COMPUTER APPLICATIONS IN THE EARTH SCIENCES:

COLLOQUIUM ON CLASSIFICATION PROCEDURES

Edited by

DANIEL F. MERRIAM

COMPUTER CONTRIBUTION 7
State Geological Survey
The University of Kansas, Lawrence
1966

Editor's Remarks

The Colloquium on Classificational Procedures hopefully will be the first of a series of such meetings on timely subjects in computer applications in the earth sciences. Interdisciplinary in nature, the colloquium is designed for maximum involvement by all participants. A colloquium, by definition, is a mutual discourse or a conversation, especially a somewhat formal one; a conference. That is exactly what we wish this meeting to be - a meeting of colleagues of approximately the same professional status to exchange ideas in an area of common interest.

The colloquium is structured to allow the latest scientific results to be incorporated into the presentations. Therefore, in some instances oral presentations may not match the written ones. This is the way of the computer.

In essence, presentations are to serve as a focal point for discussion, or a point of departure for different views; a framework on which to build or tear down. Many of the ideas will be new, others old, and some will be disguised. Regardless of presentations, however, maximum benefit of the meeting will be gained by those who actively participate.

The program treats many aspects of classification - problems, techniques, pilot studies, and proposed yet essentially untried solutions. The speakers bring to the meeting a diversity of backgrounds and experience. They are well qualified to lead discussions in their specialities. Hopefully, everyone will benefit from their interaction with these leaders.

Initial planning of the program included only those associated with the Survey or University of Kansas. Response to the announcement was surprisingly gratifying and in the future the scope of the meetings in general should be widened to include participants from other institutions.

Many people have helped with the arrangements for this Colloquium. Dr. Floyd W. Preston has been foremost in creating a favorable atmosphere in which the hosting organizations, Kansas Geological Survey, Department of Chemical and Petroleum Engineering, and Department of Entomology at The University of Kansas, could function. His efforts and encouragement are most appreciated. All of the papers presented in this Computer Contribution 7 have been read, typed, and edited by Mrs. Nan Carnahan Cocke assisted by Mrs. Alberta E. Bonnett. Mr. John C. Davis and Dr. Richard A. Reyment also assisted in reading some of the manuscripts.

Indeed, conferences of this type seemingly serve a definite purpose and fill a particular need. By co-sponsoring the Colloquium, the Survey is fulfilling yet another obligation to industry and the profession, that of disseminating information of current and immediate interest and providing the avenue of exchange of information between people with mutual problems.

Those giving formal presentations are:

Ernest E. Angino, Chief, Geochemistry Section, Kansas Geological Survey, The University of Kansas
T. P. Burnaby, Lecturer in Geology, The University of Keele (UK)
G. Dalke, Research Assistant, CRES, The University of Kansas
J. C. Davis, Geologist, Kansas Geological Survey, The University of Kansas
J. C. Griffiths, Head of Department, Geochemistry and Mineralogy, Pennsylvania State University, and Consultant to the Kansas Geological Survey, The University of Kansas
J. W. Harbaugh, Professor of Geology, Stanford University, and Consultant to the Kansas Geological Survey, The University of Kansas
John Imbrie, Professor of Geology, Columbia University, and Consultant to the Kansas Geological Survey, The University of Kansas
R. L. Kaesler, Assistant Professor of Geology, The University of Kansas
G. L. Kelly, Assistant Professor of Electrical Engineering, The University of Kansas
W. C. Krumbein, Professor of Geology, Northwestern University, and Consultant to the Kansas Geological Survey, The University of Kansas
M. N. McElroy, Geologist, Humble Oil and Refining Company
D. F. Merriam, Chief, Geologic Research Section, Kansas Geological Survey, The University of Kansas
C. O. Morgan, Geologist, Groundwater Branch, U.S. Geological Survey, The University of Kansas
M.G. Pitcher, Supervising Research Geologist, Continental Oil Company, and Associate Editor, Computer Contribution Series
F. W. Preston, Professor and Assistant Chairman, Department of Chemical and Petroleum Engineering, and Consultant to the Kansas Geological Survey, The University of Kansas
R. A. Reyment, Visiting Research Scientist, Kansas Geological Survey, The University of Kansas, and Professor of Biometrics, University of Stockholm, Sweden
F.J. Rohlf, Associate Professor of Entomology, University of Kansas
D.S. Simonett, Associate Professor of Geography, The University of Kansas
R.R. Sokal, Professor of Statistical Biology, The University of Kansas
O.T. Spitz, Chief, Operations Research Section, Kansas Geological Survey, The University of Kansas

COMPUTER APPLICATIONS IN THE EARTH SCIENCES:

COLLOQUIUM ON CLASSIFICATION PROCEDURES

Edited by

DANIEL F. MERRIAM

1966

CONTENTS Page

Introduction, by W. W. Hambleton. ii

Programming the discriminant classification function for small computers, by John C. Davis. . . 1

Homogeneity of covariance matrices in relation to generalized distances and discriminant functions, by R. A. Reyment 5

Factor analysis of a pooled cross-product matrix, by Johm Imbrie 10

Classification of map surfaces based on the structure of polynomial and Fourier coefficient matrices, by W. C. Krumbein 12

Application of color-combined multiple polarization radar images to geoscience problems, by David S. Simonett 19

Implementation of pattern recognition techniques as applied to geoscience interpretation, by George W. Dalke 24

A factor analytic scheme for grouping and separating types of fossils, by Max Pitcher 30

Classification of subsurface localities of the Reagan Sandstone (Upper Cambrian) of central and northwest Kansas, by Roger L. Kaesler and Marcus N. McElroy 42

Application of disciminant functions as a classification tool in the geosciences, by John C. Griffiths 48

Application of pattern analysis to the classification of oil-field brines, by Ernest E. Angino and C. O. Morgan 53

Classification in quantitative oil-exploration decision making, by John W. Harbaugh 57

Two-dimensional power spectra for classification of land forms, by Floyd W. Preston. 64

Distribution-free quadratic discriminant functions in paleontology, by T. P. Burnaby . . . 70

Applications of pattern recognition techniques to geoscience interpretation, by G. L. Kelly (abs.) 78

Classification of contour maps, by O. T. Spitz and D. F. Merriam (abs.). 79

INTRODUCTION

Several years ago, the Kansas Geological Survey assumed several commitments that, at the time, seemed innovative and imaginative, and presaged a completely new role for us. The first commitment was to a major effort in high-speed data processing and computation. It stemmed from recognition that we were becoming immersed in large quantities of information, that much staff time was committed to retrieval of this information, and that analysis of information was lagging. As a consequence, we have been in the forefront of program development for data storage and retrieval systems, and studies of applications of computer techniques to stratigraphic, structural, petroleum hydrologic, and economic problems. For example, we have used trend-surface studies for structural and stratigraphic analyses. The methods of power spectrum analysis are being applied to a variety of problems. Simulation models are beginning to figure prominently in our work and Geological Survey publications bear titles dealing with such subjects as coefficients of association, factor and vector analysis, and cross-associations of nonnumeric sequences.

A second commitment was to foster an interdisciplinary environment for study of earth science problems in recognition of the conclusion that although the problems of the separate disciplines are different, the patterns of solution are often the same. In implementing this commitment, we have already established a Computer Applications Laboratory in collaboration with the Department of Chemical and Petroleum Engineering in order to establish an interdisciplinary environment for study of engineering and scientific problems related to natural resources. Furthermore, we have worked closely with scientists from many other disciplines in examining tests for similarity in information and have looked at statistical methods involving correlation functions, factor analysis, analysis of variance, and other techniques. These studies, in turn, have led to greater interest in validity tests of techniques of analysis and the design of data collection systems, and the problems of classification.

This Colloquium on Classificational Procedures is a further implementation of our commitments, for it deals with an important problem of a distinctly interdisciplinary nature - a problem that depends for its solution upon computer techniques. Great significance should be attached to the fact of sponsorship by geologists, entomologists, chemical engineers, and petroleum engineers. Each of these disciplines is concerned with problems of organization of information, clustering, tests for similarity, pattern recognition, validation of techniques - all of paramount importance in classification. Recognition of the interdisciplinary character of such problems can be seen in the disciplines represented by the principal contributors to the colloquium: included are geologists with structural, stratigraphic, and paleontological leanings, electrical engineers, chemical and petroleum engineers, entomologists, biometricians, and geographers. The participants have been concerned with applications of pattern recognition, multivariate analysis, power spectrum analysis, factor analysis, orthogonal coefficients, quadratic discriminant functions, and the like to problems of oil exploration, paleontology, taxonomy, radar investigations, stratigraphic analysis and other areas of interest. In such a Colloquium they see opportunity to transfer the results of investigations from one discipline to another. Furthermore, the surpassing importance of these interdisciplinary discussions, in the framework of continuing education, can be recognized in the affiliations of other conferees. Major petroleum company research laboratories, universities, oceanographic research groups, and state and federal geological surveys, both in the United States and abroad, are represented. Hopefully, then, this Colloquium will strongly enhance our commitments to computation, giving new insights to earth science problems in an interdisciplinary environment.

Lawrence, Kansas
November 12, 1966

William W. Hambleton,
Associate Director
State Geological Survey of Kansas.

PROGRAMMING THE DISCRIMINANT CLASSIFICATION FUNCTION FOR SMALL COMPUTERS

By

John C. Davis
Kansas Geological Survey

INTRODUCTION

Many classification techniques have been developed which may be applied to geological problems. Most of these procedures originated in other fields, particularly in biology and psychology, and have only recently been used in earth sciences. The geologic profession has not been statistically oriented and geologists have not utilized the services of statisticians to the same extent as biologic and behavioral scientists. Consequently, biometricians and statistical anthropologists are not uncommon, but a geo-statistician is still a rare individual. Most are concentrated at a few major universities and research centers in the United States.

Most of these centers of concentration are equipped with extensive computing facilities, which has lead to some undesirable side effects. Because programming in general tends to follow a corollary of Parkinson's Law, which states that any given program will expand to fill all the available core of the available computer, geologic programs have become larger, more sophisticated, and increasingly complex. In many cases, they are so large and complex that they can only be run at the centers where they originated. Needless to say, this has not helped the spread of statistical techniques throughout the profession.

There has been remarkable progress in the development of geologically oriented computer programs since 1956, with little complementary progress in the application of computers in the profession, at either the industrial or academic level. This can be blamed partly on a lack of computer and statistical training in all but the most recent graduates. This lack of use can also be blamed on inadequate physical facilities needed for routine operation of most geologic programs available today.

The educational factor is gradually being solved by retraining programs and symposia such as this. The equipment factor can be attacked either by wider dissemination of large computers or by compressing existing programs to the point that they can be used on existing hardware. Considering the expense of establishing and maintaining computer centers, the second alternative is the only practical solution for the majority of smaller universities and industrial offices.

A variety of compact programs have been developed and published by the Kansas Geological Survey for small computers. These include a trend-surface program (Sampson and Davis, 1966), discriminant analysis program (Davis and Sampson, 1966), a time-series package, and a response surface or "hypersurface" program. These were developed for an IBM 1620 computer having 20K bits core storage. This is one of the most common machines available and is probably the smallest computer routinely used for scientific computation. Programs operable on this machine should be useable on most other computers having equivalent software. Computers having this or equivalent configuration are available at most small colleges, at banks and EDP firms in many small cities, and at district or regional offices of many oil companies. Machines of this size also are very common overseas.

Response to these initial small computer publications has been gratifying, suggesting that field geologists will use computers and statistical techniques if the means for their use are readily available. These same geologists rapidly lose interest if their work must be done at a distant computing center with the attendant possibilities of mistakes and delay. Computers seem to be a remote and rather academic sort of thing until the geologist actually processes his own data and obtains results from his own work.

Most statistical techniques being utilized by geologists can be programmed for small computers. This sometimes involves being content with less-than-optimum speed, simplified output, or multi-pass programming, but this is a small price to pay if the alternative is no programming at all. Because these programs must be highly efficient in terms of core requirements, they are best written by a professional programmer under the supervision of a geologist. The multiple discriminant analysis program published by the Kansas Geological Survey (Davis and Sampson, 1966) is outlined as an illustration of small computer programming.

MULTIVARIATE DISCRIMINANT ANALYSIS

Samples of unknown origin may be classified into previously defined populations by use of a multiple discriminant function. A number of these functions are available, including simple linear discriminants, curvilinear discriminants, and discriminants of more than two populations. Programs have been written for all of these. Most familiar to geologists are the programs of the BMD biomedical collection from UCLA (Dixon, 1965) and Casetti's ONR program published by Northwestern University (1964). Discriminant functions have been used to distinguish marine from fresh water sandstones, barren ground from uranium ore, to define tectonic settings of sandstones, depositional environments of limestones, structural distribution patterns of volcanics, and to examine controlling parameters in sandstone cementation. Discriminant functions have been applied by numerical taxonomists working on protozoa, ostracodes, echinoids, molluscs, frogs, coyotes, and humans. This list is by no means inclusive. Other applications are given in Krumbein and Graybill (1965, p. 365-367), and Miller and Kahn (1962, p. 277). The list of successful applications is sufficiently impressive to suggest that discriminant analysis is one of the most powerful statistical techniques available to the geologist.

A multivariate observation may be considered as a point in multi-dimensional space, just as a two-variate observation may be represented as a point defined by the intersection of an X and Y axis in a plane. A multivariate sample has the form of an ellipsoidal cloud of points, whose dimensions are defined by the amount of variance within each of the parameters. A second sample from a different population presumably would appear as another distinct cluster of points. Some of the variables may overlap, causing the two clouds to merge in certain directions, but in other directions the two should be statistically separable if the two populations are truly different. The problem of simple discriminant analysis involves finding the linear combination of variables that defines a multi-dimensional plane efficiently separating the two clusters. This is done by a least-squares procedure similar to that used in multiple regression. The distinctness of the two clusters can be analyzed by measuring the "distance" between their multivariate means. Once this distinctness has been established and the separatory plane computed, additional unknown samples can be assigned to their proper group by computing their multivariate "location." If they fall on one side of the discriminant plane, they are assigned to the population cluster that also occurs on that side. If they fall on the other side, they are classified with the second population.

The discriminant function has the form

$$R=\lambda_a A + \lambda_b B + \lambda_c C + \dots + \lambda_k K .$$

Constants in this equation may be found by solving a series of simultaneous equations of the form

$$\begin{array}{ll} SS_A\lambda_a + SS_{AB}\lambda_b + SS_{AC}\lambda_c + \dots + SS_{AK}\lambda_k & = \Delta\bar{A}(n_1+n_2-2) \\ SS_{AB}\lambda_a + SS_B\lambda_b + SS_{AC}\lambda_c + \dots + SS_{AK}\lambda_k & = \Delta\bar{B}(n_1+n_2-2) \\ \vdots & \quad\vdots \\ SS_{AK}\lambda_a + SS_{BK}\lambda_b + SS_{CK}\lambda_c + \dots + SS_K\lambda_k & = \Delta\bar{K}(n_1+n_2-2) \end{array}$$

where $SS_A \dots SS_K$ are variance estimates obtained by pooling the sums of squares of the two populations. The covariances $SS_{AB} \dots SS_{(K-1)K}$ are obtained by pooling sums of cross products. The differences between the means of variables from the two samples are $\Delta\bar{A} \dots \Delta\bar{K}$. By pooling the two groups in this manner and equating the resulting simultaneous equations to the mean differences, a plane is computed which bisects the space between the clusters. Because the plane is fitted by least squares, deviations from it are a minimum for all points. If the space between the clusters is bisected, probabilities of assigning samples from Group 1 to Group 2 or those from Group 2 to Group 1 are equal.

When programming a procedure such as discriminant analysis on a small computer, several steps must be taken to obtain as compact a program as possible. The most obvious suggestion is to never store data in core even if it is required more than once in the program. Data should be rerun or sent to auxillary storage facilities. Sums of squares and cross-products are generated and summed as read in, and are stored in a series of small matrices. A discriminant analysis program for twenty variables requires two 2 X 20 matrices for storing sums and sums of squares, and a 20 X 20 matrix for storing cross-product sums. After final summation, the sums of cross-products are converted into covariance estimates and re-inserted into the matrix. This matrix is actually set up as a 20 X 21 matrix and the final column is filled with the differences between variable means of the two populations. Sums and sums of squares from the two small matrices are used to calculate variance estimates which are inserted into the diagonal position of the large matrix. The matrix is then in a form identical to the simultaneous equation set. This process requires approximately 20 FORTRAN statements.

Matrices should be inverted by methods which do not require establishment of an identity matrix. This may be done by procedures such as the Gauss-Jordan method or similar pivotal condensation techniques. Golden's (1965) version of the Gauss-Jordan inversion requires approximately 20 additional FORTRAN statements and involves about $K^3/3$ multiplications or less than 2,700 in the case of twenty variables. Computations are, of course, performed sequentially and do not require inordinate amounts of computer space.

The discriminant function is now essentially complete, since the last column of the inverted matrix is the discriminant coefficient array. The discriminant values (R_1 and R_2) and discriminant index (R_0) can be calculated by multiplying each coefficient by the mean value of that variable from each group, and by the combined group mean:

$$R_1 = \lambda_a \frac{\Sigma A_1}{n_1} + \lambda_b \frac{\Sigma B_1}{n_1} + \ldots + \lambda_k \frac{\Sigma K_1}{n_1}$$

$$R_0 = \lambda_a \frac{\Sigma A_1 + \Sigma A_2}{n_1+n_2} + \lambda_b \frac{\Sigma B_1 + \Sigma B_2}{n_1+n_2} + \ldots + \lambda_k \frac{\Sigma K_1 + \Sigma K_2}{n_1+n_2}$$

$$R_2 = \lambda_a \frac{\Sigma A_2}{n_2} + \lambda_b \frac{\Sigma B_2}{n_2} + \ldots + \lambda_k \frac{\Sigma K_2}{n_2}$$

The "distance" between the cluster means may be calculated by inserting the mean differences into the coefficient array, giving Mahalanobis' generalized distance.

$$D^2 = \lambda_a \Delta\overline{A} + \lambda_b \Delta\overline{B} + \lambda_c \Delta\overline{C} + \ldots + \lambda_k \Delta\overline{K}.$$

Various error measures also may be calculated. For example, a test for the significance of the two multivariate means may be devised from Mahalanobis' distance.

$$F_{K,\, n_1+n_2-K-1} = \left[\frac{n_1 n_2}{(n_1+n_2)\,(n_1+n_2-2)}\right] \left[\frac{n_1+n_2-K-1}{K}\right] \times D^2$$

Computation of these additional measures is very straightforward once the discriminant coefficients have been found. Additional program features may be desirable. For example, this program contains a ten-statement option which allows additional data cards to be submitted to the computer. Variables for each additional observation are placed in the discriminant function and the discriminant value is calculated for the observation.

The remainder of the program consists of input/output instructions and FORMAT statements. Wherever possible, formats should be uniform so every input/output command does not require a separate FORMAT statement.

No doubt many desirable options and additions could be made. However, the program, as described, consists of 101 FORTRAN statements which essentially fill the entire available core of a 20K IBM 1620. In fact, the program will not compile with IBM FORTRAN, nor with conventional PDQ FORTRAN. It will compile with a version of PDQ FORTRAN which does not have reread (IBM User's Group Program 2.0.031). This illustrates the importance of adequate software when working near the mechanical limits of a computer. What may be excessive or impossible with one programming system may be entirely feasible with another.

When the program has been trimmed to the absolute limit and still will not compile, a programmer may resort to segmenting. Some types of programs may be packaged into a series of compatible independent programs, producing output that may be utilized by other programs in the package. The Survey has in preparation a time-trend analysis package of this type. It consists of Fourier, polynomial, and sliding-averages data smoothing programs, and a cross-correlation, cross-association, auto-correlation analysis program. Data may be treated by any of the preparatory programs before being submitted to the analysis program. The preparatory programs also may be used independently for function analysis and curve simulation.

Certain programs cannot be segmented into meaningful independent programs. The IBM 1620 trend-analysis program by Sampson and Davis (1966) is of this type. The program is arbitrarily terminated at the end of matrix computations, the matrix is retained in core in a COMMON field, and the first part of the program is replaced with a new program which is entered as a second pass. This procedure is not annoyingly unwieldy, as data must also be re-entered at this time for computation of residuals and error terms. The process of segmentation theoretically can be extended indefinitely. For example, Sampson contemplated reducing the rotating factor analysis procedure by Manson and Imbrie (1964) for operation on the 1620. It was estimated that the small computer version would require six program passes. In this case, the advantages of local processing have obviously become outweighed by programming complexities!

By a process of condensation and, where necessary, segmentation, many of the current computational techniques can be made available for more widespread use. Compact programming should be encouraged, because it will result in the routine application of numerical techniques by project geologists, field workers, and students. These procedures will increase the versatility of their users and free them from much routine data assimilation. As the benefits of computer applications become increasingly apparent throughout the profession, the demands for more sophisticated hardware and numerical techniques will correspondingly increase. Progress in quantitative geology will then no longer be restricted to a few practitioners.

REFERENCES

Casetti, Emilio, 1964, Multiple discriminant functions: Tech. Rept. 11, ONR Task No. 389-135, Office of Naval Research, Geography Branch, 63 p.

Davis, J. C., and Sampson, R. J., 1966, FORTRAN II program for multivariate discriminant analysis using an IBM 1620 computer: Kansas Geol. Survey Computer Contr. 4, 8 p.

Dixon, W. J., (ed.), 1965, BMD biomedical computer programs: School of Medicine, University of California at Los Angeles, 620 p.

Golden, J. T., 1965, FORTRAN IV programming and computing: Prentice-Hall, Inc., 270 p.

Krumbein, W. C., and Graybill, F. A., 1965, An introduction to statistical models in geology: McGraw-Hill, Inc., 475 p.

Manson, Vincent, and Imbrie, John, 1964, FORTRAN program for factor and vector analysis of geologic data using an IBM 7090 or 7094/1401 computer system: Kansas Geol. Survey Sp. Dist. Publ. 13, 46 p.

Miller, R. L., and Kahn, J. S., 1962, Statistical analysis in the geological sciences: John Wiley and Sons, 483 p.

Sampson, R. J., and Davis, J. C., 1966, FORTRAN II trend-surface program with unrestricted input for the IBM 1620 computer: Kansas Geol. Survey Sp. Dist. Publ. 26, 12 p.

HOMOGENEITY OF COVARIANCE MATRICES IN RELATION TO GENERALIZED DISTANCES AND DISCRIMINANT FUNCTIONS

By

R. A. Reyment
Kansas Geological Survey and University of Stockholm

INTRODUCTION

The well known multivariate techniques of discriminant analysis and generalized distances were originally developed in response to certain taxonomic problems. The discriminant function was devised to solve a classification problem. This concerns the classification of an individual or group of individuals, with two or more populations, to one of which the individual or group of individuals actually belongs. The basic theory of this method, therefore, requires that the specimen or group of specimens genuinely belong to one of the k populations. It was not designed to demonstrate relative nearness of an individual, or group of individuals, to one or more of k populations; this is not always realized in some applications. The method of discriminant functions has begun to play a significant role in taxonometric work. The generalized statistical distance was also introduced in answer to a question essentially taxonomic in nature, although in its original context, the problem was anthropomorphic. However, the step from the quantitative classification of ethnic groups of human beings to the classification of fossils is not overly great. If the problem is one of investigating the relative nearness, hence similarity, of samples, the procedure of analysis by canonical variates provides a reasonable approach. This multivariate statistical procedure may in some respects be regarded as a generalization of principal component analysis, which deals with a single sample from a single population, whereas in analysis by canonical variates, one is concerned with samples from two or more populations.

A basic theoretical requirement for all of these methods is that the covariance matrices be homogeneous. This means that the variance of, and the covariances between, the constituent variables must be sufficiently close to each other so as not to differ significantly in the statistical sense. In other words, the assumption of homogeneity in the sample covariance matrices has the underlying requirement that they be derived from the same universe. Available tests of significance for discriminant functions and generalized distances require mostly the assumption of homogeneity of covariance matrices to be validly applicable. A criticism that has been leveled against the procedure used for testing the homogeneity of variances and covariances (cf. Anderson, 1958, p. 247-268) is that it is not robust, which means that it is sensitive to departures from the multivariate normal distribution. This is naturally a rather undesirable property for a test of this kind, but provided this is realized and kept in mind, there is no insurmountable problem involved.

Figure 1 gives a schematic presentation of how one could assemble a computer program for discriminant analysis and generalized distances in which the homogeneity of covariance matrices occupies a decisive place. The decision for which sequence of computations will be required centers around the answer to this question. The problem of testing the significance of a heterogeneous D^2 has not yet been fully worked out for the Anderson-Bahadur procedure. The statistically somewhat less attractive method developed by the writer (Reyment, 1962) is, however, related to a test of significance and if it were considered necessary to test the significance of a generalized distance for means from populations with heterogeneous covariance matrices, the computer program would have to include a certain degree of duplication of steps. The presentation here given has been kept at a level that does not presuppose special knowledge of multivariate statistical analysis.

DISCRIMINATION

In the present connection, we shall only be concerned with comparisons between two groups, although comparisons may, and are readily made, for three and more groups. For two groups, the so-called likelihood ratio should be employed for discrimination where, in terms of the natural logarithms,

$$\ln\lambda = \ln P_1(I) - \ln P_2(I).$$

Here I denotes the information available for an individual and the probability of I on the hypothesis that this individual belongs to group (1) or group (2) is $P_1(I)$ (i=1,2). A comparison of $P_1(I)$ and $P_2(I)$ is necessary for classifying the individual. In an uncomplicated situation, the classification could be defined in the following terms: if $\lambda > 1$, the individual is to be assigned to group (1), if $\lambda < 1$, the individual is assigned to group (2). The most obvious complication is provided by differences in the relative abundance of the two categories. It has also been shown that appreciable differences in the sizes of the samples used for setting up a discriminant function and in finding the related D^2 and T^2 have an influence on these values. That is, these are not independent of sample size.

Moreover, where the covariance matrices are not homogeneous and the sample sizes are equal ($N_1 = N_2$), D^2 and T^2 found by the "ordinary" procedure will be hardly different from the values yielded by a theoretically more rigorous method. The effects of heterogeneity in the variances and covariances become apparent when N_1 and N_2 become substantially different.

The discriminant function and the corresponding generalized distance are closely related. Thus, if we consider a p-variate random vector variable X, the population mean of which for group (1) is μ_1 and the population mean for group (2) is μ_2, an expression for $\ln\lambda$ in the multivariate case may be written as

$$(\mu_1' - \mu_2')S^{-1}X - 1/2(\mu_1' S^{-1}\mu_1 - \mu_2' S^{-1}\mu_2). \quad (1)$$

Here, the covariance matrix of X for group (1) is S_1 and that for group (2), S_2. If S_1 and S_2 are statistically the same, $S_1 = S_2 = S$.

The linear discriminant function is defined as z which is the linear combination of variables in X,

$$z = (\mu_1' - \mu_2')S^{-1} x. \quad (2)$$

The difference in the means of z for groups (1) and (2) is:

$$(\mu_1' - \mu_2')S^{-1} (\mu_1 - \mu_2) = D^2 \quad (3)$$

The variance of z for either population is D^2 and the ratio of the difference of means of z to the standard deviation thereof is D^2/D, the generalized statistical distance between the two entities. The test of significance of the generalized distance is given by the T^2 of Hotelling

$$T^2 = \frac{N_1 N_2}{(N_1 + N_2)} D^2 \quad (4)$$

which has the critical region,

$$T^2 \geqslant \frac{(N_1+N_2-2)p}{N_1+N_2-p-1} F_{p,\ N_1+N_2-p-1}(\alpha) \quad (5)$$

with significance level α. Here, p is the number of variables and F is the variance ratio.

If the observation vectors $x^{(1)}$ and $x^{(2)}$ are correlated, there will be complications, which should be taken into account.

There are several possible theoretical approaches in discriminant analysis, including that of canonical variates, that is, by means of an equation formed by the coefficients of the eigenvectors corresponding to one of the eigenvalues obtained from the diagonalization of the two-component covariance matrices. The method often used of performing a quasi-regression analysis has the advantage that the considerable body of theory available in this connection may be applied directly to the discriminatory problem and, for example, one may test the significance of any desired coefficient of the discriminant function.

The problem of discriminating when the covariance matrices are not equal ($\Sigma_1 \neq \Sigma_2$), may be viewed in several lights. Perhaps the most intriguing way known to the writer is that of Dempster (1964), who considers the shadow property of the ellipsoids of scatter. Anderson and Bahadur (1962) considered a distance for the case of unequal covariance matrices of the following kind,

$$\frac{b'\gamma}{(b'\Sigma_1 b)^{1/2} + (b'\Sigma_2 b)^{1/2}} \quad (6)$$

where b is of the form:

$$|t\Sigma_1 + (1-t)\Sigma_2|^{-1}\gamma \quad \text{with } 0 < t < 1 \quad (7)$$

Here, γ is the vector of differences in means, b an analog of the vector of coefficients of the discriminant function and Σ_1 and Σ_2 are the covariance matrices. When $\Sigma_1 = \Sigma_2$, twice the maximum of (6) is the Mahalanobis' distance between populations. The estimation may be made iteratively by finding t from:

$$0 = b' [t^2\Sigma_1 - (1-t)^2\Sigma_2] b \quad (8)$$

and solving

$$(t\Sigma_1 + (1-t)\Sigma_2)b = \gamma \quad (9)$$

for b. In the above equation, the vector b is the vector of discriminator coefficients.

The method suggested by the author (Reyment, 1962) is based on a generalization of the Scheffe univariate test for differences of means when the

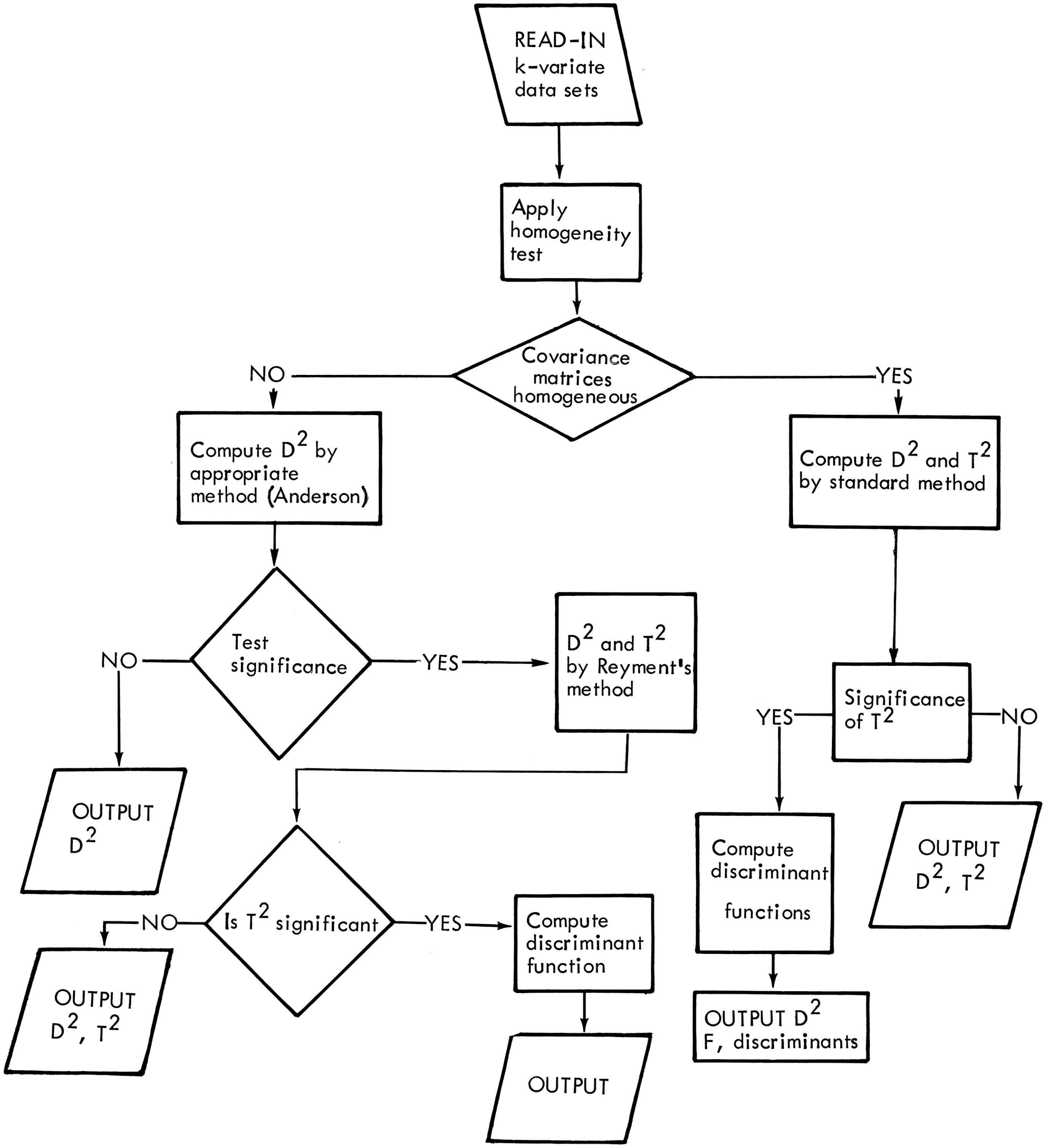

READ-IN k-variate data sets
Apply homogeneity test
Covariance matrices homogeneous
NO
YES
Compute D^2 by appropriate method (Anderson)
Compute D^2 and T^2 by standard method
Test significance
NO
YES
D^2 and T^2 by Reyment's method
Significance of T^2
YES
NO
OUTPUT D^2
Is T^2 significant
NO
YES
OUTPUT D^2, T^2
Compute discriminant function
OUTPUT
Compute discriminant functions
OUTPUT D^2, T^2
OUTPUT D^2 F, discriminants

Figure 1.- Schematic flow diagram for generalized distance and discriminant function subroutines.

variances are unequal. Computations are essentially simple and consist mainly of producing a set of "new" variables, y_i, from two sets of "original" variables, the vectors $x_i^{(1)}$ and $x_i^{(2)}$, randomly ordered:

$$y_i = x_i^{(1)} - x_i^{(2)} . \qquad (10)$$

The required generalized distance is then derived from the covariance matrix of y, T, by means of the formula:

$$D^2_{het} = 2d'T^{-1}d \qquad (11)$$

Experience shows that D^2_{het} varies considerably for random pairings of $x^{(1)}$ and $x^{(2)}$, and in practice it is advisable to take the average of several calculations for different pairings of the data. It is possible also, by analogy, to relate this distance to a vector of coefficients, but nothing is known about the properties of such coefficients from the statistical point of view.

COMPUTER APPLICATION

The application of the procedures outlined in the foregoing may be readily grouped into a single computer program where the main program computes the basic statistics and carries out the homogeneity test, the appropriate generalized distance and discriminant function calculation being available as a subroutine.

In numerical taxonomy and various other fields, one may be faced with the problem of separating two samples, each consisting of just a few individuals, with each individual having a large number of measured variables. In zoology such a situation may arise in the study of insects, and in geology this problem may occur in the analysis of recent marine sediments in connection with which a large number of variables are measured, but for various reasons, the number of stations sampled may be small. It is not here intended to take up the topic in detail but, owing to the growing significance of the problem, it might be useful to mention that several solutions are available in the literature, the most applicable of which seems to be that of Dempster (1960). It is based on a substitute for T^2, but does not appear to have been extended to the case when the covariance matrices are not equal.

ANALYSIS BY CANONICAL VARIATES

Problems involving several groups of "species" will now be considered. For three groups, there will be three distances between the sample centroids, i.e., between the first and second groups, between the first and third groups, and between the second and third groups. It is thus possible to analyze the multi-group case in terms of a two-group situation and this has been done in several biologic and geologic publications; however, this becomes very complicated when there are numerous groups. The technique of canonical analysis has the basic requirement that the covariance matrices of the groups are homogeneous.

Input quantities may be readily obtained from an appropriate analysis of variance and covariance table. If one considers the discriminant function:

$$w = tx \qquad (12)$$

where t is the vector of discriminant coefficients and x the observational vectors, the analysis of variance of w is

Among groups	t'At
Within groups	t'Bt
Total	t'Ct

Here, A is the matrix of "among-groups" variances and covariances, B is the matrix of "within-groups" variances and covariances, while C is the matrix of "total" variances and covariances.

For w to be as effective as possible in sorting out the groups,

$$R^2 = t'At/t'Ct \qquad (13)$$

has to be maximized. This is done by solving the set of linear equations:

$$(A - R^2C) = 0. \qquad (14)$$

A nonzero solution will be obtained if R^2 is a root of the equation:

$$\left| A - R^2C \right| = 0 \qquad (15)$$

Equation (15) will have p roots, R^2_1, R^2_2... ...R^2_p ($p < q$, where q is the number of groups and p the number of variables). The corresponding solutions of (14) are vectors, $t_1, t_2, \ldots. t_p$. The set of canonical variables, represented generally by,

$$w_i = t'_i\, x, \qquad (16)$$

are mutually uncorrelated. A measure of separation of the groups is given by

$$\beta = (1-R^2_1)(1-R^2_2)\ldots..(1-R^2_p) \qquad (17)$$

which is the same as the ratio of the determinants of matrices B and C; $\beta = \det B/\det C$.

This statistic derives from the theory of the generalized analysis of variance. The significance of β is gauged from the chi-squared test

$$-(n - 1/2[p+q+1]) \log_e\beta \qquad (18)$$

with pq degrees of freedom. If this is significant, the largest eigenvalue may be extracted from the total chi-squared by subtracting

$$-(n - 1/2[p+q+1]) \log_e(1-R^2_1), \qquad (19)$$

which leaves an approximate chi-square with (p-1)(q-1) degrees of freedom. This should be continued for as long as significant values of chi-square are obtained.

The canonical analysis procedure may be readily programmed and only requires a read-in routine, coupled to a subroutine for the eigenvalues of a square nonsymmetric matrix of the form A^{-1} B. Analysis by canonical variates would appear to be a method of considerable applicability in problems of classification.

FUTURE WORK

One of the major problems in quantitative work in geology, particularly multivariate studies, concerns the interpretation of results. In some types of studies, this may not pose too difficult a question, if the underlying model is straightforward and the figures issuing from the computer are readily traceable back to the input data. This is usually true in discriminant analysis and generalized distances. If a project employs the intermediary of eigen-computations, the interpretation of results becomes more complicated and may even become tainted with subjectivity, the very bugbear the researcher had hoped to avoid by using quantitative methods. Many difficulties of interpretation may be resolved by applying the statistical method to artificially prepared data, such as has been done by Imbrie (1963) and the present writer (Reyment, 1966). There is little doubt of the pressing need for orientatory experimental work in many analyses of geologic data from the spheres of paleoecology, paleontology, and sedimentology.

REFERENCES

Anderson, T.W., and Bahadur, R. R., 1962, Classification into two multivariate normal distributions with different covariance matrices: Ann. Math. Stat., v. 33, p. 420-431.

Bartlett, M.S., 1952, The goodness of fit of a single hypothetical discriminant function in the case of several groups: Ann. Eugenics, v. 16, p. 199-214.

Dempster, A.P., 1960, A significance test for the separation of two highly multivariate samples: Biometrics, v. 16, p. 41-50.

Dempster, A.P., 1964, Tests for the equality of two covariance matrices in relation to a best linear discriminant analysis: Ann. Math. Stat., v. 35, p. 190-199.

Reyment, R.A., 1962, Observations on homogeneity in covariance matrices in paleontologic biometry: Biometrics, v. 18, p. 1-11.

Reyment, R.A., 1966, Studies on Upper Cretaceous and Lower Tertiary Ostracoda from Nigeria. Part III. Stratigraphical, Paleoecological and Biometrical conclusions: Stockholm Contr. Geol., v. 14, 144 p.

FACTOR ANALYSIS OF A POOLED CROSS-PRODUCT MATRIX

By

John Imbrie

Columbia University

Other papers in this Symposium employ various multivariate techniques, including factor analysis, to the problem of constructing a meaningful classification. The purpose of this paper is quite the reverse: to employ a classification of geological samples as the basis for a meaningful R-mode factor analysis.

Consider the hypothetical data on two variables (x and y) and 26 samples given in Table 1 and plotted on Figure 1. Assume that we are interested in evaluating the relationship between x and y. If it is known, on any a priori basis, that the samples should be classified into two groups (Group A and Group B), then it is a simple matter to evaluate the statistical relation between the two variables. The samples from Group A come from a reaction realm in which variables x and y tend strongly to be linearly related. The same can be said of Group B. Clearly, however, the constants of the linear functions are quite different in the two realms.

Table 1.--Raw data for hypothetical problem: Two groups of 13 samples per group.

Group A		Group B	
x	y	x	y
4	5	8	5
5	5	9	5
3	4	7	4
4	4	8	4
5	4	9	4
2	3	6	3
3	3	7	3
4	3	8	3
1	2	5	2
2	2	6	2
3	2	7	2
1	1	5	1
2	1	6	1

Before outlining the pooled cross-product factor model, it will be helpful to review, in a somewhat unorthodox manner, the steps by which a standard R-mode factor analysis would be conducted on the data in Table 1 if membership in the two groups were unknown (or ignored):

STEP 1: Define the raw data in Table 1 as the 26 x 2 matrix X.

STEP 2: Express the elements in each column of X as deviations from the column mean. Define the resulting matrix as Z.

STEP 3: Calculate Z'Z, the minor product moment of Z. Define this result as T, the matrix of cross-products calculated on deviations from the mean of the total sample.

In the example, $$T = \begin{bmatrix} 148 & 36 \\ 36 & 44 \end{bmatrix}$$

STEP 4: Transform T into the correlation matrix R. This transformation can be compactly symoblized by defining a diagonal matrix D_t in which each element is the square root of the corresponding diagonal element in T. Then

$$R = D_t^{-1} T \ D_t^{-1}$$

In the example, $$R = \begin{bmatrix} 1.00 & .45 \\ .45 & 1.00 \end{bmatrix}$$

STEP 5: Factor R.

Note that in the example above, R contains a low correlation coefficient (r = 0.45), reflecting the nearly random relationship exhibited by the 26 points in Figure 1. If such a correlation matrix were used as the basis of a factor analysis, the number of factors, as well as the difficulty of interpretation, would be needlessly large. Is there a way of simplifying the factor analysis?

One simple solution to this problem is to factor a correlation matrix derived from a pooled cross-product matrix -- i.e., a matrix formed by summing cross-product matrices calculated separately for each group. As the constituent cross-product matrices are calculated with respect to the joint mean of each group, inter-group differences are surpressed and the resulting factor analysis will more clearly reflect intrinsic relationships among the variables.

The algebraic outline of the method may be presented as follows:

STEP 1: For each group of samples, form a matrix of cross-products of deviations from the group mean. For Group A, call this W_a; for Group B, W_b;

and so on. For the example above,

$$W_a = W_b = \begin{bmatrix} 22 & 18 \\ 18 & 22 \end{bmatrix}$$

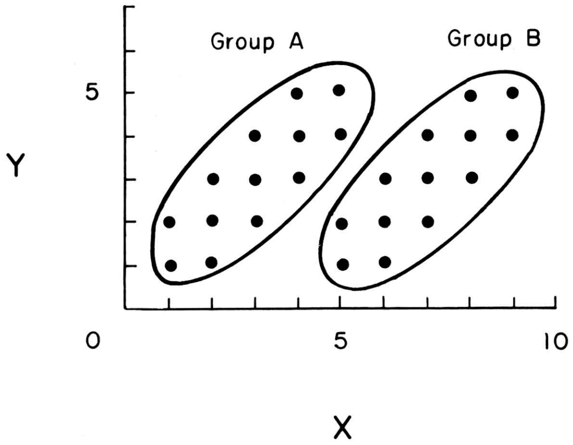


Figure 1.--Plot of data in Table 1.

STEP 2: Given g groups, form a pooled within-group cross-product matrix

$$W= W_a + W_b + \dots + W_g.$$

For the example, $W = \begin{bmatrix} 44 & 36 \\ 36 & 44 \end{bmatrix}$

STEP 3: Form the diagonal matrix D_w, where each element is the square root of the corresponding diagonal element of W. Then calculate the pooled correlation matrix

$$R_w = D_w^{-1} W D_w^{-1}.$$

For the example, $R_w = \begin{bmatrix} 1.00 & .82 \\ .82 & 1.00 \end{bmatrix}$

STEP 4: Factor R_w.

Note that the correlation coefficient in the pooled correlation matrix (0.82) is significantly higher than that in the correlation matrix for the entire sample (0.45). Under the assumptions of the method here outlined, the higher value correctly reflects the strength of the linear relationship between x and y -- a relationship that can be studied apart from differences in the linear functions characteristic of the sampled reaction realms.

The matrices W and T are identical to those calculated in multiple discriminant function analysis (see, e.g., Cooley and Lohnes, Chap. 6).

REFERENCES

Cooley, W. W., and Lohnes, P. R., 1962, Multivariate procedures for the behavioral sciences: New York, John Wiley & Sons, Inc., 211 p.

CLASSIFICATION OF MAP SURFACES BASED ON THE STRUCTURE OF POLYNOMIAL AND FOURIER COEFFICIENT MATRICES[1]

By

W. C. Krumbein
Northwestern University

ABSTRACT

Current methods of trend analysis utilize mainly the polynomial and Fourier models, both of which are derived from the general linear model. The coefficient matrices associated with the map models are conventionally structured diagonally for polynomials and in blocks for Fourier surfaces. It is possible, however, to consider the elements of such matrices as occupying points in a "coefficient space" defined by coordinate axes representing the coefficient subscripts. This space may be subdivided in various ways to yield configurations by diagonals, blocks, circles, hyperbolas, etc. Each of these classifications gives rise to a set of ranked map surfaces (sequential or cumulative) that can be used with both map models for expressing and analyzing map data in different ways, and for screening out selected components for map comparisons.

INTRODUCTION

Trend analysis of contour-type maps is currently based mainly on applications of the polynomial and Fourrier models. Both stem directly from the general linear model, but the structure of a single map observation is different in the two models, and the kinds of fitted surfaces obtained from a given set of data generally differ in the patterns of their contour lines. These differences have been described elsewhere (Krumbein, 1966) to demonstrate that both the kind of information and the amount of information (in terms of least-squares criteria) may differ markedly for the two models used on the same set of map data.

The purpose of this paper is to examine the structure of the coefficient matrices associated with the models, rather than with the structure of a single map observation. Conventional practice is to use a standard form of the coefficient matrix for each model, but these are only two interchangeable members of a larger series of matrix structures. Each structure defines an ordered set of maps, which can be produced sequentially or cumulatively. These maps generally present different kinds of surfaces depending on the particular coefficients specified by the structure, although some combinations overlap. The resulting maps are useful for substantive analysis, and as screening devices for examining selected map components under the two models.

This report emphasizes two aspects of the subject. The first is a classification of polynomial and Fourier coefficients into several classes, and the second is an illustration of a screening procedure that brings out some basic similarities in the Fourier and polynomial models. Gridded data are used for convenience.

CONCEPT OF A "COEFFICIENT SPACE"

The coefficients associated with the polynomial model are commonly shown in diagonal arrangement (Oldham and Sutherland, 1955), in which each succeeding polynomial surface (linear, quadratic, etc.) occupies a diagonal in the matrix of coefficients. James (1966) developed a block arrangement for Fourier coefficients, in which successive blocks (Fourier surfaces) contain wavelengths of diminishing magnitude.

Although map coefficients are commonly shown as occupying small squares in coefficient diagrams, they may be considered as occupying points on a plane with coordinate axes defined by the coefficient subscripts, i and j. The origin lies at the upper left, with i increasing vertically downward and j increasing to the right. In the general linear model each coefficient β_{ij} occupies a single point. For example, β_{23} has coordinates (i, j) = (2,3), and hence lies at point i = 2, j = 3. The mean value, β_{00}, has coordinates (0,0), and lies at the origin; all other coefficients lie at points defined by integer values on the plane, including zero.

The structuring of the coefficient matrices as

[1] Research conducted on this subject is supported by ONR Task No. 388-078, Contract Nonr-1228 (36), Geography Branch of the Office of Naval Research. This preliminary paper will be embodied in a Technical Report now in preparation by the author at Northwestern University.

the polynomial model the diagonals arise from a system of contours on the plane defined by the relation $(i + j) = \text{const.}$, where the constant takes on the successive values 0, 1, 2, ..., k. In the block arrangement the coefficient plane has a series of contours defined by the relation, $\max (i, j) = \text{const.}$, where the constant also assumes successive values 0, 1, 2, ..., m.

Figure 1 (upper left) shows the diagonal arrangement conventionally used for the polynomial model, in which the diagonal contours represent the usual sequence of linear, quadratic, and higher ordered surfaces. The upper right diagram, representing the block arrangement used for Fourier

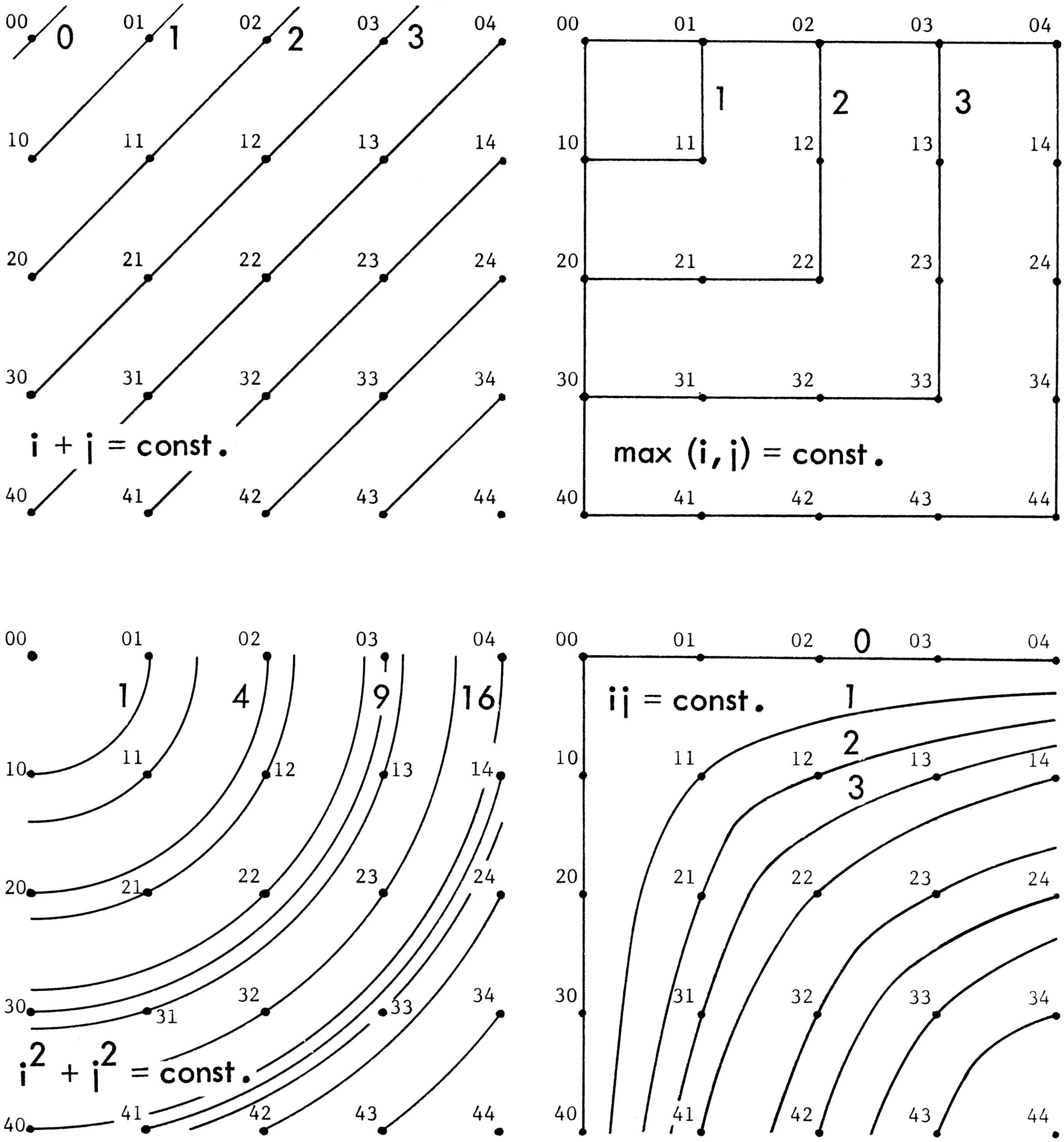


Figure 1.- Four arrangements of "coefficient space." Upper left, diagonal; upper right, block; lower left, circular; and lower right, hyperbolic.

coefficients, has reverse L-shaped contours with values that define the successive blocks.

Although diagonal contours for polynomials, and L-shaped contours for the Fourier model seem "natural", there is no reason why these structures cannot be interchanged, yielding a block arrangement of the polynomial coefficient matrix and a diagonal arrangement of the Fourier coefficient matrix. In fact, these two structures do not exhaust the possibilities, in that a variety of ranked surfaces can be defined by other combinations of i and j. For example, the relation, $(i^2 + j^2)$ = const., produces a set of concentric circular contours on the plane, with radii $(i^2 + j^2)^{1/2}$, and the product of the coordinates, (ij) = const., produces a series of hyperbolas, including the degenerate form (ij) = 0, which coincides with the axes passing through the origin. These arrangements are shown in the lower part of Figure 1.

It may be noted in Figure 1 that the several sets of contours pass through all the points on the plane that satisfy the contour equations. For example, in the diagonal arrangement the contour value 2 passes through points 20, 11, and 02. These are the three coefficients whose subscripts sum to the value 2. In the block arrangement the contour of value 2 passes through points 20, 21, 22, 12, 02, in all of which the largest subscript is 2. In the circular case the contour of value 2 passes through point 11, which is the only coefficient in which $i^2 + j^2 = 2$. The radius of this circle is 1.41, the square root of 2. Finally, in the hyperbolic case the contour of value 2 passes through points 21 and 12, whose products are 2.

COMPARISON OF RANKED SURFACES FOR POLYNOMIAL AND FOURIER MODELS

The preceding discussion is based on the coefficients in the general linear model (and also in the polynomial model), in which one coefficient occupies each integer valued point on the coefficient plane. The situation is somewhat different for the Fourier form of the general linear model, in which from one to four individual coefficients may have the same subscripts owing to the combinations of sines and cosines involved.

If we designate the equivalent of β_{ij} in the general case by F_{ij} for the Fourier model, the number of individual coefficients associated with each pair of subscripts in a 5 x 5 array is as shown in Table 1.

As Table 1 shows, the mean value is a single coefficient, and the components with i or j equal to zero each have two coefficients, whereas all other components have four. This difference in the number of coefficients associated with each (i, j) for the Fourier model means that fitted surfaces of a given rank normally have different numbers of coefficients in the successive surfaces for the two map models. There is an outstanding exception in ij = 0, however, which will be developed in an example. The coefficient plane of Figure 1 is also seen to have from one to four Fourier coefficients projected onto a given (i, j) point.

Table 1.- Number of coefficients in Fourier model for 5 x 5 array.

F_{ij}	(cos cos) cc_{ij}	(cos sin) cs_{ij}	(sin cos) sc_{ij}	(sin sin) ss_{ij}	Total
F_{00}	1	0	0	0	1
F_{01}	1	1	0	0	2
F_{02}	1	1	0	0	2
F_{10}	1	0	1	0	2
F_{11}	1	1	1	1	4
F_{12}	1	1	1	1	4
F_{20}	1	0	1	0	2
F_{21}	1	1	1	1	4
F_{22}	1	1	1	1	4
					25

Table 2 presents a classification of the ranked least-squares map surfaces based on the coefficients intersected by the various contour systems. The first column shows the rank of the surface, and the next four columns show the coordinates of the points involved, listed directly as the coefficient subscripts. The starred items represent the Fourier coefficients as F_{ij}, of which there are only 9 in a 5 x 5 grid. These starred items, and those unstarred, represent the 25 individual ij of the polynomial model. The several columns on the right indicate the number of individual coefficients in each ranked surface, yielding a total of 25 for each model.

ILLUSTRATIVE EXAMPLE

The preceding topics, covered very briefly here, are developed in greater detail in an ONR technical Report (footnote 1). It seems appropriate, however, to include at least one example of the kinds of maps obtained by the rankings of Figure 1 and Table 2. The example contrasts the block and diagonal structures for the same set of data analyzed by both the polynomial and the Fourier model. In addition, it illustrates a map screening technique used under known environmental conditions, which has interesting implications when applied to subsurface data where the controlling conditions, and hence an optimum grid orientation, must be inferred.

Figure 2 shows the areal pattern of the geometric mean diameter in mm. of sand on a Lake Michigan beach at Evanston, Illinois. The beach has a high storm berm and a lower berm that developed during declining storm conditions. This lower berm was topped by waves during its formation, so that shallow temporary lagoons developed landward of it. Sand samples were taken on a grid during ensuing quiet weather to study possible reversals in the particle size patterns that normally occur on foreshores. As anticipated, the mean grain diameter increases lakeward on the upper storm foreshore with a reversal to smaller sizes in the temporary lagoonal areas, and then increases relatively rapidly across the lower foreshore to the water line.

The map of Figure 2 was computed with James' (1966) double Fourier series program, which produces a continuous symbol map from which the contours were traced. The machine-computed surface agrees very well with a hand-contoured map. The original grid had a spacing of 25 feet between samples along shore, and a 10-foot spacing across-shore. In order to provide better visualization of the pattern, the grid is expanded landward by a factor of 2.5, to obtain a square map. This is not unlike increasing the vertical scale on a cross-section to bring out details that are otherwise overly crowded, as would be the case of contours near the water line.

The data matrix was analyzed by both the Fourier and polynomial models to obtain two sets of 25 coefficients each. These coefficients were then used under the several rankings associated with the coefficient structures of Table 2, to obtain a wide range of experimental maps. Four of the maps are selected for presentation here, inasmuch as they illustrate the diversity of patterns obtained even with relatively low-rank fitted surfaces.

Figure 3 shows the four examples. The upper left map is the linear surface containing ranks 0 and 1 of the conventional diagonal grouping of polynomial coefficients. It is also the map of rank 0 + 1

Table 2.- Polynomial and Fourier surfaces ranked by four schemes of coefficient classification for a 5 x 5 grid.

Rank of Surface	Diagonals i + j	Blocks max (i,j)	Circles $i^2 + j^2$	Hyperbolas ij	No. of Individual Coefficients							
					Polynomial				Fourier			
					D	B	C	H	D	B	C	H
0	00*	00*	00*	00*,01*,02*, 03,04,10*, 20*,30,40	1	1	1	9	1	1	1	9
1	01*,10*	01*,10*,11*	01*,10*	11*	2	3	2	1	4	8	4	4
2	02*,11*,20*	02*,12*,20* 21*,22*	11*	12*,21*	3	5	1	2	8	16	4	8
3	03,12*,21*, 30	03,13,23, 30,31,32	None	13,31	4	6	0	2	8	0	0	0
4	04,13,22*, 31,40	04,14,40, 41	02*,20*	14,22*,41	5	4	2	3	4	0	4	4
5	14,23,32, 41	None	12*,21*	None	4	0	2	0	0	0	8	0
all higher surfaces	24,33,34, 42,43,44	24,33,34, 42,43,44	22*,03,04, 13,14,23, 24,30,31, 32,33,34, 40,41,42, 43,44	23,24,32, 33,34,42, 43,44	6	6	17	8	0	0	4	0
					25	25	25	25	25	25	25	25

in the circular case. The upper right map has the same ranking (i.e., diagonal 0 + 1, and circular 0 + 1) for the Fourier model. These maps have coefficients β_{00}, β_{01}, and β_{10} for the polynomial map, and F_{00}, F_{01}, and F_{10} (i.e., cc_{00}, cc_{01}, cs_{01}, cc_{10}, and sc_{10}, as shown in Table 1). The Fourier map thus has 5 coefficients as against three for the polynomial.

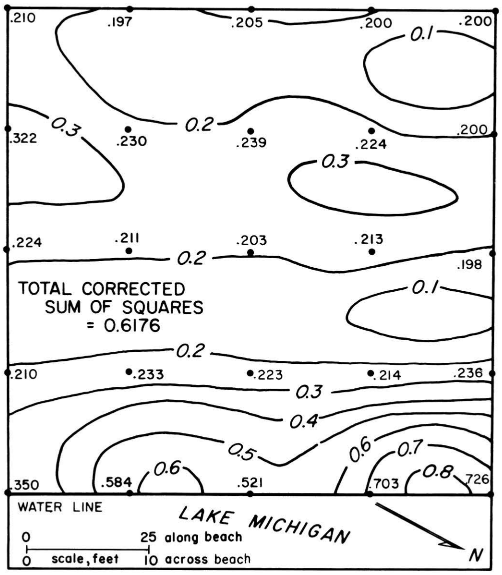


Figure 2.- Map of geometric mean diameter in mm of sand on Lake Michigan beach at Evanston, Illinois. Note across-beach scale exaggeration.

The two lower maps of Figure 3 are interesting in that they represent the Block arrangement (rank 0 + 1) for both models, each having coefficients with subscripts 00, 01, 10, and 11. Thus they both show the influence of the lowest-rank cross-product component with subscripts 11. The Fourier map on the lower right is the conventional Block (0 + 1) map which contains wave lengths of M and N respectively; introduction of the F_{11} coefficients produces parts of two elliptical areas in the upper center, and develops an additional ellipse in the lower right, where the observed grain size is largest. The polynomial map in the lower left is also a Block (1 + 2) map. Introduction of the quadratic cross-product term has the effect mainly of fanning out the linear contours of the upper left map. These maps deserve much more discussion, but space considerations limit presentation to the maps themselves, to show how different groups of coefficients under both map models change the aspects of the surface when diagonals and blocks are interchanged in the polynomial and Fourier models.

The last example concerns the hyperbolic structure of the coefficient matrix, in which ij = 0 involves 9 coefficients in both map models. These coefficients include all the grid-parallel components, which in both models account for 84.86% of the total sum of squares. Under these conditions, it can be predicted that both the polynomial and Fourier maps should be essentially the same though not identical, because one model has periodical or cyclical overtones and the other has not. Figure 4 shows two surfaces fitted to the data of Figure 2. The upper map contains all the grid-parallel components (ij = 0 in Table 2), and the lower map has all cross-product components (ij> 0), with the mean added simply to keep the values positive. If the mean value (0.29 mm.) is subtracted from the lower map, it becomes in effect a map of all residuals on the grid-parallel surface.

These maps were produced by the computer, using the Fourier coefficients. The corresponding polynomial map, hand drawn on values computed for each grid point, agrees exactly at each computed value, within normal rounding error. This agreement in sum of squares reduction and in map pattern suggests the usefulness of the hyperbolic structure (ij = const.) as a map screening device, equally suitable for both map models. That is, by separating the map components into grid-parallel and grid-angle portions, the additive aspects of the underlying data, in contrast to the interaction aspects, can be conveniently extracted from the coefficient matrices.

The grain size data used here were obtained on a grid oriented with respect to the known environmental and energy framework of a beach deposit. Normally, the main response of beach features is toward essential parallelism with the shoreline, or normal to it. That is, from a row-column analysis-of-variance viewpoint, the model is dominantly additive, with interaction effects generally playing a minor role. In the present case the additive elements account for 84.85% of the corrected sum of squares of the data (total = 0.6176), as against 15.15% for the interaction effects.

In situations where the orientation of the grid with respect to the controlling geologic conditions is not known a priori, as in many subsurface structure, thickness, and facies studies, the original map is usually oriented N-S and E-W. If the grid-parallel components are then screened out, they may represent merely the departure of the principal "grain" of the map from the arbitrarily chosen grid axes. This aspect of map analysis, as well as the use of combination Fourier-polynomial models, will be developed in the enlarged report mentioned earlier.

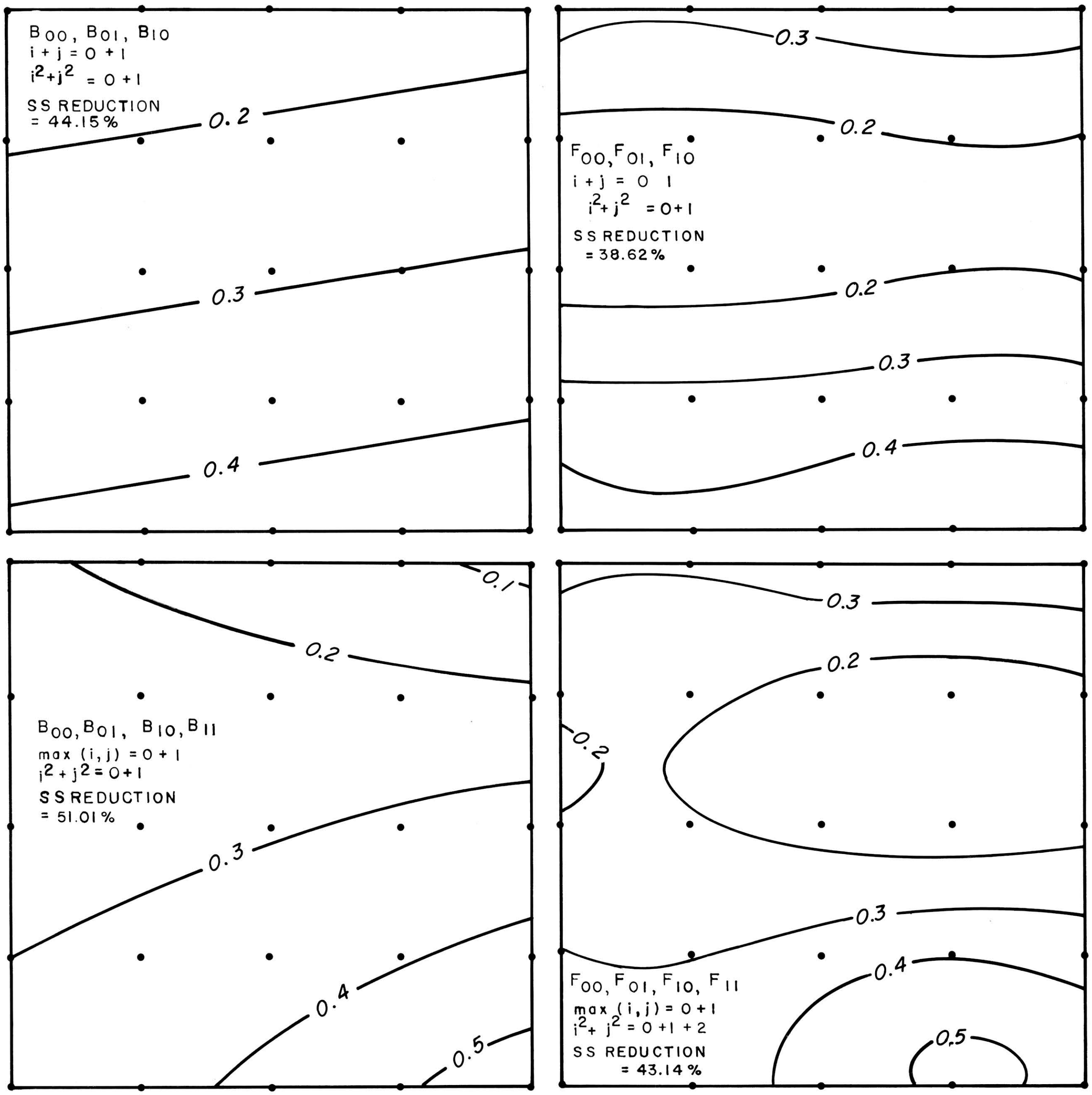


Figure 3.- Four surfaces fitted to map data of Figure 2. Upper left, conventional linear surface of polynomials, (i + j = ranks 0 + 1). Upper right, same coefficient structure in Fourier model. Lower left, block arrangement of polynomial [max (i, j) = ranks 0 + 1]. Lower right, conventional Block 0 + 1 Fourier map.

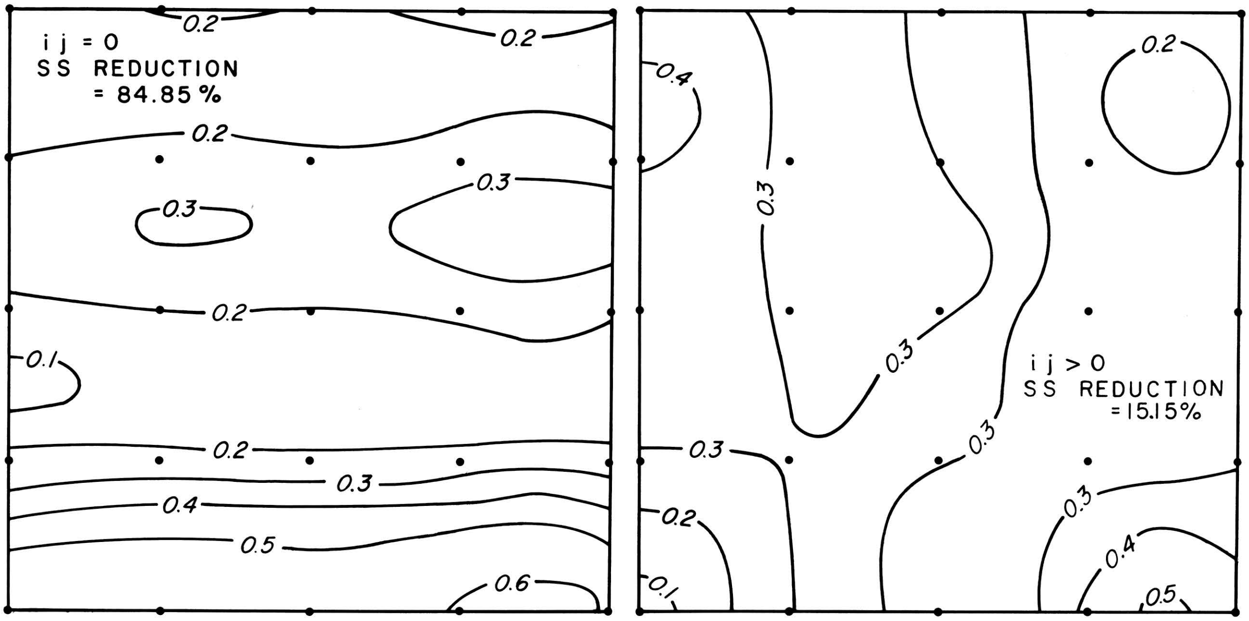


Figure 4.- Left map shows all grid-parallel components in data of Figure 2; right map shows all crossproduct components. Same map arises from polynomial and Fourier models.

CONCLUDING REMARKS

It becomes increasingly apparent, during the progress of a science, that what appear initially to be distinctly different phenomena, each characterized by its own attributes, are simply different classes of some broader, underlying whole. With further advance the classes become part of a continuous spectrum that permits examination of seemingly diverse phenomena (or models) within a basic, unifying framework. Map analysis appears to be no exception.

REFERENCES

James, W. R., 1966, The Fourier series model in map analysis: Technical Report No. 1, ONR Task No. 388-078, Contract Nonr-1228(36), Geography Branch, Office of Naval Research.

Krumbein, W. C., 1966, A comparison of polynomial and Fourier models in map analysis: Technical Report No. 2, ONR Task No. 388-078, Contract Nonr-1228(36), Geography Branch, Office of Naval Research.

Oldham, C.H.G., and Sutherland, D. B., 1955, Orthogonal polynomials; their use in estimating the regional effect: Geophysics, v. 20, p. 295-306.

APPLICATION OF COLOR-COMBINED MULTIPLE POLARIZATION RADAR IMAGES TO GEOSCIENCE PROBLEMS*

By

David S. Simonett
University of Kansas

ABSTRACT

When a number of remote sensors are used in concert they give images no one of which need give a secure diagnosis, but together these images may lead to unmistakable identification of natural or cultural objects. In order to analyze a number of such images simultaneously (an exceedingly difficult task, if done manually), an Image Discrimination Enhancement Combination and Sampling (IDECS) System has been developed by Engineering personnel of the University of Kansas primarily to work with multifrequency, poly-polarization radar images. This paper briefly describes the operation of the system in producing color combined radar images on a color television set, in producing differentiated and other modes of image enhancement, and in deriving probability density functions from the images. Examples are given of the way in which color combined multiple polarization radar images improve discrimination of lithologies and crops over that of a single radar image. Finally, a number of future modifications intended for improved texture discrimination, and data space sampling are described briefly.

INTRODUCTION

During the last three years there has been a notable surge of interest in the use of remote sensors, especially those mounted on spacecraft, as a means for mapping and grappling with geoscience and resource-oriented problems. Much of this interest has been stimulated by feasibility studies conducted under contract to the National Aeronautics and Space Administration. Many studies have adopted as a working premise, the idea that when a number of remote sensors are used in concert, they give data no part of which is diagnostic, but which together may be unmistakable. This concept, for which R.N. Colwell (1961) uses the phrase "multi-band spectral reconnaissance, " underlies much of this research.

As a portion of the NASA-sponsored feasibility studies in the use of radar, in which we are engaged at the University of Kansas, Center for Research in Engineering Science (CRES), a number of papers have been produced using single frequency radars or, at the most, two black and white radar images of different polarizations obtained simultaneously. Dellwig and Moore (1966) and Morain and Simonett (1966) report on results of geologic and geographic interest using these black and white images and employing manual inspection of the images.

A number of universities and government agencies have jointly recommended on a preliminary basis to NASA that three radar wavelengths -- 3.7, 15 and 16 cms.-- appear to be technically feasible for spacecraft use. The concept behind simultaneous data acquisition with multi-frequency radars is that this would represent the microwave equivalent of multi-band spectral reconnaissance. The different information content of the three widely spread frequencies in the microwave region should give us much more secure identification of objects of geoscience interest than the use of a single frequency. The same kind of reasoning lies behind a parallel recommendation that multiple polarization data be obtained in each frequency. In each frequency three polarizations are involved and thus a matrix of nine combinations of frequency and polarization is possible. It will immediately be recognized that this presents a classic multi-discriminant problem which would be impossible for the unaided individual to handle effectively or even to comprehend a small portion.

When a single radar image is used, it has been our experience that instrumental aids to discrimination can be helpful in image interpretation and analysis. When two or more images need to be considered simultaneously it soon becomes apparent that the eye and the mind are unable to accommodate the complexities contained in two or more data-planes. Simple devices (such as change-detection systems which alternately present one image to the left eye and the other to the right eye) have proven inadequate to develop quantitative understanding of image differences, and it has been necessary to develop more flexible equipment. This paper describes some of the approaches we have developed at CRES particularly in the color

* This paper was supported by NASA Contract NSR 17-004-003.

combining of multipolarization radar images an an aid in discrimination and sampling of multiple images.

ATTRIBUTES NEEDED IN AN IMAGE PROCESSING SYSTEM

Ideally, an image processing system needs to be able to combine in various ways a number of images, to be able to use color as a means of aiding discrimination, to enhance selected areas or edges for improved identification or discrimination, and to be able to sample portions of the data for statistical and other manipulations.

In geologic terms for example, we are interested in being able to:

1. discriminate lineaments both natural and artificial and separate them from non-lineament background, in such a manner as to speed up discrimination of an operator over that of the manual use of several images.
2. improve identification and discrimination of lithologies and,
3. improve segregation of data into manageable sets and subsets which have meaning within a geologic reference scale
4. derive relations which we are unable to obtain manually because the mind cannot comprehend vector-space with two or three images
5. rapidly delineate anomalous details as a means for focusing energy on areas of geologic interest.

It is to be recognized that such an instrument may speed identification, but not necessarily, for if discrimination is improved, the additional information also needs to be processed.

The instrumental setup I propose now to describe is one of several developed under the direction of Dr. R. K. Moore of the Electrical Engineering Department, University of Kansas. This data handling system was designed specifically to enable geologists and geographers to cope with the differences in information content available in multiple frequency, poly-polarization radar images.

THE IDECS DATA HANDLING SYSTEM

The Image Discrimination Enhancement Combination and Sampling (IDECS) Data Handling System developed for geoscience research is illustrated in Figure 1. A systems diagram for the IDECS Data Handling System is given in Figure 2.

The first operation in the IDECS Data Handling System is the scanning of up to three images in a three-channel flying-spot scanner (FSS). The electrical analogs of images placed on the faces of the FSS are fed individually or collectively to a matrixing unit and the output of the matrix are

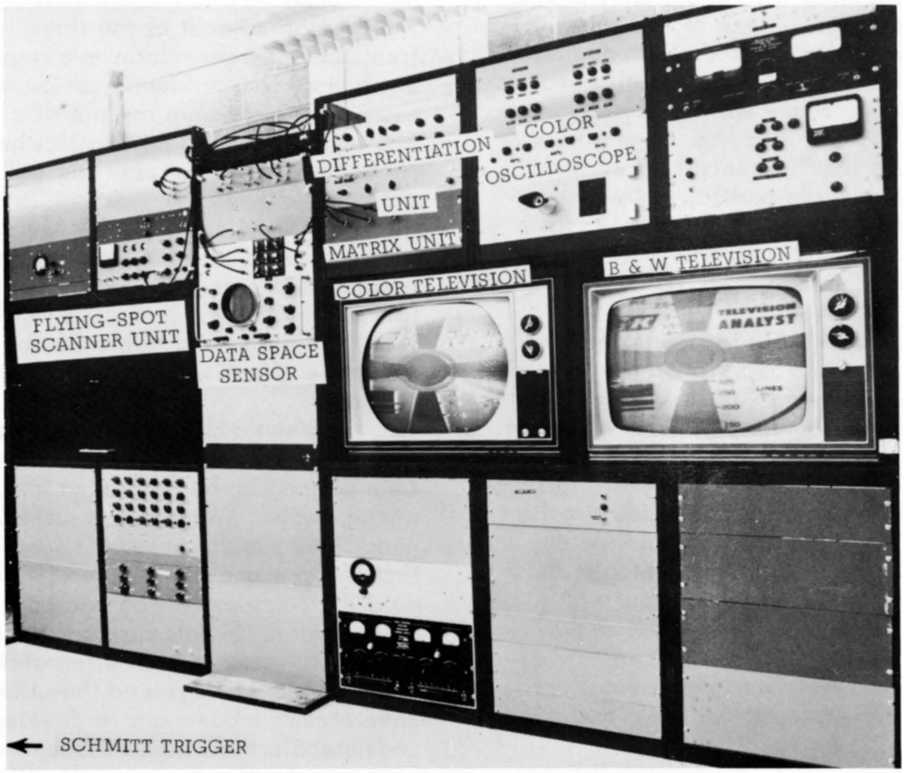


Figure 1.-The IDECS (Image Discrimination, Enhancement, Combination and Sampling) Data Handling System developed at the University of Kansas.

presented to the three (red, blue, and green) electron guns of the cathode ray tube in a color television (CTV). By this means, combinations of images can be reproduced in various colors to aid their geoscience interpretation. A black and white television (BWTV) is also used to present the output from any one of the red, blue or green channels from the matrix unit. Both the CTV and BWTV are synchronized with the FSS by synchronizing their rasters (see Figure 2). An alternative tri-color oscilloscope (TCO) may also take the outputs from the matrix or directly from the A, B, and C channels of the FSS. This device has very high color fidelity, but poorer resolution than the CTV. It may, however, be used for scanning images and handling radar scatterometer data. No further discussion of this unit will be given in this paper. The output of either the red, blue or green channels from the matrix unit or the A, B, C channels of the FSS may also be fed into a pulse-height analyzer to produce probability density functions and other statistical manipulations. The A, B, and C channels of the FSS may also be sampled with a data space sensor (DSS) or a Schmitt trigger and the outputs from these devices in turn may be fed to the CTV to selectively enhance the data sampled by the DSS and Schmitt trigger. Finally, a differentiation unit is included which enables one or more images to be differentiated and presented in different colors to the CTV. The detailed operation of each of these separate sections may now be described.

Flying-Spot Scanner

The present FSS is a three-channel instrument. Future modifications include the addition of further channels.

Matrix Unit

The matrix unit adds portions of three input signals from the FSS to form three output signals. The three outputs are produced by multiplying each input signal by a constant and summing the products. For example, signals x(t), y(t), z(t) are multiplied respectively by A_{11} , A_{21} , A_{31} , (where the first subscript denotes a row and the second subscript denotes a column) to give the output $x(t)A_{11} + y(t)A_{21} + z(t)A_{31}$. The second and third outputs are produced similarly with the summation over the second and third columns of constants. The matrix constants may be adjusted in discrete steps of 0.1 through the range $-1 \leqslant A_{mn} \geqslant 1$. The operation of the matrix unit can be represented mathematically by the multiplication of a 1x3 matrix by a 3 x3 matrix. The prime use of the matrix unit is as an aid in interpretation of imagery by fractional overlaying of imagery. The procedure is to place congruent, multifrequency or multipolarization radar images on the faces of each of the three cathode ray tubes of the FSS. The outputs of the flying-spot

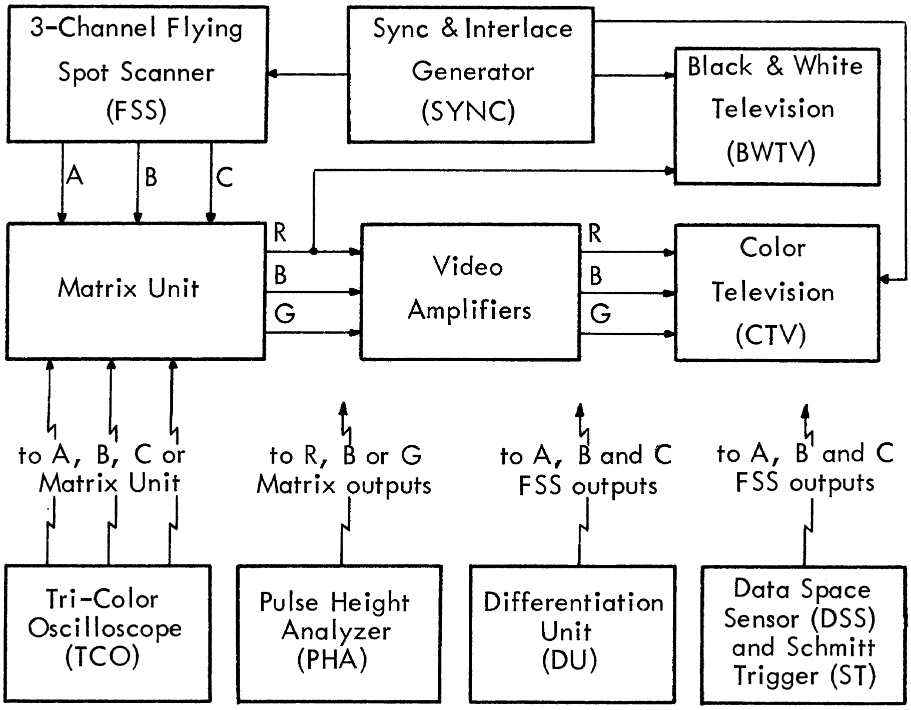

3-Channel Flying Spot Scanner (FSS)
Sync & Interlace Generator (SYNC)
Black & White Television (BWTV)
A
B
C
Matrix Unit
R
B
G
Video Amplifiers
R
B
G
Color Television (CTV)
to A, B, C or Matrix Unit
to R, B or G Matrix outputs
to A, B and C FSS outputs
to A, B and C FSS outputs
Tri-Color Oscilloscope (TCO)
Pulse Height Analyzer (PHA)
Differentiation Unit (DU)
Data Space Sensor (DSS) and Schmitt Trigger (ST)

Figure 2.- Systems diagram of the IDECS data handling system.

scanner are then applied to the inputs of the matrix unit and the outputs of the matrix unit are connected to the video amplifiers of the CTV. Matrix constant switches are then set to obtain the desired enhancement of imagery. Because it is possible to adjust the polarity of the image, it becomes feasible to use three colors with a two-image display as well as three colors with a three-image display. Through using the polarity switches to give positive and negative images in the different colors it becomes possible to level-slice and clip out the upper or lower portions of an image.

Using this system it is also possible to mix a differentiated and unaltered signal, allowing many variations in the amount and degree of differentiated and unaltered signal to be applied to the CTV. A number of color images obtained with this system are presented and discussed in the oral presentation of this paper. However, for purposes of reproduction it has been possible to use only black and white illustrations.

Differentiation Unit

It is often difficult to determine the rate of density change in multiple radar images either in black and white or in color reproduction. The density change can be abrupt or gradual depending on the rate of variation of radar scattering parameters of the area being examined. Color combination tends to point out some density differences, but this is often not emphatic enough for rapid evaluation. Also, for some combined images, some of the subtle differences at points of abrupt density changes are not immediately evident.

To overcome some of these problems, a differentiation unit has been devised for edge enhancement. This unit consists of three independent differentiation circuits. On each circuit the time constant can be adjusted over a wide range thereby detecting lines of different widths. The circuit is modified so that both positive and negative derivitives give a positive output. By using three independent differentiating circuits we can simultaneously enhance three input photographs in contrasting colors on the CTV. Also, it is possible to hook the derivitive circuits in sequence so that we can take multiple derivitives of the same signal.

Data Space Sensor

The data space sensor (DSS) allows the operator to arbitrarily choose a certain joint photographic density value from two images and these points can be indicated by an enhancement of an image presentation. It measures the value of joint probability density functions at particular points for a two image system. Figure 3 shows an oscilloscope trace of the trajectories due to simultaneously scanning two photographs and presenting these on an oscilloscope. The length of time the trajectory is in a small area on the oscilloscope is proportional to the probability that the joint photographic density is within the values defined by the small area. The DSS consists of a fiber optics rod which is physically held over the face of the oscilloscope. The area sampled is defined by the cross-sectional area of the rod and the magnification of the two-dimensional trajectory densities on the face of the oscilloscope. A photomultiplier tube connected to the rod has an output whenever the trajectory is in the sampled area. Since the output of the photomultiplier is proportional to the time spent in the sampled data space it measures the joint probability. The unique application of the DSS makes use of the fact that the output of the photomultiplier is obtained synchronously with the presentation of information in a flying-spot scanner system. This output, therefore, can be used to emphasize or enhance in a special color the image that is being presented at a particular time on the CTV.

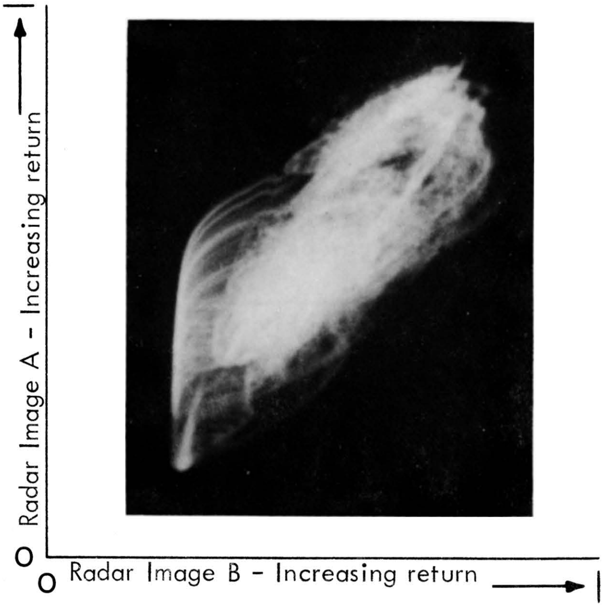
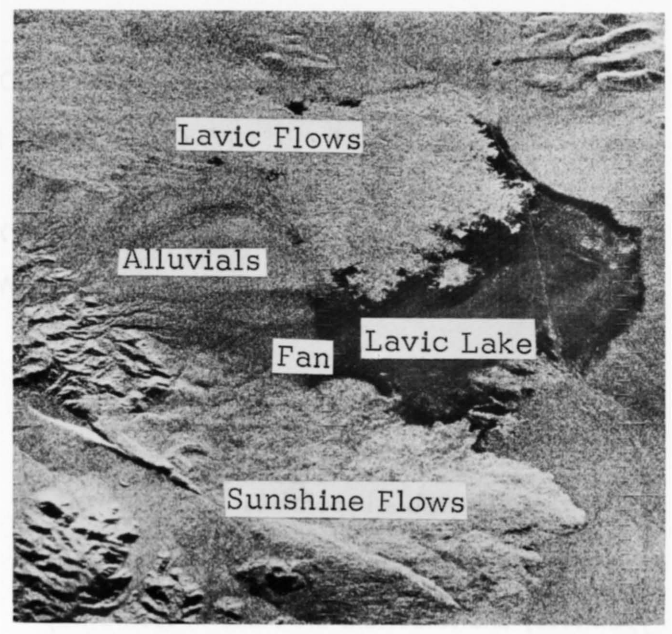


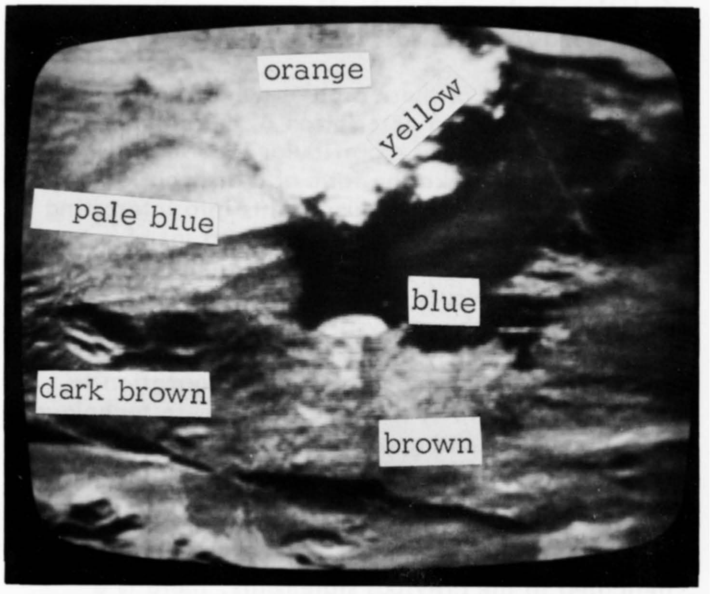
Figure 3.- Joint-probability density plot produced by scanning two radar images simultaneously.

Schmitt Trigger

The Schmitt Trigger (ST) is a device which can select any position in the dynamic range of a single radar image for selective enhancement on the CTV in color. All areas with values above or below that selected are assigned one color. Each of a number of images can be scanned in turn clipping at the same level so that later comparisons may be made.

Because the ST clips at a finite level certain areas (say lithologies) may be displayed as solid color on the CTV if all components of the probability density function of the area (lithology) lie above or below the selected clipping level. In other cases, areas which cannot be distinguished from one another on the basis of average gray scale (such as would be obtained with a microdensitometer trace) may be discriminated because the skirts of the probability density function of one natural object may lap across the zone where the ST is clipping while another may not. It therefore becomes possible to discriminate between two closely similar regions on the basis of the degree to which they acquire a speckled character on the CTV arising from the presence of small numbers of high intensity returns or low intensity returns, as the case may be. It becomes apparent, then, that the ST is a device which samples in a broader area of data space than the DSS unit, but that at the same time it may add to our ability to discriminate. At the time of the writing of this paper only one Schmitt Trigger was in operation. By the time of the symposium, however, we will have Schmitt Triggers which can slice in such a manner as to produce oblongs or squares in data space. Figures 4 and 5 may be compared to see the nature of the color combinations and enhancement possible.





FUTURE CAPABILITIES

The IDECS system as now operative gives some textural discrimination. However, this is not specifically a portion of the design, but arises incidentally from operation of the Schmitt Trigger and Data Space Sensor. A modification to the system is planned which will enable both improved texture discrimination and measurement of accutance This will be achieved by changing the sweep in the flying spot scanner to give various modes of spiral and circular trajectories.

Another technique of potentially great discriminant power is now under development, which will represent a significant expansion of capabilities of the Data Space Sensor and Schmitt Trigger. This device, for which both the theory and circuit design have been completed, is intended to sample in n-space and produce a binary color-coded output in the form of a color coded image. On this image discrete classes should be distinguishable. This latter development forms the basis for Mr. Dalke's presentation at this colloquium and is described in detail by him.

REFERENCES

Colwell, R. N., 1961, Some practical examples of multiband spectral reconnaissance: Am. Scientist, v. 49, p. 9-36.

Dellwig, L. F., and Moore, R. K., 1966, The geological value of simultaneously produced like- and cross-polarized radar imagery: Jour. Geophysical Res., v. 71, p. 3597-3601.

Morain, S. A., and Simonett, D. S., 1966, Vegetation analysis with radar imagery: Proc. Fourth Symposium Remote Sensing of Environment, Institute of Science and Technology, Univ. Michigan, p. 605-622.

IMPLEMENTATION OF PATTERN RECOGNITION TECHNIQUES AS APPLIED TO GEOSCIENCE INTERPRETATION

by

George W. Dalke
University of Kansas

INTRODUCTION

Statistical decision theory with a Bayes strategy provides a powerful technique for selecting one of a set of possible hypotheses in a statistical environment. The formulation requires a statement of conditions of profit and loss (either economic, strategic, or scientific) relative to the possible hypotheses, and selections are made in such a way to maximize the average profit.

For a problem in which a meaningful set of data is provided and the desired results are precisely formulated as a set of testable hypotheses, decision theory provides a mathematically optimum solution (relative to the strategy). However, neither of these conditions are fulfilled in most geoscience problems.

In many of these problems the final results desired are simply a description of the nature of the earth's surface at a particular location. This description may take the form of a distribution map of the location which groups points into interesting categories. Clearly, formulation of a set of hypotheses from which a typical distribution map could be constructed would be prohibitively complex. Further, unless the analysis of the location is simple, the investigator may use somewhat subtle techniques and previously obtained collateral information to deduce the nature of the location. These techniques and information are generally referred to as "insight" and "experience" and are probably too complex to include in the decision-making process.

Although statistical decision theory cannot compete with the trained investigator on the level indicated in the previous statements, there is a growing application for which the investigator is unable to use his special talents. This refers to the great amounts of data that are being obtained remotely from aircraft and satellites. This data may either be too extensive for the investigator to grasp or may be in such an unfamiliar form that the investigator's insight and experience cannot be used.

This report discusses a unique approach to this problem that is in the early stages of investigation at CRES (Center for Research in Engineering Science, University of Kansas). This approach uses statistical decision theory to select and transform the essential properties of input data into a form suitable for interpretation by the trained investigator. There is a logical twofold effort involved in this study: applying decision theory to geoscience sample data, and the investigation of effective techniques for displaying information.

Some of the display techniques are described in "Application of color-combined, multiple polarization radar images to geoscience problems" by D. Simonett in this issue of Computer Contributions. This report will outline the concept of the integrated system and the technique for including statistical decision theory.

STATISTICAL DECISION THEORY WITH A BAYES STRATEGY AND A DECISION SYSTEM

A generally accepted model for a decision system (shown in Figure 1) consists of a Receptor and Categorizer with abstractly specified collections of elements that envelop all possible inputs and outputs of the devices.

Object space is the set of all objects that can stimulate the receptor. Each object or category of objects is denoted s_k where k is an index used to identify a particular object or category. Decision space is the set of all interesting features of object space. Each feature or category of features is commonly referred to as an hypothesis and is denoted S_J where J is an index used to identify a particular hypothesis. Measurement space is the set of all possible outputs of the receptor, each of which is denoted by the n-dimensional vector $\mathbf{x} = (x_1, \ldots, x_n)$.

In operation, a particular object is exposed to the Receptor and certain members of decision space are indicated by the Categorizer.

The internal structure of a receptor is shown in Figure 2. A receptor consists of a set of tests, each of which performs a measurement on an imput object. The measurement resulting from the ith test is a number x_i. For convenience we represent the ordered set of measurements $(x_1, \ldots, x_n)$ as the components of a vector $\mathbf{x}$.

The output $\mathbf{x}$ is not uniquely determined by specification of the object k for two reasons. The first is that the receptor can be affected by extraneous signals such as thermal noise. The second is that,

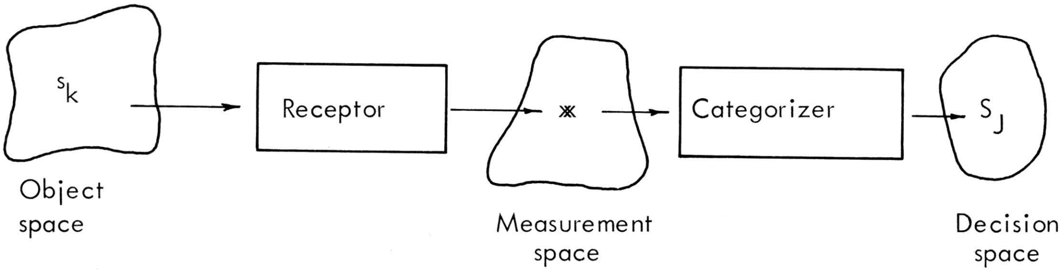


Figure 1.- Model for a decision system.

in practical cases, the index k refers not to a particular object at some juxtaposition of space and time, but rather to a group of objects that we wish to consider identical and which, nevertheless, affect the receptor differently. We therefore consider $\mathbf{x}$ as a random variable of a statistical process governed by a joint probability density function which depends on the class of objects to which the input object belongs. These conditional probability density functions are represented as $p(\mathbf{x} | s_k)$. The set of these probabilities for all values of k completely specifies the operation of the receptor.

The function of a categorizer is to examine an input measurement $\mathbf{x}$ and indicate an output decision, S_J. The general formulation for this process is called a decision rule and is denoted as $d(S_J | \mathbf{x})$. Precisely stated, the decision rule is a function giving the probability that the output is J when the input is $\mathbf{x}$. The internal structure of the categorizer depends intimately on the strategy and assumptions concerning a particular type of decision. Some common decision techniques are:

1. Classification by hyperplane
2. Multivariate discriminant classification
3. Maximum likelihood estimators
4. Linear perceptrons.

These techniques are equivalent for certain classes of problems and some are equivalent for all classes of problems. In addition to these techniques, some of the strategies used to obtain the "best" decision rule are:

1. Minimum mean square error
2. Minimum false identification
3. Minimum false dismissal
4. Maximum *a posteriori* probability of correct classification.

Each of these techniques and strategies are special cases of statistical decision theory with a Bayes strategy, which is described in the next section. All of the above techniques require an *a priori* assumption of the form of the probabilities, $p(\mathbf{x} | s_k)$. To avoid this necessity, a solution for quantized measurement vectors will be presented that requires no statistical assumptions.

FORMULATION OF AVERAGE RISK

We define a loss function $L(S_J, s_k)$ as the loss associated with deciding S_J when the object category is s_k. By convention a "profit" is designated as a negative number in constructing the loss function.

The function $\sigma(s_k)$ is a probability function designating the *a priori* probability of occurrence for each input object or object category.

The expected value of the loss for a given decision rule is called the risk and denoted R(d). The risk is given by

$$R(d) = \sum_{S_J} \sum_{s_k} \int_{\mathbf{x}} L(S_J, s_k) d(S_J | \mathbf{x}) p(\mathbf{x} | s_k) \sigma(s_k) d^n x$$

using the shorthand notation

$$\int_{\mathbf{x}} d^n x \equiv \int_{x_1} \int_{x_2} \cdots \int_{x_n} dx_1 dx_2 \ldots dx_n$$

$$\sum_0^{\mathbf{r}} \equiv \sum_0^{r_1} \sum_0^{r_2} \cdots \sum_0^{r_n}$$

Any decision rule, d, that causes R(d) to take its minimum value is a Baye's decision rule. Although a decision rule is a probability function, it is easily shown that for simple loss functions there always exists a Baye's decision rule that is deterministic. Non-random decision rules are generally

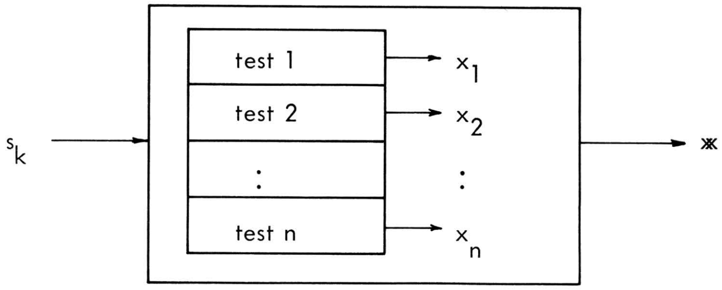


Figure 2.- The internal structure of a receptor.

advantageous in the design of equipment and will be the only ones considered below.

BAYE'S SYSTEM FOR QUANTIZED MEASUREMENTS

Consider a system for which each component, x_i, of the measurement vector can take on only certain discrete values, say 0, 1, 2, ..., r_i. Thus, a measurement vector, $\mathbb{x}$ has a lower bound 0 and an upper bound $\mathbb{r}$, where $\mathbb{r}$ has components r_i. The risk for such a system is:

$$R(d) = \sum_{S_J} \sum_{s_k} \sum_{\mathbb{x}=0}^{\mathbb{r}} L(S_J, s_k) d(S_J \mid \mathbb{x}) p(\mathbb{x} \mid s_k) \sigma(s_k)$$

where, as before, d refers to a particular decision rule and σ to a particular a priori distribution. We define $R(d,\mathbb{x})$ as follows:

$$R(d,\mathbb{x}) = \sum_{S_J} \sum_{S_K} L(S_J, S_x) d(S_J \mid \mathbb{x}) P(\mathbb{x} \mid s_n) \sigma(S_k)$$

If we choose to consider no functional constraints on the decision rule, then R(d) is a minimum when each $R(d,\mathbb{x})$ is a minimum. In order to consider individual measurements we take $\mathbb{x}$ to be a subscripted quantity $\mathbb{x}_i$ where i = 1, 2, ..., u and

$$u = \prod_{i=1}^{n} r_i$$

The algorithm for calculating a non-random Baye's decision rule is:

(1) Form the matrix $\mathbb{P} = (P_{ki})$ and $\mathbb{L} = (L_{Jk})$, where

$$P_{ki} = P(\mathbb{x}_i \mid s_k)\sigma k; \text{ and}$$

$$L_{Jk} = L(s_k, S_J).$$

(2) Define $\mathbb{R}$ by the m x u matrix (where m is the number of hypotheses and u is defined above)

$$\mathbb{R} = \mathbb{L}\ \mathbb{P}$$

(3) Examine each column of $\mathbb{R}$. If the smallest element in the ith column is R^i, and its location in the column is J^i, then the set of (J^i) is the Baye's decision rule and

$$R = \sum_{i=1}^{u} R^i \text{ is the total average risk.}$$

This completes the required algorithm. For further study a good presentation of statistical decision theory is presented in Statistical Communication Theory, David Middleton, McGraw Hill, 1960.

DATA AND DISPLAY FORMAT

Convenient and sufficiently general representations for input data and display information are matrices. For uniformity the data and information are registered so that corresponding elements on any of the input or output matrices refer to the same point on the earth's surface. Using this convention requires that the size of the matrices be determined by the "largest" input which makes it necessary to designate, say with the symbol ∅, unused matrix locations for some inputs.

INPUT DATA

The following list refers to some typical input data:

1. Image data
   a. Spectra-zonal photographs (IR and visible light)
   b. Radar imagery (cross and like polarization, various frequencies and resolutions)
2. Linear data
   a. Scatterometry radar data (various frequencies, polarizations, and inclination angles)
   b. Field investigation data (any feasible set of measurements)
3. Point data
   a. Field investigation data.

Point and linear data can generally be directly entered into the required matrix formulation. For image data a flying spot scanner system can be used to set up the required matrix (Some present f.s.s. systems will reduce 70 mm photographs to a 1024x 1024 matrix with a photographic density resolution of 64 gray levels.)

For a typical problem there may be a number of matrices for each type of data input, that is, there may be n_i image matrices, n_l linear matrices and n_p point matrices. The total number of matrices is $n = n_i + n_l + n_p$ which is the dimension of the measurement vector for the recognition system.

DISPLAY

It was pointed out that for a typical image input the size of the system matrices may be 1024x 1024. This is prohibitively large for an investigator to readily interpret. However, if the matrix elements are presented as intensities proportional to the value of the elements, then the display matrix will appear as an image. Actual systems under investigation use an image-forming display such as a cathode ray tube, or more generally, a color television tube. The matrix representation is simply an abstraction to aid system development and to allow simulation of the system on digital computers.

A color display system has two advantages.

First, the use of colors has a number of advantageous subjective properties. The human eye is able to integrate complex color patterns and allow the brain to form a simplified generalization for the patterns. Whereas changes in intensity are highly resolved by the eye for nearby points, there is little discrimination for distant points. The use of colors, on the other hand, gives little nearby discrimination but allows similarities of distant points to be sensed.

The second advantage of a color display is that each point on the display is able to present three dimensional information to the observer. For a suitably constructed system, these dimensions consist of an intensity and two independent color variables. A more detailed discussion of the properties of colors is given in CRES Technical Memorandum 61-4 (Dalke, 1964).

UTILIZING QUALITIES AS DATA

One conventionally imagines input data from some point to be measurements of a dynamic range of possible values, such as "the scattering cross section of the point at x-band is -10.4db" or "the moisture content of the soil at the point is 0.03%" where subsequent measurements could be -10.3db. or 0.05%, and so on. Some of the most important data concerning a point, however, may not be of this form. For example, to say a particular point is "limestone" is important information but we generally do not consider a dynamic measurement such that a value of 3.4 indicates limestone and 3.7 indicates basalt. In the former case two measurements that are close to the same value indicate a similarity between the points, whereas in the latter "closeness" is not particularly meaningful.

We designate properties of a point that have no associated numerical designations as qualities. Some of the important data about regions and indeed the desired results of some investigations are the qualities of the points in a region placed in their geometric context. An extremely useful property of the quantized measurement decision rule is that any set of mutually exclusive qualities can be used as an input dimension.

Because of the subjective nature of color and intensity, the display for the general processing system discussed in this report will use certain selected colors to indicate qualities and will use intensity to enhance the displayed information.

DESCRIPTION OF PROCESSING SYSTEM

A simplified flow diagram for the processing system is shown in Figure 3.

This system is a straightforward implementation of the techniques presented previously in this report. In order to describe the operation of the processor an example problem will be considered and the results of two computer simulated problems will be given.

The required circuitry for a real-time version of the processor is presented in Appendix D of CRES Technical Report 61-16 (Dalke, 1966). Computer

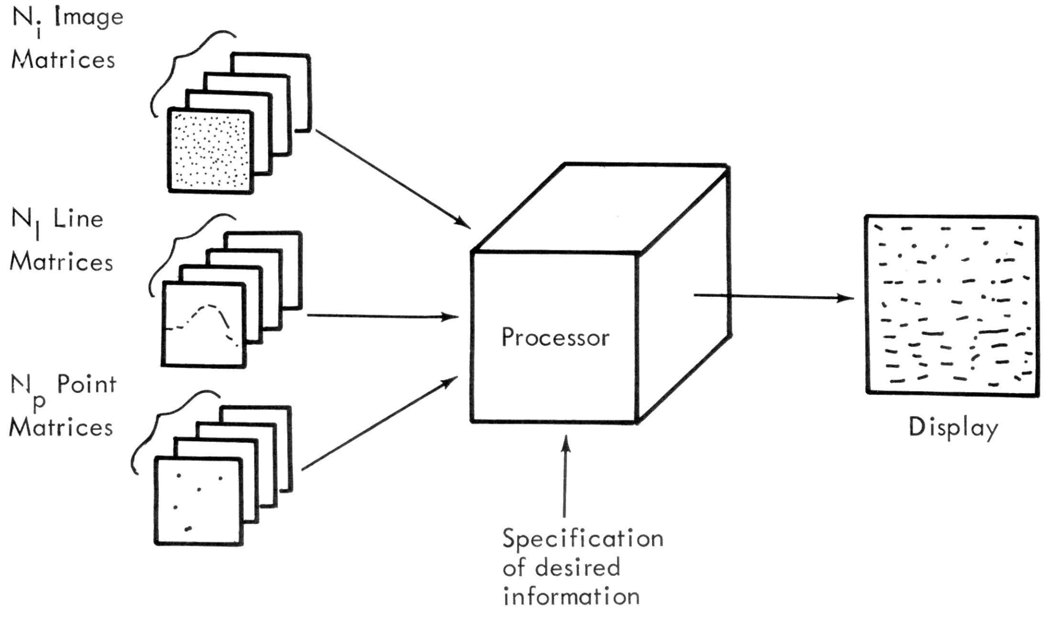


Figure 3.- Simplified flow diagram for the processing system. The processor contains a decision system to translate input data to a form suitable for display and contains enhancement devices to aid the interpretation of the displayed information.

programs simulating the operation of the processor have been prepared although the lineprinter output is an unsatisfactory representation of the color display described in this report.

EXAMPLE PROBLEMS

Problem 1.- Three 70 mm photographs selected from a set taken by an Itek nine lens multispectral camera in a flight over Phoenix, Arizona were processed relative to the hypotheses: (1) cultivated area of type A, (2) residential area, (3) streets, (4) cultivated area of type B, (5) unimproved roads, (6) multilane highways, (7) cultivated area of type C, and (8) no decision.

Comparison with standard regions identified on enlarged photographs indicated points were correctly classified about 70% of the time. The computer printout for this problem is shown in Figure 4. A complete description of this experiment is given in CRES Technical Report 61-16 (Dalke, 1966).

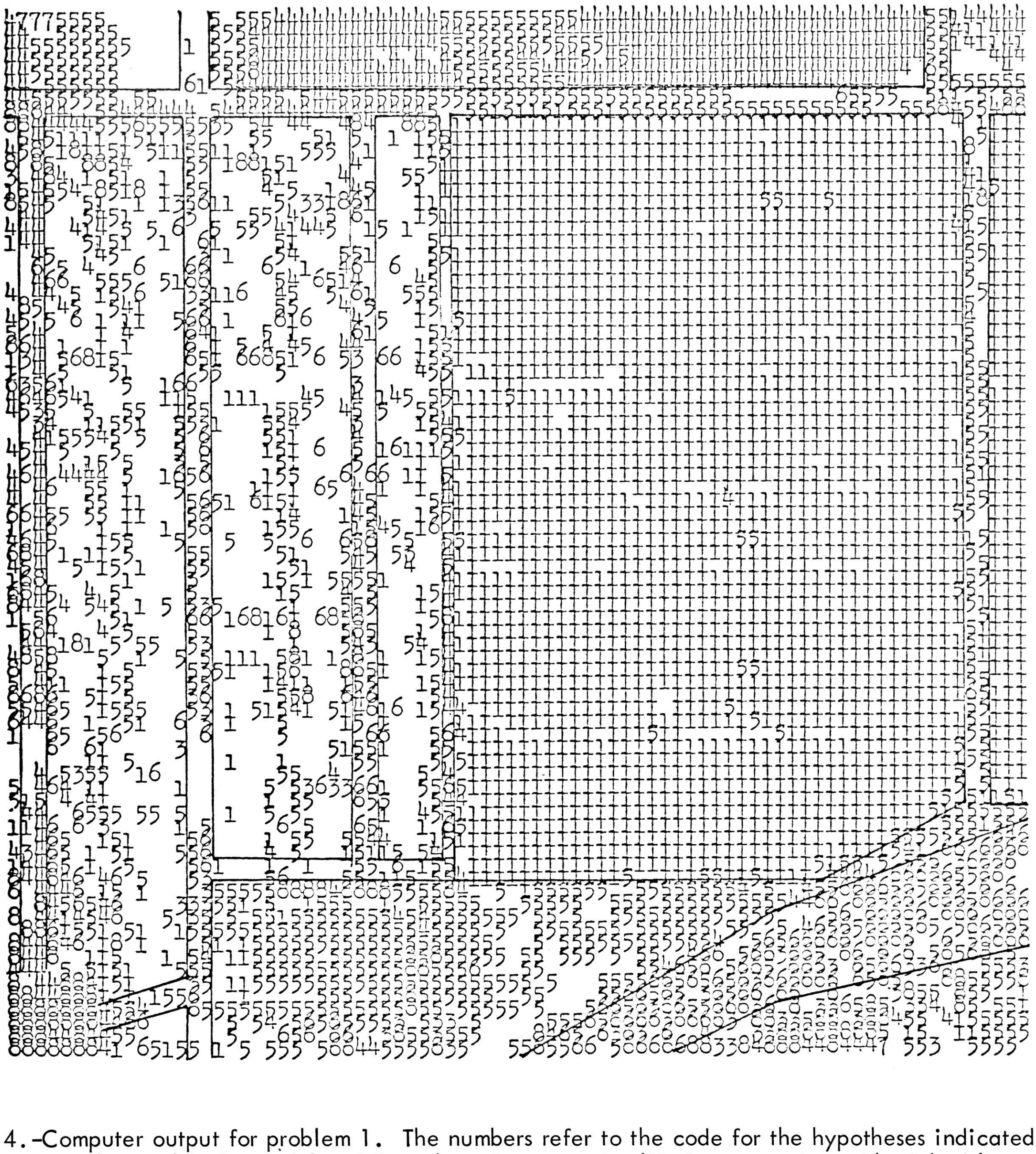


Figure 4.-Computer output for problem 1. The numbers refer to the code for the hypotheses indicated in the text. The number 2 was deleted from the printout to simplify interpretation. The inked lines indicate roughly the boundaries of the regions on the original photographs.

Problem 2.- An x-band scatterometry radar system measured the scattering cross-section of the terrain for like and cross polarization and incidence angles of 0°, 30°, 45°, and 70°, in a flight over Pisgah Crater, California. The data were processed relative to the hypotheses: (1) sloping alluvial deposits, (2) flat silt and clay deposits, (3) through (8) aa and pahoehoe lave with various structural differences, (9) and (10) two types of igneous rock outcrops. Points were classified correctly 97% of the time. The quality of the data, however, was too poor to attach any particular significance to these rather remarkable results. A complete description of this experiment will soon be published as CRES Technical Report 61-17 (Dalke, 1966).

Problem 3.- We consider the hypothetical example for which field data has provided accurate selection of certain given hypothesis at isolated points in a large region. The problem is to take remotely obtained data (radar and optical images) for the region and generalize the data from the isolated points to the entire region in a statistically optimum fashion . This problem can be performed on the processor. An interesting display of the results would be to code each of the hypotheses in a particular color and use the intensity to register a confidence level for the generalized results. The points for which the field data were obtained will be the assigned color and will be the brightest points on the display. Points that are considerably different will be dark and will, therefore, have less effect on the observer's interpretation.

CONCLUSIONS

The subjective nature of much of geoscience interpretation makes the application of decision theory techniques difficult. The processor described in this report utilizes decision theory for the intermediate problem of preparing data so that it is suitable for interpretation and utilizes techniques for effectively presenting the transformed data. This concept of operator involvement in a processing system may best be suited for effectively utilizing the power of decision theory and the capabilities of the experienced investigator.

REFERENCES

Dalke, George, 1964, Color theory: CRES Technical Memorandum 61-4, 16 p.

Dalke, George, 1966, Automatic processing of multi-spectral images: CRES Technical Rept. 61-16, 61 p.

Dalke, George, in press, Automatic processing of scatterometry data and application to Pisgah Crater: CRES Technical Rept. 61-17.

A FACTOR ANALYTIC SCHEME FOR GROUPING AND SEPARATING TYPES OF FOSSILS

By

Max Pitcher
Continental Oil Company

ABSTRACT

Fossil groups such as fusulinids, when described quantitatively, can be compared and classified by an n-dimension vector analysis. Nine species from four genera of fusulinids are grouped by a factor analytic scheme. When compared to a traditional classification specific and generic fields emerge on the two dimensional factor plots. Boundaries of groups are arbitrary. Generic relationships can be studied from the factor plots. The close similarity of Pseudoschwagerina to Triticites is shown and the gradational nature of Dunbarinella glenensis between Dunbarinella and Triticites can be inferred. Three species of the fusulinid genera Dunbarinella show specific separations when factored alone.

A modified factor score, calculated by post multiplying the transpose of the normalized raw data matrix by the square of the rotated factor matrix with sign retained, ranks the influence of different morphologic variables on generic and specific separation.

INTRODUCTION

Attempts to use only quantitative data in identifying and classifying organisms presupposes those data to embrace all of the distinguishing characters necessary for the identification and classification of the organisms. Not all organisms can thus be described. Fusulinids, however, are amenable to the use of numeric methods for their identification. For a routine description of a fusulinid some twenty to thirty characters are measured, and relatively few of the distinguishing criteria are not included in these numeric data. Several of these measurements described various growth stages of the foraminifer. Some of the characters not usually measured quantitatively are position and intensity of septal fluting, shape of the spirotheca near the poles (various intensities of concavity and convexity), chomata intensity, and development and position of cuniculae. Chomata intensity and development of cuniculae are generic characters in Triticites-Schwagerina and Schwagerina-Parafusulina respectively, but the former are criteria of specific differences. If these characters are not used, the analysis may give spurious results. Fortunately those characters which are not numeric in nature may be correlated with those that are; nonnumeric characters are not included in the analysis.

A simple bivariate analysis is not able to cope with the complex descriptive data. A multivariate scheme such as factor analysis (Imbrie, 1963) is well suited to separate morphologic groups by a simultaneous consideration of several parameters.

Data for this study was taken from M. L. Thompson's American Wolfcampian Fusulinids (1954). Representative samples were taken from his data to compare the results of a strictly numeric analysis with Thompson's identifications. The species and numbers of individuals chosen are: Dunbarinella fivensis - 8; D. americana - 6; D. glenensis - 7; Triticites creekensis - 7; T. rockensis - 6; Schwagerina vervillei - 7; S. longissimoidea - 8; Pseudoschwagerina needhami - 4; P. texana - 1. Raw data from Thompson's tables were standardized by a percent maximum transformation obtained by dividing each value of a parameter by the maximum recorded for that parameter. The 55 samples were compared to each other by a COS θ similarity coefficient and factored in the Q-mode. Nine centroid factors were extracted and seven rotated by the Varimax criterion. Only four are considered significant. Figure 1 is the resultant rotated factor matrix, and communalities. Figure 2 is a plot of representative species of each genus on two of the factors.

Acknowledgments.-I acknowledge the guidance of John Imbrie under whose direction this work was completed.

METHODS

Raw data from Thompson's tables were tabulated with samples on the rows and variables down the columns. To exaggerate the spread of the statistics, the minimum value of each variable was subtracted from each value of the variable so that there was at least one zero in each column. Then the percent of the maximum was obtained by dividing each value by the maximum of that variable. The data matrix was transposed so that the samples were in columns and variables in rows.

The transposed matrix was then subjected to a factor analysis. All of the analyses are in the Q-mode. Nine centroid factors were extracted and seven rotated. Figure 1 is the rotated factor matrix. Figures 2-6 are plots of representative species from the genera on several combinations of factors. As shown in Figure 3, when all of the 55 samples are plotted on a two dimensional display, the fields merge. Notice how adding a third dimension reduces the overlap of *D. glenensis* and *T. creekensis* (Fig. 4). The groups circled with dotted lines indicate species as identified by Thompson. Those circled with solid lines are quantitatively misclassified within the dimensions shown. Because it is so difficult to illustrate the group relations in four dimensions, only representative species are plotted on some of the figures.

There is considerable overlap on the plot of factor I and II (Fig. 3), but the areas of concentrations are well defined for each genus. Generic fields can be delineated. One of the purposes of such plots is not only to separate groups, but to see morphologic relationships. The plots show the close morphologic similarity of *Psuedoschwagerina* and *Triticites*, thus reflecting the possible phylogenetic relationship of the two genera. *Dunbarinella* and *Triticites* are well separated.

The use of several species from so many genera seemed to be too much variation for the system to extract specific groups, therefore three species from the genus *Dunbarinella* were factored. Four factors were extracted, but the first three explain most of the data. Figure 7 is the resultant rotated factor matrix. Figures 7 and 8 are plots of the specimens along the factors. The circled groups are Thompson's species. Only one specimen does not agree with the original classification. The boundaries of the species as shown would be difficult to choose without some *a priori* determination of their affinites, but their relative position facilitates morphologic comparisons. The author knows of no method available to extract groups that are not arbitrary. Grouping decisions can be tested, however, by a multivariate scheme such as discriminatory analysis.

In Figure 7, specimen 11 falls within the *D. glenensis* group and in Figure 8 within D. *americana*. It was placed by Thompson with *D. fivensis*. If a specimen remains within a group as they are plotted on various factors, it adds validity to its classification. If it changes affinity with the varying dimensions, its classification is dubious. The latter is true for specimen 11.

FACTOR SCORE

It is often difficult to evaluate the critical parameters causing separation in a Q-mode analysis. The end member compositions in the oblique projection scheme (Imbrie) and the R-mode analysis are attempts to summarize the interrelations of the variables. None of these schemes, however, is capable of summarizing outstanding contributions within factors of a particular Q-mode analysis. A factor score (Harman, 1960) is a more useful tool.

The factor score used for this study is simplified from that described by Harman (1960). It is obtained by relating the data matrix to the rotated factor matrix. Variables that are abnormally high are combined with high factor loadings so that the resultant product highlights those variables that contribute most to the sample separation. The factor score is computed in the following way:

1) Normalize each variable of the data matrix by subtracting the mean and dividing by the standard deviation.

Thus: $[Z_{ij}]_{(N,n)}$ normalized $= [Z'_{ij}]_{(N,n)}$

2) Transpose the Z'_{ij} matrix =

$$[Z'_{ij}]^{T}_{(n,N)} = [Z'_{ji}]_{(n,N)}$$

3) Square the elements of the rotated factor matrix, but retain the sign of the individual loadings.

$$[b_{if}]_{(N,m)} \longrightarrow \pm [b_{if}]^{2}_{(N,m)} = [B_{if}]_{(N,m)}$$

4) Complete the matrix product

$$[Z'_{ji}]_{(n,N)} \cdot [B_{ij}]_{(N,m)} = [F_{jf}]_{(n,m)} =$$

factor score matrix

Figure 9 is a factor score matrix for the Q-mode factor analysis of 55 samples and 27 variables.

INSPECTION OF FACTOR SCORE MATRIX

The factor score is read with the sign of the maximum factor loading carried to the same factor of the factor score. For example, most of the high loadings on factor II of the rotated factor matrix are positive (Fig. 1). A high positive number for a variable on the factor score thus indicates a greater than average reading for those samples with a high loading on factor II. Factor III is negative; therefore, high negative values on the third factor of the factor score are most significant.

Even though this is an averaging scheme, the statistical consistency is shown by examining some of the interrelated variables. Figure 1 shows *Psuedoschwagerina* and *Triticites* to have high loading on factor II. The following is an analysis of the morphologic traits of these genera from the factor score.

Factor II (Factor Score)

1. Variable 3 is the ratio; length over width. If the width is shown to be high, the ratio should be low. The loading of variable 2 is + 8.213 and variable 3 is - 4.173. This is indicating the robust nature of Tricites and Psuedoschwagerina.

2. A high loading of variable 3 indicates the large proloculus in Psuedoschwagerina and Triticites as opposed to the small proloculus in Dunbarinella.

3. The high volution height of volutions 1-6 is correlated with the high width of variable 2.

4. Variables 22-27 show the Pseudoschwagerina and Triticites have a thicker spirotheca than Dunbarinella.

Factor III (Factor Score)

Dunbarinella fivensis has the highest negative loading on the third factor of the rotated factor matrix. Its distinctive characteristics are shown by the high negative loadings of variables 11-16 on the third factor of the factor score. The dimminution in value of the form ratio from volutions 1-6 indicate that the early growth stages are the most distinct.

Factor IV (Factor Score)

Dunbarinella glenensis is dominant on factor IV of the rotated factor matrix. An examination of high negative loadings in the factor score matrix on factor IV shows D. glenensis to be long, wide, and have a large proloculus. This would bring it morphologically more similar to Triticites. This relationship is also shown by the gradational loadings of the factor matrix between D. glenensis and T. creekensis.

CONCLUSIONS

Factor analysis and factor scoring are useful tools in studying the interrelationships of fusulinid species. Exact separation of specific groups is arbitrary, but generic and specific "fields" are clearly defined. The factor score aids in isolating critical variables that contribute most to sample separation.

The author has successfully separated graptolite populations using similar methods (in press). With very close stratigraphic control of measured fossils it may be possible to discern minute evolutionary changes within species which will allow very close stratigraphic correlations.

REFERENCES

Harman, H. H., 1960, Modern factor analysis: University of Chicago Press, 337 p.

Imbrie, John, 1963, Factor and vector analysis programs for analyzing geologic data: Technical Report No. 6 of ONR Task No. 389-135, Contract Nonr 1228(26), Office of Naval Research, Geography Branch, 83 p.

MATRIX OF ROTATED FACTORS

	+	+	-	-				Commu-
Samples	1	2	3	4	5	6	7	nality
1	0.5636	0.1771	-0.6675	-0.3212	-0.1796	0.1962	0.0411	0.9702
2	0.5291	0.3736	-0.5710	-0.4151	-0.0604	0.0069	0.0820	0.9283
3	0.5416	0.1922	-0.6946	-0.3468	0.0930	-0.0937	-0.0654	0.9547
4	0.4571	0.2538	-0.6757	-0.3502	0.0765	-0.1537	-0.2177	0.9294
5	0.7521	0.2733	-0.3869	-0.4130	-0.0040	-0.0250	-0.0357	0.9625
6	0.4094	0.1709	-0.6521	-0.5025	0.0618	0.0620	-0.2490	0.9442
7	0.2914	0.4233	-0.6657	-0.4515	-0.1495	-0.0571	0.0115	0.9369
8	0.5235	0.3209	-0.4916	-0.2289	-0.4836	0.0058	-0.0507	0.9076
9	0.2254	0.1386	-0.9192	-0.1803	0.0308	0.0649	-0.0521	0.9554
10	0.2467	0.2448	-0.8997	-0.1790	0.0563	-0.0196	0.0071	0.9658
11	0.3341	0.2816	-0.5164	-0.5960	-0.2090	0.1527	0.0513	0.8825
12	0.3355	0.2568	-0.6212	-0.2006	-0.1744	-0.0412	-0.0610	0.6406
13	0.4858	0.2149	-0.7782	-0.1704	-0.0674	0.1590	0.1172	0.9605
14	0.6225	0.2231	-0.5068	-0.0458	-0.1935	0.0749	0.3288	0.8475
15	0.3207	0.5139	-0.6367	-0.3176	-0.1812	0.0435	0.0457	0.9100
16	0.2958	0.4280	-0.4454	-0.6922	0.0679	0.0990	-0.0817	0.9693
17	0.2329	0.4913	-0.2676	-0.7615	0.0478	0.0677	0.0341	0.9552
18	0.2524	0.5366	-0.3447	-0.6515	0.0785	-0.1306	-0.0676	0.9228
19	0.1521	0.5673	-0.2116	-0.7208	-0.0225	-0.1112	-0.1763	0.9533
20	0.2087	0.3512	-0.4238	-0.7214	-0.2523	0.1126	0.0438	0.9452
21	0.2396	0.5857	-0.2640	-0.6805	-0.0991	0.0080	0.0033	0.9431
22	0.4461	0.6194	-0.3081	-0.4812	0.0401	-0.0228	-0.0605	0.9150
23	0.2346	0.8755	-0.2151	-0.3187	0.0347	0.0201	-0.0325	0.9720
24	0.1757	0.8759	-0.0990	-0.3935	0.0008	-0.0338	-0.0395	0.9654
25	0.1824	0.8920	-0.1396	-0.3015	-0.1060	-0.0365	0.1267	0.9680
26	0.4202	0.8422	-0.1430	-0.2348	-0.0979	-0.0441	0.0483	0.9753
27	0.0786	0.9116	-0.1600	-0.3128	-0.0215	-0.0623	0.1053	0.9760
28	0.3661	0.8094	-0.3082	-0.2938	0.0380	0.0246	-0.0353	0.9737
29	0.5206	0.6415	-0.3686	-0.1836	-0.1498	-0.2117	-0.1796	0.9516
30	0.4311	0.6433	-0.3864	-0.2762	-0.1297	0.0104	-0.3439	0.9604
31	0.4973	0.6556	-0.1404	-0.2567	-0.1010	0.1184	-0.4000	0.9469
32	0.6928	0.5190	-0.2950	-0.1214	-0.1595	0.0395	-0.2235	0.9280
33	0.4946	0.7075	-0.3584	-0.2334	0.0302	-0.1505	-0.1332	0.9695
34	0.4574	0.8034	-0.2720	-0.1383	-0.0457	-0.1177	-0.1329	0.9813
35	0.5778	0.5434	-0.4689	-0.1975	-0.0559	0.2460	-0.1129	0.9644
36	0.5346	0.4273	-0.5696	-0.3324	-0.0856	0.1226	-0.1451	0.9467
37	0.7344	0.3531	-0.4666	-0.2172	-0.0344	-0.0290	-0.0800	0.9374
38	0.7980	0.2788	-0.4221	-0.1955	-0.0655	-0.0812	-0.2019	0.9826
39	0.8434	0.3334	-0.2976	-0.2211	-0.1055	-0.0837	-0.0719	0.9832
40	0.6822	0.4051	-0.5385	-0.1960	-0.0869	0.0901	-0.0488	0.9759
41	0.5878	0.6502	-0.3606	-0.1981	-0.0816	0.0733	-0.1132	0.9623
42	0.6973	0.3686	-0.5167	-0.1778	-0.1574	-0.0351	0.0418	0.9484
43	0.5330	0.6217	-0.4194	-0.2768	0.0876	0.1180	0.1169	0.9584
44	0.7523	0.4094	-0.3857	-0.2044	0.0647	0.2227	-0.0099	0.9779
45	0.6750	0.5221	-0.4809	-0.1485	-0.0057	0.1017	0.1130	0.9807
46	0.6607	0.5689	-0.3949	-0.2099	0.0951	0.0906	0.0104	0.9775
47	0.6451	0.6586	-0.2570	-0.2507	0.0717	-0.0132	0.0289	0.9849
48	0.5763	0.5779	-0.1911	-0.3907	0.0555	0.2518	0.1066	0.9331
49	0.7115	0.4935	-0.3930	-0.2649	0.0377	0.0372	0.0062	0.9774
50	0.6894	0.5051	-0.2084	-0.4060	0.0188	0.1389	0.1176	0.9721
51	0.4510	0.7412	-0.2955	-0.1277	-0.0792	0.2584	-0.0496	0.9319
52	0.3727	0.7811	-0.3792	-0.1920	-0.0732	0.2120	-0.0010	0.9800
53	0.4042	0.7564	-0.2755	-0.2207	-0.0153	0.3069	-0.1427	0.9749
54	0.3943	0.8257	-0.2540	-0.1906	-0.1487	0.1075	-0.0413	0.9734
55	0.4356	0.6342	-0.3862	-0.2786	-0.0767	0.2878	-0.0689	0.9121
Variance	13.9597	17.0690	11.3828	7.1026	0.7908	0.8758	0.9224	52.1032

Figure 1.- Rotated factor matrix for 55 fusulinid specimens from 4 genera and 9 species. Samples 1-6 *D. americana*; 7-15 *D. fivensis*; 16-22 *D. glenensis*; 23-28 *T. creekensis*; 29-34 *T. rockensis*; 35-42 *S. vervillei*; 43-50 *S. longissimoidea*; 51-54 *P. needhami*; 55 *P. texana*.

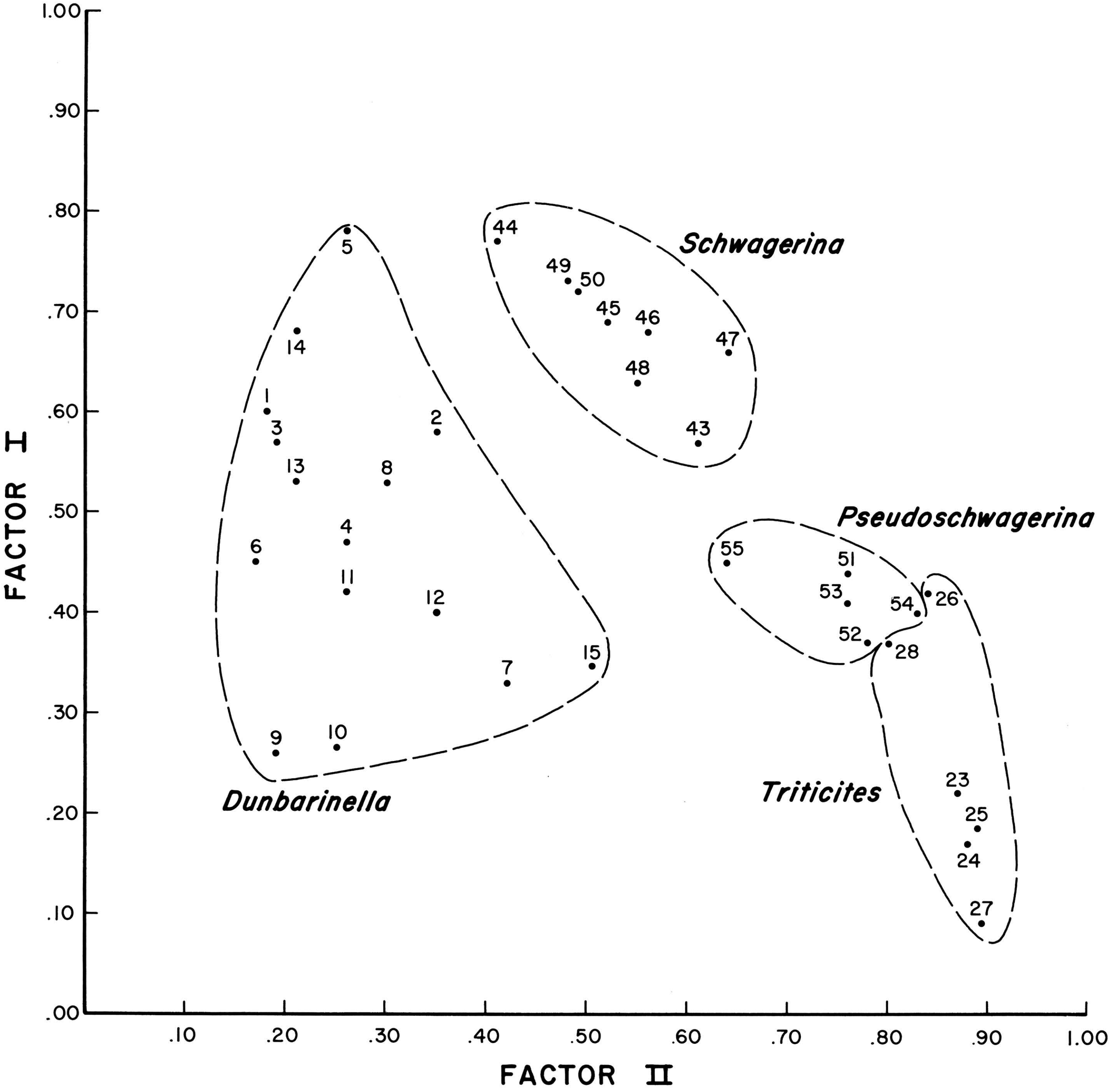


Figure 2.- Plot of two factor loadings of representative fusulinid specimens from 4 genera. Circled groups are generic identifications.

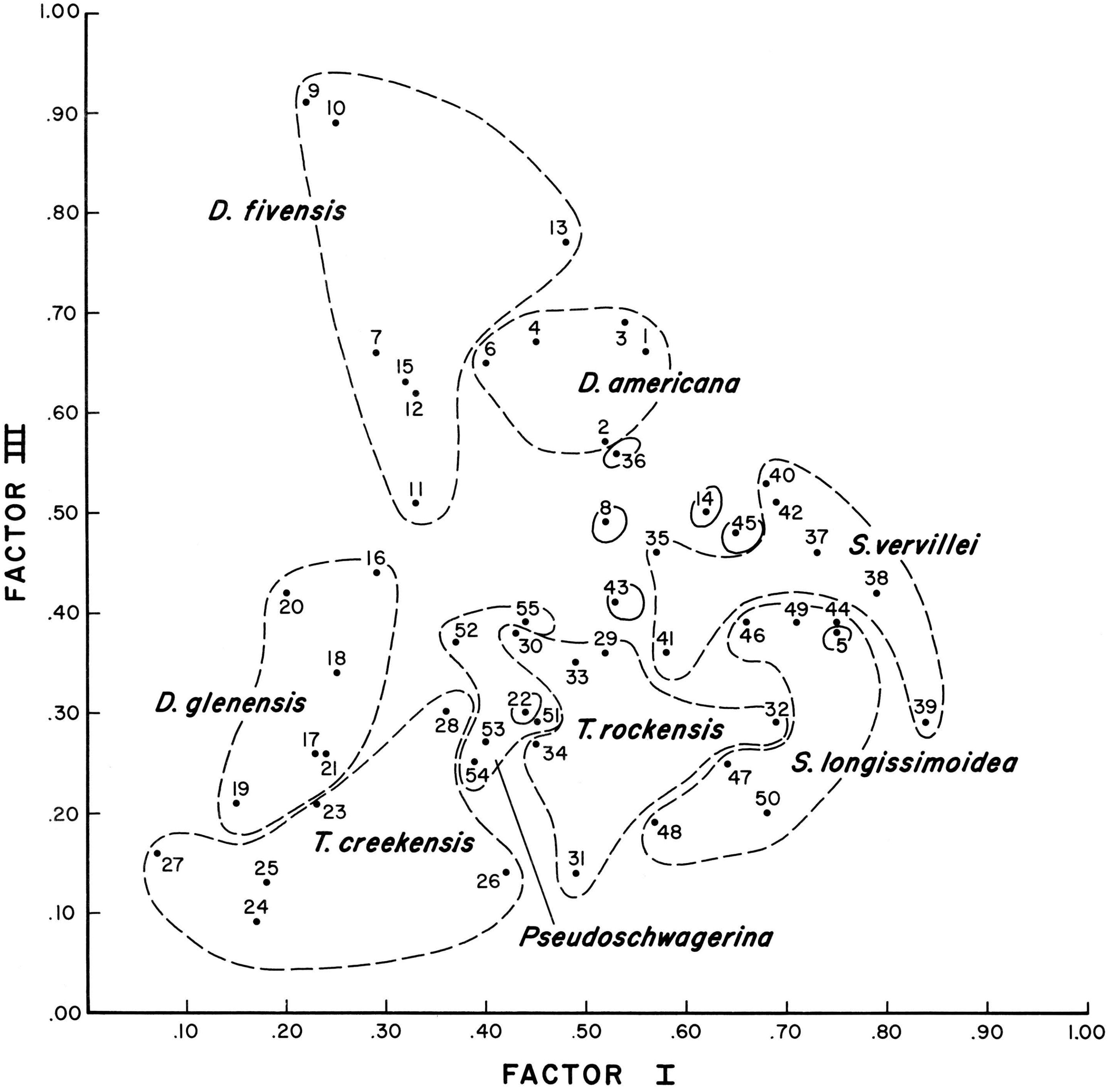


Figure 3. - Two dimensional plot of factor loadings of 55 fusulinid specimens. Dotted fields indicate Thompson's identification. Solid circles indicate specimens quantitatively misclassified in dimensions illustrated.

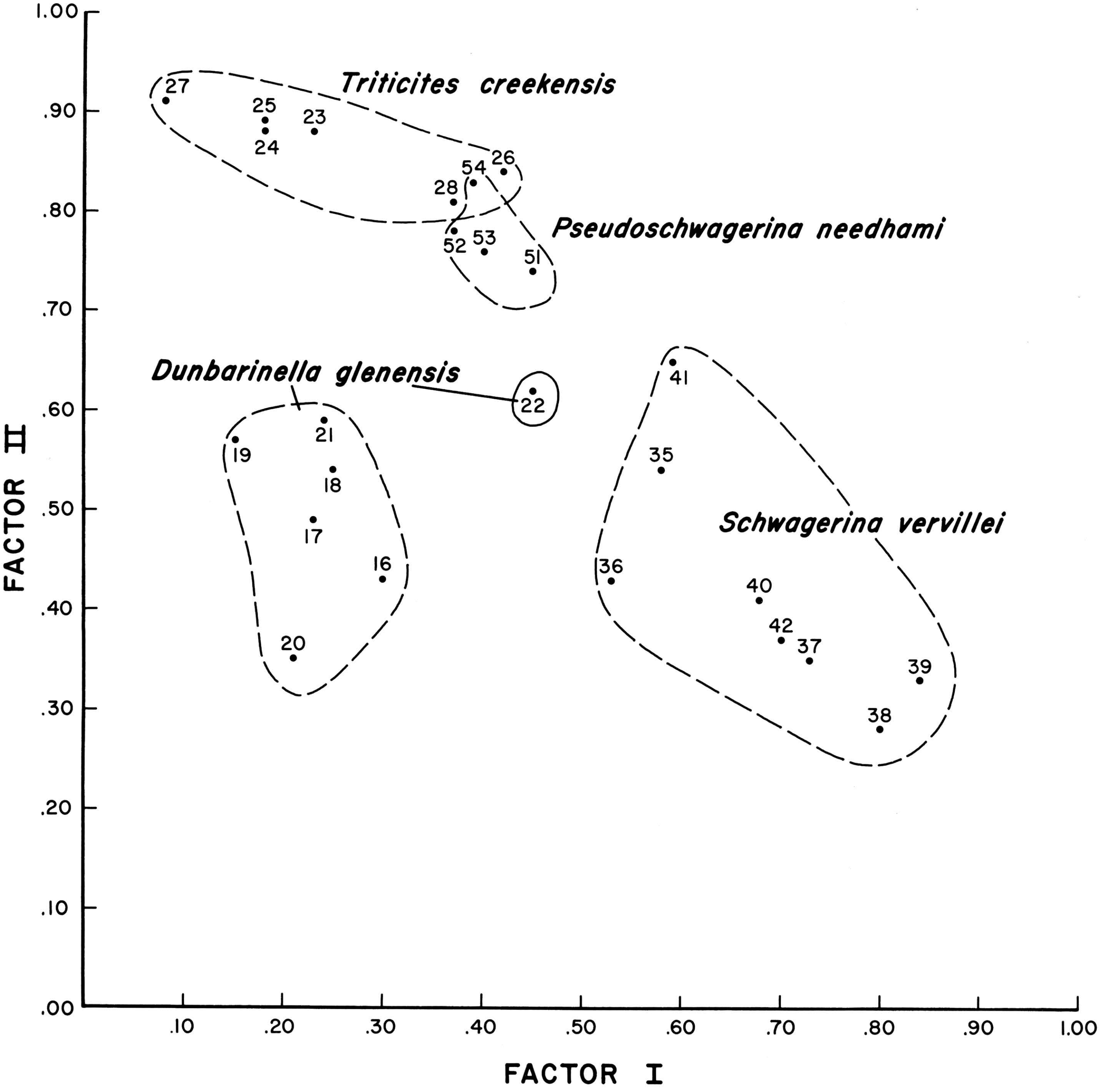


Figure 4. - Representative fusulinid species plotted in two dimensions. Solid circle indicates misclassification within these two dimensions.

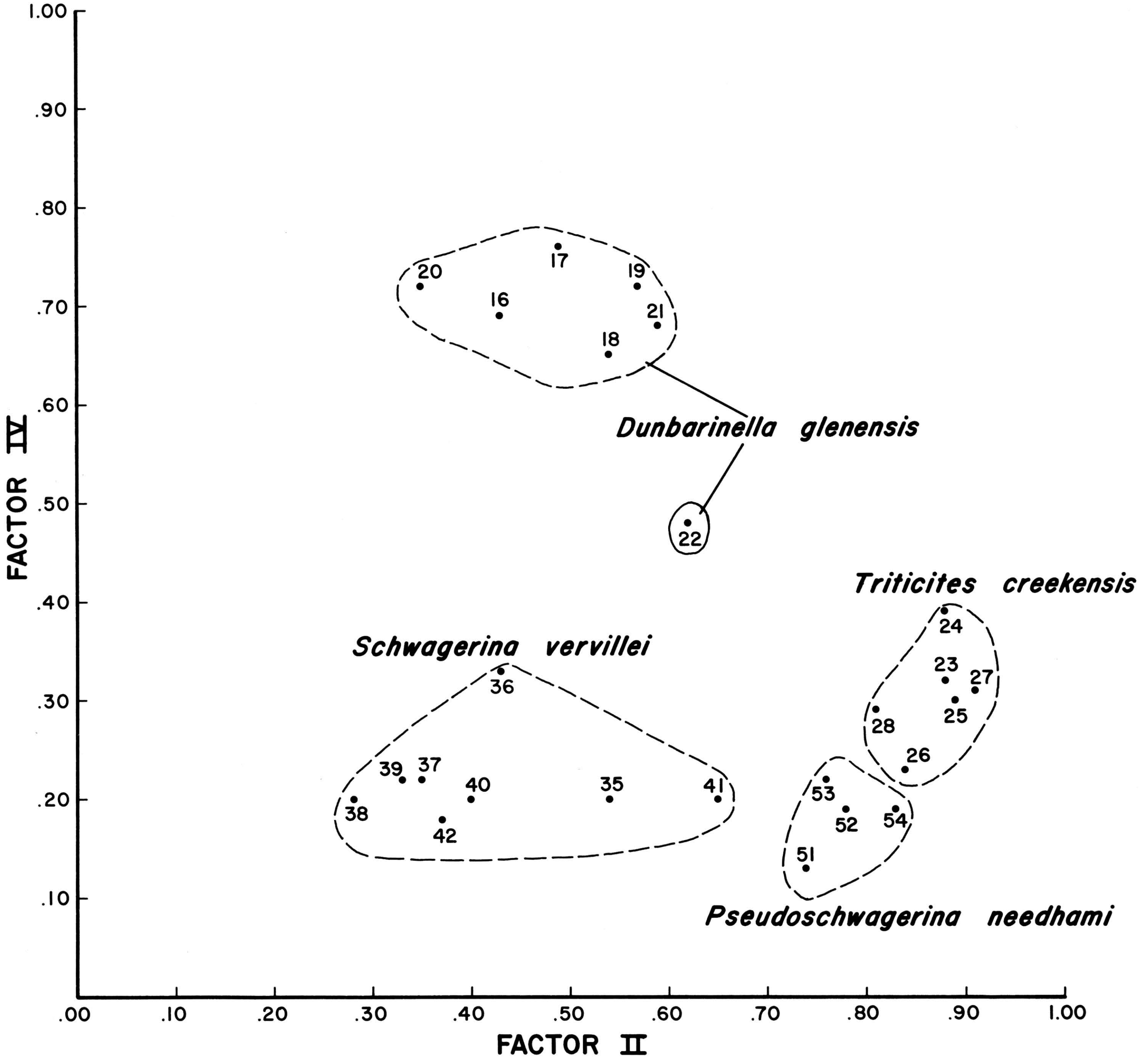


Figure 5. – Same fusulinid species as Figure 4, plotted in dimensions of factors II and IV.

MATRIX OF ROTATED FACTORS

1	-0.42153	-0.27021	0.78960	0.24790	0.93563
2	-0.29603	-0.50164	0.74580	0.14361	0.91612
3	-0.26347	-0.41128	0.80505	0.03995	0.88828
4	-0.39292	-0.46608	0.69919	0.05570	0.86358
5	-0.14128	-0.39889	0.86607	0.06992	0.93404
6	-0.19162	-0.37859	0.80219	0.25695	0.88958
7	-0.59256	-0.57951	0.37863	0.28686	0.91261
8	-0.61903	-0.37274	0.50176	0.28748	0.85654
9	-0.51218	-0.17628	0.67527	0.30318	0.84131
10	-0.63313	-0.32448	0.60814	0.17282	0.90584
11	-0.26724	-0.42100	0.51633	0.64314	0.92889
12	-0.65728	-0.47512	0.52707	0.12336	0.95078
13	-0.49084	-0.20343	0.77274	0.23356	0.93399
14	-0.50028	-0.16344	0.68321	0.28531	0.82517
15	-0.64315	-0.57819	0.44736	0.04219	0.94986
16	-0.23721	-0.76279	0.50609	0.19203	0.93112
17	-0.14348	-0.88053	0.31865	0.18073	0.93011
18	-0.14552	-0.84883	0.38120	0.08224	0.89377
19	-0.23210	-0.90476	0.14800	0.14820	0.91634
20	-0.36604	-0.64113	0.22704	0.57691	0.92941
21	-0.37465	-0.83922	0.23817	0.20596	0.94380
22	-0.28446	-0.74665	0.47813	0.07840	0.87315

Figure 6. - Rotated factor matrix for 22 fusulinid specimens belonging to 3 species of Dunbarinella. Samples 1-6 D. americana; 7-15 D. fivensis; 16-22 D. glenensis.

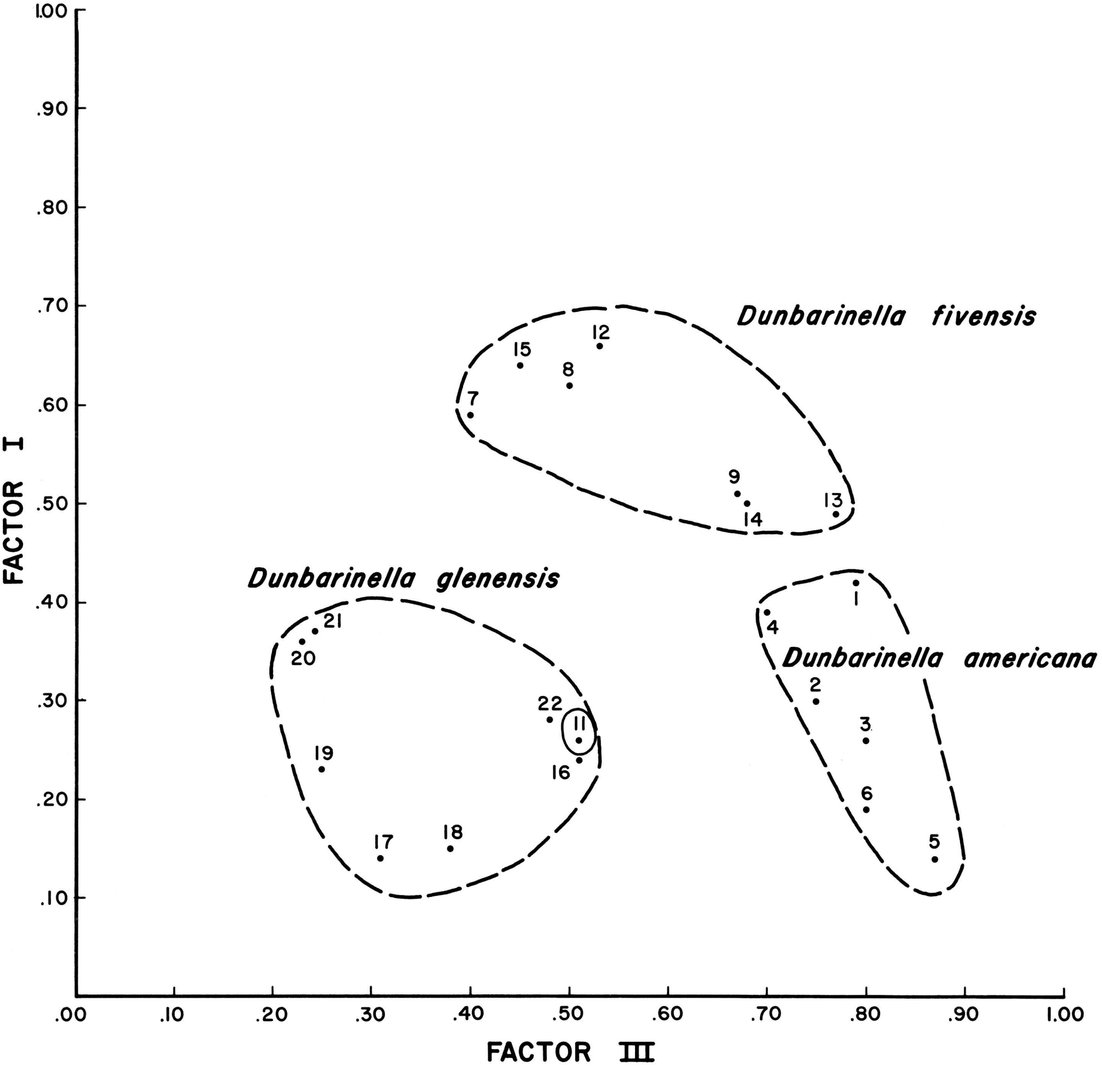


Figure 7. - Two dimensional factor plot of 3 fusulinid species of Dunbarinella.

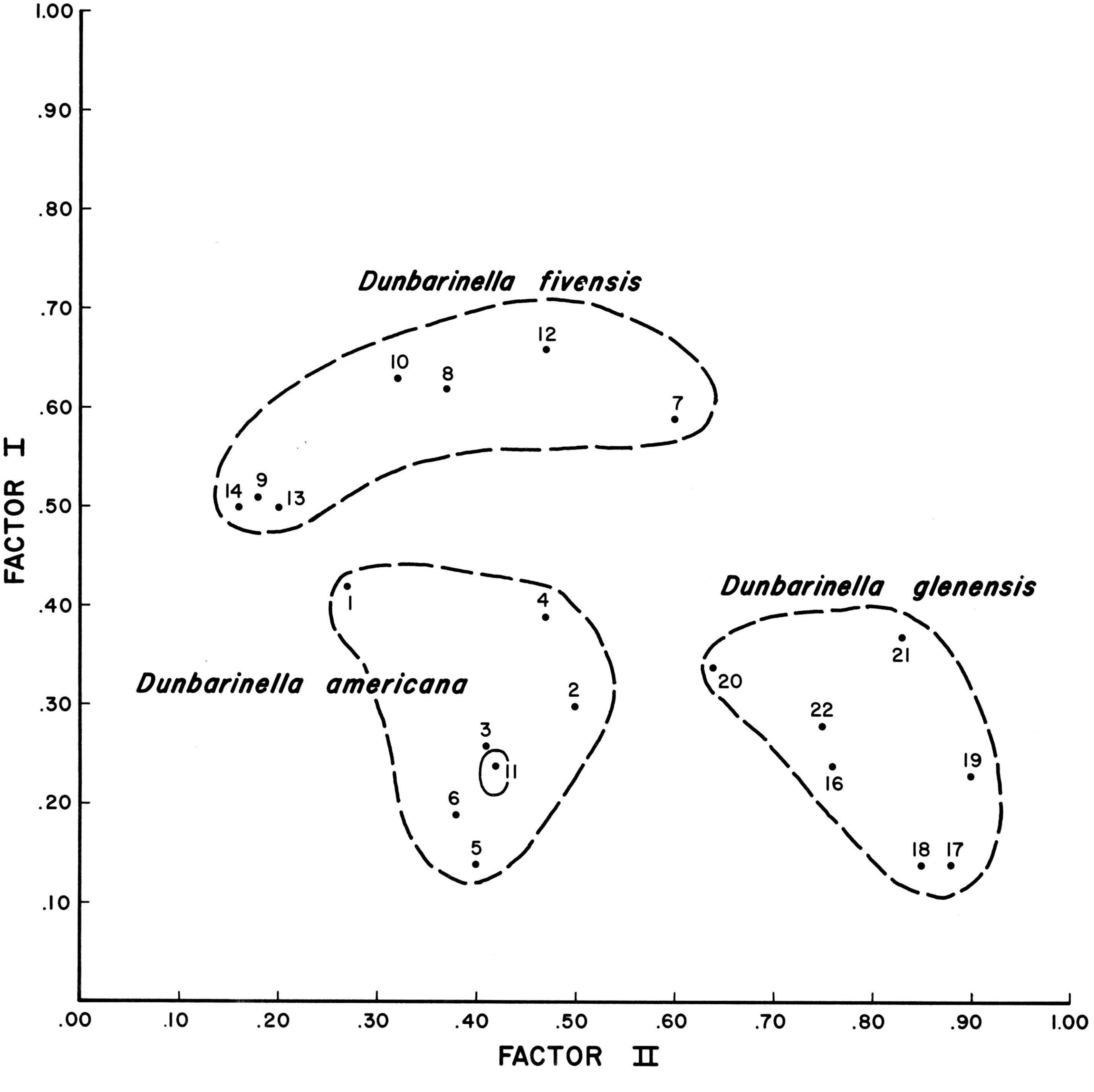


Figure 8. - Factor plot of same 3 fusulinid species as Figure 7, but in dimensions of factors I and II.

FACTOR SCORE

	Variables		Schwagerina	Triticites Pseudo-schwagerina	Dunbarinella fivensis americana	Dunbarinella glenensis			
			+	+	-	-			
1	Length		3.002	1.462	4.275	-1.514	0.639	0.264	0.165
2	Width		-4.902	8.213	5.997	-3.527	0.318	0.358	-0.306
3	Ratio		7.056	-4.173	0.374	1.183	0.465	-0.028	0.366
4	Proloculus		-2.973	8.924	5.024	-1.223	0.599	-0.173	0.279
5	V1	volution height	-2.099	9.281	5.047	0.232	0.485	-0.199	0.346
6	V2		-2.160	10.460	5.858	0.840	0.518	-0.283	0.014
7	V3		-1.681	10.164	5.363	1.266	0.550	-0.173	0.033
8	V4		-1.185	10.532	6.241	1.883	0.430	0.134	-0.117
9	V5		-0.804	8.373	5.252	1.655	0.477	0.649	-0.017
10	V6		0.141	7.739	4.843	2.536	0.360	0.864	0.072
11	V1	form ratio	-1.483	-3.628	-6.116	0.357	0.653	-0.115	-0.316
12	V2		2.508	-6.068	-6.640	2.489	0.198	-0.072	0.059
13	V3		4.599	-5.668	-3.821	2.728	-0.156	0.450	0.151
14	V4		5.413	-4.414	-2.700	3.374	0.122	0.356	0.513
15	V5		6.663	-3.787	-1.232	3.217	0.328	0.285	0.500
16	V6		1.260	-2.949	-1.447	1.569	-0.021	-0.029	0.163
17	V2	tunnel angel	5.758	1.293	3.884	3.043	-0.086	-0.369	-0.904
18	V3		6.358	1.595	2.955	3.930	0.321	-0.071	-0.321
19	V4		6.914	-0.132	1.749	3.874	0.362	0.050	-0.020
20	V5		6.091	2.104	3.070	4.082	0.469	0.432	0.079
21	V6		6.006	2.028	2.686	4.315	0.453	0.468	0.009
22	V1	spirothecal thickness	-2.618	7.813	3.640	0.492	-0.010	0.048	0.820
23	V2		-0.905	8.998	4.148	2.773	0.149	0.072	0.516
24	V3		-1.059	9.579	4.517	2.760	0.325	0.169	0.240
25	V4		-0.132	9.627	4.941	3.264	0.437	0.251	-0.039
26	V5		1.381	8.890	5.053	3.917	0.494	0.383	-0.095
27	V6		2.188	7.307	4.738	3.390	0.550	0.523	0.033

Figure 9. - Factor score matrix showing significance of variables in sample separation in Q mode analysis of 55 fusulinid specimens on the basis of 27 characters. Columns compare with factor columns in Figure 1.

CLASSIFICATION OF SUBSURFACE LOCALITIES OF THE REAGAN SANDSTONE (UPPER CAMBRIAN) OF CENTRAL AND NORTHWEST KANSAS

By

Roger L. Kaesler
University of Kansas

and

Marcus N. McElroy
Humble Oil Company

INTRODUCTION

The purpose of this paper is to present an application of the methods of numerical taxonomy (Sokal and Sneath, 1963) to clustering of subsurface geological localities. The principal advantage of the method is that it allows the geologist to consider simultaneously all quantifiable lithologic, structural, and paleontological features (characters) that he obtains from the study of cores or well cuttings. Furthermore, results of cluster analysis are presented as convenient graphs or dendrograms. Graphic presentation obviates the assimilation and evaluation by the investigator of large bodies of quantitative results.

The method used has two inherent disadvantages which necessitate that it not be used indiscriminantly. First, a dendrogram is a two-dimensional representation of a complex, multidimensional configuration expressed by the matrix of correlation coefficients or distance coefficients (hereafter called the similarity matrix). Its use results in some distortion of similarities, particularly in high-order clusters such as are used in this study. Extent of the distortion may be measured by a correlation coefficient of the similarity matrix with the cophenetic values from the dendrogram (Sokal and Rohlf, 1962; Rohlf, 1963a).

A second disadvantage of the method is that a dendrogram gives a hierarchical classification. Thus, it forces a transitional entity into the cluster with which it is most similar. Few phenomena in nature are hierarchically structured, and uncritical use of the nested clusters shown on a dendrogram can give a false impression, particularly in the very important cases where transition exists. Sokal and Sneath (1963, p. 171-174) gave a detailed discussion of the kinds of distributions that may appropriately be given nested classifications. In spite of this shortcoming of most kinds of data, many investigators have obtained meaningful results by applying these methods to the solution of geological problems (Bidwell and Hole, 1964; Bonham-Carter, 1965; Rucker, 1965; Kaesler, 1966; Maddocks, in press).

The investigator can often largely overcome the two disadvantages of the method used here by comparing the dendrogram with the similarity matrix in search of discrepancies and transition. Nevertheless, we believe that until more is known about the effects of distortion of similarities and the importance of transition, application of the methods of numerical taxonomy to clustering of subsurface geological localities should be regarded as reconnaissance geology.

Acknowledgments.-The broader research project, of which this study was a small part, was suggested by D. F. Merriam, and M. N. McElroy's work on the project was supported by the Kansas Geological Survey. We wish to acknowledge the assistance of many others who helped make the larger project possible, among them H. A. Ireland, L. F. Dellwig, P. V. Wells, W. M. Merrill, F. W. Preston, A. J. Rowell, G. F. Stewart, R. H. Lippert, R. W. Scott, V. B. Cole, M. J. Reynolds, and Rober Miller.

All computation was done at The University of Kansas Computation Center using programs developed by F. J. Rohlf, J. R. L. Kishpaugh, and R. L. Bartcher. Without their help and the use of Computation Center facilities, this study would not have been possible.

DESCRIPTION OF REAGAN SANDSTONE AND CHARACTERS USED

Lithology of the Reagan Sandstone and its stratigraphic relationships with adjacent rock units were discussed in detail by Scott and McElroy (1964) and McElroy (1965). The following brief description of the unit was taken primarily from a summary by McElroy and Kaesler (1965).

The Reagan Sandstone, which is present only in the subsurface in Kansas, has been determined to be Late Cambrian (Dresbachian) in age on the basis of meager fossil evidence. Its type section is in southern Oklahoma, and it is a lithostratigraphic equivalent of the Lamotte Formation which crops out in southeastern Missouri.

The Reagan is fairly uniform in composition throughout central Kansas. It is predominantly a quartzose sandstone, but locally it is feldspathic near the base. The Reagan becomes dolomitic near its top and grades into the overlying Arbuckle Group. Glauconite is an important accessory mineral in the northern one-third of western Kansas, but elsewhere it makes up less than 2 percent of the rock. Cementing materials include quartz, calcite, glauconite, hematite, and siderite. Unlike composition, grain size and sorting of the Reagan are highly variable over short distances.

In its type locality the Reagan overlies a very thin arkosic unit which lies directly on Precambrian igneous and metamorphic rocks. In places in the Kansas subsurface, however, the arkose or "granite wash" is as much as 100 feet thick. It is not regarded as being genetically related to the Reagan, but in places the contact between the two is transitional.

From each of eighty wells in western Kansas, seventeen characters were used, which included, for the Reagan Sandstone and adjacent units, thickness-elevation data, lithologic data, and relationships to structural geology. To facilitate handling, quantitative data were coded in equal intervals from 0 to 9 (with the exception of mean grain size, coded 0 to 6). Qualitative and semi-quantitative data were coded as appropriate. All information used in this study is on open file with the Kansas Geological Survey and may be examined upon request.

Character 1. Elevation of top of Precambrian. Range -4595 to -1110 feet (reference sea level); coded 0 to 9.

Character 2. Thickness of Arbuckle Group (above Reagan). Range 0 to 994 feet thick; coded 0 to 9.

Character 3. Thickness of Reagan Sandstone. Range 6 to 175 feet thick; coded 0 to 9.

Character 4. Thickness of sub-Reagan arkose. Range 0 to 49 feet thick; coded 0 to 9.

Character 5. Amount of quartz in Precambrian rocks. Coded 0 for schist;
1 for volcanics and granite gneiss;
2 for Precambrian sediment, except quartzite;
3 for Precambrian quartzite.

Character 6. Amount of biotite in Precambrian rocks. Coded 0 for Precambrian quartzite;
1 for Precambrian sediment, except quartzite;
2 for volcanics and granite gneiss;
3 for schist.

Character 7. Amount of feldspar in Precambrian rocks. Coded 0 for Precambrian quartzite;
1 for schist;
2 for Precambrian sediment, except quartzite;
3. for volcanics and granite gneiss.

Character 8. Relative distance from nearest known fault. Measurements were in inches from faults on the map by Cole (1962), scale 1 inch = 10 miles. Range 0.1 to 3.4 inches; coded 0 to 9.

Character 9. Percent quartz in Reagan Sandstone. Range 80 to 100 percent; coded 0 to 9.

Character 10. Percent glauconite in Reagan Sandstone. Range 0 to 9 percent; coded 0 to 9.

Character 11. Percent dolomite in Reagan Sandstone. Range 0 to 20 percent; coded 0 to 9.

Character 12. Mean grain size of Reagan Sandstone. Range 5 to -2 phi; coded 0 to 6.

Character 13. Roundness index (Dapples, Krumbein, and Sloss, 1953; Russell and Taylor, 1937; Pettijohn, 1956). Range 0.01 to 1.00; coded 0 to 9.

Character 14. Sorting measure (McElroy and Kaesler, 1965). Range 130.1 to 4; coded 0 to 9.

Character 15. Percent feldspar in Reagan Sandstone. Range 0 to 10 percent; coded 0 to 9.

Character 16. Topography of Precambrian. Coded 1 for Precambrian topographic low;
2 for intermediate;
3 for Precambrian high.

Note: This character is a measure of local topographic configuration of the Precambrian surface and is not the same as elevation of the Precambrian surface (Character1).

Character 17. Relative movement of associated fault.
Coded 0 for well on downthrown side of fault;
1 for no fault association;
2 for well on upthrown side of fault.

Ideally, the 80 wells used in this study should have been chosen at random. In practice this was not possible. Hopefully oil wells are rarely drilled at random but rather are aimed at specific targets--structural or stratigraphic traps containing petroleum. The wells used in this study are ones that penetrated Precambrian rocks and are as widely distributed areally as possible.

RESULTS

A Q-type matrix of distance coefficients was computed from the data after standardization by characters. Figure 1 is a dendrogram prepared from the distance coefficient matrix by the unweighted pair-group method using arithmetic averages (UPGMA; Rohlf, 1963b). The correlation coefficient between the original distances and the cophenetic values (distances implied by the dendrogram) is 0.730. This value, which represents the amount of distortion in the dendrogram, is only slightly smaller than results reported by Sokal and Rohlf (1962) for numerical taxonomic work.

Figure 2 shows the location of the 80 wells used in this study and some of the major structural features of Kansas. Clusters of wells at the 1.4 phenon level in Figure 1 are shown on the map by

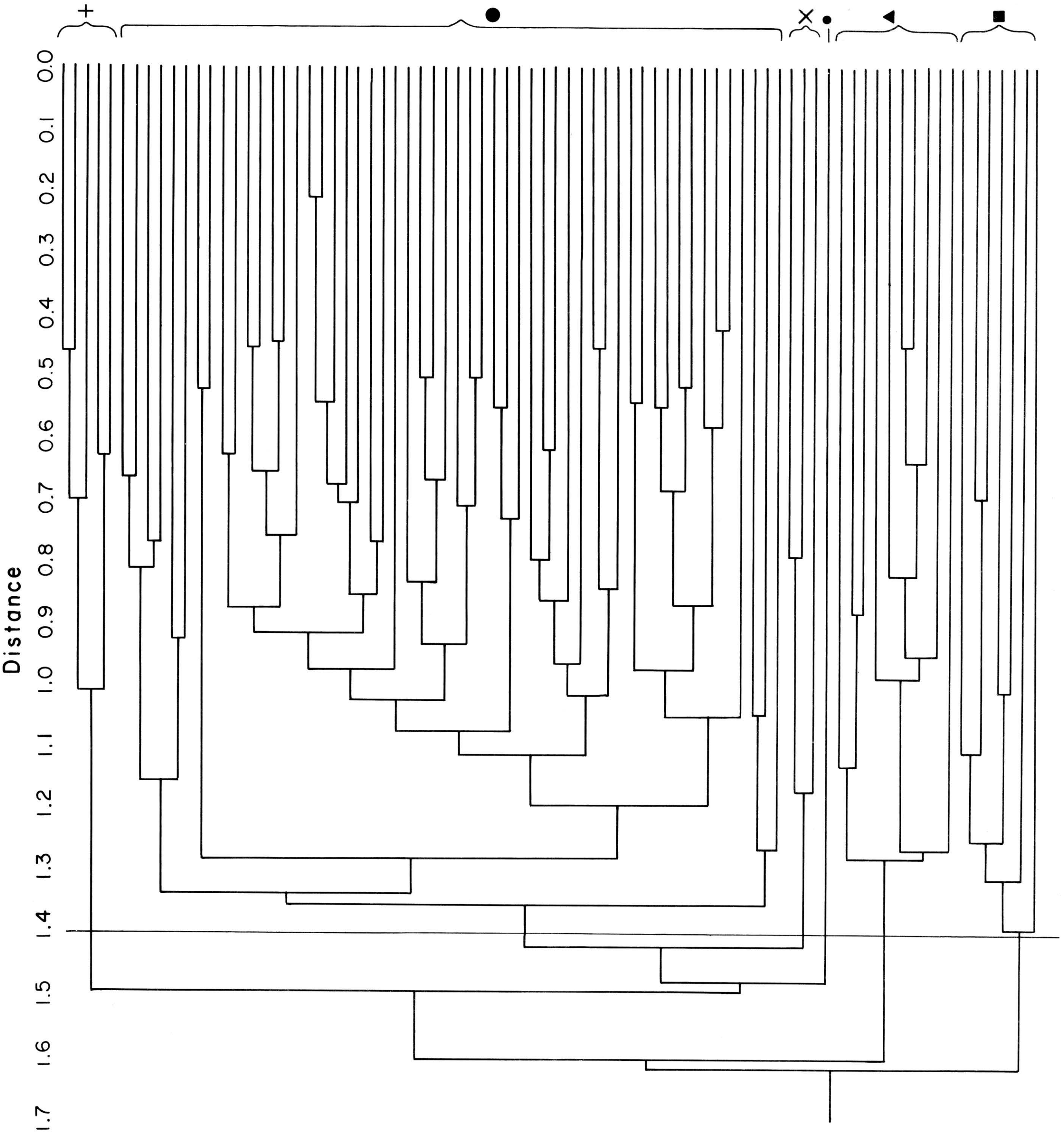


Figure 1.-Dendrogram prepared by the unweighted pair-group method using arithmetic averages (UPGMA) based on distance coefficients computed from standardized, coded characters taken from 80 wells in western Kansas. Map symbols: + indicates wells 1, 14, 12, 6, 26; ● indicates wells 2, 13, 3, 4, 22, 28, 7, 15, 9, 17, 10, 49, 52, 37, 19, 21, 61, 27, 51, 23, 24, 25, 30, 36, 62, 43, 40, 42, 58, 56, 69, 70, 73, 20, 53, 54, 38, 60, 31, 32, 33, 34, 41, 44, 55, 59, 67, 63, 64, 65, 68, 8, 16, 11; × indicates wells 47, 50, 74; • indicates well 39; ▲ indicates wells 5, 29, 45, 18, 46, 48, 72, 71, 66, 57;■ indicates wells 35, 75, 78, 79, 80, 76, 77. All wells listed in order of appearance on dendrogram.

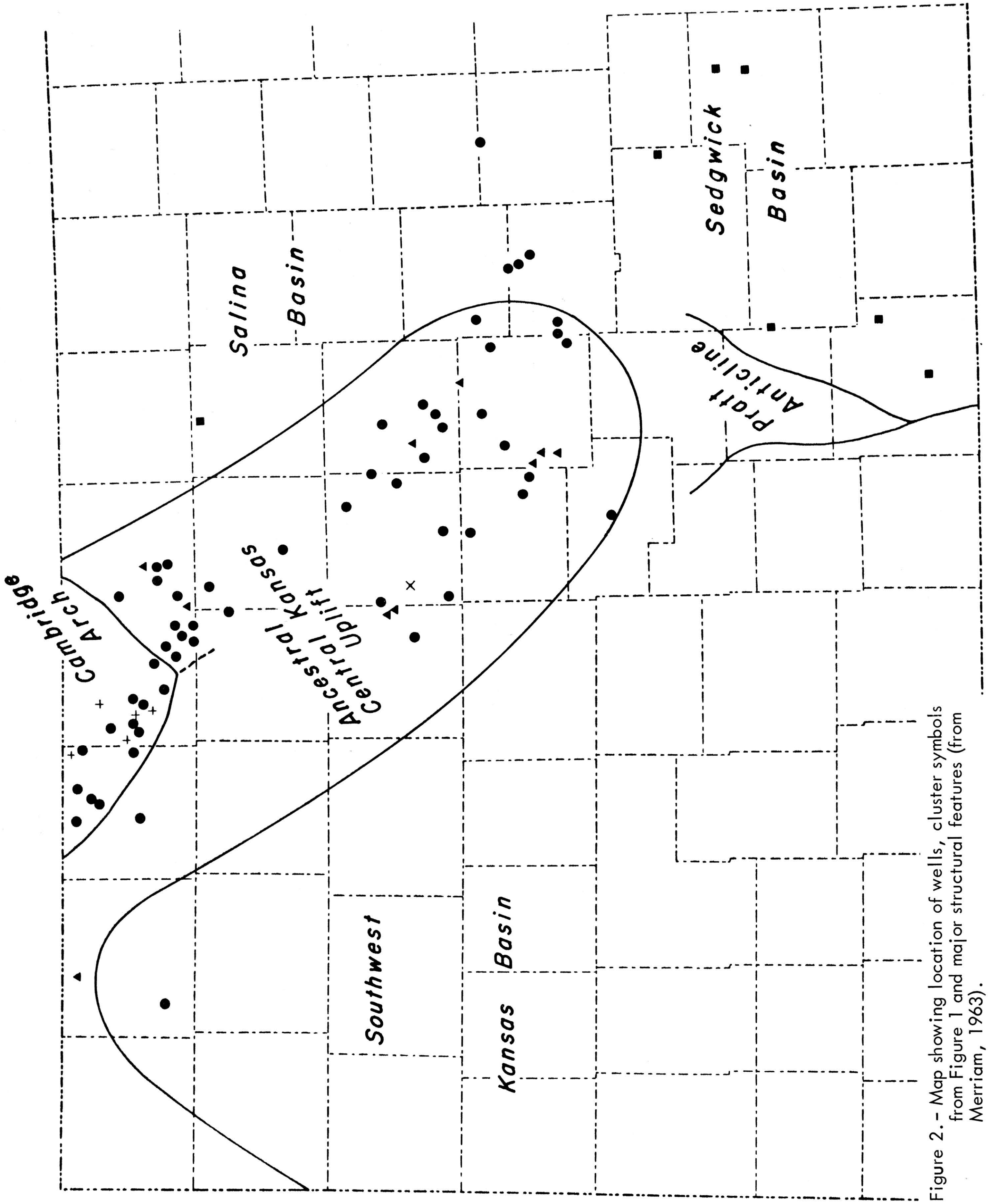


Figure 2.- Map showing location of wells, cluster symbols from Figure 1 and major structural features (from Merriam, 1963).

similar patterns. The Cambridge Arch, Pratt Anticline, Salina Basin, and Sedgwick Basin are pre-Desmoinesian post-Mississippian structures; the Ancestral Central Kansas Uplift and the Southwest Kansas Basin are pre-Mississippian post-Devonian in age (Merriam, 1963, p. 178).

Merriam (1963, p. 179) emphasized the difficulty with which pre-Mississippian structural deformation is recognized in Kansas. In spite of this difficulty, evidence exists for the presence of the Cambridge Arch during sedimentation of the Aubuckle Group; and the Ancestral Central Kansas Uplift was also "mildly active at this time" (Merriam, 1963, p. 212; see also Scott, 1966). Merriam (1963, p. 211, 212) and other workers also recorded the presence of a syncline east and southeast of the Ancestral Central Kansas Uplift, which syncline was formed before St. Peter Sandstone was deposited in the area.

This evidence of pre-Ordovician, post-Precambrian deformation is cited because we believe the method used in this study provides evidence of the same structures. Fifty-four wells used in this study form a single, large cluster. These wells are shown in Figure 2 as solid circles. All but five of these are located within the boundary of the Ancestral Central Kansas Uplift, and those five are not far outside the boundary. It would be a mistake, however, to draw far-reaching conclusions from this configuration because most of the wells used in the study were drilled on the uplift.

Geographic location of wells belonging to other clusters may be more meaningful. The wells indicated by a plus sign all lie within the boundary of the Cambridge Arch, although they are not the only wells located there. Seven of the nine wells indicated by a solid triangle are located at the southern end of the Uplift, and all nine are located toward the flanks of the structure. The small cluster with wells indicated by an X also lies to the south and on the flanks of the Uplift. Figure 1 shows this cluster as somewhat more closely similar to the major cluster than to the solid triangle cluster, and these wells may represent transition between the two larger clusters.

All wells in the remaining cluster (solid square) lie in the synclinal area east and southeast of the Uplift. Only one well (well 39, small solid circle) from the Southwest Kansas Basin was included in the study. It did not join any other clusters at the 1.4 phenon level and, thus, forms a single-well cluster.

SUMMARY AND CONCLUSIONS

Geographic distribution of clusters of subsurface localities, formed by simultaneous consideration of many characters, roughly coincides with known structural features. This congruence indicates that the methods of numerical taxonomy may provide a useful tool for subsurface geology. However, until more use is made of the method and the effects of its shortcomings are better understood, it should be regarded as a method of reconnaissance geology.

Results of application of the method indicate three areas from which more information is needed:

1. Southwest Kansas Basin, to determine if wells there form a distinct cluster.

2. Salina Basin, to see if wells there consistently cluster with those in the Sedgwick Basin.

3. Rice and McPherson Counties, to establish the reasons for similarity of wells there to the major well cluster and for differences from typical basin wells.

We plan to extend the study of these wells and of the method by using non-hierarchical clustering methods that employ principal component analysis.

REFERENCES

Bidwell, O. W., and Hole, F. D., 1964, Numerical taxonomy and soil classification: Soil Science, v. 97, p. 58-62.

Bonham-Carter, G. F., 1965, A numerical method of classification using qualitative and semi-quantitative data, as applied to the facies analysis of limestones: Canadian Petroleum Geology Bull., v. 13, p. 482-502.

Cole, V. B., 1962, Configuration map of the Precambrian basement rocks in Kansas: Kansas Geol. Survey, Oil and Gas Invest. no. 26.

Dapples, E. C., Krumbein, W. C., and Sloss, L. L., 1953, Petrographic and lithologic attributes of sandstones: Jour Geology, v. 61, p. 291-317.

Kaesler, R. L., 1966, Quantitative re-evaluation of ecology and distribution of Recent Foraminifera and Ostracoda of Todos Santos Bay, Baja California, Mexico: Univ. of Kansas Paleontological Contr., Paper 10, 50 p.

McElroy, M. N., 1965, Lithologic and stratigraphic relationships between the Reagan Sandstone (Upper Cambrian) and sub-Reagan and supra-Reagan rocks in western Kansas: Univ. of Kansas, doctoral dissertation, 173 p.

————, and Kaesler, R. L., 1965, Application of factor analysis to the Upper Cambrian Reagan Sandstone of central and northwest Kansas: The Compass, v. 42, p. 188-201.

Maddocks, R. F., (in press), Distribution patterns of living and subfossil podocopid ostracodes in the Nosy Be Area, northern Madagascar: Univ. of Kansas Paleontological Contr., Paper 12, 72 p.

Merriam, D. F., 1963, The geologic history of Kansas: Kansas Geol. Survey Bull. 162, 317 p.

Pettijohn, F. J., 1957, Sedimentary rocks: Harper and Bros., New York, 718 p.

Rohlf, F. J., 1963a, Congruence of larval and adult classifications in *Aedes* (Diptera: Culicidae): Systematic Zoology, v. 12, p. 97-117.

————, 1963b, Classification of *Aedes* by numerical taxonomic methods (Diptera: Culicidae): Annals Entomological Soc. America, v. 56, p. 798-804.

Rucker, J. B., 1965, Bryozoa distribution in Venezuela-British Guiana shelf sediments (abstr.): Geol. Soc. America, Program 1965 Annual Meeting, p. 140-141.

Russell, R. D., and Taylor, R. E., 1937, Roundness and shape of Mississippi River sands: Jour. Geology, v. 45, p. 225-267.

Scott, R. W., 1966, New Precambrian (?) formation in Kansas: Am. Assoc. Petroleum Geologists Bull., v. 50, p. 380-384.

————, and McElroy, M. N., 1964, Precambrian-Paleozoic contact in two wells in northwestern Kansas: Kansas Geol. Survey Bull. 170, pt. 2, 15 p.

Sokal, R. R., and Rohlf, F. J., 1962, The comparison of dendrograms by objective methods: Taxon, v. 11, p. 33-40.

————, and Sneath, P. H. A., Principles of numerical taxonomy: W. H. Freeman and Co., San Francisco, 359 p.

APPLICATION OF DISCRIMINANT FUNCTIONS AS A CLASSIFICATION TOOL IN THE GEOSCIENCES

By

John C. Griffiths
Pennsylvania State University

ABSTRACT

Simple linear discriminant functions have been used to classify populations of rocks and minerals into oil-bearing and barren sediments, uranium ore-bearing and barren sediments, loess and loess-like alluvium, beach and dune sands, Circumoceanic and Oceanic Island types of basalts, refractory and nonrefractory quartzites; multiple discriminants in which there are three or more classes have also been used to differentiate glacial till, delta and beach sands, and to subdivide potentially promising mining areas into six dollar-value classes on the basis of their geological environments.

These discriminants may be used either as empirical tools for classification or as one step in an attempt to understand what geological factors lead to differentiation of the classes.

Selection of the variables to achieve optimal discrimination is best performed by means of components analysis in the Q mode.

INTRODUCTION

Classification is one of the earliest and simplest steps in ordering the complex of events in the "real world"; it is always superimposed on the world of events whether it is considered a "natural" or "artificial" classification. Defining, classifying and naming objects are early steps in the procedure known as the scientific method and they form the basis of language and communication. Classification requires a set of criteria for assigning objects to classes; the criteria are usually properties of the individual objects and, if properly constructed and used, the classes which result are mutually exclusive and exhaustive. If the classes so constructed are characterized by names the nomenclature of the classes forms a nomial scale; that is, there is no necessary relationship between the classes other than those dependent on the exhaustive and mutually exclusive requirements. Quite frequently the classes are ordered so that a granite-gabbro sequence of classes is characterized by change from light to dark color and, similarly, a granite-rhyolite sequence of classes is based on a change in grain size. In this case the arrangement of the classes leads to an ordinal scale and the ordered ranking may be of interpretive value (Barnard, 1935).

More refined classification leads to class relationships which may be related to interval or ratio scales as in the classification of frequency of grain sizes where grain size is a continuous variable of interval or ratio scale in measurement level.

Natural, as contrasted with artificial, classifications usually contain some systematic content, i.e. the natural classification exhibits relations which are not evident until after the classification has been accomplished (Hempel, 1952), a typical example being the periodic classification of the chemical elements.

One of the more common misuses of classifications arises when a series of objects are arranged in classes and the classification so constructed is presumed to be a fact rather than an artifact; in other words, the search for order represented by classification is a function of the human instinct and whether it is representative of "natural order" among the items classified is debatable. The rectangular classification of igneous rocks after Rosenbusch (see Johannsen, 1939, p. 119ff.) or that of the sedimentary rocks after Krynine (1948) are useful for identification and naming rocks and also for putting rocks in rectangular draws but their originators believed that the arrangements and relations among the classes have some interpretive or predictive content. Many bitter controversies have arisen over problems of nomenclature and classification because of confusion in the degree of achievement of these two independent objectives.

The simplest form of classification consists of a subdivision of the population of individual elements into two classes; the classes should be mutually exclusive and exhaustive and the criteria forming the basis of assignment of individuals to classes should be chosen so that these requirements are fulfilled. Fisher (1936) proposed the discriminant function as a statistical tool for subdividing a set of objects or individuals into two such classes on the basis of their properties. The discriminant function is selected so that it is the most effective linear compound composed of the weighted properties of the objects to enable the two classes to be distinguished.

To set up a discriminant it is necessary, then, to define two classes which are mutually exclusive and exhaustive and to select appropriate criteria (properties of the objects to be classified) which will permit the assignment of each object to one or other of the two classes. In practice, the classes chosen are usually exhaustive but possess various degrees of overlap so that they are not mutually exclusive; the discriminant function is first established by using objects which are clearly assignable to each of the two classes and, after a suitable function is found, objects with unknown class affinities are assigned by means of the discriminant function. If the two classes are not mutually exclusive the assignment of an unknown to a class, resulting from the application of a discriminant, has an associated probability of misclassification (Rao, 1952, p. 296).

Various refinements and extensions are possible in establishing a discriminant function; for example, it is possible to test whether a discriminant based on say five properties is as effective as one based on a lesser number of properties (Rao, 1952, p. 252ff.). Again it is possible to extend the discriminant to encompass more than two classes using a multiple discriminant function (Rao, 1952, p. 307ff.). In fact the discriminant index, conventionally symbolized as Z, may be two valued, leading to a simple discriminant function, many valued, as in multiple discrimination or, if Z is a continous variable, the discriminant index is equivalent to the dependent variable in a multiple regression equation (see for example, Griffiths, 1961).

An alternative way of looking at a discriminant function is to consider the simple linear discriminant as the multivariate analogue of the univariate Student's test and a multivariate equivalent is Hotelling's T^2 statistic (Miller and Kahn, 1962, p. 248). There are also examples of the use of nonlinear discriminants both simple and multiple (Hodges, 1950, McIntyre, D. D., 1962, Harries, 1965). Discriminant analysis is, then, one tool among the varied multivariate procedures which may be used in the classification of objects (Anderson, 1958). Application of a discriminant function as a classification procedure may be based on the simple requirement of finding some objective means of subdividing a population into two or more sub-populations without expecting more than the subdivision as an outcome; or, as is more usual, it may be one step in attempting to elucidate the relationship among groups of objects with the objective of attempting to interpret the relationships both among the discriminated classes and among the properties which permit the discrimination.

DISCUSSION OF SELECTED EXAMPLES

Use of a discriminant function as a classification tool in the geosciences is quite common (see list of references) and ranges over a wide variety of subject matter problems, some of which are discussed below because they illustrate different uses of discriminants and the different roles they may play in different situations.

Simple Discriminants

Perhaps the first example in the geosciences was an attempt to distinguish oil-bearing from barren sediments (Emery, 1954; Emery and Griffiths, 1954). In this case the two sub-populations are exhaustive and appear to be mutually exclusive but in fact they are not; the discriminant function was constructed on the basis of six petrographic properties but only three were effective in discriminating the two classes (see Griffiths, 1963, p. 643). In this example the discriminant actually succeeded in separating the two classes but a more detailed examination showed that the basis for separation depended on the fact that the oil-bearing samples, taken from one Berea oil-sand core, were much more homogeneous than those of the barren sandstones representing several horizons in the stratigraphically equivalent Pocono sandstones (Griffiths, 1966, in press). Here the discriminant, while successful in achieving its immediate objective of separating the classes, also led to an understanding of the reasons for the separation.

Subsequently this investigation was extended by mapping changes in value of the discriminant index over the barren Pocono sandstones of Pennsylvania and it was demonstrated that the index could be used in this case to indicate the change in favorability from barren sandstones in the east to potentially more favorable sandstones in the west (Shadle, 1957, Griffiths, 1963, p. 645).

A second example in which a discriminant was used to separate two classes, loess and loess-like alluvium, is described by Millette (1955); here the two sub-populations could again be considered exhaustive in terms of the area sampled but they were hardly mutually exclusive. Nevertheless the discriminant could be used to successfully separate the two classes.

Hulbe (1957) attempted to differentiate some beach from dune sands by measuring the size and shape of their contained quartz grains; from a geological point of view the conclusions are interesting because a univariate comparison of the size of the grains was inconsistent; the long "a" axis of the quartz grains in the beach sand were smaller than those in the dune whereas the intermediate "b" and short "c" axes were larger. This suggests that attempts to use size are likely to yield unstable comparisons. On the other hand, when the three axes were combined into a discriminant function the two classes were clearly separated and the relationship between the short "c" and long "a" axes were the controlling factors. In other words, the shape of the quartz grains may differentiate beach from dune sand independently of their sizes; if this conclusion is affirmed by more extensive analysis then the relationship among long and short axes may be shown to be invariant across absolute changes in grain size. The process leading to the development of a dune sand from a beach sand is one of selective sorting, and on the basis of these findings it is apparently a process which selects on the basis of shape rather than size. From the statistical point of view this illustrates an example in which, while the means are not consistently different, the discriminant succeeds because of a suitable interaction among the variables.

An example of the application of a discriminant to the differentiation of uranium-bearing from barren sediments indicates some of the difficulties which may arise in the use of this tool (Griffiths, 1957); here the two sub-populations were assigned on the basis of the presence or absence of ore but the real comparison required was between rock which was minable and that which was not. The petrographic properties sufficed to discriminate but the exact implications were not clear. Two difficulties arise; first the difference between ore and barren rock is not mutually exclusive unless the definition of the classes is refined. Secondly the two sub-populations were apparently not homogeneous; for example an ore-bearing sediment in the Shinarump formation is a gravelly sandstone and the barren rock is coarser in size. Similarly the ore-bearing rock in the Salt Wash member is a very fine grained sandstone whereas the barren sandstones were somewhat coarser. In such cases local differences may suffice to differentiate the two classes and in a regional sense the two sub-populations exhibit a wide range in variation of the values of the properties and a very considerable overlap. If the sampling had been paired on a local basis then the differences would probably have sufficed to yield a meaningful discriminant.

Chayes (1965) proposes to use a discriminant function as an objective means of classifying igneous rocks plotted in a ternary diagram in terms of three normative end-members; since the three bases of classification sum to a constant a two term discriminant is adequate and optimal for the separation of the phases. He has also applied discriminant analysis to distinguish between Circumoceanic and Oceanic-Island types of basalt in terms of their differing chemical composition (Chayes and Velde, 1965).

Multiple Discrimination

Multiple discrimination was first applied by Barnard (1935) and an extensive investigation of the cranial characters of five different tribes in India is described by Mahalanobis et al (1949). McIntyre (1962) used this procedure in an attempt to differentiate three associated environments, a glacial till, fluvioglacial delta and a beach on the bais of the shapes of some of their contained minerals (quartz, garnet, and hornblende); this investigation is of especial interest because the covariance matrices of the three classes appeared to be heterogeneous and so McIntyre used quadratic discriminant functions as recommended by Fisher (1936) to separate the three classes. He also attempted to apply sequential analysis by discriminants to obtain a cleaner separation.

Harries (1965) also applied a multiple discriminant function procedure coupled with Bayesian decision functions to assign areas (called cells) on the basis of their geological properties to different dollar value classes in an attempt to subdivide a region into potentially promising targets for mining exploration. In this case the discriminant scale is ordered in a similar manner to that used by Barnard (1935).

A number of other examples of the application of discriminant functions to geoscience problems are listed in Miller and Kahn (1962), Krumbein and Graybill (1965) and in the list of references in the present article.

SELECTION OF VARIATES FOR DISCRIMINATION

Selection of the variates to be included in the discriminant function has usually been based on empirical testing of the contributions of each successive variate or sets of variates to the power of the discriminant; two approaches may be used in analogous fashion to the construction of a multiple regression equation. First it is possible to commence with as many independent (X) variates as possible and then to test successively the contribution of the last variate or last set of variates to the power of the discriminant. Alternately it is possible to build up the discriminant by commencing with one variate and upon adding a second test the additional contribution of each succeeding variate. Unfortunately these procedures do not lead to unique results since the order in which the variates are added plays a role in the results. In extreme cases a variate may be rejected when first tested and subsequently, if introduced at a later stage, may contribute significantly to the regression or discriminant function. This arises because the X variates interact and then are not strictly independent.

In order to obtain a more objective and consistent result it is possible to subject all variates, dependent and independent simultaneously, to a components analysis and then use the results of this analysis to build the regression equation or discriminant function. In general if the objective is to explain one variate, the dependent or Y variate, by the X or independent variates, the R mode is used for the component analysis (Griffiths, 1963, 1966); the component with the largest loading in the Y variate is selected and the first X is the independent variate loaded most heavily on this component. No other X will explain more so this is the optimal choice; similarly the component with the next highest loading in the Y variate is examined for its X with highest loading and this is the second independent variate to be included in the regression.equation. This second X variate is, on the basis of the component analysis, independent of the first and no other variate will contribute more to explaining the variation in the dependent variate. This escalation upwards may be continued building a stepwise regression equation which is the best that may be designed from the available matrix of observations.

A similar procedure may be used to build the "best" discriminant function by using a Q mode components analysis (Imbrie, 1963) as the basis for selection of components and variates (see for example, Griffiths, 1966).

These procedures will generally suffice to evade the equivocal character of applying multiple regression and discriminant analysis to a matrix of observations. It also permits the investigator to decide whether a regression equation or discriminant will be effective at all on the basis of the information in the observations matrix.

CONCLUSIONS

From the brief discussion of principles and the illustrative examples in the preceeding account it may be deduced that the discriminant function is a useful tool in classifying a population into two or more sub-populations. It is necessary, however, to be specific about the objective and not to confuse the multiple potential uses of this powerful tool. In general it is best to consider the discriminant function either solely as an empirical means of achieving a subdivision of a population into sub-populations, or as a single step in a number of analytical procedures aimed at understanding the relationships among the subdivisions established by the discriminant.

In both cases the discriminant is best devised by using a components analysis, in the Q mode, applied to the original data matix, or its transformed equivalent, to select the minimum number of variates or properties which may best achieve discrimination.

REFERENCES

Anderson, T. W., 1958, An introduction to multivariate statistical analysis: Wiley and Sons, Inc., N.Y., 374 p.

Barnard, M. M., 1935, The secular variations of skull characters in four series of Egyptian skulls: Ann. Eugenics, v. 6, p. 352-371.

Bateu, W. D., and DeWitt, C. C., 1944, Use of discriminant functions in the comparisons of proximate coal analyses: Ind. Engr. Chem., Anal. Ed., 16, p. 32-34.

Chayes, F., 1965, Classification in a ternary diagram by means of discriminant functions: Am. Mineralogist, v. 50, p. 1618-1633.

―――――, and Velde, D., 1965, On distinguishing basaltic lavas of Circumoceanic and Oceanic-Island type by means of discriminant functions: Am. Jour. Science, v. 263, p. 206-222.

Day, B. B., and Sandomire, M. M., 1942, Use of the discriminant function for more than two groups: Jour. Am. Stat. Assoc., 37, p. 461-472.

Emery, J. R., 1954, The application of a discriminant function to a problem in petroleum petrology: Unpub. master's thesis, Dept. of Mineralogy, The Pennsylvania State Univ., 120 p.

―――――, and Griffiths, J. C., 1954, Differentiation of oil-bearing from barren sediments by quantitative petrographic analysis: The Pennsylvania State Univ., Min. Ind. Expt. Sta. Bull., 64, p. 63-68.

Fisher, R. A., 1936, The use of multiple measurements in taxonomic problems: Ann. Eugenics, v. 7, 179 p.

Griffiths, J. C., 1957, Petrographical investigations of the Salt Wash sediments: Final Rept. RME 3151, Office Technical Services, Dept. Commerce, Washington, D. C., 38 p.

―――――, 1961, Measurement of the properties of sediments: Jour. Geology, v. 69, p. 487-498.

―――――, 1963, Statistical approach to the study of potential oil reservoir sandstones: Computersin the Mineral Industries, Stanford Univ. Pub. Proc. 3rd Ann. Conf., v. 9, no. 2, pt. 2, p. 637-668.

――――――, 1966, A genetic model for the interpretive petrology of detrital sediments: Jour. Geology, in press.

Harries, P. D., 1965, Multivariate statistical analysis - A decision tool for mineral exploration: Short Course and Symposium on Computers and Computer Applications in Mining and Exploration, College of Mines, Univ. of Arizona, v. 1, p. C1-C35.

Hempel, C. G., 1952, Fundamentals of concept formation in empirical science: Intl. Encyclop. Unified Science, v. 11, no. 7, 93 p.

Hodges, J. L., Jr., 1955, Discriminatory analysis, 1. Survey of discriminatory analysis: School of Aviation Medicine, U.S.A.F., Proj. no. 21-49-004, Rept. no. 1, 115 p.

Hulbe, C. W. H., 1957, An investigation of the size and scope of quartz grains, Pedro Beach, California: Unpub. master's thesis, Dept. of Mineralogy, The Pennsylvania State Univ., 69 p.

Imbrie, J., 1963, Factor and vector analysis programs for analyzing geologic data: Office of Naval Research, Geography Branch, Tech. Rept. 6, 83 p.

Johannsen, A., 1939, A descriptive petrography of the igneous rocks: Univ. Chicago Press., v. 1, 318 p.

Krumbein, W. C., and Graybill, F. A., 1965, An introduction to statistical models in geology: McGraw-Hill Book Co., Inc., 475 p.

Krynine, P. D., 1948, The megascopic study and field classification of sedimentary rocks: Jour. Geology, v. 56, p. 130-165.

Mahalonobis, P. C., Majumdar, D. N., and Rao, C. R., 1949, Anthropometric survey of the United Provinces 1941. A statistical study: Sankhya, v. 9, 90 p.

McIntyre, D. D., 1962, A comparison of three associated environments, glacial till, fluvioglacial delta and beach sand, in terms of the shapes of their quartz, garnet and hornblende grains: Min. Ind. Expt. Sta. Pub. , the Pennsylvania State Univ., 78 p.

Mellon, G. B., 1964, Discriminatory analysis of calcite- and silicate-cemented phases of the Mountain Park sandstone: Jour. Geology, v. 72, p. 786-809.

Miller, R. L., and Kahn, J. S., 1962, Statistical analysis in the geological sciences: Wiley and Sons, Inc., N. Y., 483 p.

Millette, J. F. G., 1955, Loess and loess-like deposits of the Susquehanna River Valley of Pennsylvania and a section of the Laurentians in Canada: Unpup. doctoral dissertation, Dept. of Agronomy, The Pennsylvania State Univ.

Rao, C. R., 1952, Advanced statistical methods in biometric research: Wiley and Sons, Inc., N. Y., 390 p.

Shadle, H. W., 1957, A petrographic study of the Pocono formation of western Pennsylvania and an areal test of discriminant function technique: Unpub. master's thesis, Dept. Mineralogy, The Pennsylvania State Univ.

――――――, and Griffiths, J. C., 1955, An attempt to establish oil-reservoir favorability criteria based on quantitative petrographic analysis: The Pennsylvania State Univ., Min. Ind. Expt. Sta. Bull no. 68, p. 61-66.

Stanonis, F. L., 1956, The petrology of the Chipmunk sandstone and its relationship to reservoir properties: Unpub. doctoral dissertation, Dept. Mineralogy, The Pennsylvania State Univ., 151 p.

Williams, E. G., and Griffiths, J. C., 1961, Applications of statistical methods in prospecting for high-alumina clay: Min. Ind. Expt. Sta. Bull. No. 77, p. 29-34.

Wood, G. V., 1961, Discriminating between refractory and nonrefractory quartzite by quantitative petrography: Jour. Sed. Pet., v. 31, p. 530-533.

APPLICATION OF PATTERN ANALYSIS TO THE CLASSIFICATION OF OIL-FIELD BRINES

By

Ernest E. Angino
Kansas Geological Survey

and

Charles O. Morgan
U.S. Geological Survey

ABSTRACT

The comparison and correlation of oil-field brines by means of pattern analysis has proven to be feasible. The use of Stiff and Piper diagrams to represent graphically the chemical characteristics of a particular brine provides a rapid, simple, and convenient method of simultaneously comparing and characterizing the chemical composition of several brines. Pattern analysis emphasizes clearly the chemical differences between brines from different formations. The techniques of pattern analysis should prove extremely valuable in water-quality studies for tracing sources of brine pollution.

INTRODUCTION

An understanding of the chemical composition of oil-field brines is vital to research on (1) formation of oil-field brines (2) stream pollution by brines, (3) treatment of water used for water-flooding projects, and (4) classification of brines as to origin. For example, one of the more obvious uses of water analysis is in the determination of the source of water infiltrating petroleum and natural gas wells. For this purpose it is necessary that both the water of the oil-producing formation and contaminating water be analyzed. By comparison of water analyses from a given formation with that of water from other geologic formations in the area, the source of the infiltrating or contaminating water sometimes can be identified.

One of the difficulties associated with research of oil-field brines is that of selecting a practical classification for the identification of brine types. Obviously, all the brines are high in sodium and chloride; hence an attempt to identify and catogorize several different brines by merely "looking" at the gross chemical analysis is often a frustrating and time-consuming job.

Several systems have been proposed or are in use in studies of brine chemistry. Three of these are: (1) the hypothetical-combination system, (2) the Palmer system, and (3) the ionic system (Reistle, 1927). The ionic system, which reports the concentration of each component in parts per million (ppm) has been satisfactory in studies of oil-field brines. For describing, identifying, or differentiating any two waters, the system cannot be misleading; however, as noted above, it is not practical to differentiate several brines using this technique. Utilization of this system has other drawbacks, but they will not be discussed here.

The main objective of this study was to examine the feasibility of identifying brines by means of the patterns (pattern analysis) generated by utilizing different graphical techniques to present the chemical data for each brine. Chemical data for brines from the Kansas City Group of Pennsylvanian age and the Arbuckle Group of Ordovician age were examined by several graphical methods to test the "uniqueness" of the pattern developed for each brine.

PROCEDURE

The graphic patterns used were the Stiff (1951) diagram, and the Piper (1953) quaternary cation-anion, trilinear-cation, and trilinear-anion-percentage diagrams. Digital computer programs for handling these diagrams (Morgan et al., 1966) were developed by the University of Kansas Computation Center and the U.S. Geological Survey office in Lawrence, Kansas. These programs, written in FORTRAN IV, manipulate the brine data, make the necessary calculations, and produce the Stiff and Piper diagrams.

The Stiff diagram is a means of graphically representing a single chemical analysis of a water sample. It allows similarities and differences in the concentration of constituents to be seen at a glance and can be used for visual comparison of analyses.

Chemical data are read into the computer in parts per million and instructions in the program convert the constituent values in parts per million to equivalents per million (epm). The constituent values used on the diagram are then compared by the computer with a set of scale values that have been previously read in, and a scale is selected by the program so that all concentrations fall within the range of the diagram. Identification and scale information are printed, and the Stiff diagram is prepared in the form shown in Figure 1. Horizontal dashed bars represent the different components, with cation values projecting to the left and anion values to the right of a central, vertical zero line. The program continues from analysis to analysis until all data are graphed.

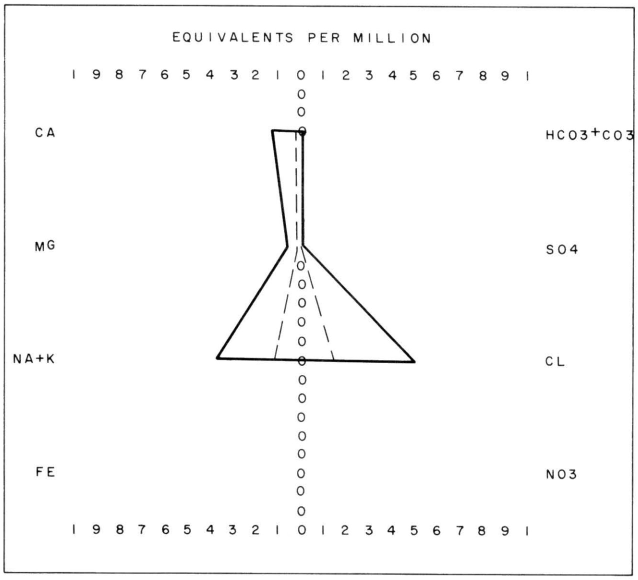


Figure 1.-Stiff diagram showing pattern developed by the average concentrations for each chemical component of brines from the Arbuckle Group (light dashed line--70 samples) and Kansas City Group (heavy solid line--78 samples). Each horizontal unit in equivalents per million equals 25. Total scale equals 500 epm.

The quaternary or Piper diagram shows the chemical composition of a water in terms of cations and anions as a percentage pattern with all analyses from a single souce or formation on a single graph. The computer output is an adaptation of the normally diamond-shaped diagram. Concentration values are converted from parts per million to equivalents per million and the percentages of the concentrations are calculated. The values in equivalents per million are plotted with the percentages of each component arranged as shown in Figure 2. Trilinear plots of cation and anion percentage concentrations are shown in Figure 3.

RESULTS AND DISCUSSION

Of the graphical schemes attempted, the Stiff and Piper plots gave the best separations of the brine data studied. The advantage of these methods is that they present clearly in one pattern an idea of the major element composition of a brine. One of the distinctive features of a Stiff diagram is the tendency for a pattern of a particular brine to maintain its characteristic shape with either dilution or concentration (Stiff 1951). As a consequence, the "type" pattern of a given brine indicates the total salt concentration as well as the chemical composition.

The Stiff plot drawn from the average analysis of 78 Kansas City and 70 Arbuckle brine samples (Fig. 1) permits the immediate comparison of the relative concentration of 10 chemical components for the brines, although only eight constituents were used in this study (Fig. 1). Note that components of negative and positive valence are grouped separately.

We have followed Stiff (1951) and developed the Stiff pattern by connecting the end points of each horizontal bar representing the appropriate concentration in equivalents per million for each component. The pattern thus developed clearly shows that the brine from the Kansas City Group has a considerably higher concentration of calcium (Ca) and magnesium (Mg) than that from the Arbuckle Group.

Both, of course, have high sodium (Na) and chloride (Cl) concentration values but the greater salinity of the Kansas City brine is emphasized. Both brines are low in sulfate (SO_4--) and bicarbonate (HCO_3-.) Similar plots for individual samples show the same pattern regardless of where the samples were collected (i.e. on a comparable scale, each brine can be differentiated and identified).

When the data for the individual brine samples are plotted on a Piper diagram, the pattern shown in Figure 2 is produced. In this pattern the rather high chloride concentrations common to brine from the Kansas City Group, cause a very tight clustering near the top of the graph. No overlap of data from the Arbuckle Group is present. Again in terms of the chemical parameters plotted a clean separation of the two brine types is noted. The Piper diagram also indicates the higher calcium and magnesium content of brines from the Kansas City Group.

For some studies, it is of interest to measure the difference in anionic and cationic character of a particular brine. A means of obtaining this information is to have the anion and cation components plotted separately on a trilinear diagram. Such a plot is shown for the anion content of brines from

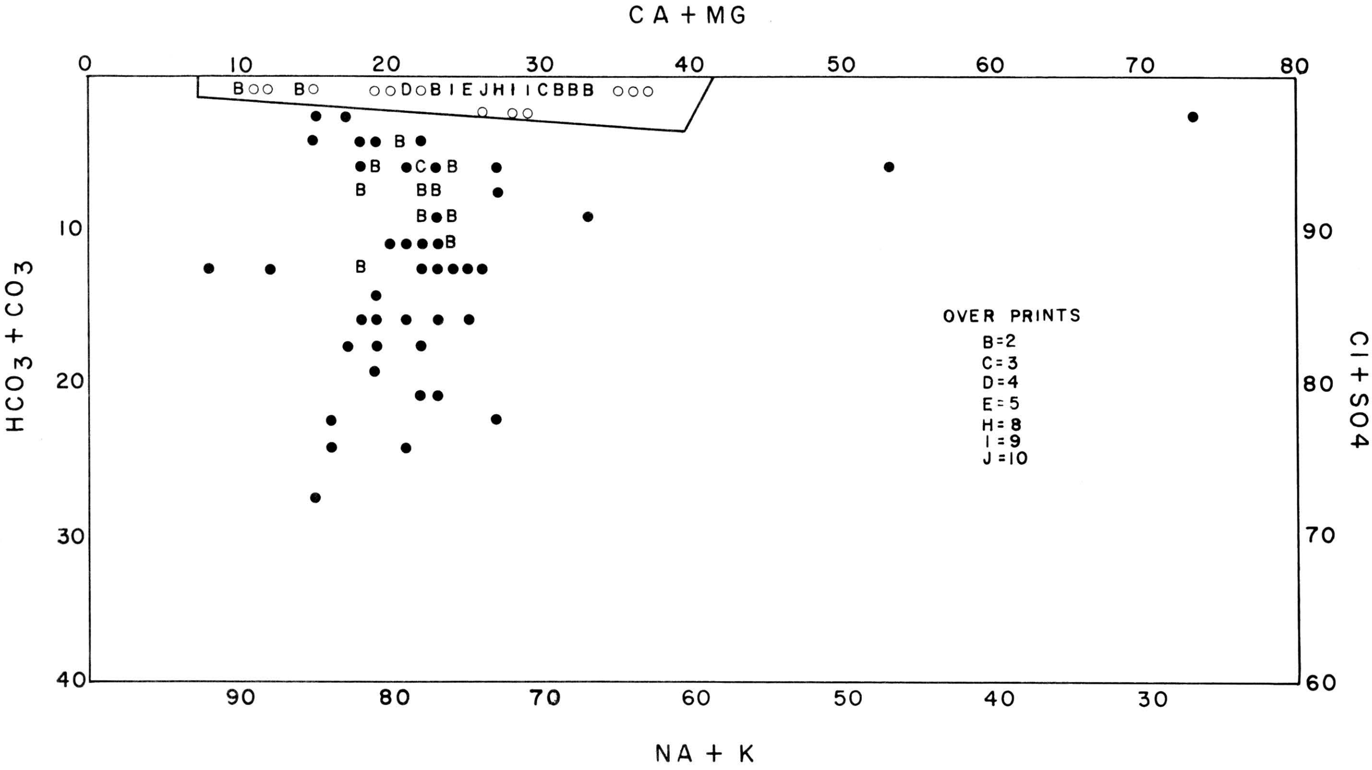


Figure 2.-Piper quaternary cation-anion pattern illustrating sharp separation of the two brine types. Open and closed dots represent samples from Kansas City and Arbuckle Groups, respectively. Note that only a portion of the diagram is illustrated.

the Kansas City and Arbuckle Groups (Fig. 3). In the pattern developed on this diagram, the high chloride content of the Kansas City brine is again emphasized, but, in addition, the higher bicarbonate content of the Arbuckle brine is now clearly shown. Concomitantly, the low concentration of sulfate in both brines is also shown. The separation pattern developed on the trilinear diagram is exceptionally clear. Similar separations have been observed on patterns developed on a trilinear plot of the cationic components.

The separation of the two brine types by the means of the graphical procedures used is extremely good. Brine data from more than two formations can be handled in the same manner, and the patterns developed on the Stiff and Piper graphs can be used to illustrate chemical differences between brines from different formations. If the amount of brine data is large, then an arithmetic mean of the chemical parameters for each formation can be calculated and the mean figures treated in the manner described in this paper.

In this way the chemical characteristics of many formations can be compared simultaneously. Brines can be easily classified as calcium-magnesium-rich, sulfate-rich, or bicarbonate-rich brines. Examination of the patterns presented on the diagrams used herein, clearly shows that brines from the Kansas City Group have a much higher chloride, magnesium, and calcium content, but a lower bicarbonate and sulfate content than those from the Arbuckle Group (Figs. 1-3).

Costs for preparing Stiff and Piper diagrams by computer for the 148 analyses (assuming that the data cards were previously prepared) are less than $60. Estimated costs for equivalent manual work are between $500 and $750. Data storage is an added consideration and advantage of these analyses. Once the information is on tape, it is readily available for future uses and need not be recompiled.

CONCLUSIONS

The comparison of oil-field brines by means of pattern analysis is feasible. The use of Stiff and Piper diagrams to represent graphically the chemical characteristics of a brine provides a rapid, simple, and convenient method of comparing and characterizing the chemical composition of several brines simultaneously. As Stiff (1951) points out, pattern analysis also allows the correlation of brines from one area to another, as the pattern would tend to maintain its basic shape upon either concentration or dilution of a given brine. It should be possible by analyzing patterns developed on the diagrams described to detect the presence of "foreign" water and to determine its source. Such a technique would be extremely valuable in water-quality studies for tracing sources of brine pollution. Although the brines described could have been differentiated by a simple comparison of the chemical composition, the pattern analysis emphasizes more clearly the basic chemical differences between the two brines. If one were to compare additional brine data from other formations, the superiority of the use of pattern analysis to differentiate the brines would be obvious.

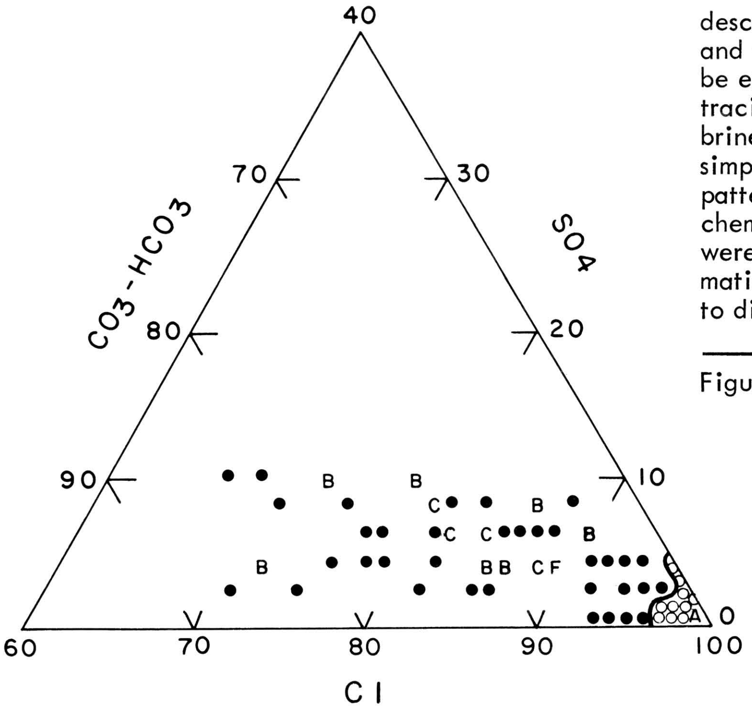


Figure 3.-Trilinear plot showing separation pattern of the two brines using the anionic content. Open circles represent samples from the Kansas City Group, solid dots from the Arbuckle Group. The A in the Kansas City pattern represents 40 overprints. The letters B, C, and F represent 2, 3, and 6 overprints, respectively. Only the one corner of the trilinear plot was needed to compare these brines. Normally, the whole field is printed out.

REFERENCES

Morgan, C. O., Dingman, R. J., and McNellis, J. M., 1966, Digital computer methods for water-quality data: Ground Water,4, p. 35-42.

Piper, A. M., 1953, A graphic procedure in the geochemical interpretation of water analyses: U.S. Geol. Survey Ground-Water Note 12, 14 p.

Reistle, C. E., Jr., 1927, Identification of oil-field waters by chemical analysis: U.S. Bureau Mines Tech. Paper 404, 25 p.

Stiff, H. A., 1951, The interpretation of chemical water analysis by means of patterns: Jour. Petroleum Technology Tech. Note 84, 3, 15 p.

CLASSIFICATION IN QUANTITATIVE OIL-EXPLORATION DECISION MAKING

by

John W. Harbaugh
Stanford University

INTRODUCTION

Decision making in oil exploration, when viewed simply, can be separated into two principal types of operations: (1) classification of prospects, and (2) assessment of prospects according to a probabilistic decision-making system. Most oil-exploration decisions are made subjectively and qualitatively with intuition and experience greatly influencing decisions. Human intuition and experience are valuable, of course, but decision makers tend to be inconsistent in moving toward a goal unless their decision-making methods are disciplined and rigorous. In oil exploration, most decision makers have not used rigorous methods because of the large degree of uncertainty in oil exploration. This is unfortunate, because quantitative probabilistic decision-making systems, with their ability to deal rigorously with uncertainty, appear potentially capable of improving efficiency when the objective is to maximize profits. Even a very small improvement in efficiency (one percent or less) would have substantial effect on the industry.

The concept of a quantitative scale of probability is inherent in most rigorous methods in dealing with uncertainty. Probability can be expressed on a scale range from zero to one, in which zero represents impossibility and one represents complete certainty, with all gradations between. Of course, probability can also be expressed in percent, 100 percent representing complete certainty and so on.

PROBABILISTIC CORE OF AN OIL EXPLORATION SYSTEM

C. Jackson Grayson (1960) has provided a clearly written introduction to the art of quantitative decision making. Through the use of quantitative probability estimates, it is possible to make exploration decisions that are objective and that are consistent with one's operating policies. I have incorporated many of Grayson's ideas in a program for IBM 7090/7094 computers, which has been used to calculate data shown in subsequent tables.

In the forthcoming examples, the general objective has been to maximize profits, consistent with risk-taking ability. Given this objective, the first step is to prepare a "payoff table" which lists possible events (such as dry hole, million-barrel discovery, etc.), versus possible acts (such as don't drill, drill with 100 percent working interest, etc.). An example of an abbreviated payoff table is shown below.*

Table 1.- Abbreviated payoff scale.

	POSSIBLE ACTS		
Possible events in bbls discovered	Don't drill	Drill with 100 percent working interest	Farmout with 1/8 override
Dry Hole	$0	-$50,000	$0
300,000	$0	$425,000	$71,000
1,000,000	$0	$918,000	$141,000
4,000,000	$0	$1,138,000	$171,000

*Assumptions: Dry hole cost is $50,000; completed well costs $70,000, discount rate is 6 percent, present price of oil is $3.00 per barrel; and producing rate is 30,000 barrels per year.

Clearly, this table indicates that the most favorable event would be to obtain a 4,000,000 barrel discovery, and that the most favorable act, if a 4,000,000 barrel discovery is to be made, is to drill the prospect with 100 percent working interest. However, so far, no consideration has been given to the probabilities of the different events. Desirable as a 4,000,000 barrel discovery is, it is more probable that a dry hole will result.

The next stage is to assign probability values to the possible events. This is illustrated in the output from the computer program, in which a payoff table (Table 2) has been printed out. In addition, however, the financial consequences for each event and act have been multiplied by the probability estimate for that event. This yields an expected monetary value for each act, and it is this value that is of prime importance in making decisions. Note, in Table 2, that the expected monetary value of the act of drilling with 100 percent interest is only $15,000, whereas the payoff, if a 4,000,000 barrel discovery were made, would be over a million dollars. But, the probability estimate of a 4,000,000 barrel discovery is very small (0.002 or 1 in 500), whereas the probability estimate of a dry hole (0 barrels) is large (0.60).

In calculating the payoff table (Table 2), an additional refinement has been made by making

Table 2.- Output from prototype single-well decision program relating possible acts and possible events in the form of a payoff table in thousands of dollars. Producing rate is assumed to be 30,000 barrels per year, discount rate is 6 percent, dryhole cost is $50,000, and producing well cost is $70,000. For example, act of choosing 100 percent working interest has an expected monetary value of $15,000.

```
TRIAL, DISCOUNT RATE = 6 PERCENT, DRYHOLE=COST 50,000 , COMP =70,000

PAYOFF TABLE IN THOUSANDS OF DOLLARS
                                                                                POSSIBLE ACTS
POSSIBLE EVENTS     PROB                                WORKING INTEREST                              FARMOUT                 FARMOUT,BACK AFTER PAYOUT
TOTAL PROD 1000'S BBLS                     100 PERCENT    75 PERCENT    50 PERCENT   25  PERCENT  1/8 OVERRIDE 1/16 CVERRIDE   25 PCT W.I.   50 PCT W.I.
                             0      .600           -50           -37           -25           -12             0             0             0             0
                           30.      .150            -7            -5            -3            -1             9             4             0             0
                           60.      .100            51            38            25            12            17             8            12            25
                           90.      .070           107            80            53            26            25            12            26            53
                          150.      .040           211           158           105            52            40            20            52           105
                          300.      .020           425           318           212           106            71            35           106           212
                          600.      .010           712           534           356           178           112            56           178           356
                         1000.      .005           918           688           459           229           141            70           229           459
                         2000.      .003          1099           824           549           274           166            83           274           549
                         4000.      .002          1138           853           569           284           171            85           284           569
EXPECTED MONETARY VALUE                             15            11             7             3            10             5            11            23
```

varying assumptions as to the future price of oil and assigning probability estimates to them. If the assumptions are made that (a) the probability that oil prices will remain stable is 0.50, (b) that oil prices will rise by 2 percent a year is 0.25 and (c) that oil prices will decline by 1 percent a year is 0.25, then the expected future price of oil (Table 3) fifty years hence will be $3.94 per barrel if today's price is $3.00 and so on. Estimates of the future value of oil are implicit in every oil exploration decision, whether estimated quantitatively or not.

The payoff table (Table 2) fails to take into consideration the risk policy of the investor. For example, a small operator's risk-taking ability is generally much less than that of a major oil company. The individual firm's risk policy is conveniently expressed in terms of the utility concept advanced by Daniel Bernoulli, and expanded into the individual utility concept by Von Neumann and Morgenstern (1947).

The utility concept provides a means of describing the consequences of a given event and act in terms of its utility to the individual operator. This may be done by developing a function (conveniently represented by a curve) which relates consequences in dollars to utiles. Utiles are arbitrary units and are a measure, positive or negative, of the consequences to an individual operator. The utility curve of a hypothetical oil operator is shown in Figure 1. Using this curve, the values in the payoff table (Table 2) may be converted to utiles and the expected utility value of each act calculated and presented in tabular form (Table 4). This provides an objective means of choosing the most appropriate act by selecting the one that yields the highest expected utility value.

OBTAINING PROBABILITY ESTIMATES FOR PROSPECTS

It is obvious that obtaining the quantitative probability estimates for various outcomes of individual prospects is a vital part of a quantitative oil-exploration decision system. This is basically a problem in classification. At the moment, it is the Achille's heel, because we have not yet learned very much about applying rigorous classification methods to geologic data. In view of this challenge, my remarks below are directed toward suggesting avenues that may be profitable to follow.

Step I: Selecting Data

The first step in developing such an objective classification system is to select the types of data that bear on the problem of oil exploration. These data may be segregated into (1) geologic data, (2) production data, (3) economic data, and (4) psychological data. These categories may, in turn, include the following forms of data:

Geologic Data:

Formation tops (which in turn yield structural data)

Lithology (from electric logs)

Petrology (lithology, and core analysis data, including porosity and permeability)

Oil field water chemistry

Geophysical data (particularly seismic data that can be interpreted structurally)

Lithology and structure of basement rocks.

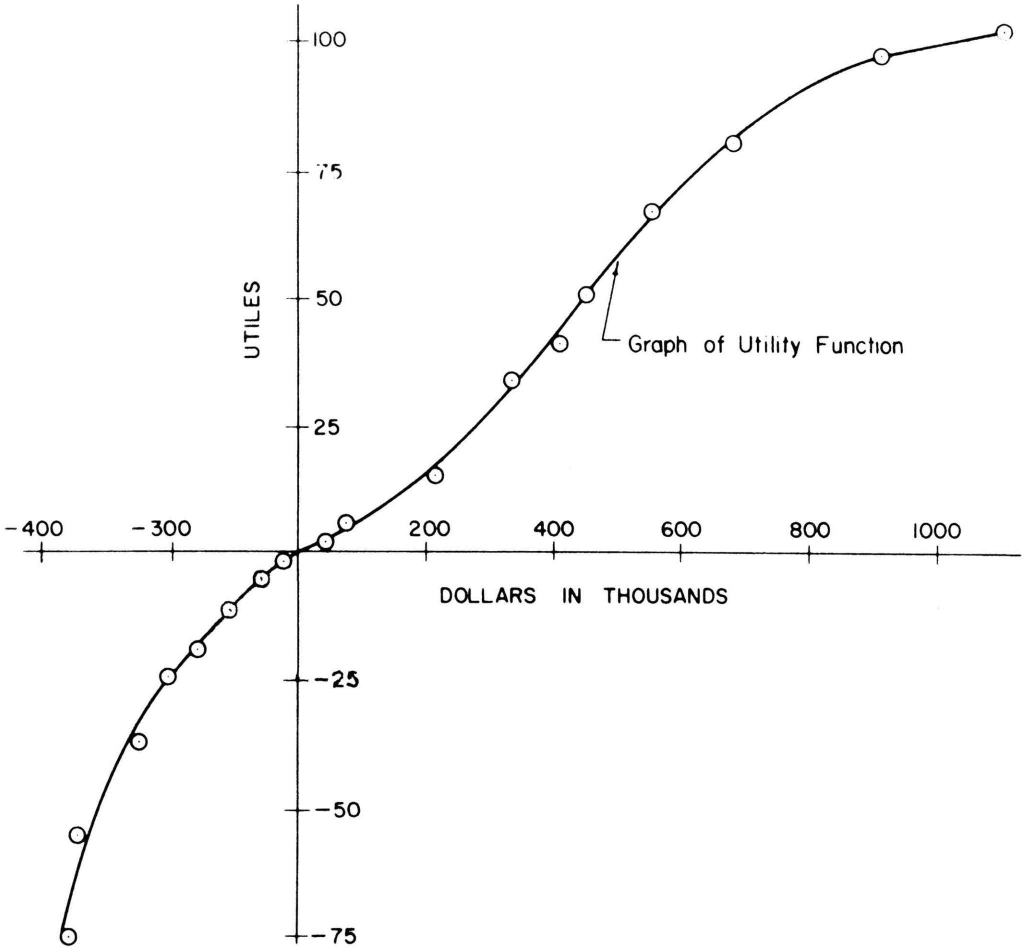


Figure 1.- Graph of utility function in which losses (shown with negative signs) and gains in terms of dollars are related to their utility values expressed as utiles. Curve is hypothetical and would pertain to a single individual or single organization. Shape varies from one utility curve to the next reflecting differences in risk-taking ability. In general, negative part of curve shapes more steeply than the positive part.

Production Data:

Well-test data (drill stem test data, shut-in pressures, production test data, etc.)

Presence of shows, mud-log records, etc.

Oil gravity and oil chemistry data.

Economic Data:

Lease prices

Lease expiration dates

Drilling costs

Road-building and other drill site costs

Current oil and gas prices

One's utility function (economic influences that bear on it)

Psychological Data:

Degree of "hotness" or "coldness" of exploration activity.

Reputation of firm (bears on ability to turn deals, secure favorable farmouts, etc.)

Degree of optimism or pessimism as far as general economic future is concerned, including outlook toward future oil and gas prices.

Government regulatory climate

One's utility function (psychological influences that bear on it).

Step II: Analysis of Data to Obtain Numerical Parameters for Classification Systems

The second step is to devise analytical methods for obtaining numbers or coefficients that

Table 3.-Output from oil-exploration decision program listing expected oil prices for next 100 years using assumptions given.

Trial, discount rate = 6 percent, dryhole=cost 50,000, comp = 70,000
Expected future prices for next 100 years making following assumptions

Present price of oil is $3.00 per barrel
Probability that prices will remain stable is .50
Probability that prices will rise by 2.0 percent is .25
Probability that prices will fall by 1.0 percent is .25

Year	Price	Year	Price	Year	Price	Year	Price	Year	Price	Year	Price
2	$3.01	19	$3.20	36	$3.53	53	$4.04	69	$4.76	85	$5.78
3	$3.02	20	$3.21	37	$3.55	54	$4.08	70	$4.82	86	$5.86
4	$3.02	21	$3.23	38	$3.58	55	$4.12	71	$4.87	87	$5.93
5	$3.03	22	$3.24	39	$3.60	56	$4.16	72	$4.93	88	$6.01
6	$3.04	23	$3.26	40	$3.63	57	$4.20	73	$4.98	89	$6.09
7	$3.05	24	$3.28	41	$3.66	58	$4.24	74	$5.04	90	$6.18
8	$3.06	25	$3.30	42	$3.69	59	$4.28	75	$5.10	91	$6.26
9	$3.07	26	$3.31	43	$3.71	60	$4.33	76	$5.16	92	$6.35
10	$3.08	27	$3.33	44	$3.74	61	$4.37	77	$5.23	93	$6.43
11	$3.09	28	$3.35	45	$3.77	62	$4.42	78	$5.29	94	$6.52
12	$3.10	29	$3.37	46	$3.81	63	$4.46	79	$5.36	95	$6.62
13	$3.12	30	$3.39	47	$3.84	64	$4.51	80	$5.42	96	$6.71
14	$3.13	31	$3.41	48	$3.87	65	$4.56	81	$5.49	97	$6.81
15	$3.14	32	$3.43	49	$3.90	66	$4.61	82	$5.56	98	$6.90
16	$3.15	33	$3.46	50	$3.94	67	$4.66	83	$5.63	99	$7.00
17	$3.17	34	$3.48	51	$3.97	68	$4.71	84	$5.71	100	$7.10
18	$3.18	35	$3.50	52	$4.01						

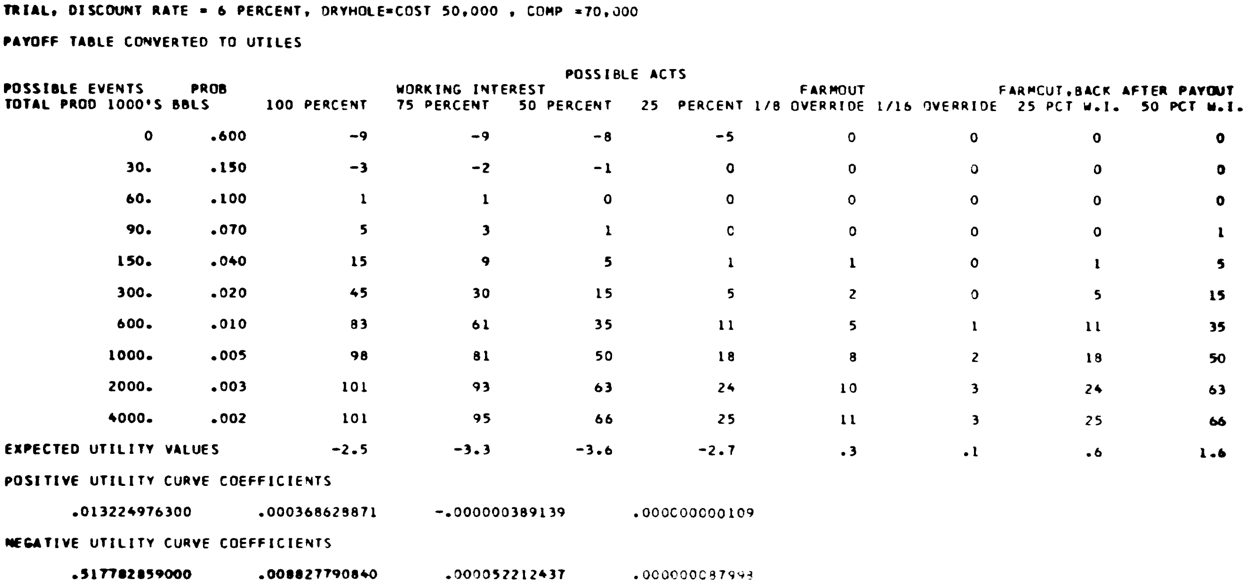
serve to describe the different factors in numerical terms. The immediate purpose of using numerical descriptive methods is to reduce complex relationships to sets of numbers or coefficients which can, in turn, be used for numerical taxonomy purposes. Techniques are available for numerically describing surfaces (such as structure contour maps, lithofacies maps, etc.), fault and fold systems, and for representation of quantitative and qualitative relationships which may be thought of as existing in a kind of multidimensional space. Some techniques are available at present; others remain to be developed.

The techniques that are most advanced at the moment are those which may be used to analyze surfaces, such as structure contour maps, isopach maps and lithofacies maps. These techniques include trend-surface analysis (using power series) and harmonic analysis (using double Fourier series).

Power-series trend-surface analysis is useful in describing gross or regional configuration of structural surfaces and of stratigraphic contour maps, such as lithofacies maps. The coefficients of terms in the power series of various degrees provide a numerical description of the gross configuration of surfaces and may be used for classification purposes. Power-series trend surfaces are fitted to satisfy the least-squares criterion. Residual values, obtained by subtracting trend-surface values from actual data, are also useful in discerning relationships between oil accumulation and structural features (Merriam and Harbaugh, 1963). Specialized applications of power-series trend analysis include mapping of structural closure with respect to inclined oil-equipotential surfaces resulting from hydrodynamic conditions (Harbaugh, 1964). Through the use of such techniques, it may be possible to discern relationships which bear strongly on oil entrapment but which are not readily apparent when more traditional, subjective methods of interpretation are used.

In spite of many previous applications of power-series trend-surface analysis in geology (see bibliography in Merriam and Harbaugh, 1964), additional work is needed to (1) devise numerical means of describing the configurations of residuals, (2) assess the significance of the shape of confidence

Table 4.- Output from prototype simple program listing consequences of possible acts versus possible events in terms of utiles in accordance with the utility function of Figure 1. The act with the greatest positive utility value to the operator is to attempt to farmout the prospect and come back in for a 50 percent working interest after payout. On the other hand, this table shows that even though the acts involving direct working interest have positive expected monetary values (Table 1), they have negative expected utility values to this operator because of his risk position and he should avoid them if he is unsuccessful in farming out the prospect.



```
TRIAL, DISCOUNT RATE = 6 PERCENT, DRYHOLE=COST 50,000 , COMP =70,000

PAYOFF TABLE CONVERTED TO UTILES

                                                     POSSIBLE ACTS
POSSIBLE EVENTS    PROB                 WORKING INTEREST                         FARMOUT                  FARMCUT,BACK AFTER PAYOUT
TOTAL PROD 1000'S BBLS    100 PERCENT  75 PERCENT   50 PERCENT   25  PERCENT 1/8 OVERRIDE 1/16 OVERRIDE  25 PCT W.I.  50 PCT W.I.

          0      .600          -9          -9           -8          -5           0            0            0            0

        30.      .150          -3          -2           -1           0           0            0            0            0

        60.      .100           1           1            0           0           0            0            0            0

        90.      .070           5           3            1           C           0            0            0            1

       150.      .040          15           9            5           1           1            0            1            5

       300.      .020          45          30           15           5           2            0            5           15

       600.      .010          83          61           35          11           5            1           11           35

      1000.      .005          98          81           50          18           8            2           18           50

      2000.      .003         101          93           63          24          10            3           24           63

      4000.      .002         101          95           66          25          11            3           25           66

EXPECTED UTILITY VALUES      -2.5        -3.3         -3.6        -2.7          .3           .1           .6          1.6

POSITIVE UTILITY CURVE COEFFICIENTS

     .013224976300     .000368628871    -.000000389139     .000C00000109

NEGATIVE UTILITY CURVE COEFFICIENTS

     .517782859000     .008827790840     .000052212437     .000000C87993
```

surfaces (Krumbein, 1963), and (3) devise more effective methods of deciding on the number of terms to be included in the power series. Finally, it should be pointed out that there are many other types of series potentially useful in trend-surface analysis. These include Taylor's series and McLaurin's series, and power series (using positive data) that use powers which are represented by decimal fractions rather than by integers (as in conventional power series analysis).

Double Fourier series analysis offers an effective means of analyzing and numerically describing complex surfaces that are oscillatory in their configuration (Preston and Harbaugh, 1965). Fourier series, which may be fitted so as to satisfy the least-squares criterion, provide a means of smoothing and of obtaining residuals similar to that of power-series trend analysis, but, in addition, they permit the underlying harmonics in the data to be isolated and identified. The amplitudes of the harmonics of various period lengths may be represented by arrays of Fourier coefficients, or the contributions of the individual harmonic terms may be represented in somewhat more compact form by power spectrum values. Complex surfaces can be effectively represented by double Fourier series. This means that complexly folded (and even faulted) oil-field structures may be represented by double Fourier series and in turn, permits objective, quantitative comparison between different structures.

A second, principal geologic use of double Fourier series is description and extrapolation of stratigraphic features, such as rhythmic linear sand bodies (Off, 1963), and carbonate banks and reefs whose geographic distribution may be complex but is characterized by underlying periodicities.

Power-series hypersurface analysis is similar to power series trend analysis except that an additional dimension (or dimensions) is involved. The technique may be effectively used to study the distribution of a geologic variable in three-dimensional space, as for example, the distribution of permeability values in a reef reservoir, or distribution of oil-gravity values with respect to depth and geographic location (Harbaugh, 1965). Power-series hypersurfaces are fitted so as to satisfy the least-squares criterion. The spatial distribution of residual values obtained by subtracting hypersurface values from observed values may be particularly revealing. The coefficients and other numerical properties of power-series hypersurfaces are also potentially useful in numerical classification systems.

Step III: Applying Numerical Classification Methods

Once methods of numerically describing geological relationships are available, it is possible to bring numerical classification methods to bear on the problem. Some of the methods of numerical classifications are already at hand (Sokal and Sneath, 1963; Cooley and Lohnes, 1962), and include factor analysis, discriminant-function analysis, the selective use of transformations to aggregate points into clusters in n-dimensional Euclidean space (Jizba, 1964), and maximization of the distance between point clusters (Sebestyen, 1962). These techniques are part of the general realm of pattern recognition techniques, whose applications in automated oil well-log analysis have been discussed by Daskam. The numerical classification methods would be background for the third stage where they could be used objectively to classify oil prospects in terms of success probabilities. The principal problem in the use of numerical classification systems is to devise transformations which will yield classifications that are found useful in practice. In other words, the idealized objective in numerically classifying oil prospects is to devise a system which will tend to distinguish prospects that are eventually proven to be productive, from those that turn out to be unproductive. The usefulness of numerical taxonomic system in distinguishing cherts or jasperoids that form potential host rocks for ore deposits, from those that are not, has already been demonstrated by Howd (1964). Consequently, there appears to be no inherent reason why numerical taxonomic methods should not be equally valid when applied to petroleum prospect data, particularly through the use of methods that are capable of using both qualitative and quantitative data, such as the general classification program of Tanimoto and Loomis (1960).

Step IV: Development of Learning Model for Estimating Exploration Prospect Success Probabilities

The fourth step would consist of developing a computer learning system for objectively appraising the probabilities of success for individual oil prospects. The proposed system would be conditioned by experience. Accordingly, it is proposed to build into the system a method by which performance is improved by experience, using some of the numerical adaptive learning techniques that are currently being developed at Stanford University (Fralick, 1964; Hu, 1963; Koford, 1964; and Specht, 1964) and elsewhere. The approach would be to select one or more areas that have been densely drilled, that contain many oil fields, and that have witnessed oil-field development more on less continuously over the past 10 years. These would then be studied in detail on a historical basis. For example, starting in 1956, utilizing the geological data available at that time, predictions would be made for 1957. Then, these predictions would be compared with the actual results in 1957. Similarly, year by year (or on some suitable time-increment basis), comparisons of predictions with actual results would be made. For predictions that prove to be correct, the system would be mathematically "rewarded;" for predictions that do not accord with observed results, the system would be mathematically "punished." Thus, in time, the system's performance should improve. One important insight that should be gained in the process is how it was improved.

Probability values of the various degrees of success (dry hole, 200,000 barrel discovery, 50 million barrel discovery, etc.) obtained in such a manner could be effectively used by both human decision makers, and with computerized, numerical oil-exploration decision systems which might be subsequently developed to monitor a company's entire spectrum of oil exploration activities over a large area.

REFERENCES

Cooley, W.W., and Lohnes, P.R., 1962, Multivariate procedures for the behavioral sciences: Wiley and Sons, Inc., New York, 211 p.

Fralick, S.C., 1964, The synthesis of machines which learn without a teacher: Tech. Rept. no. 6103-8, Systems Theory Lab., Stanford Electronics Laboratories, 19 p.

Grayson, C.J., 1960, Decisions under uncertainty: Drilling Decisions by oil and gas operators: Harvard Business School Press, 402 p.

Harbaugh, J.W., 1964, Trend-surface mapping of hydrodynamic oil traps with the IBM 7090/94 computer: Colorado School Mines Quart., v. 59, no. 4, p. 557-578.

Harbaugh, J.W., 1965, A computer method for four-variable trend analysis illustrated by study of oil-gravity variations in southeastern Kansas: Kansas Geol. Survey Bull. 171, 58 p.

Harbaugh, J.W., and Preston, F.W., 1965, Fourier series analysis in geology: International Symposium on Computers and Computer Applications in Mining and Exploration, Tucson, Arizona, v. 1, p. R-1-R-46.

Howd, F.H., 1964, The taxonomy program - A computer technique for classifying geologic data: Colorado School Mines Quart., v. 59, no. 4, p. 207-222

Hu, Michael Jen-Chao, 1963, A trainable weather-forecasting system: Tech. Rept. no. 6759-1; Systems Theory Laboratory, Stanford Electronics Laboratories, 19 p.

Jizba, Z.V., 1964, A contribution to statistical theory of classification; in Computers in the Mineral Industries, Stanford Univ. Publ., Geological Sciences, v. 9, no. 2, p. 729-756.

Koford, J.S., 1964, Adaptive pattern dichotomization: Tech. Rept. no. 6201-1; Systems Theory Laboratory, Stanford Electronics Laboratories, 114 p.

Krumbein, W.C., 1963, Confidence intervals on low-order polynomial trend surfaces: Jour. Geophys. Res., v. 68, p. 5869-5878.

Merriam, D.F., and Harbaugh, J.W., 1964, Trend-surface analysis of regional and residual components of geologic structure in Kansas: Kansas Geol. Survey Sp. Dist. Publ. 11, 27 p.

Merriam, D.F., and Harbaugh, J.W., 1963, Computer helps map oil structures: Oil and Gas Jour., v. 61, no. 47, p. 158-159, 161-163.

Off, T., 1963, Rhythmic linear sand bodies caused by tidal currents: Am. Assoc. Petroleum Geologists Bull., v. 47, no. 2, p. 324- 341.

Preston, F.W., and Harbaugh, J.W., 1965, BALGOL programs and geologic application for single and double Fourier series using IBM 7090/7094 computers: Kansas Geol. Survey Sp. Dist. Publ. 24, 71 p.

Sebestyen, G.S., 1962, Decision-making processes in pattern recognition: ACM Monograph series, The Macmillan Co., 162 p.

Sokal, R.R., and Sneath, P.H.A., 1963, Principles of numerical taxonomy: W.H. Freeman and Co., 359 p.

Specht, D.F., 1964, Vectorcardiographic diagnosis utilizing adaptive pattern-recognition techniques: Systems Theory Laboratory, Stanford Electronics Laboratories, 54 p.

Tanimoto, T.T., and Loomis, R.G., 1960, A taxonomy program for the IBM 704: Math. and Applications Dept., International Business Machines Corp., New York.

Von Neumann, John and Morgenstern, Oskar, 1947, Theory of games and economic behavior: Princeton Univ. Press.

TWO-DIMENSIONAL POWER SPECTRA FOR CLASSIFICATION OF LAND FORMS

by

Floyd W. Preston
The University of Kansas

ABSTRACT

The methods of two-dimensional power spectrum analysis are suggested for the numerical description of land forms and possibly for their classification.

LAND FORM DESCRIPTION

Description of land forms has always been of interest to geologists and geographers. Early descriptions, though useful, were verbal and representations of the viewer's visual experience. For communication purposes and for efficiency of information transmittal, most of the verbal descriptive phrases are neglected and maps are used to convey the essential characteristics of the land form. The failure of verbal descriptions to adequately represent the surface even when the descriptions are voluminous is evident if one performs the simple experiment of transmitting the description to a competent geologist or geographer and asks that the description be the basis for construction of a two-dimensional picture or a scaled three-dimensional model of the prototype. Of course, carefully drawn maps would allow adequate surface reconstruction. The problem with map representation is that no really general methods exist for map comparison. Surface classification becomes a highly arbitrary, subjective process. If map comparison merely implies point to point differencing of the dependent variable portrayed at each coordinate location on two maps being compared, then these differences can be plotted to give a difference map which then becomes but an additional map with the same interpretation and generalization difficulties of the first map. This is not to say that such maps are without meaning or utility. Indeed, for certain purposes, difference maps for first, second, third, and even higher orders of differences are very useful. However, it is not our purpose to consider the many possible derivative maps but rather to suggest one possible means for map or rather surface characterization. Those who are concerned with the concepts and problems of map generalization are referred to the excellent discussion of this subject by Tobler (1966).

QUANTIFICATION OF SURFACE GENERALIZATION

The objectives of land form numerical characterization influence the means used for characterization. If one is interested primarily in the grossest simplification of the land surfaces in terms of slope and direction, then one might seek to extract the "linear trend." For such a characterization, the model which one would use is the simplest form of the "general linear model" of Krumbein and Graybill (1965). The observed elevation at coordinate location $x_i y_i$ is $S(x_i, y_i)$ and is considered to be composed of a trend $T(x_i, y_i)$ and a random component, $\epsilon_{i,i}$, uncorrelated with x_i or y_i. This model is represented by:

$$S(x_i, y_i) = T(x_i, y_i) + \epsilon_{i,i} \tag{1}$$

The linear trend is given by

$$T(x_i, y_i) = A_0 + A_1 x_i + A_2 y_i \tag{2}$$

where the coefficients A_0, A_1, A_2 are most generally determined from the data by the method of least squares. Higher order trends may be extracted by the same methods if there exists suitable evidence that the appropriate generalization of the surface should be these higher order surfaces. The method of trend-surface analysis has been used by geologists and geophysicists for almost a decade to generalize observations represented by maps or to examine deviations of the observation from the generalized surface.

Conventional trend surface analysis uses power series polynomials of the form:

$$T(x_i, y_i) = \sum_{n=0}^{n=N} \sum_{m=0}^{m=M} A_{m,n} x_i^n y_i^m \tag{3}$$

As a method of surface characterization, this method suffers from the limitation that for such polynomials, the coefficients $A_{m,n}$ of low order do not remain constant as higher order polynomial terms are added to the series.

This well known limitation of such series is overcome by use of orthogonal polynomials. Such polynomials are described by Oldham and Sutherland (1955). One member of the class of orthogonal polynomials that has been used with increasing frequency has been the Fourier series extended to two dimensions. Bhattacharyya (1965) has applied the series to magnetic data and Tsuboi (1959) used the series to study gravity anomalies. Bennion (1965) and Bennion and Griffiths (1965) used Fourier series as one method of describing two-dimensional variation in reservoir rock properties. The same methods have been shown by Harbaugh and Preston (1965) to be useful for describing surface variation in mineral content and description of land forms. Computer programs for double Fourier series and additional examples of fitting of land surfaces are given by Preston and Harbaugh (1965). More recently, James (1966) has shown how the coefficients of the Fourier series can be computed when data are not equally spaced on a rectangular grid. Krumbein (1966) has compared the relative merits of orthogonal polynomials of the power series and Fourier series type for purposes of surface interpolation and extrapolation. He states that a power series representation is superior for extrapolation and the Fourier series for interpolation. The Fourier series, because it has fewer extrema for a given number of terms, gives the better fit to the surface in the region fitted. However, beyond the region fitted, extrapolation is not reasonable because the pattern within the measured boundary is replicated exactly outside the boundaries.

The use of such orthogonal polynomials for surface representation implies the possibility of using the coefficients of the polynomials as a set of numerical descriptors of the surface. Such coefficients could then be used in the various classification techniques well known to numerical taxonomists and summarized by Sokal and Sneath (1963). Indeed, such a proposal was made by Fara and Scheidegger (1961) using one-dimensional Fourier series for characterization of porous media. The coefficients were stated to be numerical descriptors of a porous medium represented in one dimension by a Fourier series. This method was applied by Preston and Henderson (1964) to electric log analysis. Coefficients obtained by fitting a one-dimensional Fourier series to resistivity logs of certain limestone formations could be used in a discriminant function analysis to discriminate one formation from another. The method still needs further work to test its generality. It would seem logical to extend this same type of analysis to coefficients of two-dimensional Fourier series fitted to geologic and topographic surfaces.

If surface representation alone is the primary goal, then the Fourier or other orthogonal polynomial series suffices and the coefficients for such series can be obtained by the method of least squares.

SPECTRAL ANALYSIS OF SURFACES

The above methods are based upon the "general linear model." An important assumption inherent in the model when applied to surfaces is that, as mentioned in equation (1), the surface is considered to be composed of a random component, $\epsilon_{i,j}$ and a series of terms either (a) representing integral powers of the independent variables (power series) or (b) representing harmonics of the fundamental wave lengths (measured intervals in the two independent directions). For some surfaces, this assumption may be perfectly valid either theoretically or even from the pragmatic viewpoint of empirical surface fitting. However, another model does exist that is not so restrictive in this sense. This model considers the observed surface $S(x_i, y_j)$ to be a random variable.

Statistically, a random variable may be either discrete or continuous. We shall concern ourselves only with continuous random variables. A continuous random variable has a continuous range of variation but its value at any time (if it is considered to be time variant) or at any location in space (if it is considered to be space variant) is not given by any deterministic relation. Instead its value at any time (for time variant systems) or space coordinates (for space variant systems) must be described by a probability distribution. It is this probability distribution function that is the essential characteristic of the random variable. Most of the literature of the application of statistical random functions is in terms of time variant systems and therefore the method is often considered to be an aspect of time-series analysis.

One cannot perceive by visual inspection whether an oscillatory series has been generated by a random variable or by a function having discrete harmonics. However, it is well recognized in the theory of time series analysis that harmonic analysis, which is appropriate for the phenomenon with discrete harmonics, yields misleading results when applied to a phenomenon whose generating process can be considered a random variable. This circumstance was first indicated by Bartlett (1948). Later Weiner (1949) developed a comprehensive theory for the description of such time series. More recent expositions of the theory and applications are by Bendat and Piersol (1966) and Blackman (1965). The most extensive use of this methodology has been in the statistical theory of communication. Blackman and Tukey (1958), Lee (1960), and Davenport and Root (1958) are definitive texts in this field. The method has also been applied in economic time series analysis by Granger and Hatanaka (1964). Recently an excellent bibliography by Wold (1965) has appeared on all aspects of time series analysis and stochastic processes. Very little work has been done applying the concept of a random variable to spatially variant data. A pioneering effort was that of Pierson. A computer program for this analysis has been prepared by Esler and Preston (1966). Simultaneously, similar work is being done by Tobler (1966).

In the Fourier harmonic analysis methodology the numerical description of the surface is the set of coefficients for the Fourier series. In the power spectrum approach the characterizing relation is the "spectrum" which more properly should be considered to be a "variance spectrum." However, for historical reasons, the term "power" spectrum is applied because of its use in the determination of the electrical frequency variation of energy (or power) dissipation within alternating current circuits.

Although a sizeable body of theory exists for the analysis of time series (here considered analogous to "distance" series) where such a series is available as a continuous signal or as a continuous analytical function, we shall concern ourselves only with the data that represent discrete equally spaced sampling.

COMPUTATION OF TWO-DIMENSIONAL SPECTRA FOR SURFACES

The power spectrum is obtained by first extracting from the data an auto correlation function and then obtaining the power spectrum by computing the finite Fourier transform of the auto correlation function. The fact that the power spectrum is indeed this transform of the auto correlation function is well known (Lee, 1960; Weiner, 1949).

The data are considered to be an array $S(x_i, y_j)$, here shortened because of the equal spacing of the data to $S(i,j)$. The auto correlation function $Q(p,q)$ for such an array $S(i,j)$ is computed at the lags p and q. The terms p and q represent the displacements between pairs of values of x_i and y_j respectively where $i = 1,2,\ldots n.,\ j = 1,2,\ldots m.$ The auto correlation function $Q(p,q)$ is computed from:

$$Q(p,q) = \frac{1}{(n-p)(m-|q|)} \sum_{j=|q|}^{m} \sum_{i=1}^{n-p} S_{i,j} S_{i+p,j+q} \tag{4}$$

for:

$p = 0,1,2,\ldots,\tau$

$q = -\tau,\ -(\tau-1),\ldots,1$

and

$$Q(p,q) = \frac{1}{(n-p)(m-|q|)} \sum_{j=1}^{j=m-|q|} \sum_{i=1}^{n-p} S_{i,j} S_{i+p,j+q} \tag{5}$$

for:

$p,q, = 0,1,2,\ldots,\tau$

Because the data are taken on a finite grid, approximations must be made in the above equations at grid boundaries. The following weighting relations are therefore assumed. Weighted values of $Q(p,q)$ are designated $wQ(p,q)$.

$$\begin{array}{ll}
wQ(p,q) = 2Q(p,q) & p = 1 \text{ to } (\tau-1) \\
 & q = -(\tau-1) \text{ to } (\tau-1) \\
wQ(0,q) = Q(0,q) & q = -(\tau-1) \text{ to } (\tau-1) \\
wQ(\tau,q) = Q(\tau,q) & q = -(\tau-1) \text{ to } (\tau-1) \\
wQ(p,\tau) = Q(p,\tau) & p = 1 \text{ to } (\tau-1) \\
wQ(p,-\tau) = Q(p,-\tau) & p = 1 \text{ to } (\tau-1) \\
wQ(0,\tau) = 1/2\ Q(0,\tau) & \\
wQ(0,-\tau) = 1/2\ Q(0,-\tau) & \\
wQ(\tau,\tau) = 1/2\ Q(\tau,\tau) & \\
wQ(\tau,-\tau) = 1/2\ Q(\tau,-\tau) &
\end{array} \tag{5}$$

The unsmoothed estimates of the power spectrum are then given by:

$$SR(r,s) = \frac{1}{2\tau^2} \sum_{q=-\tau}^{\tau} \sum_{p=0}^{\tau} Q(p,q) \cos \frac{\pi}{\tau}(rp+sq) \tag{6}$$

$r = 0,1,2,\ldots\tau$

$s = -\tau, -(\tau-1),\ldots,-1,0,1,\ldots\tau$

The boundary conditions used in the smoothing process are:

$$\begin{array}{l}
SR(-1,b) = SR(1,b) \quad b = -\tau \text{ to } \tau \\
SR(\tau+1,b) = SR(\tau-1,b) \quad b = -\tau \text{ to } \tau \\
SR(a,\tau+1) = SR(a,\tau-1) \quad a = 0 \text{ to } \tau \\
SR(a,-\tau-1) = SR(a,-\tau+1) \quad a = 0 \text{ to } \tau \\
SR(-1,\tau+1) = SR(1,\tau-1) \\
SR(\tau+1,\tau+1) = SR(\tau-1,\tau-1) \\
SR(\tau+1,-\tau-1) = SR(\tau-1,-\tau+1) \\
SR(-1,-\tau-1) = SR(1,-\tau+1)
\end{array} \tag{7}$$

Because the data are taken at discrete, equally spaced points, the spectrum can only be computed at discrete lags r,s. The resulting function would, if plotted, represent a spike or comb type function. In reality, what is wanted however is the representation that in a statistical sense is the envelope of this comb function. One method for obtaining a closer approximation to this smooth, continuous function is to employ a smoothing function (often called a filter) to the raw spectrum. Considerable effort has gone into the appropriate and optimal designs of such filters for time series. The one used here is that developed by Pierson (1960) as the two-dimensional generalization of the Tukey-von Hann filter (see Blackman and Tukey, 1958, p. 98). The weighting function is as follows:

Weighting Function Values
at point $S(x_i, y_j)$

		y_{j-1} 0.25	y_j 0.50	y_{j+1} 0.25
x_{i-1}	0.25	0.0625	0.125	0.0625
x_i	0.50	0.125	0.250	0.125
x_{i+1}	0.25	0.0625	0.125	0.0625

The smoothed spectrum will then be given by:

$$\begin{aligned} SS(r,s) = {} & 0.0625[SR(r-1,s-1)+SR(r-1,s+1) \\ & + SR(r+1,s-1)+SR(r+1,s+1)] \\ & + 0.125\,[SR(r-1,s)+SR(r+1,s) \\ & + SR(r,s-1)+SR(r,s+1)] \\ & + 0.250\,[SR(r,s)] \end{aligned} \tag{8}$$

The final values for the smoothed power spectrum, SS(r,s), represent the contributions to the total variance made by frequencies between $2\pi(r-1/2)/2\tau\Delta t$ and $2\pi(r+1/2)/2\tau\Delta t$ in the r direction, and $2\pi(s-1/2)/2\tau\Delta t$ and $2\pi(s+1/2)/2\tau\Delta t$ in the s direction.

REMOVAL OF LINEAR TREND

Often data will contain a linear trend that will badly distort the power spectrum. Such a trend can easily be removed, however, by using standard least-squares techniques. This involves subtracting a plane of the form $Z(i,j) = ai+bj+c$ from the data points $S_{i,j}$ with $i = 1,2,\ldots m$, $j = 1,2,\ldots n$. Pierson (1960) has shown that the constants a, b, and c can be determined by the matrix equation:

$$\begin{vmatrix} \frac{mn(n+1)\ (2n+2)}{6} & \frac{n(n+1)m(m+1)}{4} & \frac{mn(n+1)}{2} \\ \frac{n(n+1)m(m+1)}{4} & \frac{mn(m+1)\ (2m+2)}{6} & \frac{mn(n+1)}{2} \\ \frac{mn(n+1)}{2} & \frac{mn(m+1)}{2} & \frac{mn}{1} \end{vmatrix} \begin{vmatrix} a \\ b \\ c \end{vmatrix} = \begin{vmatrix} \sum_{i=1}^{n} \sum_{j=1}^{m} jD_{i,j} \\ \sum_{i=1}^{n} \sum_{j=1}^{m} iD_{i,j} \\ \sum_{i=1}^{n} \sum_{j=1}^{m} D_{i,j} \end{vmatrix} \tag{9}$$

Solving for a, b, and c, and subtracting Z(i,j) from the data matrix will then yield a new data matrix $S_{i,j}$ better adapted to the power spectrum analysis method.

CHOICE OF PARAMETERS

The accuracy and usefulness of the power spectrum depends largely on the choice of certain parameters, such as the sampling interval Δt and the maximum value of lag τ. Proper choices for these parameters will minimize aliasing and distortion, as well as maximizing the number of degrees of freedom and the resolution.

Bendat has suggested that the sampling interval Δt should be given by:

$$\Delta t = \frac{1}{2f_c} \tag{10}$$

where f_c is the lowest frequency of interest in the record along either of the two axes. This gives two points per cycle at the cutoff frequency f_c. Bendat recommends that, wherever possible, more points be used in practice for improved results. An accurate correlation function can be formed by taking

$$\Delta t = \frac{1}{4f_c} \tag{11}$$

If the power spectrum is of prime concern, a spacing of

$$\Delta t = \frac{2}{5f_c} \tag{12}$$

is sufficient. Values of Δt as close as possible to $1/2\,f_c$ are, of course, most economical, as fewer points are needed.

The maximum value of lag τ should be limited by the smaller of the two dimensions, say m, of the data matrix. Several limits have been suggested for τ in the one-dimensional case. For instance, Granger and Hatanaka (1964) recommend use of $\tau < m/3$. Blackman (1965) recommends the use of $\tau < m/10$, as does Crowson (1963), who performs a very thorough error analysis. With the larger number of data points in the two-dimensional problem, a value of $\tau < m/4$ should provide adequate results.

Another important aspect to consider in the choice of τ is the equivalent resolution bandwidth, B_c, desired for the power spectrum calculations. Bendat has determined B_c to be given by:

$$B_c = \frac{1}{\tau\Delta t} \tag{13}$$

For a given Δt, B_c will decrease as τ increases.

The degrees of freedom, f, of each spectral estimate in the one-dimensional case has a Chi square distribution, where f is determined by

$$f = 2\left(\frac{n}{\tau} - \frac{1}{4}\right) \tag{14}$$

Pierson (1960) has expanded this formula to two dimensions to give:

$$f = 1.58\left(\frac{n}{\tau} - 1/2\right)\left(\frac{m}{\tau} - 1/2\right) \tag{15}$$

Equation (15) may give an underestimate of the true number of degrees of freedom. Values of 1/4 instead of 1/2 may therefore weight equation (15) more properly.

PORTRAYAL OF SPECTRAL ANALYSIS

The final portrayal of the spectral analysis would be that used to portray the function SS = f(r,s) where r and s are the two independent (orthogonal) variables. This could be presented as a three-dimensional model or as a computer produced map for machine or computer contouring. However, these representations contain within them the very pitfalls the spectral analysis procedure is trying to avoid. A more appropriate representation would be as a two-dimensional array of variables, analyzable by existing techniques of numerical classification.

ACKNOWLEDGMENTS

The author would like to acknowledge the contributions of Jim Esler who under an N.S.F. Undergraduate Research Program, developed the computer programs for this analysis, and furnished some material on sampling and confidence intervals. Computer time was also made available (August 15, 1965) under project grant 0570 from The University of Kansas Computation Center.

The author also wishes to express his appreciation to Professor W.R. Tobler of the University of Michigan for making his paper on spectral analysis available in advance of publication.

REFERENCES

Bartlett, M. S., 1948, Smoothing periodograms from time-series with continuous spectra: Nature, v. 161, p. 686-687.

Bendat, J. S., and Piersol, A. G., 1966, Measurement and analysis of random data: John Wiley and Sons, New York, 390 p.

Bennion, D. W., 1965, A stochastic model for predicting variations in reservoir rock properties: The Pennsylvania State Univ., doctoral dissertation.

————, and Griffiths, J. C., 1965, A stochastic model for predicting variations in reservoir rock properties: Soc. Petroleum Engineers Paper SPE-1187, Presented at the Denver Ann. SPE(AIME) Meeting Oct., 1965.

Bhattacharyya, B. K., 1965, Two-dimensional harmonic analyses as a tool for magnetic interpretation: Geophysics, v. 30, p. 829-857.

Blackman, R. B., 1965, Linear data-smoothing and prediction in theory and practice: Addison-Wesley Co. Reading, Mass. 182 p.

————, and Tukey, J. W., 1958, The measurement of power spectra from the point of view of communication engineering: Dover Publincations Inc., New York, 190 p.

Crowson, H. L., 1963, Error analysis in the digital computation of the autocorrelation function: AIAA Jour., v. 1, p. 488-489, 1968-9F.

Davenport, W. B., Jr., and Root, W. L., 1958, An introduction to the theory of random signals and noise: McGraw-Hill Book Co., 393 p.

Fara, H. D., and Scheidegger, A. E., 1961, Statistical geometry of porous media: Jour. Geophysical Research, v. 66, no. 10, p. 3279-3284.

Granger, C. W. J., and Hatanaka, M., 1964, Spectral analysis of economic time series: Princeton Univ. Press.

Harbaugh, J. W., and Preston, F. W., 1965, Fourier series analysis in geology: College of Mines, Arizona Univ., v. 1.

James, W. R., 1966, The Fourier series model in map analysis: Technical Report ONR Task No. 388-078, Contract Nonr-1228(36), Dept. of Geology, Northwestern Univ.

Krumbein, W. C., 1966, A comparison of polynomial and Fourier models in map analysis: Technical Report No. 2, ONR Task No. 388-078, Contract Nonr-1228(36), Dept. of Geology, Northwestern Univ.

————, and Graybill, F. A., 1965, An introduction to statistical models in geology: McGraw-Hill Book Co.,475 p.

Lee, Y. W., 1960, Statistical theory of communication: John Wiley and Sons, New York, 509 p.

Oldham, C. H. G., and Sutherland, D. B., 1955, Orthogonal polynomials and their use in estimating the regional effect: Geophysics, v. 20, no. 2, p. 295-306.

Pierson, W. F., Jr., and others, 1960, The directional spectrum of a wind generated sea as determined from data obtained by the stero wave observation project: Metrological Papers, v. 2, no. 6, New York Univ., College of Engineering.

Preston, F. W., and Esler, J. E., in preparation, A FORTRAN program to compute two dimensional power spectra: Kansas Geol. Survey Computer Contr.

————, and Harbaugh, J. W., 1965, BALGOL programs and geologic application for single and double Fourier series using an IBM 7090/7094 computer: Kansas Geol. Survey Sp. Dist. Pub. 24, 72 p.

————, and Henderson, J. H., 1964, Fourier series characterization of cyclic sediments for stratigraphic correlation: Symposium on cyclic sedimentation, D. F. Merriam ed., Kansas Geol. Survey Bull. 169, v. 2, p. 415-425.

Sokal, R. R., and Sneath, P. H. A., 1963, Principles of numerical taxonomy: W. H. Freeman and Co.

Tobler, W. R., 1966, Numerical map generalization: Michigan Inter-University Community of Mathematical Geographers, Discussion Paper No. 8, January, 1966.

————, 1966, Spectral analysis of spatial series: Paper delivered at Fourth Ann. Conf. on Urban Planning, Information System and Programs, Berkeley, California, Aug. 1966.

Tsuboi, C., 1959, Application of double Fourier series to computing gravity anomalies and other gravimetrical quantities at higher elevations from surface gravity anomalies: Report No. 2, Institute of Geodesy, Photogrammetry and Cartography, The Ohio State Univ.

Wiener, N., 1949, The extrapolation, interpolation, and smoothing of stationary time series: John Wiley and Sons, New York.

Wold, H. O. A., 1965, Bibliography on time series and stoichastic processes: The M.I. T. Press, 516 p.

DISTRIBUTION-FREE QUADRATIC DISCRIMINANT FUNCTIONS IN PALEONTOLOGY

By

T. P. Burnaby
University of Keele

ABSTRACT

It is often supposed that conventional methods of multivariate analysis lack robustness with respect to departures from the usual assumptions, i. e., parent populations are multivariate normally distributed with identical dispersion matrices. However, recent work by P. W. Cooper has shown that, subject to mild restrictions, the D^2 generalized distance statistic is robust, almost to the extent of being distribution-free, when used as a discriminant function to assign an unknown vector $\underline{x}$ to one of several parent populations. The main requirement is that the determinants of the dispersion matrices of the distributions of the parent populations should be equal. It is natural to inquire whether it is possible to select a simple type of variate transformation which will bring the determinants as near to equality as possible. A method for doing this is described, and illustrated using paleontological data. The possibility of implementing a completely nonmetric approach to the problem is also discussed.

INTRODUCTION

An observation vector $\underline{x}$, of p rows and 1 column, originates from one of K populations. We wish to determine the value of k, where $1 \leqslant k \leqslant K$, such that the statement "$\underline{x}$ is a member of the k^{th} population" has the highest possible probability of truth. We shall assume equal prior probabilities for each of the K populations, and equal costs of misclassification.

Practical applications may require greater generality. Thus for example, $\underline{x}$ may consist of p measurements on an unidentified fossil specimen, and we wish to assign it to the k^{th} species from a genus for which K - 1 species have been described: since there is always a nonzero probability that the fossil may belong to a hitherto-undescribed species, there must exist (at least) one population for which no data is available with regard to the location of its mean vector. It would clearly be misleading to assume equal prior probabilities, and equal costs of misclassification, in this case.

NOTATION

x, X denote scalars, $\underline{x}$, $\underline{x}'$, $\underline{X}$ denote a column vector, a row vector, and a square p x p matrix respectively. Let $\underline{u}_k$ be the mean vector of the k^{th} population in p-dimensional space, estimated by $\underline{\bar{x}}_k = n^{-1} \sum_{i=1}^{n} \underline{x}_{ik}$. Let the dispersion matrix of the k^{th} population be $\underline{W}_k$, with inverse $\underline{W}_k^{-1}$. If the second moment of the population distribution exists, $\underline{W}_k$ is identical with the covariance matrix $\underline{\Sigma}_k$, estimated by $\underline{S}_k = [s_{ij}]$ such that

$$s_{ij} = (n - 1)^{-1} \sum_{1}^{n} (x_i - \bar{x}_i)(x_j - \bar{x}_j)$$

The generalized squared distance (Mahalanobis' D^2) between a vector $\underline{x}$ and the mean point $\underline{u}_k$ of the k^{th} population we shall denote by

$$Q_k(\underline{x}) = (\underline{x} - \underline{u}_k)' \underline{W}_k^{-1} (\underline{x} - \underline{u}_k) = \sum_{i=1}^{p} \sum_{j=1}^{p} w^{ij}(x_i - u_{ik})(x_j - u_{jk}) \qquad (1)$$

We shall consider unskewed, unimodal, multivariate distributions having frequency densities of the form

$$dF_k = A_k |\underline{W}_k|^{-1/2} f_k [Q_k(\underline{x})^{1/2}] \, dx_1 dx_2 \ldots dx_p \qquad (2)$$

where f_k is a functional form integrable over p-dimensional space. Thus for example, for the multivariate normal frequency density function, we would have

$$dF_k = (2\pi)^{-1/2p} |\underline{\Sigma}_k|^{-1/2} \exp[-1/2 Q_k(\underline{x})] \, dx_1 dx_2 \ldots dx_p$$

The corresponding density functions for the multivariate Pearson Types II and VII distributions are given in Cooper (1963). Together with the multivariate normal, they form a continuous family, Type II being platykurtic with finite range, and Type VII leptokurtic with infinite range. The normal is the least leptokurtic distribution with infinite range. The complete family of Type II distributions passes through the rectangular into U-shaped distributions, but these last we shall not consider further.

We shall now define a general quadratic discriminant function by the expression

$$G_k(\underline{x}) = a_k Q_k(\underline{x}) + b_k \qquad (3)$$

the constants a_k and b_k being determined so that for any vector $\underline{x}$, the statement "$\underline{x}$ is a member of the k^{th} population" has the smallest probability of error for that value of k for which the value of $G_k(\underline{x})$ is minimized, given that the prior probabilities of the K populations are equal.

PROPERTIES OF QUADRATIC DISCRIMINANT FUNCTIONS

It is well known (Anderson, 1958) that when f_k is the multivariate normal frequency density function, the quadratic discriminant function $G_k(\underline{x})$ is given by

$$G_k(\underline{x}) = (\underline{x} - \underline{u}_k)'\underline{\Sigma}_k^{-1}(\underline{x} - \underline{u}_k) + \ln|\underline{\Sigma}_k|$$

and that this discriminant is optimal, in the sense that the total probability of misclassification is minimized by assigning $\underline{x}$ to the population for which the value of $G_k(\underline{x})$ is least. If the $\underline{\Sigma}_k$ are all identical, we have, for the h^{th} and k^{th} populations

$$\begin{aligned} G_{h,k} &= G_h(\underline{x}) - G_k(\underline{x}) \\ &= 2\underline{x}'\underline{\Sigma}^{-1}(\underline{u}_k - \underline{u}_h) + \text{a constant depending} \end{aligned}$$

on $\underline{u}_h$ and $\underline{u}_k$, whence we get

$$G_{h,k} = c_0 + c_1x_1 + c_2x_2 + \dots + c_px_p$$

so that in this case the difference of the two quadratic discriminants is identical to the linear discriminant function, and this is optimal in the sense stated above.

It has been shown by Cooper (1965) that when the distributions for all K populations have the same functional form f, the quadratic discriminant function defined in (3) above, is optimal for the following situations:

(1) $\underline{\Sigma}_k$ not identical in the K populations: f is the frequency density function for the multivariate unimodal Pearson Type II distribution.

(2) $\underline{\Sigma}_k$ not identical: f is the f.d.f. for the multivariate Pearson Type VII distribution with finite second moment.

(3) As for (2), but f is the f.d.f. of the Type VII distribution without finite second moment. In this case, $\underline{\Sigma}_k$ is replaced by $\underline{W}_k$, where $\underline{W}_k^{-1}$ is a scaling matrix transforming the distribution from an ellipsoidal into a spherically symmetrical shape.

(4) For any unimodal unskewed multivariate distribution whatever, having scaling matrices $\underline{W}_k$ not identical for each of the K populations, but having equal determinants, i. e.$|\underline{W}_h| = |\underline{W}_k|$ for all h, k. In this case, the constants a_k and b_k in (3) are the same for all K populations and can therefore be omitted. The matrices $\underline{W}_k$ need not be estimable by computing variances and covariances; but provided they can be estimated by some other method, and their determinants can be shown (or assumed) to be equal, the only remaining information needed is the set of estimates $\underline{\bar{x}}_k$ of the population means $\underline{u}_k$. The optimality of the discriminant procedure then depends only upon the precision of estimation of the $\underline{u}_k$ and $\underline{W}_k$. The actual error of misclassification that will be achieved with fresh observation vectors $\underline{x}$ can be estimated by a method due to P. A. Lachenbruch (M. Hills, 1966). This consists in treating each individual from each of the K samples as an "unknown" and classifying it by using the remainder of the data to compute the $G_k(\underline{x})$.

LINEAR VS. QUADRATIC DISCRIMINANT FUNCTIONS

It is often assumed that the simplest method of setting up discriminant functions is to take the K populations two at a time and to compute a linear discriminant function for each of the 1/2K(K - 1) pairs of populations of the form

$$G_{h,k} = \underline{x}'\underline{c} + c_0$$

This yields a set of 1/2K(K - 1) hyperplanes having equations of the form $G_{h,k} = 0$ which partition the observation space into K distinct regions, one for each population.

However, the identification of the vector $\underline{x}$ involves computing 1/2K(K - 1) linear discriminant function values, which requires a total of 1/2pK(K - 1) multiplications. On the other hand, to compute K values of the quadratic discriminant function requires no more than 1/2pK(p + 3) multiplications (write $\underline{x}'\underline{W}^{-1}\underline{x}$ as $\underline{x}'\underline{T}'\underline{T}x = \underline{z}'\underline{z}$ where $\underline{T}$ is triangular. See for example Rao, 1952, 1965), so that the labor of computing the linear and quadratic discriminant function values for a given $\underline{x}$ is the same if K = p + 4. For larger K, the quadratic discriminant requires less arithmetic.

SELECTING A VARIATE TRANSFORMATION

Cooper's result, showing that the quadratic discriminant function is optimal for a very wide class of distributions provided that the determinants of the dispersion matrices are equal, makes it natural to consider whether it is possible to select a variate transformation which will have the effect of bringing a set of data having unequal determinants as near as possible to the desired condition.

We consider the variate transformation $y = x^{\theta}$. This can be generalized as follows:

$$y = [(x+\theta_2)^{\theta_1} - 1]/\theta_1 \qquad \text{if } \theta_1 \neq 0$$

$$y = \ln (x+\theta_2) \qquad \text{if } \theta_1 \to 0 \qquad (4)$$

θ_1 can take any value from +1 to -1.

(Values outside this range are permissible but are seldom needed: Tukey, 1957.) The value of θ_2 must be such that $(x + \theta_2)$ is always greater than zero, but experience suggests that the value chosen is not critical (Box and Cox, 1964). If there is no likelihood of zero or negative variate values in the data, it will usually be sufficient to take $\theta_2 = 0$. We shall assume that the values of $(u_{ik} + \theta_2)$ are in all cases large in comparison with their standard deviations, so that we have

$$\text{var } y_i = \left(\frac{dy_i}{du_i}\right)^2 \text{var } x_i \text{ approximately}$$
$$= (u_i + \theta_2)^{2(\theta_1 - 1)} \text{ var } x_i$$

$$\text{cov}(y_i, y_j) = \frac{dy_i}{du_i}\frac{dy_j}{du_j} \text{ cov } (x_i, x_j) \text{ approximately}$$
$$= (u_i + \theta_2)^{\theta_1 - 1}(u_j + \theta_2)^{\theta_1 - 1} \text{ cov } (x_i, x_j)$$

whence it follows that, to the first-order approximation, the matrix of correlation coefficients remains unchanged under transformation.

Let $\underline{Z}_k$ be the covariance matrix of the transformed data $\underline{y}$ for the k^{th} population. Then we have

$$|\underline{Z}_k| = |\underline{\Sigma}_k| \prod_{i=1}^{p} (u_{ik} + \theta_2)^{2(\theta_1 - 1)} \qquad (5)$$

Bearing in mind that the ratios of the determinants, rather than their absolute values, are what matters (the variance ratio being the univariate analogue), we adopt a logarithmic least-squares procedure to minimize the dispersion of the K values of log $|\underline{Z}_k|$.

From (5) we get

$$\log |\underline{Z}_k| = \log |\underline{\Sigma}_k| + 2(\theta_1 - 1) \log [\prod_{i=1}^{p} (u_{ik} + \theta_2)]$$

Write $d_k = 1/2 \log |\underline{\Sigma}_k|$

and $m_k = \log [\prod_{i=1}^{p} (u_{ik} + \theta_2)]$

and we require to minimize the sum of squares

$$\sum_{k=1}^{K} [d_k + (\theta_1 - 1)m_k - \bar{d} - (\theta_1 - 1)\bar{m}]^2$$

by differentiating with respect to θ_1 and equating to zero. We obtain

$$\theta_1 = \frac{\sum_{k=1}^{K} m_k(m_k - d_k) - K\bar{m}(\bar{m} - \bar{d})}{\sum_{k=1}^{K} m_k^2 - K\bar{m}^2}$$

which is simply the slope of the regression of d_k upon m_k, subtracted from unity:

$$\theta_1 = 1 - \frac{\text{cov }(m,d)}{\text{var } m} \qquad (6)$$

To apply this result, the first step is to check whether values of θ_2 different from zero are likely to be needed. Provided that the data does not contain any negative or very small values, it should be safe to take $\theta_2 = 0$ throughout. We next compute the means and covariance matrices of the untransformed data, and compute θ_1 from equation (6). If the value found for θ_1 does not lie somewhere between -1 and +1, there is likely to be something wrong; either a computational error or data which are genuinely heterogenous with respect to within-sample variance. If the value of θ_1 seems reasonable, we may put $y = x^{\theta_1}$, or $y = \ln x$ if θ_1 is almost zero, and compute the means and covariance matrices of the transformed data. As a check, we can now re-compute θ_1 using the transformed means and covariance matrices: its value should now of course be approximately 1.0.

If there are only two populations, it is possible to find a value of θ_1 which will bring the determinants of the two covariance matrices to exact equality: it is (to the first order)

$$\theta_1 = 1 - \frac{d_2 - d_1}{m_1 - m_2} \qquad (7)$$

A NUMERICAL EXAMPLE

As an illustration, we shall determine the appropriate value of θ_1 for some unpublished data relating to seven samples of measurements of fossil shells of Upper Carboniferous fresh-water bivalves. Each sample is from a single locality and horizon and is believed to consist of conspecific individuals: the five species represented are members of two closely related genera (Carbonicola and Anthracosia). Six shell measurements were made upon each specimen, all linear dimensions in mm.

In Table 1 are shown, in successive columns, (1) the number of specimens per sample, (2) the product of the mean values for the six measurements, (3) the determinant of the dispersion matrix, (4) the logarithm of the product of means, and (5) half the logarithm of the determinant of the dispersion matrix. Columns (4) and (5) thus give the values of m_k and d_k . We obtain

$$\frac{\text{cov}\,(m,d)}{\text{var } m} = \frac{8.4841}{9.0937} = 0.9330$$

$$\theta_1 = 1.0 - 0.9330 = 0.0670$$

Since this value is quite close to zero, it seems reasonable to choose the transformation y = ln x.

Table 2 shows the same quantities as Table 1, but computed from the natural logarithms of the data values instead of from the untransformed values. We have

$$\frac{\text{cov}\,(m,d)}{\text{var } m} = \frac{-0.3095}{2.9893} = -0.1035$$

$$\theta_1 = 1.0 + 0.1035 = 1.1035$$

showing that the logarithmic transformation is overcorrecting the heterogeneity of dispersion of the raw data, although only very slightly. The ratio of the largest to the smallest determinant has been reduced from about 2.0×10^4 to around 4.0×10^1, which corresponds to a univariate variance ratio of roughly 1.8, as compared with about 12.2 for the untransformed data.

In Table 3, the means, standard deviations, and correlation matrix for the first of the seven samples (Anthracosia concinna) are shown computed from the untransformed and from the transformed data. Although the untransformed standard deviations are rather large in proportion to their means for this sample, the correlation matrix is little changed on transforming the data.

DEGREES OF FREEDOM LOST THROUGH TRANSFORMING DATA

In selecting a transformation, we are treating θ_1 as an unknown parameter whose true value may not exactly coincide with the value estimated from the data. Box and Cox (1964) examine the question of what reduction in the degrees of freedom of the covariance matrices should be made in order to allow for this fact. They conclude that one degree of freedom should be deducted from the total for the error variance for each θ parameter fitted. Although their method of determining θ is not the same as the one described here, there will probably be no harm in adopting a similar rule. In the case of the example considered in the previous section, the loss of one degree of freedom deducted from the overall total for the seven samples is small enough to ignore.

SKEWNESS

We have assumed throughout, as does Cooper, that the parent distributions are unskewed. However, the use of a transformation will not leave the symmetry of the distributions unaffected, and it would be very useful to be able to test whether the same transformation that serves to minimize heterogeneity of variance is also the one best calculated to minimize skewness. Box and Cox (1964) have investigated this point in some detail, mainly with reference to the univariate case, although the methods appear to be capable of extension to at least the simpler multivariate cases. It is interesting to note that in practice, it often turns out that the same transfromation does indeed minimize both skewness and heterogeneity of variance.

However, one might expect that in general, a multivariate procedure would be more robust with respect to skewness than the corresponding univariate procedure. If we consider skewness with respect to a single odd-order cumulant only (say the 3rd) then it will be possible to select an orthonormal transformation which will remove skewness from all the marginal distributions except one. (It is of course possible to suggest pathological distributions, having probability density contours which are non-ellipsoidal, for which this would not be true.) Thus, skewness of this type can affect only one of the p dimensions of a multivariate distribution.

Among the commoner causes of skewness in practice are the presence of rogue observations and heterogeneity of the populations sampled. It is consequently unwise merely to ignore it. I have elsewhere (Burnaby, 1966) described a projection-matrix technique for dealing with heterogeneous multivariate data.

NONMETRIC METHODS

There appears to be no reason why quadratic

Table 1. - Upper Carboniferous bivalves: untransformed data.

	(1)	(2)	(3)	(4)	(5)
	n	$\pi(u_{jk})$	$\lvert \underline{S}_k \rvert$	m_k	d_k
A. concinna	21	$6.8629_{10}3$	$2.2063_{10}-3$	3.8364	-1.3282
C. fallax	21	$5.3579_{10}4$	$3.3315_{10}-3$	4.7289	-1.2387
Anthracosia sp.	30	$1.3954_{10}6$	$3.2525_{10}1$	6.1447	0.7561
C. cristagalli 4	18	$2.1612_{10}6$	$4.4907_{10}1$	6.3347	0.8262
C. cristagalli 2	11	$2.5765_{10}6$	$1.3010_{10}1$	6.4110	0.5571
C. cristagalli 3	27	$2.6650_{10}6$	$2.0971_{10}1$	6.4257	0.6608
C. pseudorobusta	35	$3.4427_{10}7$	$7.1155_{10}3$	7.5369	1.9261

Explanation of Table 1. n = number of specimens per sample; $\pi(u_{jk})$ = product of mean values for six shell measurements (for details see Table 3); $\lvert \underline{S}_k \rvert$ = sample estimate of 6 x 6 covariance matrix determinant; m_k = common log of entry in column (2); d_k = half the common log of entry in column (3).

Table 2. - Upper Carboniferous bivalves: data transformed $y = \ln x$.

	(1)	(2)	(3)	(4)	(5)
	n	$\pi(u_{jk})$	$\lvert \underline{S}_k \rvert$	m_k	d_k
A. concinna	21	3.9224	$4.7203_{10}-11$	0.5936	-5.16302
C. fallax	21	19.6553	$1.2630_{10}-12$	1.2935	-5.94930
Anthracosia sp.	30	122.835	$2.7109_{10}-11$	2.1084	-5.28344
C. cristagalli 4	18	152.906	$1.0558_{10}-11$	2.1844	-5.48821
C. cristagalli 2	11	159.263	$3.0944_{10}-12$	2.2021	-5.75471
C. cristagalli 3	27	171.060	$3.4848_{10}-12$	2.2331	-5.72891
C. pseudorobusta	35	464.377	$8.0594_{10}-12$	2.6669	-5.54685

Explanation of Table 2. As for Table 1, except that all means and covariances have been computed from the data after transformation to logarithms.

Table 3.- Measurements of 21 specimens, untransformed and transformed data.

Anthracosia concinna: sample of 21 specimens, untransformed data.

Variate	L	H	C_P	D_P	C_A	A
Means (mm)	17.29	7.907	1.940	3.062	3.095	2.730
St. devs.	5.205	2.471	0.4491	0.8909	1.451	0.8935
Correl.	1.0000	0.9819	0.6215	0.8966	0.9448	0.7789
matrix		1.0000	0.6465	0.9296	0.9638	0.7738
			1.0000	0.6303	0.6167	0.4258
				1.0000	0.8633	0.7221
					1.0000	0.6706
						1.0000

The same sample: data transformed $y = \ln x$

Variate	L	H	C_P	D_P	C_A	A
Means	2.811	2.027	0.638	1.083	1.043	0.956
St. devs.	0.2814	0.2847	0.2283	0.2711	0.4087	0.3207
Correl.	1.0000	0.9774	0.6448	0.8941	0.9210	0.7955
matrix		1.0000	0.6712	0.9205	0.9423	0.7687
			1.0000	0.6922	0.5880	0.5034
				1.0000	0.8206	0.6830
					1.0000	0.6746
						1.0000

Explanation of Table 3 The six shell measurements are:-

L Max overall length, parallel to long axis of profile.

H Max overall height, perpendicular to L.

C_P Radius of posterior extremity of shell (90° arc).

D_P Humping of postero-dorsal shell margin.

C_A Radius of anterior shell extremity (90° arc).

A Distance of umbo from anterior extremity, parallel to L.

discriminant functions should not be constructed using the ranks of the observed variate values transformed to equivalent normal scores. It would be necessary to re-score the vector $\underline{x}$ to be identified for each of the K populations: this would make the procedure somewhat laborious. However, by assigning nonintegral ranks to the elements of $\underline{x}$, the necessity for re-computing the dispersion matrix for each new $\underline{x}$ for each sample could be eliminated. I have not attempted any trials with actual data.

CONCLUSIONS

Until recently, quadratic discriminant functions have been rather neglected in favor of the linear discriminant, probably on account of the latter's apparent simplicity. As we have seen, however, the quadratic discriminant does not necessarily entail more arithmetic, if the number of parent populations is not small. Cooper's studies, demonstrating the use of the quadratic discriminant in a wide variety of situations involving nonnormal distributions, point to the desirability of adopting the quadratic form as the basic concept and of regarding the linear discriminant merely as a convenient computational device for use in special circumstances.

The distance interpretation of the quadratic form is of special importance in evolutionary studies in paleontology, as pointed out by Lerman (1965a, b). The amount of within-population variance is the natural unit of measure for comparing small morphological changes affecting the mean values of p different characters in a group of fossil organisms. The rate of change of the generalized distance per unit of geological time should help to throw light on the evolutionary process: for example, the question of whether the speed of evolutionary advance in complexity of biological organization is limited by the nature of the genetic mechanism, or by factors external to the organism itself.

The effective implementation of the generalized distance depends to a large extent upon the possibility of establishing an effectively constant metric for data exhibiting a very marked heterogeneity of variance. The present contribution illustrates a relatively simple method of approaching this problem. More sophisticated techniques, such as those studied by Box and Cox (1964), are now becoming available, although at present these require very much more computational work, and have not so far been tested in multivariate problems.

REFERENCES

Anderson, T.W., 1958, An introduction to multivariate statistical analysis, Wiley and Sons, New York.

Box, G.E.P., and Cox, D.R., 1964, An analysis of transformations: Jour. Roy. Statist. Soc. B, v. 26, p. 211-252.

Burnaby, T.P., 1966, Growth-invariant discriminant functions and generalized distances: Biometrics, v. 22, p. 96-110.

Cooper, P.W., 1963, Statistical classification with quadratic forms: Biometrika, v. 50, p. 439-448.

Cooper, P.W., 1965, Quadratic discriminant functions in pattern recognition: IEEE Trans. on Information Theory, v. IT 11, p. 313-315.

Hills, M., 1966, Allocation rules and their error rates: Jour. Roy. Statist. Soc. B, v. 28, p. 1-31.

Lerman, A., 1965a, On rates of evolution of unit characters and character complexes: Evolution, v. 19, p. 16-25.

Lerman, A., 1965b, Evolution of *Exogyra* in the late Cretaceous of the southeastern United States: Jour. Paleontology, v. 39, p. 414-435.

Mahalanobis, P.C., 1949, A historical note on D^2 statistic: Sankhyā, v. 9, p. 237-240.

Rao, C.R., 1952, Advanced statistical methods in biometric research: Wiley and Sons, New York.

Rao, C.R., 1965, Linear statistical inference and its applications: Wiley and Sons, New York.

Tukey, J.W., 1957, On the comparative anatomy of transformations: Ann. Math. Statist., v. 28, p. 602-632.

APPLICATION OF PATTERN RECOGNITION TECHNIQUES TO GEOSCIENCE INTERPRETATION

By

Gerry L. Kelly
University of Kansas

ABSTRACT

Everything in nature has certain spectral characteristics which can be used to distinguish one entity from another or to obtain information about shape, size, or other physical properties if proper sensing devices are utilized for measurement. The pattern recognition process is the examination of the data obtained in measurements of the environment to determine invariant patterns. The pattern recognition process basically groups together similar events from the organized environment. The criteria used to determine the similarity of various patterns within a particular pattern group will in general depend upon the interests of the investigator.

Geoscientists usually are confronted with vast amounts of data to be analyzed. It becomes exceedingly difficult to determine the salient features of the data unless some automatic method is utilized. Pattern recognition techniques can be particularly helpful to geoscientists who have several different categories to distinguish. A particular pattern recognition method may be implemented with actual "hardware" or simulated on a digital computer.

One type of pattern recognition process is when the statistics of sample patterns are known in advance. The sample patterns, sometimes called the training set, are used to develop discriminant functions which can be used to classify patterns not in the original training set. Another type of pattern recognition process is when no a priori knowledge is known about the categories being considered. In this case measurement space is partitioned into similarity sets on the basis of conditional and marginal probability density functions obtained from the data.

CLASSIFICATION OF CONTOUR MAPS

by

Owen T. Spitz

and

Daniel F. Merriam

Kansas Geological Survey

ABSTRACT

Comparison of contour maps and their classification have been of interest for some time. Until recently, however, most of the work has been of a qualitative nature, and only with advent of the computer have more rigorous and sophisticated methods been available.

Considerable effort now is being directed toward pattern discernment and representation. Geologists, long accustomed to reading maps, are especially involved with map interpretation. This use has been based mainly on visual inspection and for the most part has been highly subjective.

Trend analysis is a quantitative method of evaluating spatial data and essentially allows a large-scale component to be separated from remaining small-scale components. Thus, a complex situation can be broken down into simple components, which are then analyzed. For purposes of comparison or classification, the (1) raw data, (2) trend component, or (3) residual information have been used with varying degrees of success.

The comparison and classification of maps entail using numerical descriptors. Mirchink and Bukhartsev (1959) compared the distribution of features on one map to another by examining the absolute depth (the raw data) of the two surfaces at many localities. Coefficients of power-series polynomials of well-fitted third-degree surfaces were used by Merriam and Sneath (1966) to group structural maps. A comparison of residual values from trend analyses of structural data were made by Merriam and Lippert (1966). Nothing, to our knowledge, has been published using the matrix of a trend analysis, although this technique was suggested by Miller (1964). Other possible numerical descriptors would include coefficients of orthogonal polynomials (Spitz, 1966) and Fourier coefficients. To simplify the problems of size, shape, and orientation for comparison, structural data on several horizons in different parts of Kansas were used. The availability of this data presented a unique opportunity for such a study.

REFERENCES

Merriam, D.F., and Lippert, R.H., 1966, Geologic model studies using trend-surface analysis: Jour. Geology, v. 74, no. 3, p. 344-357.

Merriam, D.F., and Sneath, P.H.A., 1966, Quantitative comparison of contour maps: Jour. Geophys. Res., v. 71, no. 4, p. 1105-1115.

Miller, R.L., 1964, Comparison-analysis of trend maps, in Computers in the Mineral Industries: Stanford Univ. Publ., Univ. Ser., Geol. Sci., v. 9, p. 669-685.

Mirchink, M.F., and Bukhartsev, V.P., 1959, The possibility of a statistical study of structural correlations: Dokl. Earth Sci. Sec., English Transl., v. 126, no. 5, p. 1062-1065.

Spitz, O.T., 1966, Generation of orthogonal polynomials for trend surfacing with a digital computer, in Computers and Operation Research in Mineral Industries: Pennsylvania State Univ., 6th Ann. Symposium, p. NN-1-NN-6.

COMPUTER CONTRIBUTIONS

Kansas Geological Survey
University of Kansas
Lawrence, Kansas

Computer Contribution

1.	Mathematical simulation of marine sedimentation with IBM 7090/7094 computers, by J. W. Harbaugh, 1966	$1.00
2.	A generalized two-dimensional regression procedure, by J.R. Dempsey, 1966	$0.50
3.	FORTRAN IV and MAP program for computation and plotting of trend surfaces for degrees 1 through 6, by Mont O'Leary, R. H. Lippert, and O.T. Spitz, 1966	$0.75
4.	FORTRAN II program for multivariate discriminant analysis using an IBM 1620 computer, by J. C. Davis and R. J. Sampson, 1966.	$0.50
5.	FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data, by W. J. James, 1966	$0.75
6.	FORTRAN IV program for estimation of cladistic relationships using the IBM 7040, by R. L. Bartcher, 1966	$1.00
7.	Computer Applications in the Earth Sciences: Colloquium on classificational procedures, D. F. Merriam, editor, 1966	$1.00

EDITORIAL STAFF