REVIEW AND ANALYSIS OF SEVERAL STATISTICAL METHODS IN GEOLOGY

Submitted to: Dr. Larry Krissek

Submitted by: R. Douglas Wise

Date: November 22, 1983

Abstract

Computers are increasingly being used to solve problems in geology. This report discusses several of the more popular methods of statistical analysis, determines the feasibility of each method, and makes a recommendation as to the practicality of computer use in geology.

Table of Contents

List of Figures

Glossary

Amphiboles - a family of minerals in the mafic group. These minerals are dark in color, crystallize at high temperatures, and contain abundant iron and magnesium.

Batholith - large igneous intrusion of magma resulting in large irregular bodies of rock, usually granite.

Igneous rocks - rocks derived from molton magma.

Metamorphic rocks - rocks derived from previously formed rocks through the processes of heat, pressure, and chemical agents.

Sedimentary rocks - rocks formed at surface temperature and conditions via the accumulation of weathered material from other rocks and organic material.

List of Symbols

Al - Aluminum

Ca - Calcium

CO_2 - Carbon dioxide

Fe^{+2} - Ferrous ion of iron

Fe^{+3} - Ferric ion of iron

H_2O - Water

K - Potassium

Mg - Magnesium

Mn - Manganese

Na - Sodium

P - Phosphorous

Si - Silicon

Ti - Titanium

Foreward

Computers have increasingly been used to solve problems in geology in the last thirty years. They have particularly been used to analyze the large amounts of data obtained during research and experiments. This report discusses the practicality and feasibility of using statistical analysis methods not only for research, but also in terms of companies using these methods as an accurate and effective measure on which to base important business decisions.

Summary

In order to determine the practicality and feasibility of using computers in geology, I reviewed and analyzed several of the more popular statistical analysis methods applied to geological studies. The methods reviewed include the following: principal components analysis, Q-mode and R-mode factor analysis, cluster analysis, and discriminant function analysis.

Computers were determined to be effective and accurate in analyzing the large quantities of data encountered in geological problems. They prove to be very practical in that they allow geologists to make more quantitative studies, without sacrificing quality.

However, I must stress that the results gathered from the computer are only as accurate as the interpretation is correct. The computer is capable of handling mass quantities of data, and the statistical methods applied can greatly reduce the number of pertinent factors, but the accuracy of the solution depends largely on the soundness of the interpretation.

Literature Review

In order to gain some insight as to the relative importance of the use of computers in geology, I reviewed some of the geological literature which was published over the last twenty-three years. As could be expected, the review showed a considerable increase in the use of computers from 1960 to present.

The Journal of Sedimentary Petrology was reviewed. In making the survey, I considered articles from this journal which were published in the following years: 1960 through 1963, 1970 through 1973, and 1980 through 1983--with the exception of Volume 53, Number 4, which was not published at the time of this review.

In revewing the literature, I read the abstracts of the articles and looked for the following "key" words: discriminant function, factor analysis, principal components analysis, cluster analysis, spectral analysis, trend function analysis, Markov chain analysis, multivariate analysis, multiple regression analysis, analysis of variance, and linear regression analysis. Any study which involved the use of a computer, but which didn't contain one of the above key words in its abstract, was unfortunately overlooked. Likewise, studies which used computers merely to speed up basic mathematical calculations were not considered.

The review showed that computers and statistical analysis methods are commonly used in sedimentology studies which involve a large number of samples on which many variables can be measured. These variables are measures such as

grain size, grain shape, and chemical composition, and can be measured on a variety of different sediment samples.

The statistical methods discussed in this paper, principal components analysis, cluster analysis, Q- and R-mode factor analysis, and discriminant function analysis, are five of the more commonly used methods. The importance of these methods, with respect to other statistical methods used in geology, is shown in Fig. 1, which illustrates the findings of this review.

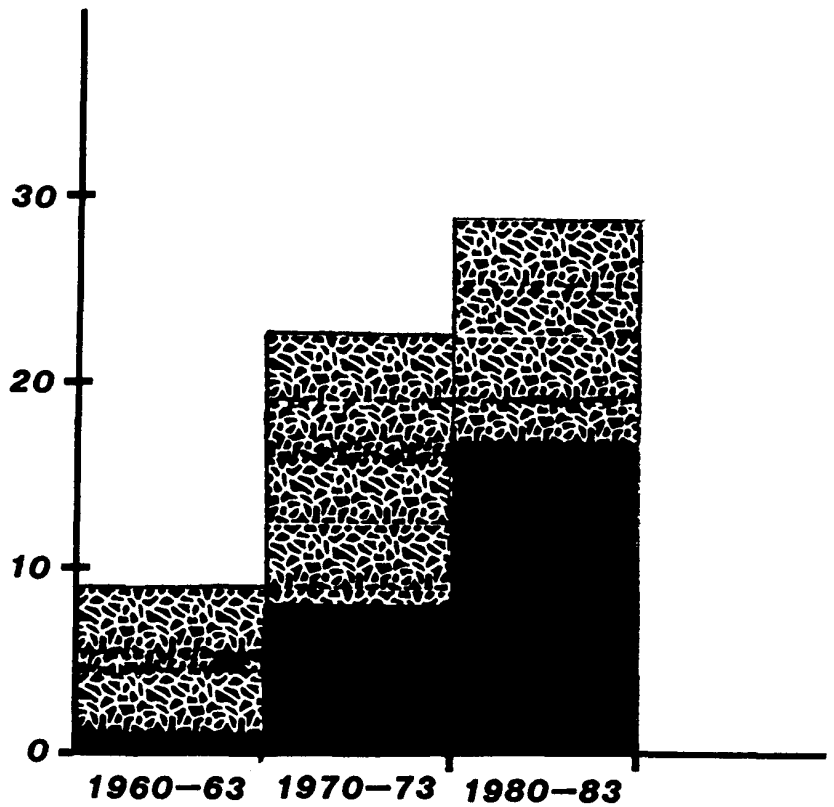


Figure 1- Results of literature review showing the number of articles published in the Journal of Sedimentary Petrology in which computers and statistical analyses were used.

Introduction to Discussion

Geologists have incorporated many statistical analysis methods into their studies over the past few years. This is largely due to the fact that computers have become readily available to most researchers.

Some areas in geology, such as sedimentology, geochemistry, and paleontology, are more appropriate for statistical analyses since they consist of studies involving large amounts of data. Thus, a more quantitative approach can be used in these fields.

This report presents brief discussions and applications of several of the more commonly used statistical methods in geology: principal components analysis, Q-mode factor anlysis, R-mode factor analysis, cluster analysis, and discriminant function analysis. It does so in hopes of giving geologists a general understanding of the methods so that they may determine the appropriateness of a method for their study. Geologists interested in applying one of these methods are advised to first consult one of the several texts published on this subject.

The following discussion of statistical analysis methods was largely taken from Davis (1976).

Factor Analysis

Factor analysis is a widely used multivariate procedure which allows experimenters to work with and analyze large amounts of data while having little insight to the structure of the data. It enables experimenters to reduce the number of variables into factors which represent interrelationships among variables.

It was originally developed by psychologists who were interested in analyzing the results of intelligence tests. Although the tests could not measure mental ability directly, they believed factor analysis could extract the mental factor of intelligence when the test scores were analyzed with other factors such as educational background and testing conditions.

When applying factor analysis in areas in which no or very little knowledge exists as to the relationships between variables, one must deduce the meanings of the resultant factors. This is often impossible since not enough of a pattern exists in the factor representation of the original variables. However, several methods have been introduced which allow the experimenter to express these factors in terms of the original variables which may or may not yield a better interpretation.

These methods require expertise, however, and most researchers should conclude their problem inappropriate for factor analysis if factor analysis alone doesn't yield a plausible interpretation.

Principal Components Analysis

Principal components analysis transforms a number of variables to an equal number of new variables such that the new variables are linear combinations of the original variables. These new variables represent the relative importance of the original variables, and allow the analyst to address the problem with fewer, more significant variables.

Variables are first measured for variance and then compared to each other in the form of a data matrix. Variables which show equally high variances are considered to be positively associated while variables of dissimilar variances are negatively associated. Thus, in a given set of data, the positively associated variables should represent a common underlying factor which accounts for a large percentage of the total variance.

As an example, we can consider a problem in which two variables exist. A component is chosen such that it represents the maximum amount of variance possible between the two variables. This component is represented by a straight line oriented in the data set in such a way that it maximizes the variance (Fig. 2). If the data points are projected into this line at right angles, they show greater dispersion along it than along either of the two original variables. In order to account for the remaining variance, a second line is drawn normal to the principal component axis. The principal component is then compared to the original variables to determine which variable is more closely associated with it and which, therefore, accounts for most of the variance.

This is better illustrated in a problem with many variables. The relationships between variables are determined by means of a matrix of variance. In this case, more than one variable will account for the variance of a principal component. Also, two or three principal components may constitute equally

large amounts of the total variance. In either case, any and all of the original variables which contribute to the variance of the principal components must be considered when interpreting the results.

For example, we can make several measures of shape on any number of objects and use these measurements as variables. Upon analyzing the variables, we may find one transformed variable, or factor, which accounts for seventy percent of the total variance. If this factor is heavily weighted by three variables which all contain a particular measure of shape, such as length or width, this factor could be considered as representing this measurement. Therefore, in the above case, the length or width is the measure which accounts for seventy percent of the difference among the objects.

Similarly, each factor can be treated as a variable representing its percentage of variance in the data set. Although most of the variation can be expressed in terms of a few factors with large magnitudes of variance, the remaining components may give very important detail to the analysis.

R-mode Factor Analysis

R-mode factor analysis is a method used to analyze interrelationships between variables in a data set. It requires the investigator to have some insight as to the nature of the relationships between variables in order for him to reduce the number of factors necessary to solve the problem. The common assumption is that the number of factors is less than the number of variables.

R-mode analysis uses principal components, but the data are standardized such that all the variables are equally weighted. Standardization converts the data to a form where all variables extend essentially over the same ranges.

As a result of this, each variable has a mean of zero and a variance less than or equal to one. This allows the principal components to be converted into factors.

Here again, we can think of a factor as a line oriented in a data set so that it maximizes the variance. The variance being a measure of the deviation from the mean. Factors with variances greater than one, thus greater than the original variables, represent relationships between the variables and are retained for further analysis.

In cases where the raw data are uncorrelated or show very weak correlation, the number of factors with variances greater than one will be over half the number of original variables. This really doesn't reduce the number of variables to a desirable amount of factors for interpreting the results. Since the theory behind factor analysis is the correlation between variables with underlying factors, any problem which requires that a large number of factors be retained to account for much of the variance is probably inappropriate for factor analysis.

Once a representative number of factors has been determined, attempts are made to show better correlation between them. The factor axes are rotated through a plot of the original variables in an attempt to further maximize the variance. The position of the variables is unchanged with respect to each other, but the factor axes are rotated to a position such that the projections of the variables onto the axes are near the mean or near the maximum variance. This rotation results in the factors having few significant variables, variables which are near the extremities, and many insignificant variables which are near the mean. This is illustrated in Fig. 3.

An application of R-mode factor analysis in geology is its use in determining the economic worth of coal. A number of variables which affect the value of coal can be measured on samples from different coal beds. Factor analysis may reduce the number of variables into a few factors which could then be used to rank the coals according to quality.

Q-mode Factor Analysis

Q-mode factor analysis is a method used to arrange a group of samples into a meaningful order so that relationships between samples can be determined. It is particularly valuable in cases where there are a large number of objects and also when there is little a-priori knowledge of the interrelationships between samples.

Q-mode analysis is designed to show interrelationships between objects, as compared to R-mode analysis which analyzes interrelationships between variables. Although R-mode analysis does provide a certain degree of similarity between objects, it does not adequately measure inter-object similarity. In other words, the correlation of variables may not be the best criterion for determining the similarity between two objects.

Measurements, such as chemical composition or measures of shape, are made on all objects. A measure of similarity is computed between every pair of samples, and the results are arranged in a standardized matrix form.

To simplify, we can represent the objects as points plotted graphically on the basis of two measured variables. The degree of similarity can be expressed in terms of the cosine of the angle between the object vectors (vectors drawn from the intersection of the variables, zero, to their respective points,

objects). The value of similarity can range from zero for object vectors ninety degrees apart to one for object vectors which are aligned. Thus, objects which differ in the amounts of constituents, but which are proportionally similar, are regarded as being identical. This is illustrated in Fig. 4.

In Q-mode analysis, principal components are determined through interrelationships between the objects. These principal components, or factors, are rotated through the data set in order to achieve maximum variance. The rotated factors can be interpreted as "idealized" end members which represent the most dissimilar objects. Other samples in the set can then be classified in a gradient from one end member to another. Once the samples have been classified, the researcher must interpret the geological meaning of the classification.

Q-mode analysis can be applied to geology in a case where one desires to classify igneous rock samples. Chemical analysis of the rocks will yield variables to be used in the construction of the data matrix. Q-mode analysis will separate the most different samples and group together the most similar samples on the basis of their chemical composition. This gives a gradational classification of the rocks from one end member to another.

Cluster Analysis

Cluster analysis is a method which, like Q-mode analysis, measures the similarity among samples and places them into more or less homogeneous groups. No or little a priori knowledge of the interrelationships between samples is required since the relationship between groups will be revealed through the classification of cluster analysis.

Several techniques for clustering samples have been developed; however, a description of each is beyond the scope of this paper. I will consider a general technique of clustering based on the similarity coefficient.

In this method, samples are compared and their similarities are measured by a technique which closely resembles the one used in Q-mode factor analysis. The similarity measurements are commonly standardized into a matrix form of variance and covariance.

Samples with the highest mutual similarity are placed into groups. Once two samples have been grouped together, their similarity measures are recalculated into a single similarity coefficient, and the group is treated as one sample. Groups are then associated with mutually similar groups until all samples have been classified.

A problem arises with the number of different clustering methods available in that they all yield slightly different results. This may lead experimenters to choose the method which yields the most satisfactory results with their data. Thus, a certain degree of subjectivity may be introduced to an objective study.

Cluster analysis can be applied equally as well as Q-mode analysis in the previously discussed example of classifying igneous rocks.

Although cluster analysis and Q-mode analysis are essentially the same, it is worth noting that the former is much less costly in terms of computing time.

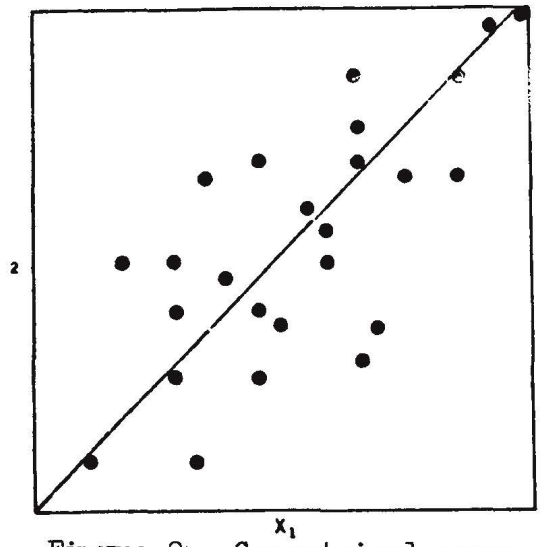
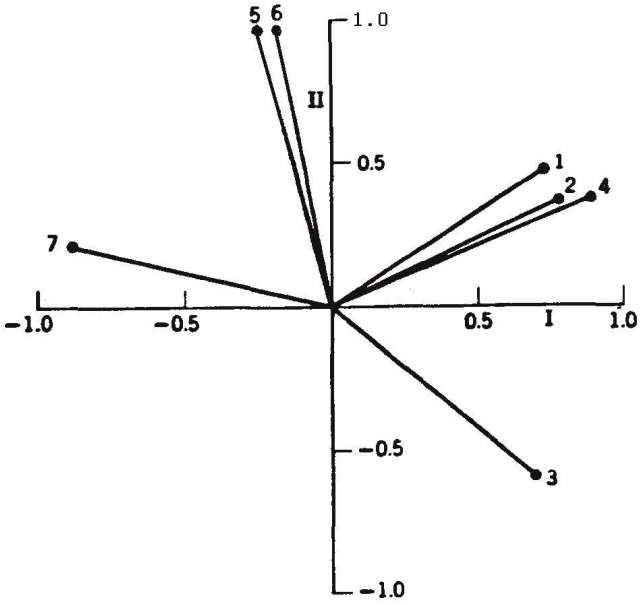


Figure 2 - Geometrical representation of a principal component. Variance of the data set is greater along this line than along either of the original variables.(Davis,1976)



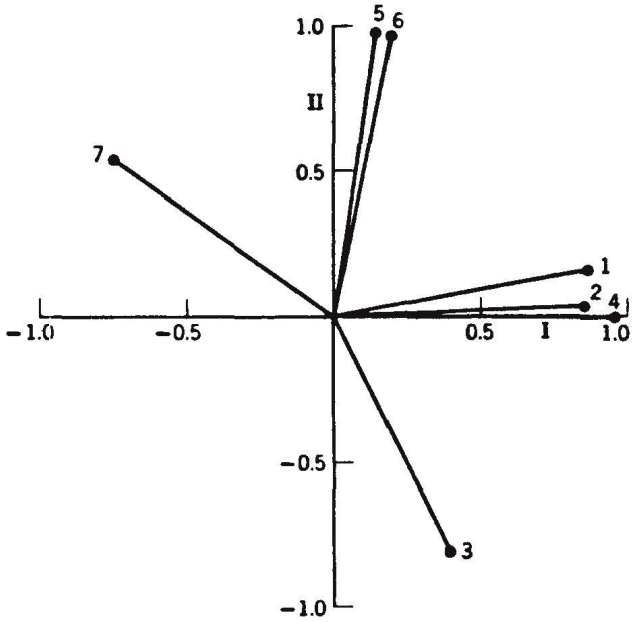


Figure 3- Plot of seven original variables on two factors (top), and after factor rotation (bottom). (Davis,1976)

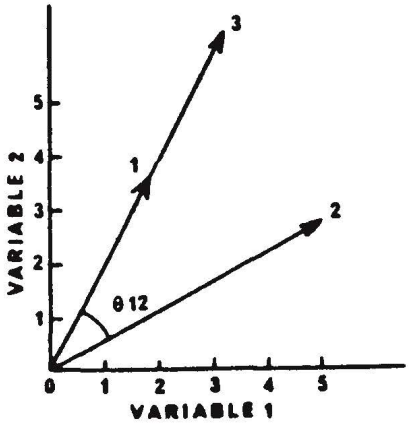


Figure 4- Plot of three objects on two variables. Degree of similarity is measured by the cosine of θ. Cosine θ for objects 1 and 2 is .838. Cosine θ for objects 1 and 3 is 1.00. (Joreskog, et. al, 1976)

Discriminant Function Analysis

Discriminant function analysis is one of the most widely used multivariate procedures in geological sciences. It breaks down a multivariate problem into a problem which involves only one variable. It requires a priori knowledge of the interrelationships between samples since the number of different groups to be used must be set prior to analysis.

In discriminant function analysis, one new variable is determined which maximizes the variance between groups. Each original sample is defined as belonging to one of the set groups, and no new groups are established for slightly different samples.

As seen in Fig. 5, the linear discriminant function defines a line oriented in a multivariate data set such that it maximizes the variance between groups while minimizing the dispersion within each group. This line is called the transform, and the position on it where a sample is projected is the discriminant score, or transformed variable.

The transform determined in discriminant function analysis is usually not a variable which can be measured on a sample, but a discriminating way of presenting the original variables. After the original measurements have been made on new samples, the samples are placed into a group depending on their discriminant scores.

Discriminant function analysis can be used in geology in a problem which attempts to group new samples on the basis of similarity with previously analyzed samples. An example of this is when an analysis of trace elements is made on stream sediments from several areas. One area could be known to contain abundant heavy metal deposits, while another area may have been

mined and determined not to contain any valuable deposits. Discriminant function analysis may yield a transform which separates the two areas. Samples taken from streams which drain unprospected areas can then be tested against the discriminant function to determine the economic worth of their deposits.

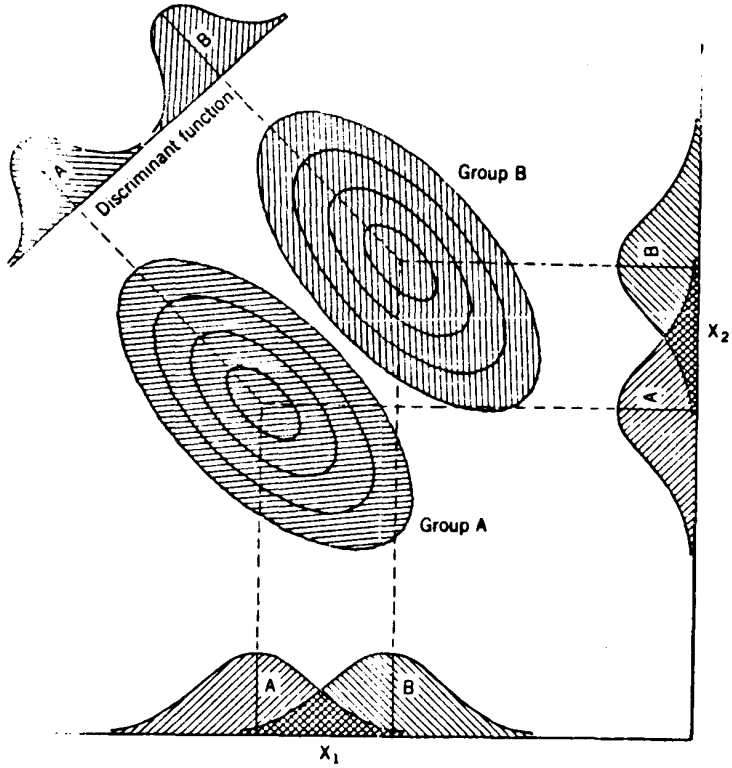


Figure 5- Discriminant function line showing maximum variance between groups A and B although both groups overlap on the two original variables.(Davis,1976)

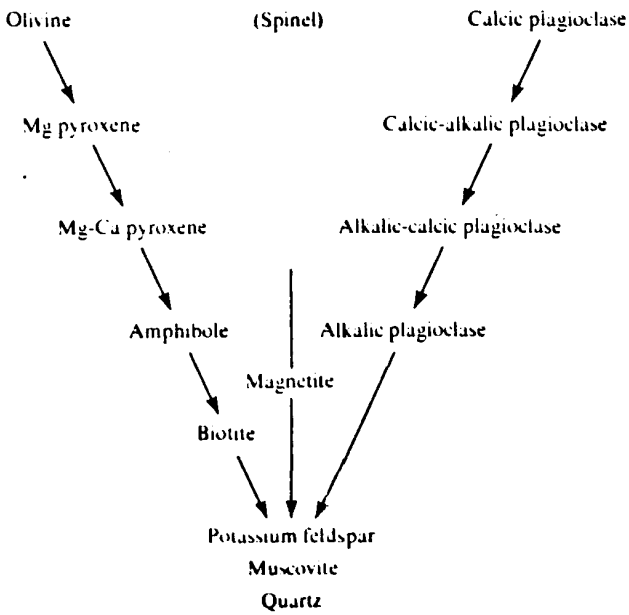


Figure 6- Bowen's reaction series. Minerals near the top crystallize at higher temperatures and are rich in Mg and Ca, while minerals near the bottom crystallize at lower temperatures and are rich in K and Si.(Ehlers and Blatt,1982)

Examples of Methods

This section discusses some actual geological applications of factor analysis methods. The number of applications of geological factor analysis extends far beyond the scope of this section, and I am merely presenting those examples which illustrate the practicality of using factor analysis to solve problems in geology; however, the following examples do represent methods which have been used in various fields of geology. Joreskog et al. (1976) gives a more detailed description of these and other examples in his text.

Problem

Igneous rocks are often chemically complex. Amphiboles may consist of up to twelve different ingredients. These ingredients are chemical elements, ions, and molecules. In addition to the number of variables, some of the ions and elements are capable of substituting for each other within the crystal structure. Thus, minerals with the same chemical composition may vary in structure, and structurally similar minerals can vary in chemical composition as well. The problem being that chemical analysis alone won't suffice in identifying the mineral structure, on which many properties are dependent.

Solution

Saxena (1969) applied principal components analysis to this problem in amphiboles. Using the ions, elements, and molecules present in amphiboles as variables, his analysis yielded five principal components which comprised eighty percent of the total variance. All five variables were interpreted as reflecting substitutional relationships between certain elements. Saxena succeeded in reducing the number of variables, necessary to differentiate between amphiboles, from twelve to five, less than half of the original number.

Problem

Determining environments of deposition is a major concern to sedimentologists, and establishing sedimentary criteria for making these determinations is vital. For clastic sediments, various criteria based on grain size characteristics have been used.

Klovan (1966) and, later, Solohub and Klovan (1970) applied Q-mode factor analysis to this problem in an attempt to separate sediment samples into distinct environments based on the grain sizes contained in the samples.

Solution

Each sediment sample was sieved into ten size classes. The size classes were variables plotted against the sediment samples. Q-mode factor analysis yielded three eigenvalues to account for 97.5% of the total variance.

The three eigenvalues were plotted as end members of the sediment samples, each representing a different characteristic of the samples: Factor I - bimodal frequency distribution, Factor II - poorly sorted, fine-grained sediments, and Factor III - well sorted, coarse-grained sediments. The plot of

the entire sample revealed a cluster of samples near each Factor, but also a continuous plot of samples between each Factor. Although each Factor represents a type of sediment deposited under the influence of a different depositional process, the continuous trend of samples between Factors fails to distinguish any "grain-size facies", environmentally distinct groups, within the samples considered. It does, however, allow one to identify the type and the roles of certain depositional processes which have formed a particular sediment sample.

Problem

Detailed chemical analysis of metamorphic rocks is often made difficult because of the metamorphic alteration of primary composition. Metamorphic petrologists need to unravel these complex effects of alteration and better understand the processes which determine the ultimate composition of metamorphic rocks so as to gain a fuller knowledge of this field.

Solution

Davis et al (1974) used correspondence analysis, a combination of R-mode and Q-mode methods, in an attempt to solve this problem. They analyzed seventy-five rocks collected from a meta-diorite batholith near Val d'Or, Quebec for twenty-two major, minor, and trace elements.

An analysis made using thirteen major and minor elements yielded the following results. Two factors accounted for eighty percent of the total variance. The factors represented positive and negative relationships between elements in the rock samples. Factor II consisted of an arrangement of elements--Mg, Ti, Fe^{+2}, Ca, and Mn at one end, through Fe^{+3}, Al, P, Si, Na, and

K at the other end--which closely resembles the expected order in a differentiation series as seen in Fig. 6.

Rock samples plotted on this axis are accordingly grouped into their positions in the mafic to silicic differentiation series (Joreskog, 1976). Factor I was heavily weighted with respect to CO_2 and H_2O. Rocks samples plotting towards the H_2O and CO_2 end are interpreted as having suffered higher degrees of metamorphism than those tending along the K and Mg line. Thus, with the relative degree of metamorphism known, relationships between the extent of alteration and the ultimate chemical composition can be better investigated.

Application of Methods

In the last section, I presented a somewhat detailed discussion of a few examples in which statistical methods have been used. This section will discuss some applications of statistical analysis which were not encountered in the literature, but seem very appropriate for this type of analysis.

An interesting use of R-mode factor analysis or principal components analysis would be investigating the underlying factors of the degree of compaction of sandstones. Sandstones from many different sources could be measured for the following variables: grain size, grain shape, composition of the matrix, composition of grains, and the precentage of matrix. A relative scale of sorting could be established through thin section inspection of the samples prior to anaysis.

These variables are all very important factors in the degree of compaction, but statistical analysis may yield factors which were previously overlooked. Perhaps a relationship exists between grain shape and matrix composition such that rocks with irregularly shaped grains are more compact when their matrix is of a different composition than the grains. This could possibly be interpreted as being the result of reworking of the sediment by the agent supplying the matrix. This reworking would tend to rearrange the grains into a more compact state.

A practical application of Q-mode analysis or cluster analysis would be in determining the ultimate source of the sediment supplied to a particular depositional environment.

Suppose we took core samples from a particular sandstone bed, and measured the following variables: grain size, degree of sorting, chemical composition of matrix, and chemical composition of grains. If we took our

samples from a circular area of radius one mile, then any sample taken should represent a combination of the processes which acted on that area during deposition.

Statistical analysis will ideally yield a classification of the samples based on similarities between them. Perhaps a classification will be produced which recognizes three end members. For instance, End Member I may consist of coarse grained, well-sorted sands with minor heavy minerals; End Member II may consist of fine grained, well-sorted sands; and End Member III may consist of poorly sorted, coarse to fine grained sands.

By placing these end members in their respective sample positions on the map of the drilled area, and noting the position of the other samples according to the classification, some idea as to the depositional environment may be presented. If the end members were to plot in a more or less straight line across the map area, from End Member I to End Member III to End Member II, this may be interpreted as a nearshore environment. End Member I would represent a beach environment, End Member II would be an offshore deposit, and End Member III could be indicative of an area close enough to shore to be affected by periodic storms which deposit the coarser sediments. By noting the position of the end members and the samples on the map, an idea as to the size of the depositional shelf could be gathered.

The number of applications of statistical analysis methods in geology is unlimited. Any study which involves an appreciable amount of data can probably be easily solved, or at least broken down into fewer factors, through the use of statistical analysis.

Conclusion

Five statistical analysis methods, principal components analysis, Q-mode factor analysis, R-mode factor analysis, cluster analysis, and discriminant function analysis, were reviewed, and the following conclusions were made:

1. Statistical analysis is a useful technique for solving problems involving large amounts of data in that it allows one to break down a multivariate problem into one of fewer factors.

2. Many problems exist in geology such that the use of computers and statistical analysis is both a feasible and practical means of solving them. These problems are encountered in fields such as sedimentology, geochemistry, and paleontology, and involve variables ranging from size measures of sediment grains or fossils to chemical constituents of rocks.

References

1. Davis, John C., 1976, Statistics and Data Analysis in Geology: John Wiley & Sons, Inc., New York, pp. 442-533.

2. Davis, M., Campiglio, C., and Darling, R., 1974, Progress in R- and Q-mode analysis: Correspondence analysis and its application to the study of geological processes, Can. J. Earth Sci., 11:131-146.

3. Ehlers, Ernest G., and Blatt, Harvey, 1982, Petrology: igneous, sedimentary, and metamorphic., W. H. Freeman and Company, San Francisco, pp. 148-149.

4. Joreskog, K. G., Klovan, J. E., and Reyment, R. A., 1976, Geological Factor Analysis: Elsevier Scientific Publishing Company, Amsterdam.

5. Saxena, S. K., 1969, Silicate solid solutions and geothermometry, 4. Statistical study of chemical data on garnets and clinopyroxene, Contrib. Mineral. Petrol, 23:140-156.