CONF-791016--6

MASTER

APPLICATION OF DISCRIMINANT ANALYSIS AND GENERALIZED DISTANCE MEASURES TO URANIUM EXPLORATION

J. J. Beauchamp
C. L. Begovich
V. E. Kane
D. A. Wolf

October 1979

Computer Sciences Division
Union Carbide Corporation
Nuclear Division*
Oak Ridge, Tennessee 37830

*UCC-ND is operated for the Department of Energy under contract W-7405-eng-26

ABSTRACT

The National Uranium Resource Evaluation (NURE) Project has as its goal estimation of the nation's uranium resources. It is possible to use discriminant analysis methods on hydrogeochemical data collected in the NURE Program to aid in formulating geochemical models which can be used to identify the anomalous regions necessary for resource estimation. Discriminant analysis methods have been applied to data from the Plainview, Texas Quadrangle which has approximately 850 groundwater samples with more than 40 quantitative measurements per sample. Discriminant analysis topics involving estimation of misclassification probabilities, variable selection, and robust discrimination are applied. A method using generalized distance measures is given which enables assigning samples to a background population or a mineralized population whose parameters were estimated from separate studies. Each topic is related to its relevance in identifying areas of possible interest to uranium exploration.

Keywords: discriminant analysis, variable selection, uranium favorability, generalized distance measures, regional variables.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES, Continued

INTRODUCTION

Multivariate statistical methods provide a natural framework for studying the interrelationships of geochemical parameters considered in mineral exploration. Typically, samples of some media are collected over wide geographic areas and analyzed for numerous geochemical parameters. The interpretative phase consists of separating typical background samples from anomalous samples which are possibly associated with mineralization. The following presentation has as its objective the use of discriminant analysis methodology to: (1) identify the geochemical parameters which may be important in formulating regional geochemical models, (2) validate the geologic origins of the samples, and (3) identify possible mineralization related samples based on either the background population or known mineralized populations. Groundwater data from the Plainview NTMS Quadrangle, collected as part of the National Uraninum Resource Evaluation (NURE) Program, are used for illustration.

Prior to application of discriminant analysis methods it is important to consider preprocessing the data. Treatment of censored laboratory data and evaluation of distributional considerations of the variables are important aspects of preprocessing. Additionally, it is assumed that the samples are preliminarily assigned to a geologic unit representing the geologic origin of the sample. This assignment will be assessed by discriminant analysis as part of the preprocessing.

DESCRIPTION OF STUDY AREA

The Plainview NTMS Quadrangle in Texas was selected from the NURE program because it had both published hydrogeochemical data and a published evaluation of potential uranium mineralization using geologic, radiometric, drilling, and hydrogeochemical data (Amaral, 1979). The Plainview Quadrangle is an area of approximately 20,350 km^2 (7,860 mi^2) located in the Great Plains Province between lat. 34° and 35° N and long. 100° and 102° W. It is divided into Rolling Plains to the east and the Llano Estacado of the southern High Plains to the west by the generally north-south trending Caprock Escarpment.

Although the subsidence of the Palo Duro Basin exerted an influence on depositional systems throughout the end of the Permian, units exposed at the surface are relatively flat lying, creating a relatively simple geology. The San Andres (Blaine) Formation (coded as PGEB), Whitehorse Group (coded as PGWC), and Quartermaster Group (coded as POQ) dip generally 25 ft/mi to the west. The Dockum Group and Ogallala Formation (coded as TPO) overlie the Permian section at an uncomformable surface and dip generally to the southeast at approximately 10 ft/mi. These are relatively shallow dips and the lack of major faulting result in relatively predictable geology in the subsurface with the surficial outcrop pattern complicated by the relatively deep erosion in the Palo Duro Canyon along with the Caprock Escarpment.

The uniform dips of the bedrock units result in relatively simple groundwater flow patterns for those aquifers that are penetrated by domestic water wells. The water table surface in the Permian section east of the Caprock Escarpment follows the general topographic slope. West of the Caprock Escarpment, groundwater in the Permian units may flow along regional dip which is approximately 25 ft/mi to the west. Groundwater flow in the Dockum Group appears to be to the southeast with a regional dip of the water table of approximately 10 ft/mi. The Ogallala Formation is the major aquifer in the western half of the quadrangle and groundwater flow appears to be relatively consistent with a regional dip of the water table to the southeast at approximately 10 ft/mi. It is recognized that although regional topography and dip are relatively consistent, there may be many variations in groundwater flow direction resulting from changes in permeability, local structures, and overpumping of the aquifer.

DESCRIPTION OF DATA AND PREPROCESSING

Well and spring water samples were collected by field personnel and shipped to Oak Ridge, Tennessee, where chemical analyses consisting of about 40 measurements were performed. A field form with approximately 30 items of information was completed for each sample in the field. The information on the form includes the assignment of a geologic producing horizon of the groundwater, and the measurement of the total alkalinity, pH, and conductivity (converted to specific conductance) of the water.

Chemical analysis procedures included fluorometry for uranium, atomic absorption for arsenic, spectrophotometry for sulfate, and plasma source emission spectrometery for barium, boron, calcium, lithium, magnesium, molybdenum, sodium, vanadium, and zinc. Complete details of field and laboratory procedures are described by Arendt (1978) and appear in the "Reports Procedures Manual for Groundwater Reconnaissance Sampling" (Uranium Resource Evaluation Project, 1978). Basic data analyses and displays of both the groundwater and stream sediment data appear in the report "Hydrogeochemical and Stream Sediment Reconnaissance Basic Data for Plainview NTMS Quadrangle, Texas" (Uranium Resource Evaluation Project, August 1978). (A computer tape of all data can be obtained from Dalton Atkins, GJOIS Project, UCC-ND Computer Applications Dept., 4500N Building, ORNL P. O. Box X, Oak Ridge, Tennessee 37830.)

Since the discriminant analysis methods to be used in subsequent analyses require noncensored (above the laboratory detection limit) data whose distribution may be reasonably approximated by a multivariate normal, the preprocessing portion of the data analysis is an initial screening method to reduce the number of variables under consideration and to evaluate the statistical distribution of the selected variables. The following steps are involved: (1) check and note the samples within each geologic unit for missing data; (2) plot histogram and probability plots (these plots may be used as an initial screen for obvious non-normality); (3) determine the proportion of noncensored data; (4) test for normality for each of those variables that are not affected by missing data or a large proportion of censored observations.

An examination of the data for the original set of observed variables for each of the five geologic units of interest revealed that problems of missing data or censored observations were non-existent or minimal for the 12 variables: uranium, boron, barium, calcium, lithium, magnesium, sodium, zinc, sulfate (SO_4), specific conductance (SP), total alkalinity (TAK), and pH. These were the variables used in the distribution evaluation portion of the preprocessing.

Summary statistics for the 12 variables, as well as three additional uranium-related variables (arsenic, vanadium, and molybdenum) to be used later, were examined for each geologic unit. Robust measures, such as the median, were examined to evaluate the influence of any censored data. After the deletion of observations with missing values, the sample sizes for the different geologic groups were: 345 for TPO, 118 for PGEB, 267 for PGWC, and 73 for POQ. Groundwater samples from the Dockum Group were not considered because of the small sample size (16).

Figure 1 is a histrogram of the calcium variable of the Quartermaster (POQ) Group. An examination of this figure reveals an apparent bimodal distribution. The observations associated with each mode were clustered together in distinct geographic regions. Therefore, the POQ samples were partitioned into two separate subgroups, denoted by POQE (39 samples) and POQW (34 samples) for subsequent analyses.

The next phase of the data preprocessing was to determine if the distribution of the observed variables could be reasonably approximated by the normal or lognormal distribution. A combination of different techniques was used to evaluate the adequacy of the log-transformation: probability plots (Sinclair, 1976), histrograms, sample skewness and kurtosis measures (Snedecor and Cochran, 1967), Shapiro-Wilk test statistics (Shapiro and Wilk, 1965), and a modified version of the Kolmogorov-Smirnov D-statistic (Stephens, 1974). The tests of normality for all five geologic groups indicated that the logarithmic transformation was appropriate to achieve marginal normality for each of the observed variables, except calcium, within each of the groups. The pH variable was not transformed since it already represents the log of a concentration and was approximately normally distributed. Therefore, in all subsequent analyses all variables, except calcium and pH, were transformed using the log transformation.

DISCRIMINANT ANALYSIS

Discriminant analysis provides a criterion for classifying a collection of observation vectors into one of a specified number of groups (Agterberg, 1974). In this section we want to determine whether or not a subset of the observed variables could be used to adequately discriminate between the five geologic groups (TPO, PGWC, PGEB, POQE, and POQW). The chemical concentrations of individual samples will be used to determine if the prior assignment of samples to one of the 5 groups is tenable. Additionally, the chemical parameters which most accentuate the differences in the five units will be determined, enabling formulation of possible geochemical models to characterize the regional geochemistry.

The discriminant scores used to classify an observation vector are known linear (equal population covariances) or quadratic (unequal population covariances) functions of the observed variables if the observation vectors follow a multivariate normal distribution. The observation vectors from each geologic unit were used to estimate the mean vector and covariance matrix of the assumed multivariate normal distribution. The estimated linear or quadratic discriminant score was used to classify each observation depending upon the results of a test of the homogeneity of the within-group covariance matrices (Kendall and Stuart, 1961).

Different methods are available for reducing the number of variables used in discriminant analysis and are similar to the variable selection procedures of regression analysis. Criteria based upon a measure of the differences between groups or upon minimizing the probability of misclassification are intuitively appealing and relatively easy to apply. The Wilks Λ-statistic (Rao, 1965) defined as the ratio of the determinant of the within sum of cross product matrix (W) to the total sum of cross products matrix (T)is a common measure used to evaluate differences between groups. The estimated probability of misclassification is the measure used to evaluate the performance of the discriminant function and is estimated by

$$\hat{Pr}(e) = \sum_{i=1}^{5} \pi_i \sum_{\substack{j=1 \\ j\neq i}}^{5} \hat{P}(j|i)$$

where π_i is the probability an observation vector comes from group i (for this example $\pi_1 = \ldots = \pi_5 = 1/5$) and $\hat{Pr}(j|i)$ is the observed proportion of vectors known to come from group i that are incorrectly classified in group j by the sample discriminant function. Variations in the procedures arise when equal covariance matrices are not assumed for all groups. The applied procedures are described as follows:

1. DISCRIM (McCabe, 1975) is a procedure which chooses the subset of variables based on the minimum value of the Wilks Λ-statistic, $\Lambda = |W|/|T|$, where $|A|$ is the determinant of the matrix A. This selection procedure assumes equal covariance matrices for the different geological groups.

2. BMDP7M (Dixon and Brown, 1977) is a stepwise discriminant analysis procedure which uses the F-statistic as a default to select the best subset of variables. This procedure also assumes equal covariance matrices.

3. Modified DISCRIM (McCabe, 1979) is a procedure which allows for unequal covariance matrices and makes use of the modified Wilks Λ statistic $\Pi_{i=1}^{5} |W_i|/|T_i|$ where $|W_i|$ and $|T_i|$ are the determinants

of the individual within and total sums of cross products matrices for each group.

4. Forward selection procedure first considers each variable separately, tests for equality of the covariance matrices over the different groups, calculates the discriminant function, and estimates the probability of misclassification by reclassifying the original data using the estimated discriminant function. The variable with the smallest overall probability of misclassification is chosen as the best single discriminatory variable. The remaining variables are individually considered with the best single variable and the above process is repeated to produce the best pair of discriminating variables in terms of the estimated probability of misclassification. The above process is repeated until all of the variables are included in the discriminant function. This procedure has the property that once a variable has been included in the discrimant function it will always be included in subsequent stages.

5. Backward selection procedure is similar to the forward selection procedure except that it starts with the complete set of variables and drops each variable separately at the first stage. The smallest estimated probability of miscalssification determines the first variable to be deleted. The procedure is repeated with the reduced set of variables to determine the second variable to be deleted. Once a variable is deleted it is excluded from further

consideration. The SAS procedure DISCRIM (Barr, et al, 1976) was used to do the necessary calculations for the forward and backward selection procedures.

Figure 2 summarizes the results of these preliminary analyses and displays the estimated criteria values as a function of the number of variables in the model. An examination of this Figure indicates that there is more than one feasible subset of variables based upon the Wilks Λ, the modified Wilks Λ statistic, or the estimated probability of misclassification. Figure 2(a) is a plot of the estimated probability of misclassification for the different variable selection procedures considered. Figure 2(b) is a plot of Wilks Λ and modified Wilks Λ from DISCRIM. The major reduction in these criteria occurs as the number of variables in the model increases from 1 to 3. A plot of the change in the estimated probability of misclassification going from a p to p + 1 (p = 1, 2,..., 11) variable model is shown in Figure 3 for the modified DISCRIM procedure. Small values of this change would indicate possible stopping points for the number of variables to be included in the model. Low values of the change in misclassification probability for p of 4, 7, and 10 correspond to three possible candidate models. An examination in this same figure of the corresponding changes in the modified Wilks Λ statistic shows the major reduction in this statistic occurs at p = 7. Therefore, the seven variable model seems appropriate since both the probability of misclassification and the group separation show only small changes for p > 7. The seven variables included in this model are: ln(U), ln(SP), ln(B), CA, ln(LI), ln(MG), and $\ln(SO_4)$; these

Figure 2

Figure 3

variables will be denoted as the regional variables. Using the standard default options the BMDP stepwise procedure selected a ten variable model as its final choice. However, this procedure did not allow for unequal covariance matrices and resulted in an estimated probability of misclassification greater than that for the seven variable model using the modified DISCRIM procedure.

For most reconnaissance geochemical data, the differences in geology will cause varying means and covariances between the geologic populations. The different covariance matrices, as observed in the Plainview data, necessitate use of variable selection procedures based on quadratic discrimination. Therefore, the modified DISCRIM, forward, or backward procedures were found to have an advantage over the unmodified DISCRIM or stepwise BMDP7M procedures which assume equal covariance matrices.

It is possible that additional geologic considerations might motivate the choice of a different set of variables. Table 1 displays some of the alternative model choices for the different variable selection criteria. Many of these alternative sets of variables have values of the optimization criteria that differ only slightly from the minimum value. When the geochemistry of the region is considered, one of the alternative seven variable sets may provide a more parsimonions model for the data. Selection of a set of variables from Table 1 based on a geochemical model, would probably improve the analyses in the next section.

Table 1

ALTERNATIVE MODELS FROM DIFFERENT VARIABLE SELECTION METHODS

METHOD	Elements[a]												Function Value[b]
	U	SP	B	BA	CA	LI	MG	NA	SO_4	$_T$AK	ZN	PH	
DISCRIM	X	X			X	X	X		X	X			5.40
	X	X			X	X		X	X	X			5.43
	X	X		X	X	X			X	X			5.52
	X	X			X	X			X	X		X	5.53
	X	X	X		X	X			X	X			5.54
	X	X			X	X			X	X	X		5.61
Modified DISCRIM	X	X	X		X	X	X		X				1.09
	X		X	X	X	X	X		X				1.33
	X	X	X	X	X	X			X				1.35
	X	X	X		X	X			X	X			1.40
		X	X	X	X	X	X		X				1.40
	X		X		X	X	X	X	X				1.40
Forward	X	X			X	X	X	X		X			0.232
		X			X	X	X	X		X	X		0.235
		X			X	X	X	X		X		X	0.236
		X		X	X	X	X	X		X			0.246
		X	X		X	X	X	X		X			0.255
		X			X	X	X	X	X	X			0.258
Backward		X	X	X		X	X	X		X			0.204
		X	X	X		X		X		X		X	0.228
		X	X			X	X	X		X		X	0.245
		X	X	X		X	X			X		X	0.245
			X	X		X	X	X		X		X	0.248
		X	X	X		X	X	X				X	0.249

(a) All observed element values have been transformed, except CA and PH.

(b) Function values are Wilks $\Lambda(\times 10^{-2})$ for DISCRIM, modified Wilks $\Lambda(\times 10^{-8})$ for modified DISCRIM, and estimated probability of misclassification for forward and backward selection procedures.

Let $\underline{X}$ and $\underline{Y}$ be p-dimensional column vectors then

$$D^2\ (\underline{X}_{;}\underline{Y},\ C) \ = \ (\underline{X} - \underline{Y})' C^{-1} (\underline{X} - \underline{Y})$$

is a general functional form of the squared multidimensional distance from $\underline{X}$ to $\underline{Y}$ where C is a p x p positive definite matrix so the $D^2 > 0$. If $\underline{X} = \underline{\mu}$ and $\underline{Y} = \underline{\mu}_2$, where $\underline{\mu}$'s are the mean vector from y-dimensional multivariate normal distributions with C equal to the common covariance matrix Σ, then $D^2\ (\underline{\mu}_1;\ \underline{\mu}^2,\ \Sigma)$ is the Mahalanobis distance (Rao, 1965). The D^2 distance from group i to group j may be estimated by the sample value $D^2(\bar{\underline{x}}_i,\ \bar{\underline{x}}_j,\ S)$ for the case when the covariance matrices are assumed to be equal, where $\bar{\underline{x}}_i(\bar{\underline{x}}_j)$ is the sample mean vector for group i (j), and S is the pooled sample covariance matrix from all groups. When it is not reasonable to assume equality of the group covariance matrices, the D^2 distance from group j to group i, is estimated by $D^2(\bar{\underline{x}}_j,\ \bar{\underline{x}}_i,\ S_i)$ where S_i is the estimated covariance matrix from the observation vectors of group i.

Table 2 shows values of the generalized squared multidimensional distance matrix for three different seven-variable models. The D^2 distance between the different groups indicates a reasonable separation between TPO, POQW, and the remaining groups PGEB, PGWC, and POQE. Since the generalized squared distance was used in classifying the observations to the different geologic groups, it is not surprising that the estimated misclassification probabilities also revealed an overlapping of the three Permian units PGEB, PGWC, and POQE. This overlap suggest that the available geochemical data cannot distinguish samples in these

Table 2

GENERALIZED SQUARED DISTANCES FOR THREE
DIFFERENT SEVEN-VARIABLE MODELS

(A) Best seven-variable model from DISCRIM $[D^2 (\bar{\underline{x}}_i; \bar{\underline{x}}_j, S)]$

FROM/TO	TPO	POQW	POQE	PGEB	PGWC
TPO	0	14	42	41	40
POQW	14	0	14	16	14
POQE	42	14	0	3.4	0.85
PGEB	41	16	3.4	0	4.8
PGWC	40	14	0.85	4.8	0

(B) Best seven-variable model from modified DISCRIM $[D^2 (\bar{\underline{x}}_i; \bar{\underline{x}}_j, S_j)]$

FROM/TO	TPO	POQW	POQE	PGEB	PGWC
TPO	0	16	730	95	47
POQW	12	0	180	19	11
POQE	270	96	0	2.9	3.6
PGEB	340	160	1.9	0	3.0
PGWC	240	86	0.87	0.17	0

(C) Best seven-variable model from backward procedure $[D^2 (\bar{\underline{x}}_i; \bar{\underline{x}}_j, S_j)]$

FROM/TO	TPO	POQW	POQE	PGEB	PGWC
TPO	0	13	81	57	35
POQW	17	0	33	11	9.5
POQE	81	21	0	1.2	1.3
PGEB	63	23	2.6	0	4.6
PGWC	78	23	1.6	3.3	0

3 groups; combination of the groups simplifies the analysis and increases the samples sizes which improve estimation. This could be evidence for possibly considering these groups as a single population rather than three different populations. In fact, the overall estimated probability of misclassification when these three groups were combined was reduced from 24 to 6% using the seven regional variables. Hence we may conclude that the discriminant function using a reduced set of the original variables is appropriate for separating the three or five different geologic populations.

A prerequisite for the use of the previous discriminant analysis procedures is the preliminary assignment of samples to the five geologic unit groups. This assignment of samples can be verified by standard discriminant analysis methods. Considering only samples near the geologic contacts, three samples were reassigned (12131, 11896 from PGWC to TPO; 11886 from TPO to PGWC) in the Plainview data. These reclassifications will, of course, have negligible influence on the Plainview analyses. However, field classification of the geologic origin is often unavailable or more complex geology could increase the misclassification of samples.

A primary concern in any statistical analysis is the robustness, i.e. sensitivity to the underlying statistical assumptions, of the procedures that are employed. To evaluate the robustness of the linear and quadratic discrimination procudures used, the robust discriminant analysis methods described by Randles, et al, (1978a, b) were used on six pairwise discrimination problems from the five geologic groups using the seven regional variables. The 15 various possible versions of the robust discriminant analysis methods did not appreciably improve upon the standard linear and quadratic discrimination methods. The two standard methods only differed by a few misclassified samples from the best robust method, which typically made use of the Huber (1977) estimate of the covariance matrix.

INTERPOPULATION DISTANCE MEASURES

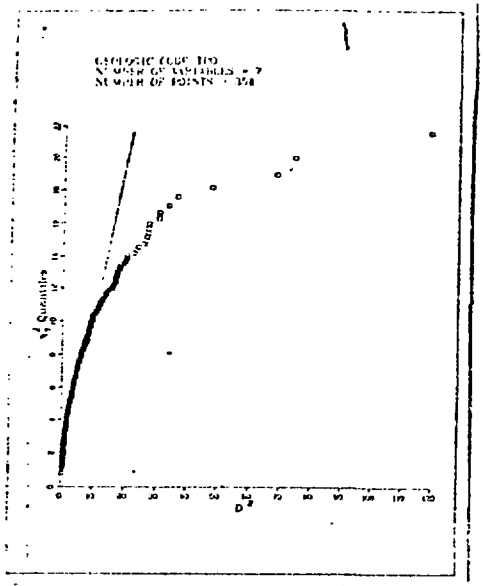
The distribution of $D^2(\underline{X};\ \underline{\mu},\ \Sigma)$ is χ^2_p (Anderson, 1958, Theorem 3.3.3) if $\underline{\mu}$ and Σ are known. It will be assumed that the large sample sizes enable the sample estimate $\underline{\bar{x}}$ (of $\underline{\mu}$) and S (of Σ) to be considered as the known quantities $\underline{\mu}$ and Σ. The presence of unusual geochemical samples will be determined by examining the fit of $D^2(\underline{X};\ \underline{\bar{x}},\ S)$ to a χ^2-distribution. Standard Q-Q plots (Gnanadesikan, 1977, pp 198-199) are used to evaluate the distributional fit and determine the D^2 threshold for unusual values. If the plotted values are on a line of slope one, only a single population is present. However, if more than a single line appears, several geochemical populations may be present. Alternatively, nonlinearity in the Q-Q plot could represent nonnormality of $\underline{X}$ or poor estimates of $\underline{\mu}$ and Σ. Samples with values of D^2 above the threshold will be geographically plotted. Geochemical subpopulations will be identified by a contiguous group of samples with unusual D^2 values.

Figure 4 shows four points which are equal distance from the mean $(\bar{x}_U, \bar{x}_{As})'$ for uranium and arsenic. In fact, any point on the ellipse has the same D^2 value. However, the simple Euclidean distance between the four points is quite variable. Figure 4 illustrates how the positive correlation between U and As would alter what might be considered an unusual sample. In general, because of the geochemical interrelationships in nature, it is meaningful to use the intervariable correlations to weight the observed deviations from the mean in defining anomalous samples. Unusual samples defined by this procedure in some cases may be relevant to uranium mineralization. However, it is important to note that samples with very low or moderate uranium values may have unusually high D^2 values (Figure 4). While these samples may be meaningful in detailed analyses, attention is primarily restricted here to samples with uranium values above the median.

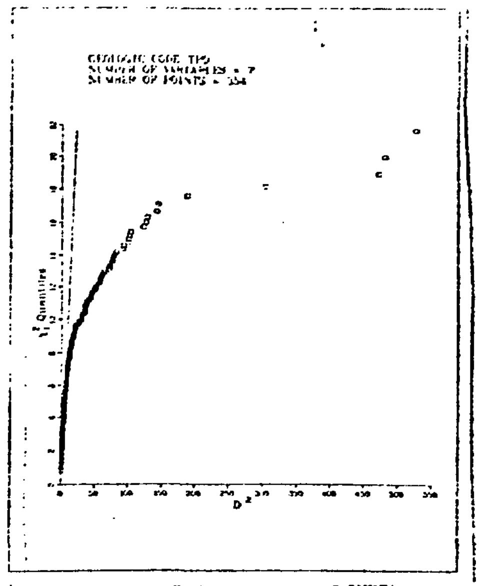
I. REGIONAL SUBPOPULATIONS

The selection of the seven regional variables that enable discrimination in the Ogallala Formation and Permian units suggests that these variables in some way characterize the regional geochemistry. Figure 5 (a) and (c) show the Q-Q plot for $D^2(\underline{x}_i; \bar{\underline{x}}, S)$ in the Ogallala and modified Permian units. The main body of the graph ($D^2 < 12$) in Figure 5(a) is reasonably linear, suggesting the expected χ^2 distributional fit is appropriate. However, 48 samples (14%) have values of D^2 above the threshold ($D^2 > 12$), suggesting some lack-of-fit in the tail of the chi-square distribution. Similarly, in Figure 5(c) 71 samples (16%) in

Figure 4

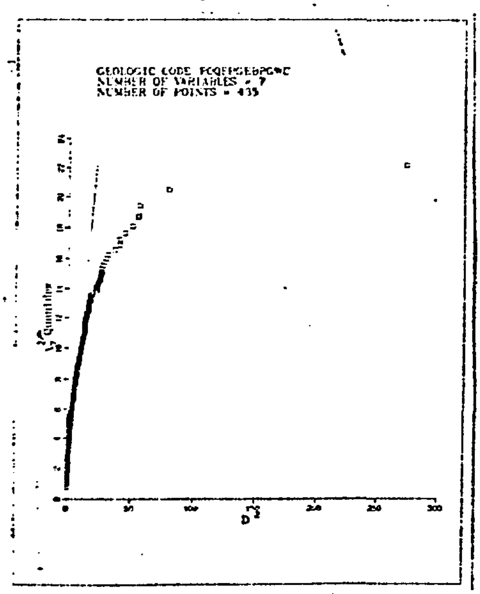


a. Standard Covariance Matrix

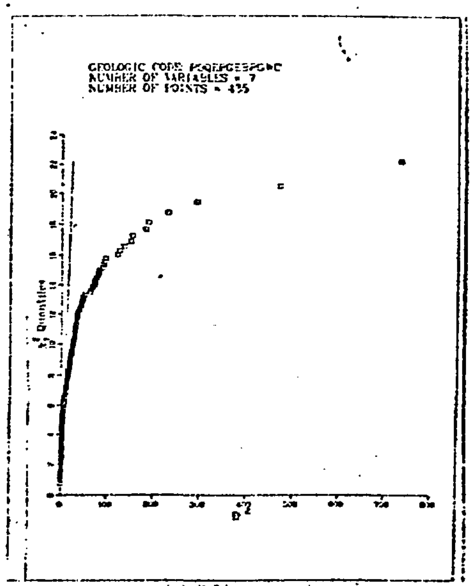


b. Robust Covariance Matrix

Ogallala Formation



c. Standard Covariance Matrix



d. Robust Covariance Matrix

Modified Permian Units

Figure 5

Q-Q PLOTS USING STANDARD AND ROBUST ESTIMATES
OF THE COVARIANCE MATRIX OF THE REGIONAL VARIABLES

the modified Permian Units have values of D^2 above the threshold ($D^2 > 12$), showing a similar lack-of-fit.

The lack-of-fit in the regional variables may be due to outlying samples causing poor estimates of $\underline{\mu}$ and Σ. Figure 5(b) shows that using a Huber (1977) robust estimator for $\underline{\mu}$ and Σ does improve the fit somewhat for the low D^2 values, but accentuates the lack of fit for the large D^2 values. Figure 5(d) shows the same general characteristic for the modified Permian samples. A possible interpretation of the accentuation of the nonlinearity is that the robust estimates minimize the influence of atypical samples in the estimates of $\underline{\mu}$ and Σ, and this makes the atypical samples have even more unusual D^2 values.

Figure 6 is a geographic plot of the samples having $D^2 > 12$ in the Ogallala Formation. Two areas I-A and I-B standout as being somewhat contiguous regions with unusual D^2 values. Region I-A is an area with very low concentrations in many elements. Region I-B consists of only seven samples, but these samples are very unusual in that the D^2 values are extremely large. Both regions were identified using an alternate set of variables (specific conductance, B, Ba, Li, Mg, Na, and total alkalinity) selected by another discriminant analysis variable selection method. Large subpopulations may influence estimates of the main population parameters that are used in the remaining analyses. Thus, samples in region I-A were deleted from the remaining Ogalalla analyses; the remaining geographic area is called the modified Ogallala.

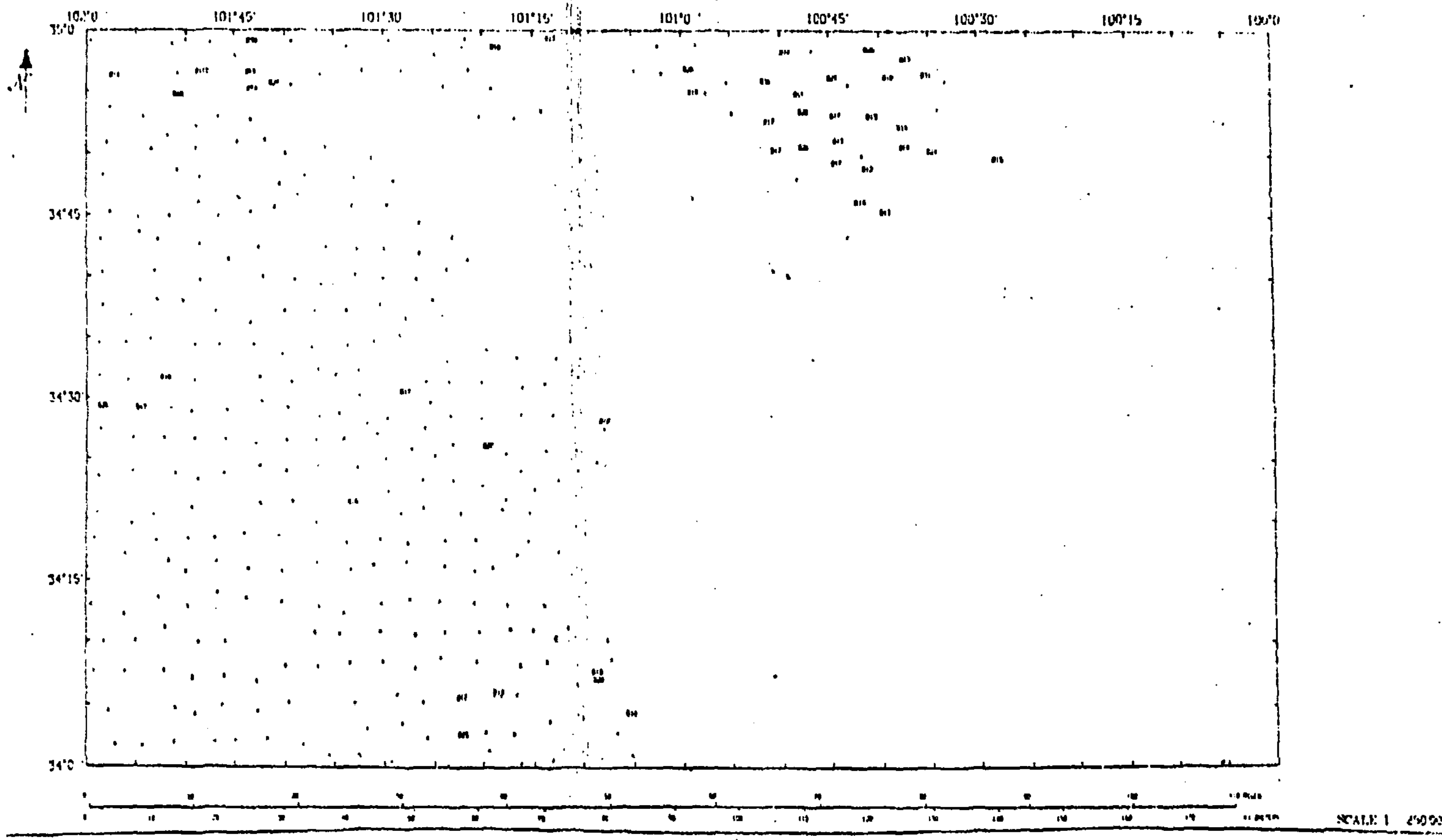


Figure 6

OGALLALA FORMATION SAMPLES HAVING EXTREME D^2 VALUES FOR THE REGIONAL VARIABLES

Figure 7 is a geographic plot of samples having $D^2 > 12$ in the modified Permian units. No large contiguous group of samples is apparent. Although several small unusual groups appear in the lower part of Figure 7. Also, many unusual samples appear along the Whitehorse Group, Blaine Formation contact.

II. A TYPICAL URANIUM SUBPOPULATIONS

Consider the uranium related pathfinder variables U, As, Mo, and V as characterizing sandstone uranium mineralization geochemistry. Hypothetically, if there are areas having potential interest for uranium exploration within a geographic area, there will be at least two uranium populations (and two lines on the Q-Q plot). One population having the smaller D^2 values would represent the uranium geochemistry of the background population. A second subpopulation with larger D^2 values represents uranium-related values differing from the background population; these samples may be of interest in exploration.

It is necessary to estimate the elements of Σ using the pairwise non-censored data for the uranium-related variables since there is a large amount of censoring due to the very low concentrations. One-half the laboratory detection limit is used for censored values in order to compute D^2 values. Both of these procedures may cause non-χ^2 variation to be exhibited in the Q-Q plot.

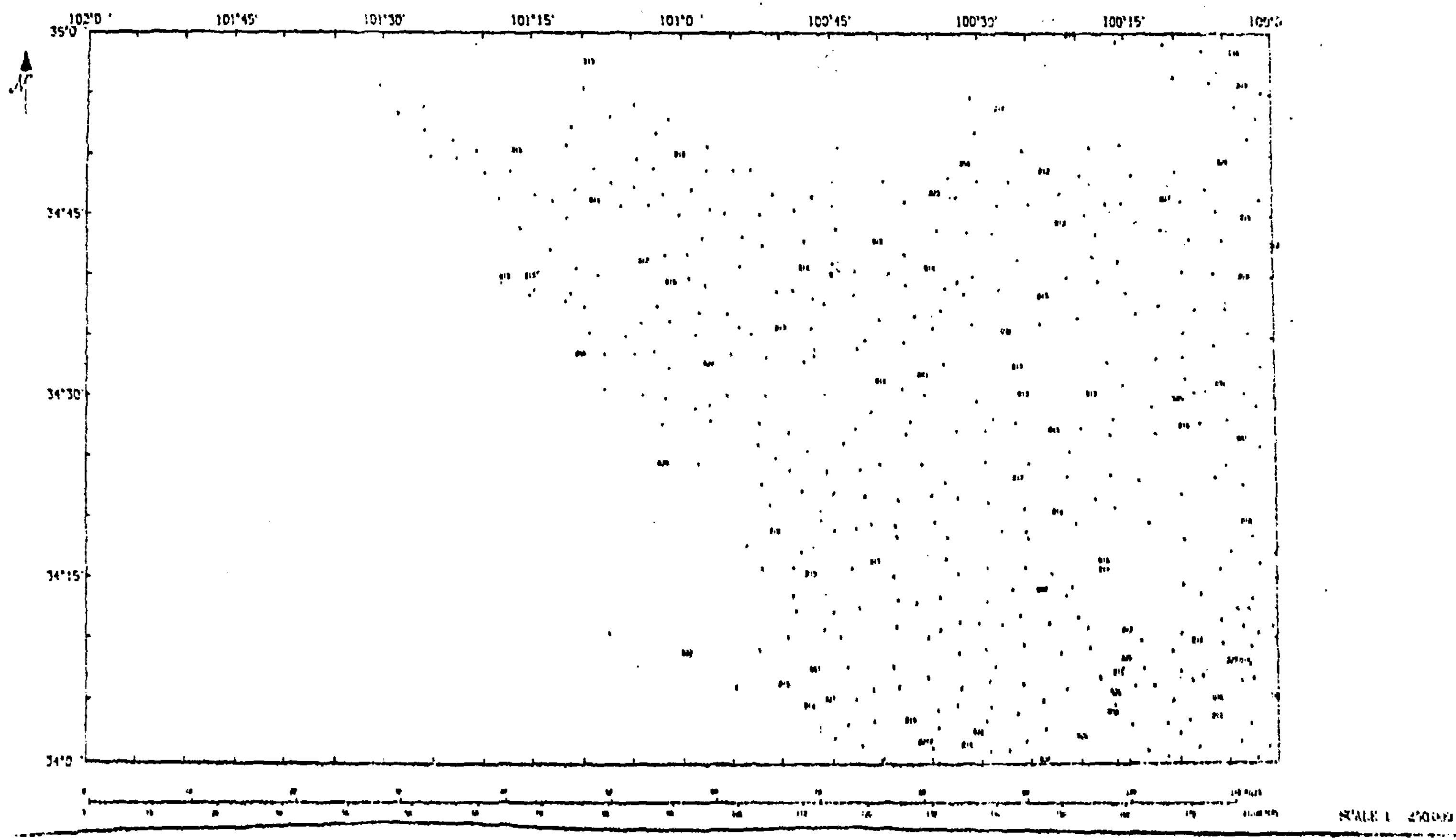
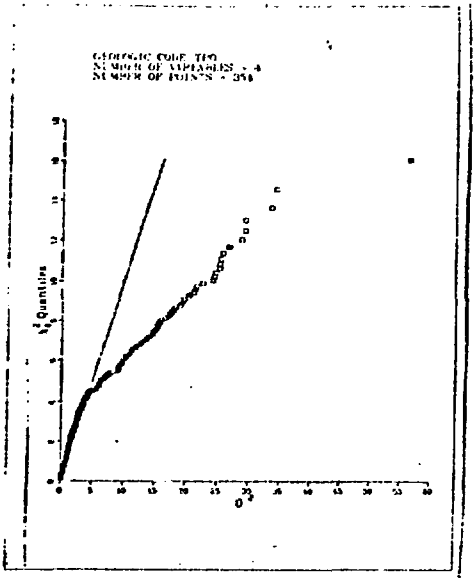


Figure 7

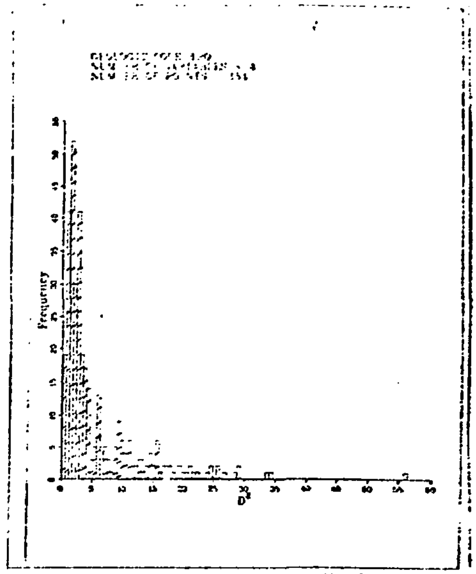
MODIFIED PERMIAN UNIT SAMPLES HAVING EXTREME D² VALUES FOR THE REGIONAL VARIABLES

Figure 8(a) shows the Q-Q plot for the Ogallala Formation. The χ^2-distribution appears to fit well for $0 < D^2 < 5$, but at D^2 of abput 5 there appears to be a break in the plot. The frequencies in Figure 8(b) also exhibit a separate population of large D^2 values. Figure 9 is a geographic plot where the samples with $D^2 > 5$ are noted. An "H" indicates high uranium samples (>80-th percentile), an "E" indicates elevated uranium samples (50 to 80-th percentile), a small "o" indicates moderate to low uranium. For the H and E samples the D^2 value is displayed to the right of the plotted letter. Three contiguous regions of unusual samples are indicated (II, A, B, C) the discussion of each region follows the method III analyses.

The Q-Q frequency plots in Figures 8(c) and (d) show that the modified Permian units exhibit unusual uranium geochemistry in that three populations appear to be present. Figure 10 displays samples where $D^2 > 15$ which is the most extreme of the three populations. The "M" indicates moderate uranium samples (20 to 50-th percentile) and an "L" indicates low uranium samples (<20-th percentile). Two three sample areas having low uranium are indicated (IID, E). In Figure 10, a third area (IIF) having high uranium is indicated and was determined by plots of the second population with $5 < D^2 < 15$. Discussion of these regions follows the method III analysis.

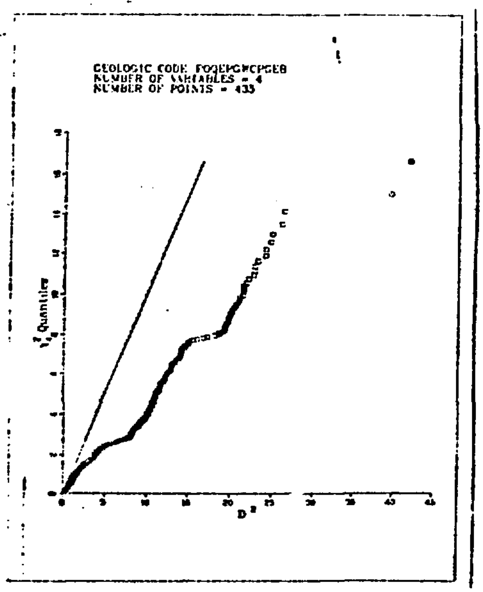


a. Q-Q Plot

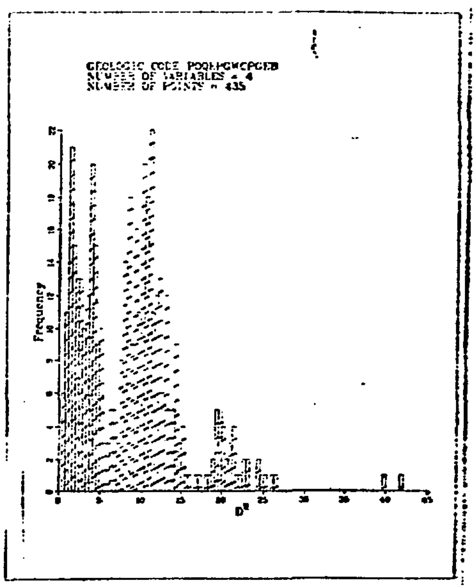


b. Frequency Plot

Ogallala Formation



c. Q-Q Plot



d. Frequency Plot

Modified Permian Units

Figure 8

Q-Q AND FREQUENCY PLOTS USING THE INCOMPLETE DATA COVARIANCE MATRIX OF THE URANIUM-RELATED VARIABLES

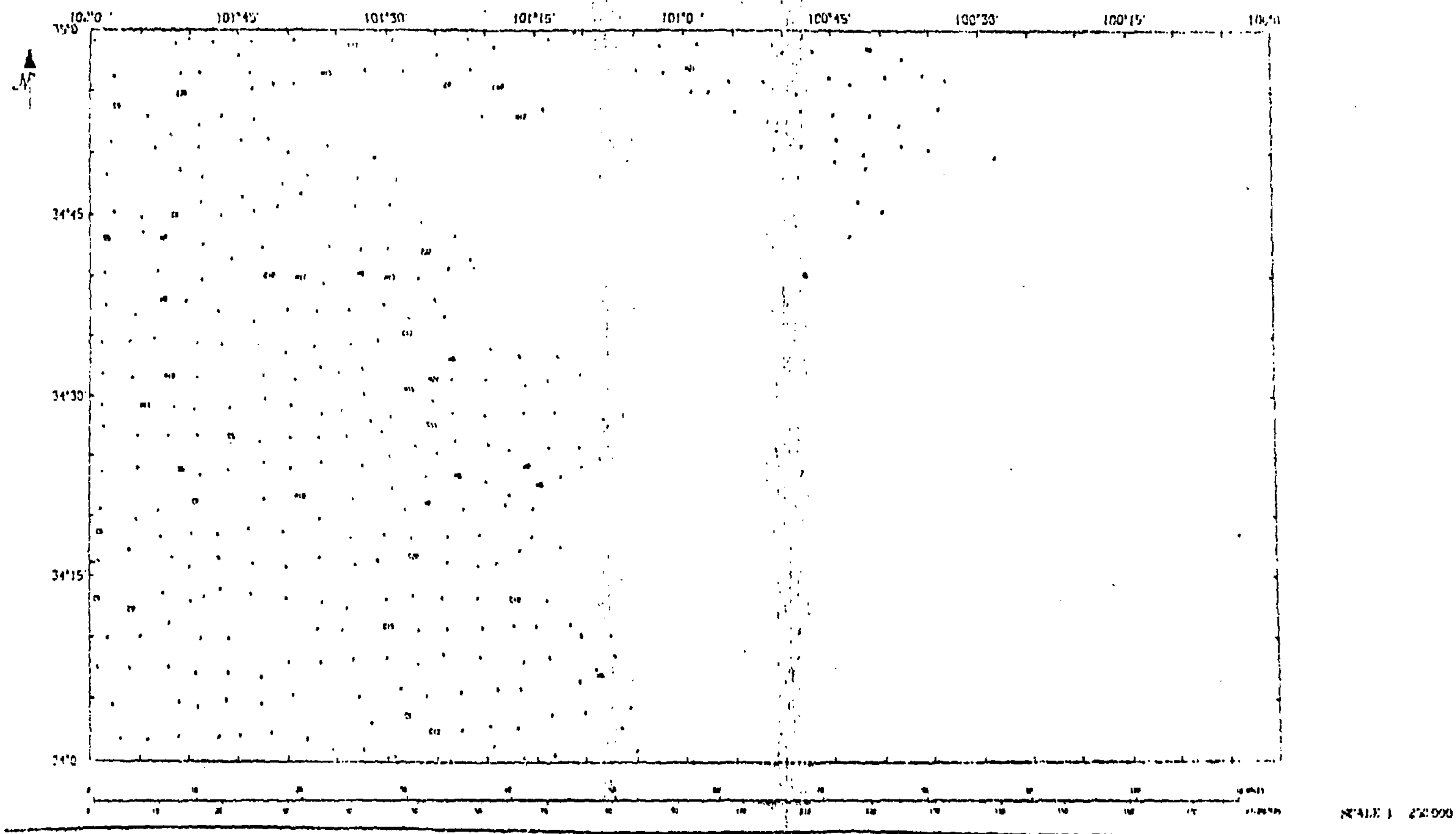


Figure 9

OGALLALA FORMATION SAMPLES HAVING EXTREME D² VALUES FOR THE URANIUM-RELATED VARIABLES

Figure 10

III. A PRIORI URANIUM POPULATIONS

In most exploration applications, it is of interest to analyze unknown areas by making an analogy to known areas of mineralization. Analysis by analogy is a common geologic tool, but it is often subjective in nature. It would be desirable to use the interelement relationships in Σ from a known mineralized region and attempt to identify samples in the unknown area that exhibit the same geochemical patterns. However, an unknown area could be expected to have both different concentration levels and variability in the parameters of interest. These differences could be due to variation in the strength of the geochemical signal which could depend upon the depth and size of the deposit in addition to groundwater flow patterns. Only the interelement relationships in an unknown area would hopefully remain similar to those in the known area. Using generalized distance measures, an approach satisfing the above criteria is given below and illustrated on the Plainview data.

Let Σ_A and Σ_B denote the covariance matrices for the anomalous and background populations. The estimate of the background covariance matrix is $\hat{\Sigma}_B = S_B$ where $S_B = (s_{ij})$, the sample covariance matrix of the background population. Now assume that the sample interelement correlation matrix, R_A, is available from a known anomalous area (or R_A could be from a hypothesized geochemical model). Seperate geochemical studies in known mineralized areas could be used to estimate R_A. To adapt R_A to the unknown area, let $\hat{\Sigma}_A = QR_AQ$ where $Q = \text{diagonal} (s_{ij}^{-\frac{1}{2}})$. Thus, $\hat{\Sigma}_A$ reflects the interelement correlations of the anomalous region and the expected variation in the background region. The matrix $\hat{\Sigma}_A$ is our *a priori* estimate of anomalous covariances in the background region.

It is possible to use $\hat{\Sigma}$, $\hat{\Sigma}$, and $\hat{\underline{\mu}}$ where $\hat{\underline{\mu}} = \bar{\underline{x}}$, the sample mean in the background region, to identify samples that have covariance patterns more similar to $\hat{\Sigma}_A$ than $\hat{\Sigma}_B$. Consider the difference

$$G^2 (\underline{X};\ \hat{\underline{\mu}}_B,\ \hat{\Sigma}_B,\ \hat{\Sigma}_A) = D^2(\underline{X};\ \hat{\underline{\mu}}_B,\ \hat{\Sigma}_B) - D^2(\underline{X};\ \hat{\underline{\mu}}_B,\ \hat{\Sigma}_A)$$

for an arbitrary sample $\underline{X}$. If the distance from $\underline{X}$ to $\hat{\underline{\mu}}_B$, weighted by $\hat{\Sigma}_B^{-1}$ is greater than the distance from $\underline{X}$ to $\hat{\underline{\mu}}_B$, weighted by $\hat{\Sigma}_A^{-1}$, then $\underline{X}$ is more likely to be from the anomalous population. Figure 11 illustrates an example where the U, As correlation is 0.2 in the background population and 0.6 in the anomalous population. The shaded area represents values of U and As which would yield $G^2 > 0$. Notice that in Figure 11 there is a large overlap of the two populations since the correlations are somewhat similar. The overlap will be reduced if the correlations are quite different or more variables having different correlations in the two populations are used.

As an approximation to the sandstone uranium correlations that may be appropriate for the Plainview data, the sample correlations for samples having uranium values above the median were computed from data in the South Texas mineralized belt in the Fleming, Catahoula, and Jackson Groups from lat. 28°-29° N., long. 97°30' to 98°30' W. The 84 samples result in the correlations in Table 3(c); Table 3(a), and Table 3(b) are

Figure 11

TABLE 3

SAMPLE CORRELATIONS FOR URANIUM-RELATED VARIABLES

(a) Modified Ogallola

	U	As	Mo	V
U	1.0			
As	.3	1.0		
Mo	0	0	1.0	
V	.2	.6	.2	1.0

(b) Modified Permian

	U	As	Mo	V
U	1.0			
As	-.1	1.0		
Mo	-.1	-.2	1.0	
V	.1	.1	.5	1.0

(c) South Texas Anomalous Region

	U	As	Mo	V
U	1.0			
As	.2	1.0		
Mo	.5	.1	1.0	
V	.3	.5	.6	1.0

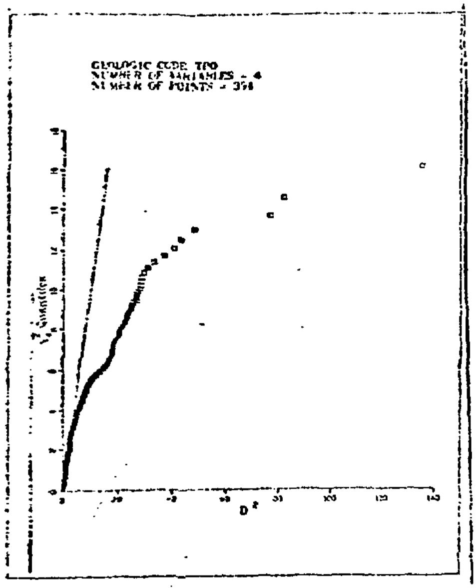
the correlations in modified Ogallala and modified Permian Units in the Plainview Quadrangle. There are obvious differences in the correlation structures of the three regions.

Figure 12(a) and (b) shows the empirical probability and frequency distribution of the G^2 values in the modified Ogallala. It is apparent that, as expected, the preponderance of sample have $G^2 < 0$, i.e. the fit for most samples to the modified Ogallala background population is preferable to the anomalous population. A geographic plot of the sample having $G^2 > 0$ appears in Figure 13 which is coded in a similar manner to Figure 9. The three contiguous regions (IIIA, B, C) indicated on the plot are discussed in the next section.

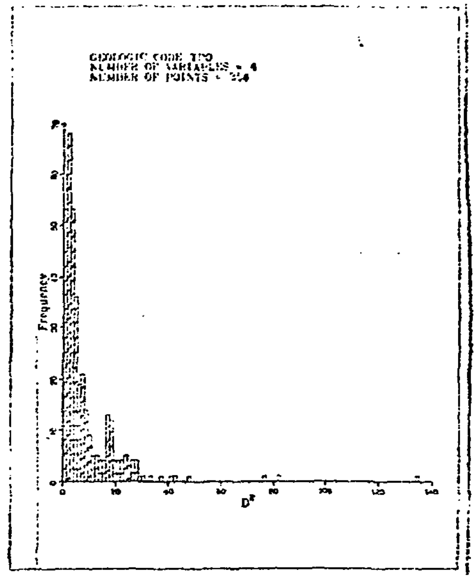
The distribution of the G^2 values for the modified Permian appear in Figure 8(c) and (d). Samples having $G^2 > 1$ are geographically plotted in Figure 14 which is coded in a similar manner to Figure 10. The contiguous region IIID as well as other regions are discussed in the next section.

INTERPRETATION OF RESULTS

A contiguous group of samples obtained in the method II or III analysis must be evaluated with respect to the actual concentration values and percentiles of the samples. Recall from Figure 4 that extreme D^2 may be obtained from any concentration level of uranium. Also, from Figure 11

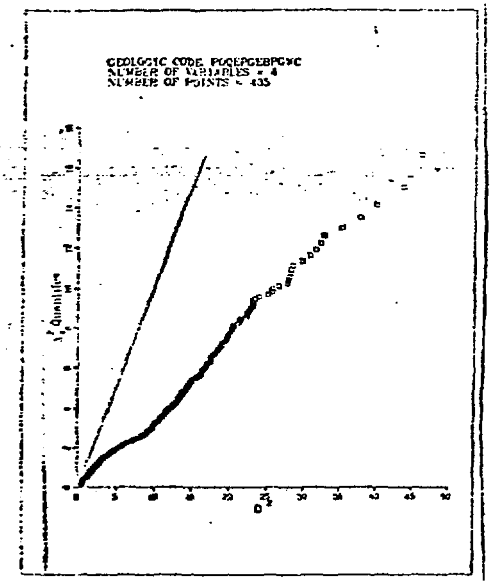


a. Empirical Probability Plot

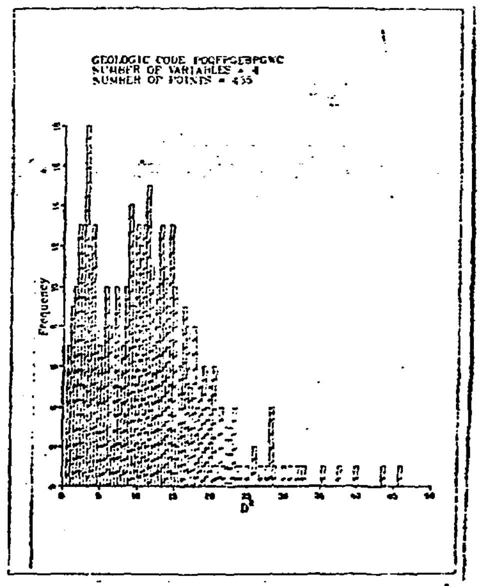


b. Frequency Plot

Modified Ogallala Formation



c. Empirical Probability Plot



d. Frequency Plot

Modified Permian Units

Figure 12

EMPIRICAL PROBABILITY AND FREQUENCY PLOTS OF G^2 VALUES
FOR THE URANIUM-RELATED VARIABLES

Figure 13

Figure 14

the shaded area where $G^2 > 0$ encompasses a region within the central portion of the background population. Additionally, extreme D^2 or G^2 values may result from samples having censored values which were assumed one-half the censoring point. Only when a group of contiguous samples display patterns of geochemical significance should areas be considered of interest. Table 4 gives the concentrations (ppb) of the pathfinder element for the areas identified by methods II and III. Specific conductance (µmhos/cm) is also given to evaluate the importance of dissolved solids. Region IIIC is a good example of an area that appears to be of little interest. While the percentiles of V, Mo, and As generally correspond [which matches the elevated correlations in Table 3(c)] the pattern of concentrations is neither consistently high or low.

Figure 15 indicated regions II A-F and III A-D. Region IIA has elevated values in the pathfinder elements and is supported by Region IIIA. The two areas exhibit very high D^2 and G^2 values. Region IIB is south of IIA and IIIA and also exhibits elevated D^2 values, but the concentration pattern is somewhat less favorable than IIA. The above three areas may result from leakage from the Dockum Formation into the Ogallala in Area B (Figure 15) which was found to be favorable for uranium by Amaral (1979). Southeast of the three above regions is the IIC and IIIC region. Notice that IIIC encompasses IIC. The D^2 values are slightly lower than IIA and IIB, but are elevated. Additionally, the pathfinder concentrations in Table 4 are elevated. The above regions (IIA, B, C; IIIA, C) are encompassed within the large area identified (Amaral, 1979) as anomalous by factor analysis. The southern most area identified as anomalous by factor analysis was not found atypical in these analyses.

Table 4

CONCENTRATION LEVELS AND POPULATION PERCENTILES FOR SAMPLES HAVING ATYPICAL URANIUM GEOCHEMISTRY

		Observed Value						Percentile					
Area Designation	Sample No.	D^2	U	As	Mo	V	Sp	D^2	V	As	Mo	V	Sp
IIA	11111	13	11.	7.0	44	39	940	90	95	85	>95	>95	85
	11114[a]	9	12.	11.	23	73	1000	85	95	>95	>95	>95	95
	11126	10	8.7	5.8	<4	12	830	90	60	80	<10	30	65
	11128	17	38.	6.0	19	50	860	95	>95	80	95	>95	70
	11156[a]	21	8.6	3.2	59	64	920	>95	80	25	>95	>95	85
IIB	9362	11	7.5	8.	<4	26	1000	90	50	90	<10	55	90
	9364	15	10.	15.	<4	24	970	90	85	>95	<10	55	90
	10900[a]	7	15.	9.1	21	46	740	75	>95	95	>95	95	40
	10902	28	20.	7.9	7	270	690	>95	>95	90	20	>95	30
	11151	13	7.0	2.8	20	4	860	90	50	15	>95	5	70
IIC	11159[a]	8	19.	5.6	11	54	850	80	>95	75	85	>95	70
	11616[a]	8	13.	5.9	28	32	1100	80	>95	80	>95	90	95
	11619[a]	5	9.3	4.1	22	37	1000	70	85	45	>95	95	90
	11620[a]	8	11.	14.	20	40	820	80	95	>95	>95	>95	65
IID	11735	23	2.4	<.5	<4	<4	1900	>95	5	<20	<20	<20	10
	11744	25	1.5	<.5	<4	<4	1500	>95	5	<20	<20	<20	5
	11745	22	2.8	<.5	<4	<4	2000	>95	5	<20	<20	<20	10
IIE	11569	42	<.2	<.5	6	12	6000	>95	<5	<20	35	60	90
	11705	23	.38	.7	22	>4	5200	>95	5	25	90	<20	85
	11737	26	.35	<.5	33	37	7300	>95	5	<20	>95	>95	95
IIF	11738	9.2	22	<.5	10	6	3700	55	90	<20	65	30	40
	11741	12	40	<.5	27	26	6700	80	>95	<20	95	90	90
	11743	9.2	19	<.5	22	11	5000	55	85	<20	90	60	75
IIIA	11109	1.1	11.	7.5	15	54	850	90	90	90	95	>95	75
	11112	0.1	7.7	6.5	14	44	930	85	65	85	95	95	90
	11113	1.8	12.	4.4	14	36	870	95	95	55	95	95	75
	11114[b]	2.3	12.	11.	23	73	1040	>95	95	>95	>95	>95	95
	11123	0.9	9.5	3.1	11	16	1200	90	85	25	75	35	>95
	11156[b]	2.3	8.6	3.2	59	64	920	>95	80	25	>95	>95	85
IIIB	10751	1.3	8.9	1.5	5	10	700	95	80	5	15	15	30
	11577	2.9	14.	1.4	7	14	770	>95	>95	5	35	25	45
	11585	4.0	17.	2.2	13	11	910	>95	>95	15	85	15	85
	11586	0.2	10.	5.7	10	19	780	85	85	75	70	55	50
	11589	0.4	7.1	2.3	7	19	840	85	60	15	35	55	70
	11592	0.0	11.	4.8	9	26	750	85	90	60	65	75	45
	11593	0.5	9.6	7.2	14	39	770	90	85	90	90	95	45
IIIC	10766	0.5	12.	7.7	12	24	750	90	95	95	85	65	40
	11159[b]	0.1	19.	5.6	11	54	850	85	>95	75	85	>95	70
	11616[b]	0.8	13.	5.9	28	32	1100	90	>95	80	>95	90	95
	11619[b]	1.5	9.3	4.1	22	37	1000	>95	85	45	>95	95	90
	11620[b]	0.3	11.	14.	20	40	820	85	95	>95	>95	>95	65
	11621	0.7	14.	5.7	12	49	900	90	>95	75	85	>95	80
	11633	2.2	11.	3.8	16	30	680	>95	95	40	95	85	25

(a) Sample is found similar to uranium-related population from Method III analysis.

(b) Sample is found to atypical from Method II analysis.

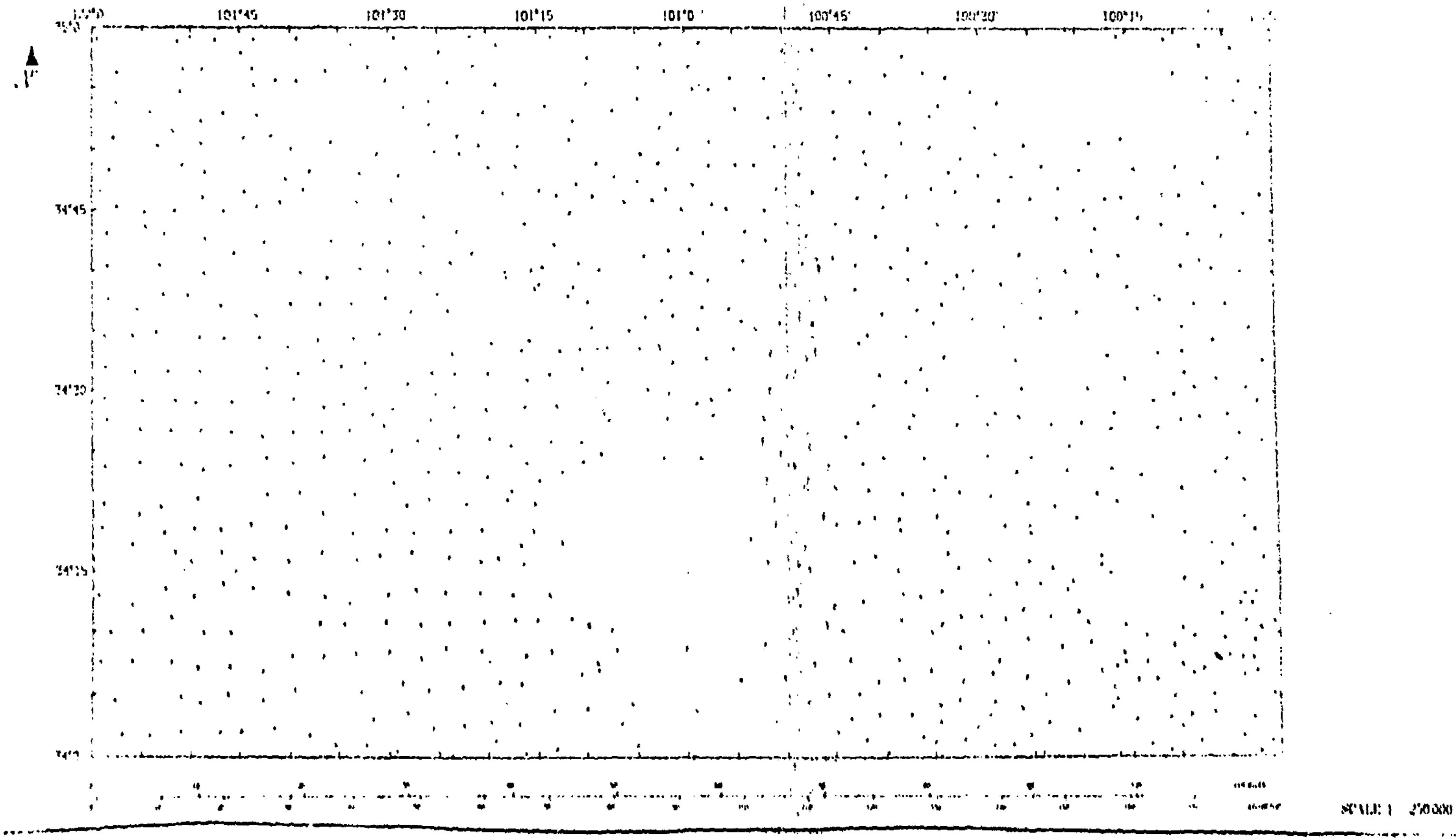


Figure 15

SUMMARY MAP FROM MULTIVARIATE DISTANCE ANALYSES

The interpretation of the three possible uranium-related populations exhibited in Figure 8(c) and (d) is unclear. When each group is plotted there does not appear to be an overall spacial relationship separating the three populations. However, several unusual characteristics appear from the analyses. Region IID has extremely depressed concentrations of the pathfinder elements. Region IIE exhibits elevated Mo and V where IIF has elevated U, Mo and V (Table 4). These abrupt changes in uranium-related elements over a small geographic area may be of interest. It should also be noted that a 10 sample selenium anomaly encompass part of Regions IIE and F, and extends to the south. In contrast to the modified Ogallala method III analysis, the unusual G^2 samples include very few high uranium samples (Figure 14). Many of the samples with $G^2 > 0$ have censored data, including those in IIID. The censored values tend to artifically inflate the G^2 values.

CONCLUSIONS

The methodology suggested here may prove useful in exploration for (a) in identifying regionalized variables that distinguish between the geologic units in a region, (b) assigning samples of unknown origin to a geologic unit or verifying preliminary assignments, (c) identifying regions of unusual geochemistry for pathfinder elements, and (d) associating samples with either mineralization models or background populations. The analysis of the Plainview data suggests the following steps for accomplishing the above:

1. Perform adequate preprocessing of the data to ensure reasonably distinct geologic populations and approximate normality of the variables.

2. Compute the Mahalanobis distance between all populations and combine those that are close as judged by small distances (large theoretical misclassification probabilities).

3. Use appropriate variable selection methods (e.g., McCabe, 1975) to identify the variable sets that are candidates for the regional variables; select the regional variables from the candidate subsets based on the geochemistry of the region.

4. Identify regional subpopulations for separate analysis from samples having extreme D^2 values for the regional variables.

5. Similarly, identify unusual regions, possibly important to exploration, from samples having extreme D^2 values for the mineralization pathfinder elements.

6. Identify contiguous groups of samples that are associated with known mineralized populations rather than the background population (i.e., $G^2 > 0$).

7. Evaluate concentration patterns of the pathfinder elements used in (5) and (6) to determine which areas may be of interest to exploration.

Application of the above methodology to the Plainview Quadrangle ground-water data indicated areas that were consistent with previous analyses and other new areas of unusual uranium geochemistry.

ACKNOWLEDGMENTS

This work was supported in part by the Oak Ridge Gaseous Diffusion Pl nt NURE Project. The authors greatly appreciate geologic input received from T. R. Butz of the NURE Project.

REFERENCES

Agterberg, F. P., *Geomathematics*, Elsevier Scientific Publishing Co., New York (1974).

Amaral, E. J., *NURE Plainview Quadrangle, Texas*, Bendix Field Engineering Corporation, Grand Junction Operations, Grand Junction, Colorado, GJQ-001 (March 1979).

Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*, John Wiley and Sons, New York (1958).

Arendt, J. W., Butz, T. R., Cagle, G. W., Kane, V. E., and Nichols, C. E., *Hydrogeochemical and Stream Sediment Reconnaissance Procedures of the Uranium Resource Evaluation Project*, Union Carbide Corporation, Nuclear Division, Oak Ridge Gaseous Diffusion Plant, Oak Ridge, Tennessee, K/UR-100 (In Press).

Barr, A. J., Goodnight, J. H., Sall, J. P., and Helwig, J. T., *A User's Guide to SAS76*, SAS Institute, Inc., Raleigh, North Carolina (1976).

Dixon, W. J. and Brown, M. B., BMDP-79, *Biomedical Computer Programs P-Series:* University of California Press, Berkeley, California (1979).

Gnanadesikan, R., *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley and Sons, New York (1977).

Huber, P. J. "Robust Covariances," *Statistical Decision Theory and Related Topics II*, Academic Press, New York (1977).

Kendall, M. G. and Stuart, A., *The Advanced Theory of Statistics, Volume 3:* C. Griffin and Company, Limited, London, p. 264-284 (1966).

McCabe, G. P., Jr., "Computations for Variable Selection in Discriminant Analysis," *Technometrics*, Vol. 17, p. 103-109 (1975).

McCabe, G. P., Jr., *Personal Communication* (1979).

Randles, R. H., Broffitt, J. D., Ramberg, J. S., and Hogg, R. V., "Discriminant Analysis Based on Ranks," *Journal of the American Statistical Assn.* Vol. 73, p. 379-384 (1978a).

Randles, R. H., Broffitt, J. D., Ramberg, J. S., and Hogg, R. V., "Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates," *Journal of the American Statistical Assn.* Vol. 73 p. 564-568 (1973b).

Rao, C. R., *Linear Statistical Inference and its Applications:* John Wiley and Sons, Inc., New York, p. 435-513 (1965).

Shapiro, S. S. and Wilk, M. B., "An Analysis of Variance" Test for Normality (Complete samples)," *Biometrika,* Vol. 52, p. 591-611 (1965).

Sinclair, A. J., *Applications of Probability Graphs in Mineral Exploration, Special Vol. 4,* The Association of Exploration Geochemists, Richmond Printers, Richmond, Canada (1976).

Snedecor, G. W. and Cochran, W. G., *Statistical Methods,* The Iowa State University Press, Ames, Iowa, p. 86-88 (1967).

Stephens, M. A., "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association,* Vol. 69, p. 730-737 (1974).

Uranium Resource Evaluation Project, *Hydrogeochemical and Stream Sediment Reconnaissance Basic Data for Plainview NTMS Quadrangle, Texas,* Union Carbide Corporation, Nuclear Division, Oak Ridge Gaseous Diffusion Plant, Oak Ridge, Tennessee, K/UR-101 (December 29, 1978). United States Department of Energy, Grand Junction, Colorado [GJBX-92(78)].

Uranium Resource Evaluation Project; *Procedures Manual for Groundwater Reconnaissance Sampling,* Union Carbide Corporation, Nuclear Division, Oak Ridge Gaseous Diffusion Plant, Oak Ridge, Tennessee, K/UR-12 (March 1978). United States Department of Energy, Grand Junction, Colorado [GJBX-62(78)].

APPENDIX A

APPENDIX

LIST OF TABLES

Table A-9

SUMMARY FROM TESTS OF NORMALITY

		U	SP	B	Ba	Ca	Li	Mg	Na	SO_4	TAK	Zn	pH	As	Mo	V
TPO	n(a)	345	345	345	345	345	345	345	345	345(7)	345	345(14)	345	345(1)	345(51)	345(33)
	Skewness(b)	2.649	3.383	7.394	2.882	11.204	0.225	1.152	3.902	15.316	0.043	4.347	-2.358	1.087	3.948	7.977
		1.738	0.271	-0.606	0.300	-2.079	-1.562	-2.422	-0.026	0.444	-2.872	0.068		-0.888	-0.207	-0.82[illegible]
	Kurtosis(b)	17.255	30.498	88.113	14.028	171.593	2.834	11.740	21.768	261.997	7.771	29.950	26.575	5.196	28.840	108.276
		7.815	5.608	5.840	4.095	21.097	5.441	11.445	6.047	5.456	25.491	2.394		4.755	3.083	3.591
	Test for Normality(b,c)	2.292	2.417	3.840	3.780	4.599	0.932	1.605	4.016	6.822	1.437	5.440	2.239	1.348	3.081	3.207
		3.157	1.547	3.005	1.466	3.067	2.756	3.630	1.584	0.830	2.065	1.140		1.297	2.099	2.470
PGEB	n	118(1)	118	118	118	118	118	118	118	118	118	118(3)	118	118(64)	118(18)	118(32)
	Skewness	2.943	0.912	3.286	5.695	-1.038	2.563	1.869	4.741	-0.348	1.012	3.050	-0.458	2.272	1.365	2.000
		-1.058	-1.014	-0.280	1.737	-2.512	-0.552	-0.802	0.233	-3.147	-0.781	0.369		0.873	-0.463	0.119
	Kurtosis	16.183	4.960	18.367	41.954	5.320	13.655	8.377	29.306	4.249	5.825	12.684	5.726	8.565	5.721	7.936
		5.089	5.492	3.258	8.787	9.415	4.005	5.334	3.148	16.231	5.250	2.533		2.595	2.337	1.956
	Test for Normality	2.111	1.450	2.276	3.577	2.311	1.248	1.488	3.107	1.791	1.172	3.412	1.530	2.991	1.461	2.308
		1.453	1.388	0.607	2.062	3.320	0.912	1.205	0.458	3.233	1.099	1.403		3.659	1.273	1.913
PGWC	n	267(1)	267	267	267(4)	267	267(1)	267	267	267	267	267(4)	267	267(37)	267(89)	267(64)
	Skewness	4.282	1.870	4.817	7.897	-0.350	2.557	2.098	3.125	-0.215	0.934	3.892	-0.803	1.359	2.960	4.183
		-0.592	-0.809	-0.001	1.437	-4.054	-0.644	-2.074	0.398	-2.660	-0.080	0.018		-0.511	0.178	-0.074
	Kurtosis	28.540	11.654	36.792	71.798	3.815	10.768	8.977	13.405	3.637	3.312	20.463	4.624	5.116	18.628	33.558
		5.419	5.640	4.650	10.250	30.374	9.590	15.707	3.093	10.745	2.264	2.697		2.533	2.044	2.103
	Test for Normality	3.058	2.697	4.205	5.887	3.196	3.367	2.891	4.267	1.969	1.810	4.873	1.959	2.161	3.587	3.669
		1.119	2.601	1.025	2.896	4.647	1.312	2.017	1.097	4.210	0.879	0.472		1.797	3.597	2.680
POQ	n	73	73	73	73	73	73	73	73	73	73	73(1)	73	73(2)	73(35)	73(19)
	Skewness	3.264	1.140	4.189	2.715	-0.066	2.050	1.791	4.407	0.961	0.614	4.284	-0.036	3.852	4.019	2.153
		0.490	-1.044	0.513	0.503	-0.824	-0.515	-0.161	0.422	-0.835	-0.720	0.416		-0.147	0.778	-0.072
	Kurtosis	14.685	4.344	24.755	11.233	1.407	9.413	6.045	26.858	4.173	2.889	23.805	3.400	23.156	23.683	9.201
		3.507	4.894	3.282	2.930	2.614	3.871	2.656	3.276	2.752	3.312	2.384		3.514	2.767	1.976
	Test for Normality	2.199	1.189	2.318	2.085	1.530	1.155	1.355	2.751	1.233	0.610	2.844	0.990	1.899	2.541	1.686
		0.744	1.131	0.645	1.044	1.920	0.672	0.629	0.916	1.636	0.845	0.865		0.835	2.537	1.525
POQE	n	39	39	39	39	39	39	39	39	39	39	39	39	39(2)	39(15)	39(10)
	Skewness	2.483	1.388	3.129	2.916	-0.885	1.327	1.429	2.320	2.030	0.241	3.324	-0.250	3.445	3.308	1.687
		0.622	0.836	0.246	0.484	-4.432	-0.417	0.242	0.335	0.890	-0.745	0.669		-0.155	0.527	-0.332
	Kurtosis	8.805	3.960	14.549	12.447	10.127	5.611	4.103	7.567	7.911	2.088	14.096	2.688	17.466	15.527	6.310
		2.968	2.744	3.176	5.338	25.403	3.687	3.049	2.800	4.083	3.137	2.631		3.228	2.536	1.992
	Test for Normality	0.651	0.809	0.642	0.674	0.859	0.911	0.813	0.650	0.812	0.964	0.489	0.971	0.658	0.614	0.830
		0.960	0.900	0.984	0.926	0.553	0.982	0.962	0.970	0.949	0.945	0.936		0.981	0.871	0.887
POQW	n	34	34	34	34	34	34	34	34	34	34	34(1)	34	34	34(20)	34(9)
	Skewness	2.536	1.154	1.149	1.187	0.875	2.468	1.192	3.863	0.628	0.261	1.448	0.187	2.513	1.751	1.335
		-0.009	-0.891	-0.048	-0.248	-0.541	-0.149	-0.285	0.497	-0.649	-0.817	0.146		-0.101	0.951	-0.051
	Kurtosis	11.395	4.094	3.694	6.199	2.811	10.338	5.229	18.748	2.375	2.472	3.696	3.852	12.076	5.100	5.112
		3.046	4.249	2.360	3.462	3.039	3.325	2.364	3.621	2.512	3.820	1.811		2.823	2.499	1.835
	Test for Normality	0.767	0.902	0.881	0.785	0.906	0.754	0.909	0.499	0.922	0.964	0.713	0.939	0.773	0.668	0.860
		0.973	0.934	0.979	0.969	0.966	0.990	0.964	0.962	0.933	0.940	0.935		0.949	0.734	0.908

(a) Value in parentheses equals number of observations < detection.
(b) First value corresponds to untransformed data, second value to log transformed data.
(c) For n ≤ 50 Shapiro-Wilk statistic is calculated, for n < 50 modified Kolmogorov-Smirnov D-statistic is calculated.

Table A-10

RESULTS FROM VARIABLE SELECTION PROCEDURES

No. of Variables in Model	DISCRIM (Equal Covariance Matrices Assumed)			BMDP[b]			Modified DISCRIM (Unequal Covariance Matrices Assumed)			Forward Selection Procedure			Backward Selection Procedure		
	Variables Included[a]	Estimated Prob. of Misclass.	Wilks x 10^{-2}	Variables Included	Estimated Prob. of Misclass.	Wilks x 10^{-2}	Variables Included	Estimated Prob. of Misclass.	Modified Wilks x 10^{-8}	Variables Included	Estimated Prob. of Misclass.	Modified Wilks x 10^{-8}	Variables Included	Estimated Prob. of Misclass.	Modified Wilks x 10^{-8}
1	9	0.484	12.7	9	0.404	12.7	9	0.488	4,216	5	0.452	4,500	2	0.512	66,570
2	9,10	0.388	10.3	9,10	0.338	10.3	5,9	0.450	53.8	5,8	0.309	3,792	2,10	0.366	20,370
3	5,9,10	0.341	8.3	5,9-10	0.341	8.3	3,5,9	0.406	5.95	5,6,8	0.346	2,140	2,8,10	0.343	>2,555
4	5,6,9,10	0.329	7.3	5,6,9,10	0.329	7.3	3,5,6,9	0.333	3.76	5-8	0.305	>259	2,3,8,10	0.303	>259
[illegible]	2,5,6,9,10	0.332	6.4	2,5,6,9,10	0.332	6.4	3,5-7,9	0.327	2.61	2,5-8	0.290	>13.1	2,3,6,8,10	0.258	>13.1
5	1,2,5,6,9,10	0.288	5.7	1,2,5,6,9,10	0.288	5.7	2,3,5-7,9	0.292	1.72	2,5-8,10	0.253	>3.9	2-4,6,8,10	0.248	>3.9
7	1,2,5-7,9,10	0.267	5.4	1,2,5-7,9,10	0.267	5.4	1-3,5-7,9	0.243	1.09	1,2,5-8,10	0.232	>2.3	2-4,6-8,10	0.204	>2.3
8	1,2,5-10	0.264	5.2	1,2,5-10	0.264	5.2	1-7,9	0.250	0.86	1-3,5-8,10	0.224	>1.8	2-4,6-8,10,12	0.195	>1.8
9	1,2,4-10	0.260	5.1	1-2,4-10	0.260	5.1	1-7,9,10	0.204	0.68	1-8,10	0.195	>1.8	1-4,6-8,10,12	0.183	>1.8
10	1-10	0.266	4.9	1-10	0.266	4.9	1-7,7-11	0.194	0.60	1-8,10,12	0.173	12.7	1-8,10,12	0.173	12.7
11	1-10,12	0.259	4.8				1-7,9-12	0.194	0.54	1-10,12	0.165	0.57	1-10,12	0.165	0.57
12	1-12	0.254	4.8				1-12	0.163	0.50	1-12	0.163	0.50	1-12	0.163	0.50

[a] 1=U, 2=SP, 3=B, 4=Ba, 5=Ca, 6=Li, 7=Mg, 8=Na, 9=SO_4, 10=TAK, 11=Zn, 12=pH (All variables transformed except Ca and pH).

[b] BMDP stopped with 10-variable model.