologic and environmental processes, or which have been recognized in data sets from similar geological and physical environments

4. Ceasing the factor-extraction when factors begin to appear in which only one input variable is heavily loaded.

5. Randomly splitting the data set into two equal parts, applying factor analysis to both, and discontinuing the extraction of factors when factor characteristics differ significantly between the two splits. This method is known as "Cross-Validation" and it has an intuitive, logical appeal. It may appear tedious but methods exist for expediting the process and are described at www.dirtbagger.com/methods.html.

In practice, an approach which has proven satisfactory is the application of Method 5, with confirmation by Method 4. A particularly lucid description of the principles and application of factor analysis in geology is given by Klovan (1968).

8.3.5 Factor Scores

For each factor, a series of factor scores can be calculated for every sample in the data set, indicating the strength of that factor, and by implication the strength of the influence of the process that it represents, on the sample. The coefficient for each input variable, in each factor, is related but not identical to that variable’s loading because the component of the original value, that is not explainable by the chosen factor model, is filtered out (Davis, 1973, p. 515.).

Factor scores can be plotted and contoured like any continuous geochemical variable and it is often in the examination of their areal distribution that the most important insights into their significance can be made.

8.3.6 The Importance of Data Integrity

It was stated above that factor analysis is based upon the extraction of the eigen-vectors of the correlation matrix. The successful and reliable calculation of the correlation coefficient is dependent on the input data being normally distributed and free of outliers; as was demonstrated in Section 6.2, the presence of a single outlier results in gross exaggeration of the strength of the correlation between two elements. In factor analysis, this will result not only in the creation of an erroneous factor, strongly weighted in the element or elements whose analyses created the outlier, and probably no other, but also, since they are by definition constrained to be mutually perpendicular, in all the other factors in the model as well.

8.3.7 Application of Factor Analysis to Geochemical Data

8.3.7.1 Stream Sediment Data, New Brunswick

One of the earliest applications of factor analysis to geochemical data in Canada is, nevertheless, convenient for demonstrating several important aspects of its use (Nichol, 1971, 1972). Factor analysis was carried out on multi-element data from a geochemical reconnaissance of the Bathurst-Jacquet River area in New Brunswick, carried out by the Geological Survey of Canada (Boyle *et al*, 1966;). The 2000 km^2 area is underlain by a variety of Lower Paleozoic sediments and volcanics, with granitic intrusions of possibly Devonian age (Figure 8.6d). Mineralization consists of massive, vein and disseminated deposits containing a number of types of mineralization, including volcanic-exhalative massive sulphide deposits, Cu-Pb-Zn-As and Ag vein deposits and, of lesser importance, Mo, Mn and Fe occurrences. None of the large volcanic-exhalative massive sulphide deposits near Bathurst (Brunswick No.12, Heath Steele, Caribou, etc.) occurs within the study area.

The initial survey involved the collection of drainage-sediment samples at quarter-mile intervals and their analysis for Pb, Zn, Cu, As, Sb, Mo, Sn, W, Ag, Ni, Co, Cr, Ba, Mn and cold-extractable heavy metal (cxHM). The minor-element distributions showed a number of associations, apparently related to mineralization, bedrock type and surface environment. The strongest associations were easily identified from a visual examination of the data. It appeared, however, that by means of factor analysis it might be possible to identify less obvious associations and to draw attention to sites where associations were present but constituted a small component and were therefore not identifiable from a visual examination of the data.

The correlation matrix upon which the factor analysis was based is shown in Figure 8.5.

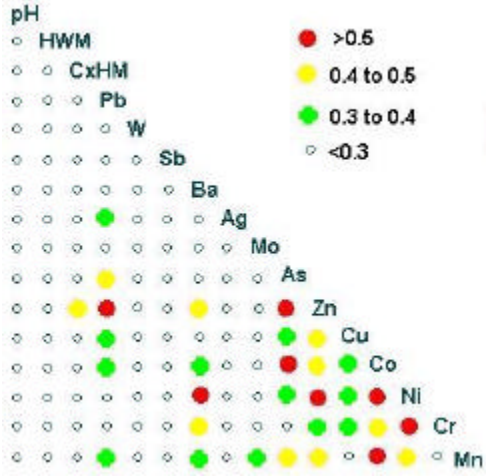

Figure 8.5: Correlation matrix for stream-sediment analyses, Bathurst District, New Brunswick

Factor analysis carried out on the GSC data indicated that several of the elements were of mixed provenance. The zinc distribution, shown in Figure 8.6a, was shown to be related to three main associations:

- Ni, Cr, Ba, Co, Zn, Cu, Mn (not shown here);
- cxHM, Zn, As, Cu and Pb (Figure 8.6b); and
- Ag, Pb, As, Zn, Cu (Figure 8.6c)

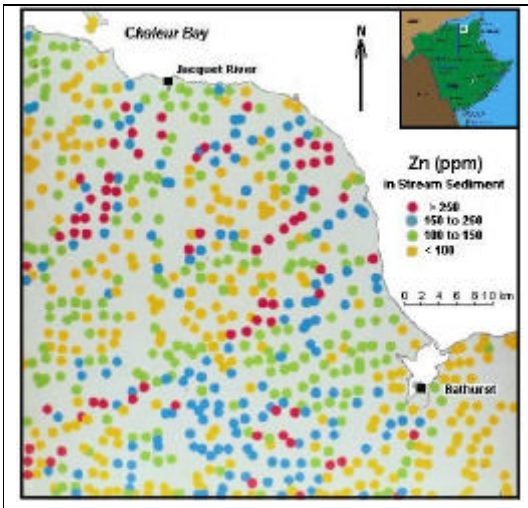

Figure 8.6a: Distribution of zinc in stream sediments, Bathurst District, New Brunswick

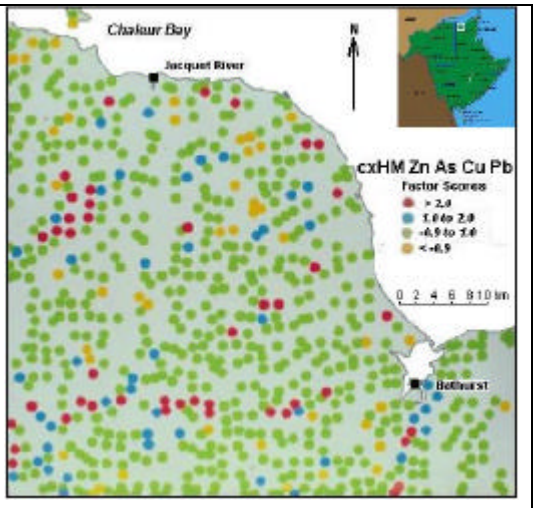

Figure 8.6b: Distribution of cxHM (cold-extractible heavy metal)-Zn-As-Cu-Pb factor in stream sediments, Bathurst District, New Brunswick

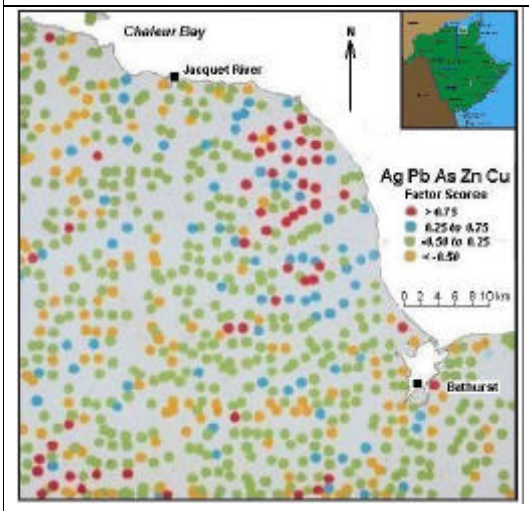

Figure 8.6c: Distribution of Ag-Pb-As-Zn-Cu factor in stream sediments, Bathurst District, New Brunswick

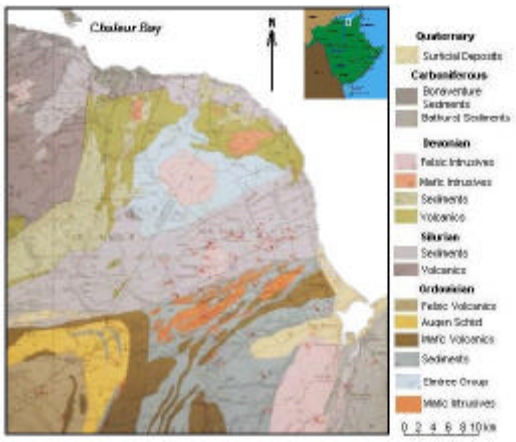

Figure 8.6d: Geological Map, Bathurst District, New Brunswick

The three associations are considered to represent bedrock geochemistry (the Ni-Cr factor) and two types of mineralization, of which high factor scores representing the first (cxHM, Zn, As, Cu and Pb) show a spatial relationship with the Orvan Brook, Rocky Turn and Armstrong volcanic-exhalative massive sulphide deposits in the southwest, as well as showing a concentration in the northwest. High factor scores representing the second (Ag-Pb-As-Zn-Cu) are associated with the Nigadoo River, Beresford and Millstream vein and replacement occurrences in the northeast, which themselves are related to a small granitic boss of Devonian age. Comparable zinc contents relate to all three associations, and from a consideration of the zinc data alone it is not possible to establish the provenance of the zinc concentrations. Using factor analysis it was possible to redistribute the zinc distribution in terms of these zinc-bearing associations.

In this case, factor analysis serves as a method for identifying the provenance of the zinc concentrations in the stream sediments. Where concentrations of an element of interest are related to more than one source, the provenance of the metal in the samples may be identified by factor analysis if the sources have diagnostic metal associations. However, it is important to bear in mind that variable surface conditions may modify the associations recognizable in the unweathered bedrock source.

8.3.7.2 Lake Sediment Data, Northern Manitoba

The NGR data set for the Granville Lake 1:250 000 sheet (NTS 64C) in northern Manitoba consists of analyses for 16 elements on lake-sediment samples from approximately 1,300 sites covering an area of 12,500 square kilometres. The area is underlain in the extreme north by rocks of the South Indian Gneiss Belt; in the north by two approximately east-west striking greenstone belts of Proterozoic age (the Wasekwan Group, making up the northern and southern Lynn Lake Greenstone Belts); in the centre by a belt of metamorphosed arenaceous rocks (the Sickle Group) and in the south by rocks of the Kisseynew Gneiss Belt. Felsic and felsic-intermediate intrusions, of varying areal extents, are common throughout the map area (Figure 8.7d). Lode and BIF-associated gold occurrences are common in the greenstone belts and have been mined at the MacLellan and Farley Lake mines, in the northern belt, and at BT in the southern. Additionally, nickel, cobalt and copper were mined from a group of magmatic-type deposits in the northern belt at and around town of Lynn Lake itself, and copper and zinc from a volcanic-exhalative massive sulphide deposit at Fox Lake, at the western extremity of the northern Lynn Lake greenstone belt.

Square-root transformations were selected for Zn, Cu, Hg, F and V; log-transformations for Ni, Co, Mn, Fe and U; and L.O.I. (loss on ignition) values were left untransformed. The following elements, also included in the database, were not included in the factor analysis because of heavy censoring (in other words, too many of their values fall below the analytical detection limit): Ag, As, Cd, Mo and Sb. Lake waters were also analysed for Ca, F, pH and total Alkali but these results are also not considered here.

The correlation matrix for the admissible data is shown in Table 8.1

Table 8.1: Correlation matrix for lake sediment data, Granville Lake Sheet

LOI									
-0.20	Zn								
-0.07	0.54	Cu							
0.42	0.21	0.24	Hg						
-0.60	0.46	0.43	-0.18	F					
-0.47	0.68	0.63	0.01	0.69	V				
-0.30	0.61	0.70	0.02	0.74	0.72	Ni			
-0.40	0.73	0.56	0.02	0.66	0.72	0.75	Co		
-0.38	0.59	0.34	-0.04	0.31	0.52	0.36	0.62	Mn	
-0.41	0.68	0.36	0.05	0.39	0.70	0.38	0.61	0.67	Fe
-0.38	0.35	0.52	-0.09	0.53	0.56	0.55	0.43	0.33	0.36 U

Examination of the results of factor analysis on these data resulted in the adoption of a 3-factor model which accounts for a total of 77% of the total data variance. The areal distribution of scores for the three factors is shown in Figures 8.7a, 8.7b and 8.7c. The first factor, which accounts for 34% of the total data variance, is dominated, in decreasing order of importance, by Ni, Cu, F, U, V and Co. It is interpreted as being related to the depth of clay deposited in glacial Lake Agassiz and highest scores for this factor are encountered in the southeast where glaciolacustrine clay several metres thick has been intersected in Sonic drill core from the floors of the deeper lakes (Figure 8.7a).

The second factor, accounting for 15% of the total data variance, is inversely related to the contents of L.O.I. and Hg and appears to represent the contribution of clastic, non-organic sediment. This also displays its highest scores in the southeast (Figure 8.7b), over a more restricted area, where many samples were collected in lakes that form part of the Churchill River system and where the supply of clastic sediment would be expected to be high. The third factor, accounting for 28% of the total variance, is dominated by Fe, Mn, Zn, Co and V and is probably related to co-precipitation with oxides and hydroxides of the first two elements. It does not display any regional-scale features (Figure 8.6c); in particular, it does not indicate the presence of the east-west-trending Lynn Lake greenstone belts, which is why it is not believed to be related to the content of mafic minerals in bedrock.

An analogous situation to that observed in the Bathurst district data is that two elements (V and Co) are relatively strongly loaded in more than one factor, *viz*. the “Lake Agassiz” and the “Coprecipitation” factors. In this case, the factor analysis has not revealed patterns that were not apparent by consideration of the individual elements (witness the similarity between the distribution of Factor 1, in Figure 8.6a, and Cu, in Figure 8.8a) but it serves to identify, by virtue of the elements that are co-associated in each factor, the principal environmental controls on composition of the lake sediment.

This serves as a platform for isolating the more subtle features of the single-element compositions that cannot be explained by these major controls; this will be dealt with in the next section.

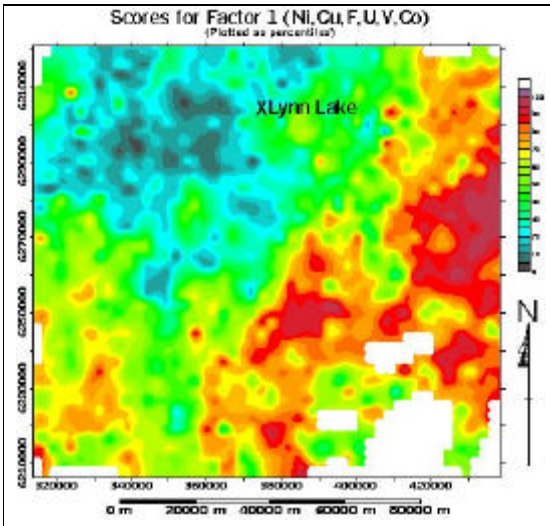

Figure 8.7a: Areal distribution of Lake Sediment Factor 1, Granville Lake Sheet, N. Manitoba. Factor is believed to be related to thickness of Lake Agassiz clay. Data from GSC Open File 1288.

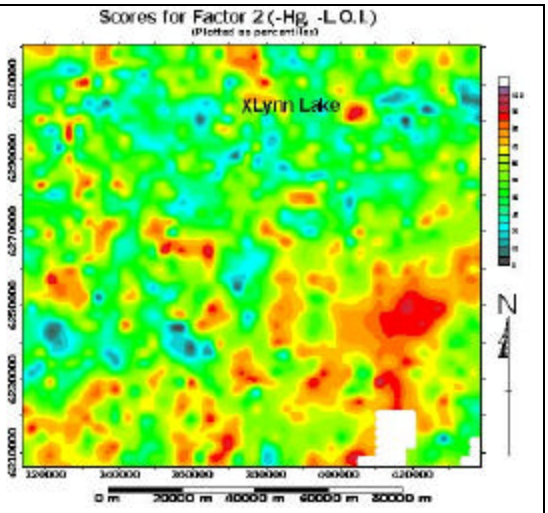

Figure 8.7b: Areal distribution of Lake Sediment Factor 2, Granville Lake Sheet, N. Manitoba. Factor is believed to be related to contribution of clastic sediment.

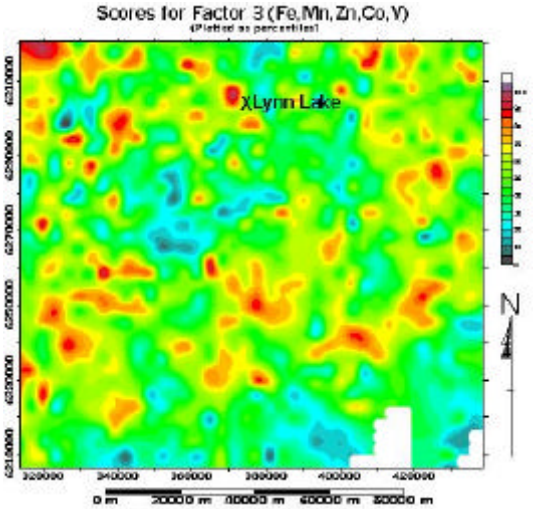

Figure 8.7c: Areal distribution of Lake Sediment Factor 3. Factor is believed to be related to coprecipitation with Mn and Fe oxides

Figure 8.7d: Geological Map, Granville Lake Sheet. **Reds:** felsic and felsic-intermediate intrusives; **greys:** metamorphosed argillaceous sediments (Kisseynew Gneiss Belt); **yellows:** metamorphosed arenaceous sediments (Sickle Group); **greens:** metavolcanic rocks (Wasekwan Group)

8.3.8 Adaptation of Factor Analysis to the Identification of Anomalies

As was stated above in Section 5.1, mineralization, whether economically significant or not, usually constitutes a statistical rarity. In other words, and in the context of factor analysis, unless the area covered by a geochemical survey is focused very closely around a zone of mineralization, it is unlikely that the mineralization process will exert sufficient influence on the data for it to be modelled as a factor. The data from the Bathurst District, if the interpretation based on their factor analysis is correct, are an exception to this. Factor analyses of regional geochemical data sets, at least, that purport to include a "mineralization factor" tend to have been derived from data that are strongly influenced by a few highly anomalous outliers. As such, their conclusions are often suspect.

Even when the response to mineralization constitutes a statistical rarity, as described above, it is sometimes possible to quantify the component of each element's composition that cannot be explained by the action of the dominant agencies, modelled numerically by the factor scores. It is preferable, and more convincing, that these agencies are recognisable as well-documented natural processes; in this

way the technique is less likely to be viewed as statistical hocus-pocus. This **residual** component is more likely to be related to an unusual situation of which the presence or proximity of mineralization constitutes one example.

This method was first applied in geochemistry by Closs and Nichol (1975) on stream-sediment data from Newfoundland; in the study reported here, the data set from the Granville Lake sheet of northern Manitoba are used once again to illustrate its application. Figure 8.8a shows the contoured distribution of raw Cu values in the lake sediment and it is clear that the composition of the latter is heavily influenced by the agency modelled by Factor 1, interpreted as being related to sedimentation in glacial Lake Agassiz (see Figure 8.7a). If the effects of this factor, and the lesser effects of Factors 2 and 3, are removed by regressing them against Cu, and the regression residuals for Cu plotted, the results are as shown in Figure 8.8b. A strong linear, east-west trending feature, scarcely noticeable in the map of raw Cu values, is apparent to the southwest of the town of Lynn Lake; this corresponds to a segment of the southern Lynn Lake Greenstone Belt. The large, irregularly-shaped feature in the southwest, which is apparent from the raw Cu values but not as prominently, is underlain by the "Transition Zone" between the Kisseynew Gneiss Belt and overlying Sickle Group rocks and corresponds rather closely to an area where numerous subeconomic Cu occurrences of possible red-bed affinity (Baldwin, 1980) have been described. In the northeast, the local feature of elevated Cu residuals may be related to several subeconomic Cu occurrences of volcanic-exhalative massive sulphide affinity, south of Barrington Lake.

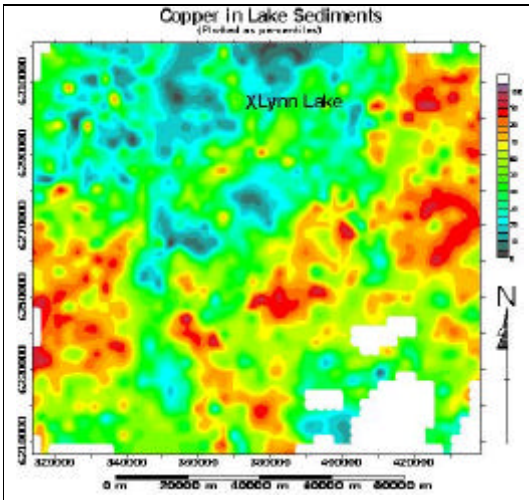

Figure 8.8a: Distribution of Cu in lake sediments, Granville Lake Sheet

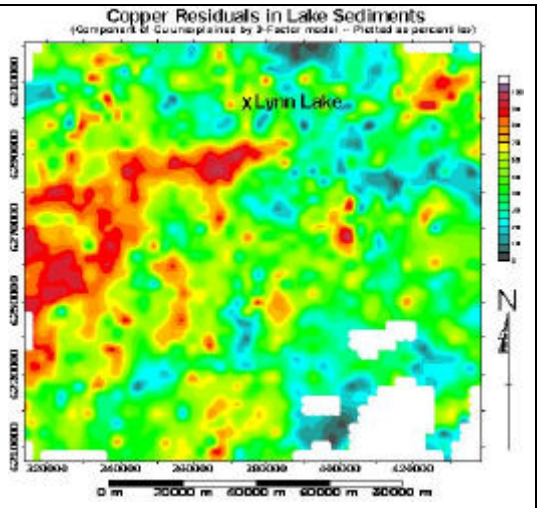

Figure 8.8b: Distribution of Cu residuals in lake sediments, Granville Lake Sheet

8.3.9 Conceptual Problems in Geochemical Factor Analysis

To summarize what was stated above, r-mode factor analysis involves treating a series of multielement analyses as a single, hyperellipsoidal cluster of points in **k**-dimensional space, and extracting **m** mutually-perpendicular axes from the hyperellipse , where **m** is considerably less than **k**. Two words that are worth examining in more detail are "single" and "axes". The first implies that however many geological or geomorphological environments the sample material is derived from, the points representing the samples form only one cluster when plotted (the method of cluster analysis, which will be described later, is aimed precisely at detecting more than one such grouping). The word "axes" implies that what is being extracted from the data are not points, representing positions in hyperspace, but a series of lines, whose directions are defined by the angles they make with the axes representing the elements in the analysis. The process does not identify specific positions on those lines, extreme or otherwise. Consequently it is problematical, at least in theory, to identify an "ultramafic" factor, even if the elements Ni, Cr, Mg (etc.) are highly loaded and if high scores of the factor coincide with the occurrence of ultramafic rocks in the protolith of the samples in question. What is being identified is a process, one end member of which is represented by ultramafic rocks, or their derivatives, and the other by something else (possibly, felsic rocks – but what if there are none in the study area?). Similarly, high scores of factors heavily loaded in Ca, Sr and Mg may indicate the presence or proximity of calcareous rocks, but the question then arises as to what low scores for such a factor would indicate – calcareous rocks do not lie at the end of a compositional continuum.

Despite this conceptual difficulty, factor analysis has proven more successful than cluster analysis in identifying specific lithologies, including calcareous and ultramafic rocks, in areas where geological information was lacking.

8.4 Discriminant Analysis

8.4.1 Applicable Situations

Most geochemical variables are measured on a continuous ratio scale. However, their ultimate function for the explorationist is as an aid to converting them to categorical variables, which may have as few as two values (do we follow this up, or not?) or more (from which lithology does this sample appear to be derived?).

These problems can be dealt with by discriminant analysis, which is a multivariate method used to treat problems of classification. This method is applicable to situations where there are previously-defined "training sets" representing classes which differ in some important, observable and important characteristic. From the multivariate observations that make up these training sets, a series of **discriminant functions** are derived, one per defined class. These are **data-dependent**, unlike the user-defined indices described in Section 7. Solution of the functions for the data on a single sample yields a series of indices, one for each classification, known as **discriminant scores**. For a sample of unknown or unspecified provenance, the class whose discriminant score has the greatest magnitude is the one to which that sample is assigned.

8.4.2 Graphical Depiction

To illustrate the underlying principles of discriminant analysis, a fictional, but realistic two-class situation is postulated in Figure 8.9 where the composition of felsic volcaniclastic rocks is known to reflect the presence of massive-sulphide mineralization, on a semiregional scale, in the form of lower sodium and higher magnesium contents. Ordinarily, discriminant analysis involves the use of many more variables than in this bivariate example, but the principles involved can be adequately demonstrated in two dimensions. In other words, rock samples collected from the vicinity of such mineralization are generally higher in MgO, and lower in Na_2O, than those collected remote from such mineralization. This relationship is illustrated by the "Na_2O" and "MgO" histograms for the "mineralized" (black) and "unmineralized" (white) groups in the following figure.

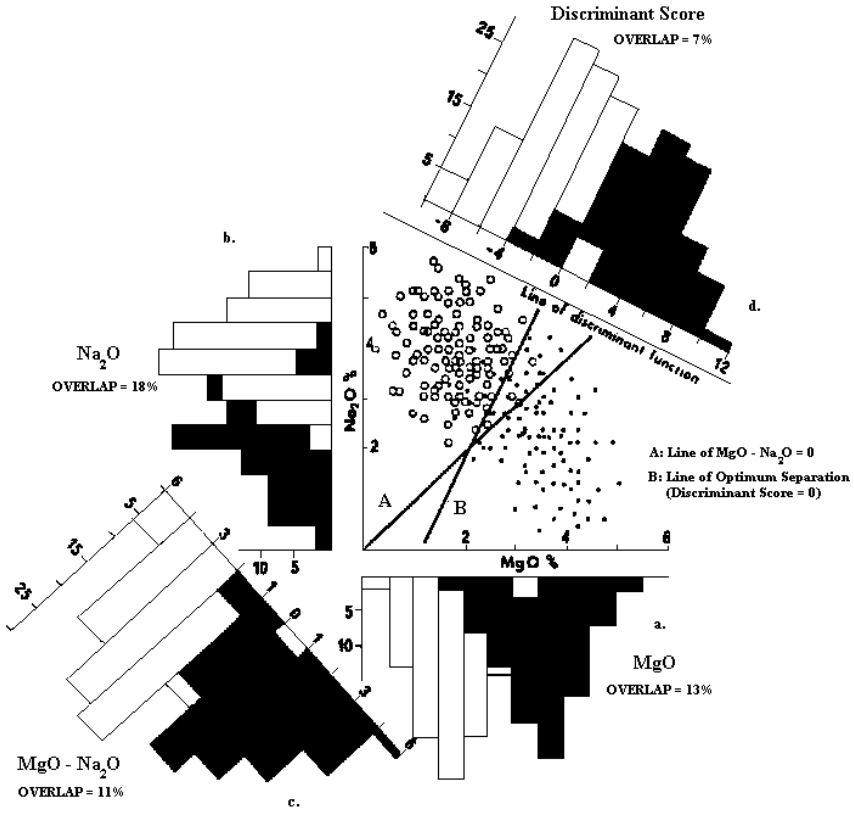

Figure 8.9: Graphical demonstration of discriminant analysis. From Amor (1983).

However, there is considerable overlap (13%) between the two populations, so that samples of unknown relationship to mineralization, whose MgO values range between 1.5 and 3.5%, could conceivably belong to either of the two groups. A similar situation exists if the Na_2O values are examined in isolation; "mineralized" samples have generally lower Na_2O content than the "unmineralized " samples, but the overlap between the groups is 18%, and the "grey" area extends from 2.0 to 4.5% Na_2O.

If, however, the MgO and Na_2O values are plotted against one another on an X-Y plot, the combined effect of the two elements causes overlap to be reduced. Samples of unknown affiliation can then be assigned to the “mineralized” or “unmineralized” groups with less ambiguity.

The combined effect of MgO and Na_2O in classifying samples as to their affiliation, as shown in the bivariate plot, can be expressed numerically by creating a two-element index $\mathbf{z_0}$, where z_0 = MgO-Na_2O. This is analogous to a multielement "index" as described in Section 7. The histograms of z_0 values for the “mineralized” (black) and “unmineralized” (white) groups are shown in the lower left of the figure. One would expect that the “mineralized” samples would be associated with the highest values of z_0, and the “unmineralized” values with the lowest. This is, of course, true of the individual analyses also, but in combination the overlap between the two groups is reduced to 11%: an improvement on either of them individually.

While the combination of the elements in their unweighted form (that is, with coefficients of 1.0), only serves to improve the separation between the two training sets, it does not necessarily maximize it. This maximization of separation (and minimum of overlap: in this case, only 9%) is the role of discriminant analysis. In this case the equal weightings of MgO and Na_2O (or coefficients of 1.0) in z_0 are transformed to coefficients of 2.917 and -1.520, in the discriminant function:

$$z_1 = 2.917MgO - 1.520Na_2O$$

When a sample's MgO and Na_2O values are inserted into this function, the result is known as the sample's discriminant score. As it is convenient to assign negative values of the discriminant score to one set, and positive values to the other, the discriminant score is transformed linearly by subtracting the discriminant score of the midpoint between the centre of the two groups. The discriminant function then becomes: $z_1 = 2.917MgO - 1.520Na_2O - 2.97$. A sample of unknown provenance, whose MgO and Na_2O values are 3.0% and 2.5% respectively, would be classed ambiguously by either of these elements individually, or by their unweighted combination, but its discriminant score assigns it unequivocally to the "unmineralized" group.

It is not hard to visualize how, if a third dimension were added to the figure, for example a CaO axis, a further reduction in overlap might result, although this does not automatically follow (see Section 8.4.3).

Clear expositions of the principles of discriminant analysis are provided by Klovan and Billings (1967) and Davis (1973, p. 450). There are examples of its use in the classification of gossans (Bull and Mazzuchelli, 1975); and identification of alteration zones around VHMS deposits (Govett, 1972; Nichol *et al.*, 1977; Amor and Nichol, 1983).

Discriminant analysis is most commonly applied to situations where there are only two previously-defined "training sets", so it is useful in two-group situations where it is necessary to discriminate and classify "mineralized" and "unmineralized" or "altered" and "unaltered" samples, where other potential inhomogeneities of the sample medium are insignificant. Where more than two groups have been identified (for example, when multiple lithologies are present within the area of a regional stream- or lake-sediment survey), it is possible to develop a series of discriminant functions, each one specific to a particular group. A sample of unknown provenance is assigned to the group whose function yields the highest score.

8.4.3 Stepwise Discriminant Analysis

Although discriminant analysis can, in theory, be carried out with as many measured variables as are available and satisfy the assumptions described above, this does not necessarily mean that every variable added to the function will improve the distinction between the training sets; such inclusion may even result in inferior discrimination to what was achieved with fewer variables. For example, if Al_2O_3 proved to be immobile during the process that resulted in MgO enhancement and Na_2O depletion, described in the example above, then the inclusion of Al_2O_3 in the discriminant function would almost certainly result in inferior discrimination to that provided by just MgO and Na_2O.

Stepwise discriminant analysis seeks to derive a function that optimizes the distinction between the training sets with the minimum number of variables necessary.

Although they may be equally important from a geochemical or metallogenic point of view, some of the input variables may fulfil essentially the same role, in terms of the samples that they correctly (and incorrectly) assign. A function derived by stepwise discriminant analysis would probably not include both of such variables; this phenomenon is known as **redundancy**. On the other hand, variables that initially appeared unimportant as discriminators may acquire added significance because of their ability to remove residual overlap between the sets. Therefore, at least some of the variables that appear in a stepwise function will probably seem unfamiliar, in terms of what is known of the geochemical processes involved.

8.4.4 Application to Rock Geochemical Data

A combination of forced and stepwise discriminant analysis was applied to compare and contrast alteration styles at seven volcanic-exhalative massive sulphide deposits from the Superior Province of the Canadian Shield (Amor, 1983; Amor and Nichol, 1983). The data were the same as those used by McConnell (1976) to derive a multielement index termed the Standardized Net Residual (see Section 7.3, above).

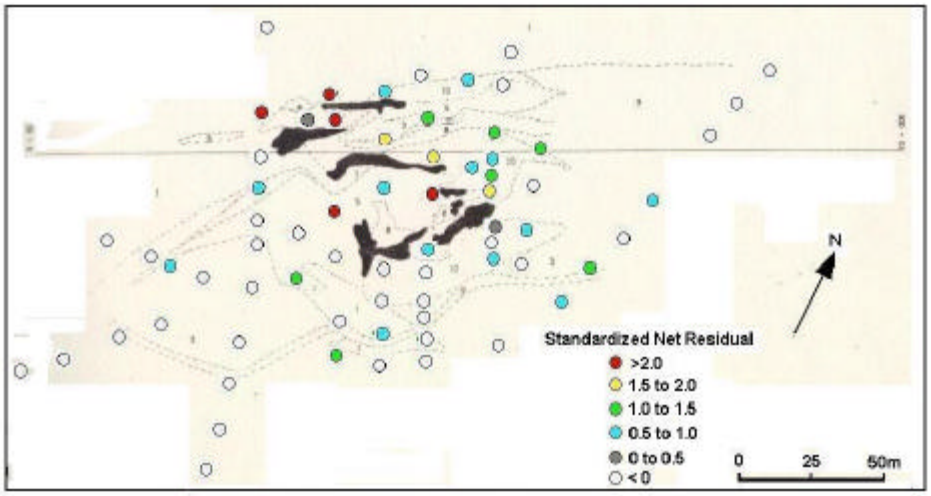

Figure 8.10a: Distribution of Standardized Net Residual (SNR) in wallrocks, South Bay deposit. Anomalous zone, as identified by red circles, is less extensive than that defined by Type II discriminant score (below). Solid markings indicate massive sulphide orebodies (from McConnell, 1976).

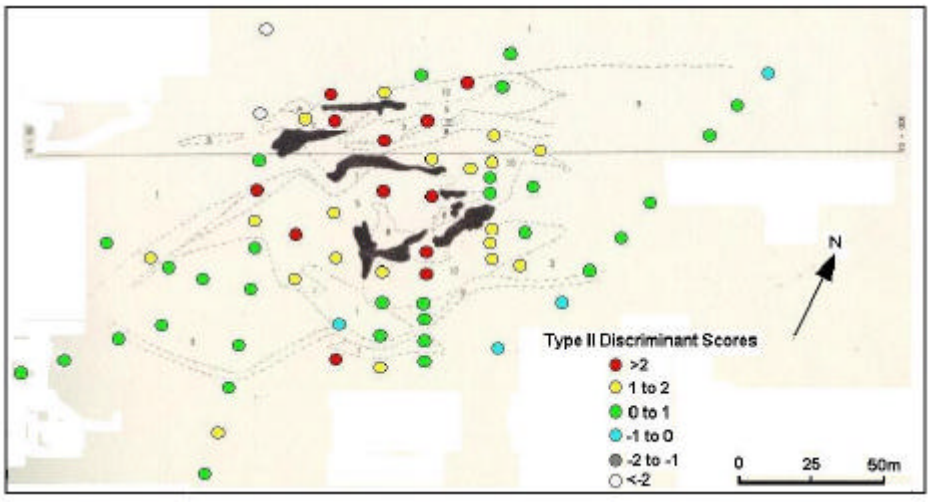

Figure 8.10b: Distribution of Type II discriminant scores in wallrocks, South Bay deposit. Red, yellow and green circles indicate positive discriminant scores (i.e. greater affinity to "altered classification"). Solid markings indicate massive sulphide orebodies (from Amor, 1983).

At the outset, a discriminant function of regression residuals (see Section 8.2) of Fe_2O_3, MgO, CaO and Na_2O was derived between wallrocks of each of the seven deposits, and local, ostensibly unmineralized and unaltered background. The scores yielded by each of these functions were compared by means of the correlation coefficient and it was concluded that the effects of two distinct alteration types could be discerned: the first (Type I) was characterized by the familiar addition of Fe_2O_3 and MgO, and removal of CaO and Na_2O, while the second (Type II) was characterized by removal of Na_2O and, less markedly, of Fe_2O_3 and MgO, and the addition of K_2O. Type I alteration was interpreted as a combination of chloritization and sericitization specific to the footwall "alteration pipe" characteristic of this deposit type, while Type II alteration, characterized by sericitization only, was encountered both in footwall alteration pipes formed under rather specific physicochemical conditions and volcanostratigraphic environments, and also as an alteration envelope possibly formed by electrochemical reaction between the massive sulphide deposit and its wallrocks.

Subsequently, alteration-specific discriminant functions were derived for the two alteration types. The Type II function was as follows:

$$D_2 = 1.0065(Fe_2O_{3\mathbf{R}}) - 3.0216(MgO_{\mathbf{R}}) - 4.7516(Na_2O_{\mathbf{R}}) - 1.1425(P_2O5_{\mathbf{R}}) - 3.9281$$

The subscript $_{\mathbf{R}}$ denotes the regression residual of the oxide in question. Positive scores of this discriminant function indicate a greater affinity to rocks altered in the Type II style, while negative scores show more affinity to unaltered rocks. The alteration zone defined by scores of the function is compared to that defined by the Standardized Net Residual, at the South Bay deposit in northwestern Ontario, in Figures 8.10a and 8.10b. The zone defined by discriminant scores can be seen to be much more more extensive and as such, it constitutes a more easily-detectable target in mine-scale exploration.

8.5 Cluster Analysis

8.5.1 Relation to Discriminant and Factor Analysis

In the previous section, discriminant analysis was described as a method of using information regarding the affiliation of samples to one or other of a series of previously-defined groups, and the multi-element composition of those samples, to optimise the distinction between those groups in multi-element space and use the knowledge gained to classify samples of unknown affiliation.

Cluster analysis resembles discriminant analysis inasmuch as the distribution of the sample points in multielement space is also considered. However, no information is incorporated regarding subdivisions of the data set as it is the role of this method to identify such subdivisions. The difference between the two methods is analogous to that between "supervised" and "unsupervised" classification in Remote Sensing.

The additional knowledge that discriminant analysis incorporates (of previously-known classifications) adds tremendously to the power of that method over cluster analysis, however carried out. Cluster analysis would not, for example, identify and distinguish the two groups of points, representing altered and unaltered rocks, that were used to illustrate discriminant analysis (Figure 8.8), because there is no "empty space" between them.

8.5.2 Mathematical Methods

8.5.2.1 Agglomerative Clustering

Clustering methods can begin with the assumption that every case in the data set represents a single cluster of points in multidimensional space; they are then joined progressively, based on their mutual separation. This type of clustering is known as agglomerative clustering and was the first method to be applied in geochemistry (e.g. Obial, 1970). Its most familiar product is the **dendrogram** in which the linkages between individual samples are expressed graphically in a tree-like structure. The size of a data set that can be meaningfully interpreted by a dendrogram is rather limited; furthermore, because it involves the calculation of the distance, in hyperspace, between every sample and every other sample in a data set (the size of the matrix of intersample distances being proportional to the square of the number of samples) and is therefore relatively computer-intensive, agglomerative clustering fell out of favour in the late 1970s, although with the current generation of powerful computers that is scarcely a reason not to use it now.

8.5.2.2 Divisive Clustering

Alternatively, the initial assumption can be made that the data form a single cluster, which is then modified by progressively splitting the data up into smaller clusters, to one or other of which each sample is assigned based on the distance between its plotted points and the centroid of the cluster. The process is continued until the required number of clusters has been created and all samples assigned to

one or other of the clusters, based on their distance from cluster centroids. This method is known as **k-means clustering** and it is much more rapid and less computer-intensive than the first method described. However, as with factor analysis, the user must decide at the outset how many clusters are to be extracted, and the experimentation required to arrive at an optimal cluster model is a time-consuming business that is not suitable for routine geochemical exploration.

8.5.3 Graphical Methods

It has been demonstrated experimentally that the human eye is more efficient at detecting clusters of points, and the discontinuities between them, and other features of data structure, than any mathematical algorithm.

Unfortunately the dimensionality of plots in which visual clustering can be accomplished is limited to 2, or under certain special circumstances, 3. Since geochemical data generally have a much greater dimensionality, it is useful to create a 2-dimensional rendition of an array of points in hyperspace, whereby the intersample distances in the higher-dimension plot are preserved as closely as possible in the lower-dimensioned plot, in which the identification of clusters and other features can be made visually. A variety of techniques have been developed to achieve this end, with such names as Planing, Non-linear Mapping and Multidimensional Scaling.

The potentiality, and limitations, of visualizing multidimensional data in two dimensions may be judged by viewing the result of applying Systat 9's multidimensional scaling module to whole rock data for a variety of previously-defined volcanic rocks in Figure 8.11. The Tholeiitic, Calc-Alkaline and Alkali Olivine Basalt trends are distinguishable, while unusual rocks such as peralkaline rocks and komatiites lie off these trends, but it is not clear in what respects a high-Al basalt of the Calk-Alkali Series differs from, or is similar to, a tholeiite.

8.5.4 Interpretation of Results

The clusters which are extracted by any of the above methods can be interpreted in terms of the elements that are relatively elevated or depleted in them, the areal distribution of the samples assigned to them, and any observable geological, geomorphological or environmental observations that characterize them. This can be a lengthy process without the aid of custom-designed software.

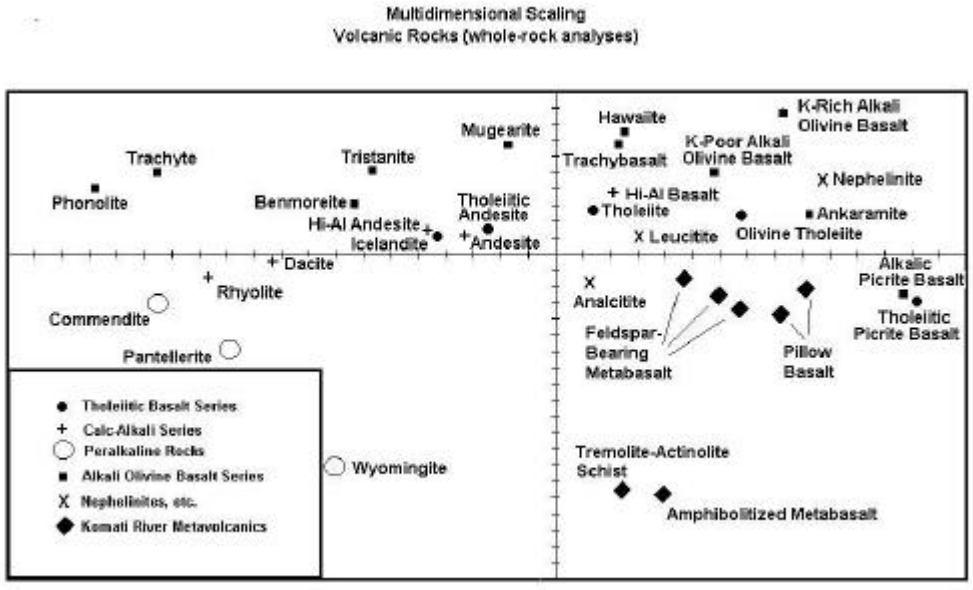

Figure 8.11: Multidimensional Scaling of whole-rock composition (SiO_2, Al_2O_3, Fe_2O_3, FeO, MgO, CaO, Na_2O, K_2O, TiO_2) of various volcanic rocks. Data from Irvine and Baragar (1971); Viljoen and Viljoen (1970).

Despite the conceptual difficulty in explaining that the chemical composition of naturally-occurring material varies on a continuous scale, rather than in a series of discrete concentrations (see Section 8.3.9), cluster analysis has not been used in geochemical exploration as widely, or with as much claimed success, as factor analysis. One reason for this may be the difference in number type between factor scores, which are continuous, and measured on an interval scale, and cluster allocations, which

are discrete, and measured on a nominal or categorical scale. In other words, whereas it is meaningful to interpolate the factor scores between two sample points in order to create a contour map (a node halfway between two sample locations, of factor scores 1.0 and 2.0, can be assigned a value of 1.5), this cannot be done with cluster allocations (a point between two sample locations, one of which is assigned to Cluster 1 and one to Cluster 2, cannot meaningfully be assigned to Cluster 1.5). As was stated above, the interpretation of the significance of factors is greatly facilitated by plotting their scores on a map.

Methods have been suggested for creating smoothed geographic realizations of categorical variables and if they are incorporated into widely-available software packages, cluster analysis may be applied more successfully in geochemical exploration.

9 CONCLUSION

The careful analysis of geochemical data, using standard computer software packages, is an important and affordable way of adding value to an exploration company's assets. The cost of a thorough data review is normally less than ten per cent of total geochemical program costs.

The analysis of a digital geochemical data set should begin with a thorough check of the data's **integrity**. This is particularly important if the data have passed through the hands of more than one individual prior to the interpretation stage. A number of time-saving techniques are available to facilitate this apparently tedious task, but a visual scanning of the data should also be carried out.

If more advanced techniques of data analysis are to be used effectively, the frequency distributions of all of the input variables should conform, at least approximately, to a **Normal distribution**; they should be free of outliers, and less than 30% of them should fall below the analytical detection limit.

One of the commonest and most serious pitfalls in the interpretation of geochemical data is the failure to utilize all previously-acquired information regarding the inhomogeneity of the data set. The frequently subtle features that characterize the presence or proximity of mineralization are more readily identified if they are not drowned out by the stronger signatures of variable host lithology, variable position in the soil horizon, or various environmental effects, many of which can be observed and incorporated into the database if the will exists to do so.

The most useful statistic that can be calculated for single-element data is the **percentile**. Percentiles are non-parametric, readily calculated and easily understood. They facilitate the comparison, and if necessary the combination, of identifiable subsets of a geochemical population, thereby alleviating the "apples and oranges" problem to some extent. They can be used to identify single-element anomalies in a consistent way, and they are easy to plot.

The most useful **graphical techniques** for summarizing the characteristics of a data set are histograms, cumulative frequency plots and probability plots. Box-and-whisker plots provide a convenient visual and intuitive means of comparing concentration levels of the same element in identifiable subsets of the data, for example, different soil colour or source lithology.

As stated above, single-element **geochemical anomalies** are most effectively identified using percentiles. The 97.5-percentile is equivalent, in a Normal distribution, to the value of the mean plus two standard deviations, used as a "threshold" value in the past; however, in non-Normal distributions, which are very common in geochemical variables, the former is more reliable. Its use is, nevertheless, merely a convenient way of highlighting the highest values in the range and it has no fundamental connection with mineralization. Consequently, values that approach the threshold, without exceeding it, are worthy of interest if the sample sites in question are closely spatially associated. Therefore, plots of the anomalous samples are important as an interpretational aid as well as a means of summarizing the findings of the study. It is sometimes possible to identify subpopulations, of differing properties, within a single frequency distribution using probability paper. Separate criteria for anomaly identification can then be evaluated for each subpopulation so identified.

When time does not permit a detailed multivariate analysis, the calculation of **multi-element** indices offers a method of combining the individual propensities of certain elements to be enriched in mineralization, or key lithologies, based on published information or analysis of mineralized rock of the type sought. Unsuspected deviations from the model may reduce their effectiveness, however.

Of the **multivariate methods**, **regression analysis** enables the establishment of a global relationship between a predictor or independent variable (or variables), and a predicted or dependent variable. Local

deviations from this model for the latter are thus thrown into sharper focus and subtle responses to mineralization may be detected.

R-mode factor analysis enables the behaviour of a large number of geochemical variables to be condensed into a smaller number of linear combinations of those variables, that account for most of the total data variance and are often relatable to recognizable geological and surficial processes. Separation of the contribution of each of these processes to the composition of individual samples, from that component that cannot be attributed thus, sometimes enables the detection of subtle, mineralization-related features that would otherwise have been swamped by the stronger signal. The method is based on the correlation coefficients between the variables in the data set, and very sensitive to variations in it; therefore, it is important to calculate the correlation coefficient in a meaningful way and avoid its distortion by outliers.

Discriminant analysis uses all of the analyses in a data set (or, in the case of stepwise discriminant analysis, no more than are necessary) to maximize the distinction between two or more previously-defined groups and to enable the subsequent allocation of samples of unknown provenance, based on analyses of the same elements. It has found application in the classifying of gossans, the identification of weak hydrothermal alteration, and reconnaissance geological mapping. The related technique of **cluster analysis** looks for groupings of points, representing individual geochemical samples, in multi-element space, without prior knowledge that they exist or what their compositional characteristics may be. It has not been used as extensively in geochemical exploration but that situation may change with the availability of software.

To sum up: powerful, relatively-inexpensive tools for detailed data analysis and display are now available to almost every explorationist, and not to make careful and informed use of at least some them is as inappropriate as going into the field without a geological hammer.

10 REFERENCES

Amor, S.D., 1983. *Application of Discriminant Analysis to a Study of Geochemcal Dispersion around Massive Sulphide deposits in Superior Province*: Unpubl. Ph.D thesis, Queen's University, 404 p.

Amor, S.D. and Nichol, I., 1983: Identification of Diagnostic Geochemical Alteration in the Wallrocks of Archean Volcanic-Exhalative Massive Sulphide Deposits: *J. Geochem. Explor*., 19, pp. 543-562.

Amor, S.D. and Elliston, H.A., 1993. Open File Report No. 4 – Evaluation of a Regional Geochemical Anomaly, Back Rio Grande Area, Portland Parish, Jamaica: Geological Survey Division (Jamaican Ministry of Production, Mining & Commerce), 365 p.

Baldwin, D.A., 1980. Disseminated Stratiform Base Metal Mineralization along the Contact Zone of the Burntwood River Metamorphic Suite and the Sickle Group: Manitoba Dept. Energy & Mines, Economic Geology Report ER79-5, 20p.

Bondar Clegg & Co., 1988. Jamaica Metallic Minerals Survey – Phase I. Report and Appendices for CIDA Project No. 504/0012280, 244 p.

Boyle, R.W., Tupper, W.M., Lynch, J., Friedrich,G., Ziauddin, M., Shafiqullah, M., Carter, M. and Blygrave, K., 1966. Geochemistry of Pb, Zn, Cu, As, Sb, Mo, Sn, W, Ag, Ni, Co, Cr, Ba and Mn in the Waters and Stream Sediments of the Bathurst-Jaquet River District., New Brunswick: Geol Surv. Can., Paper 65-42, 50p.

Bull, A.J. and Mazzuchelli, R.H., 1975. Application of Discriminant Analysis to the Geochemical Evaluation of Gossans: in *Geochemical Exploration 1974*, Elsevier Publishing Co., Amsterdam, pp. 219-226.

Closs, L.G. and Nichol, I., 1975. The Role of Factor and Regression Analysis in the Interpretation of Geochemical Reconnaissance Data: *Can. J. Earth Sci.*, 12, pp. 1316-1330.

Davis, J.C., 1973. Statistics and Data Analysis in Geology: John Wiley & Sons, 550 p.

Friske, P.W.B., Day, S.J.A., McCurdy, M.W. and Durham, C.C., 1999. Reanalysis of 1775 Lake Sediments from Regional Surveys on Central Baffin Island: *Geol. Surv. Canada* Open File 3716.

Govett, G.J.S., 1972. Interpretation of a Rock Geochemical Exploration Survey in Cyprus: Statistical and Graphical Techniques: *J. Geochem. Explor*., 1, pp. 77-102.

Irvine, T.N. and Baragar, W.R.A., 1971. A Guide to the Chemical Classification of the Common Volcanic Rocks: *Can. J. Earth Sci.*, 8, pp. 523-548.

Klovan, J.E., 1968. Selection of Target Areas by Factor Analysis: Paper Presented at the Symposium on Decision-Making in Exploration, January 26th 1968; reprinted in *The Western Miner*, February 1968, pp. 44-54).

Klovan, J.E. and Billings, G.K., 1967. Classification of Geological Samples by Discriminant Function Analysis: *Bull. Can. Petr. Geol.*, 15, pp. 313-330.

Koch, G. S. and Link, R.F., (1970), *Statistical Analysis of Geological Data*, Volume 2, Dover Publications Inc., New York, 438 p.

Lavin, O.P., 1976. *Lithogeochemical Discrimination between Mineralized and Unmineralized Cycles of Volcanism in the Sturgeon Lake and Ben Nevis Areas of the Canadian Shield*, Unpubl. M.Sc thesis, Queen's University, 249 p.

Massey, F.J., 1951. The Kolmogorov-Smirnov Test for Goodness of Fit: *J. Am. Stat. Assoc.*, 46, pp. 68-78

McConnell, J.W., 1975. *Geochemical Dispersion in Wallrocks of Archean Massive Sulphide Deposits* : Unpubl. M.Sc thesis, Queen's University, 230 p.

Nichol, I., 1971: Future Trends of Exploration Geochemistry in Canada: *Trans. 3rd Int. Geochem. Explor. Symposium*, C.I.M.M. Spec. Vol. 11, pp. 32-38.

Nichol, I., 1973. The role of computerized data systems in geochemical exploration: *C.I.M.M. Bull.*, 66, pp. 59-68.

Obial, R.C., 1970. Cluster Analysis as an aid in the interpretation of multi-element geochemical data: *Trans. I.M.M.*, 79, pp. B175-B180.

Rose, A.W., Dahlberg, E.C. and Keith, M.L., 1970. A Multiple Regression Technique for Adjusting Background Values in Stream sediment Geochemistry: *Econ. Geol.*, 65, pp. 156-165.

Sinclair, A.J., 1974. Selection of threshold values in geochemical data using probability graphs: *J. Geochem. Explor*., 3, pp. 129-149.

Siriunas, J.M., 1992. Open File Report No. 1 – Geochemical Atlas of Jamaica: Geological Survey Division (Jamaican Ministry of Production, Mining & Commerce), 4 p. + maps.

Sopuck, V.J., Lavin, O.P. and Nichol, I., 1980. Lithogeochemistry as a Guide to Identifying Favourable Areas for the Discovery of Volcanogenic Massive Sulphide Deposits: *C.I.M.M. Bull.*., 73, pp. 281-315.

Stanley, C.R., 1987. PROBPLOT, An Interactive Computer Program to Fit Mixtures of Normal (or Log Normal) Distributions with Maximum Likelihood Optimisation Procedures. Association of Exploration Geochemists

Viljoen, R.P. and Viljoen, M.J., 1970. The Geology and Geochemistry of the Lower Ultramafic Unit of the Onverwacht Group and a Proposed New Class of Igneous Rocks: *in* The Upper Mantle Project (*Geol. Soc. S. Afr. Spec. Pub. 2*), pp. 55-85

Wilkinson, L., 1999. Density Charts : Chapter 3 in SYSTAT 9 -- Graphics : User's Manuals. SPSS Inc., 459 p.