BNL-21693
CONF-750355--2

Brookhaven Procedures for Statistical Analyses of Multivariate Archaeometric Data*

Edward V. Sayre

Chemistry Department, Brookhaven National Laboratory, Upton, New York 11973

The accumulation in various laboratories of large numbers of multi-component analyses of archaeological artifacts has required the development of increasingly more sophisticated methods for intercomparing these data and analyzing them statistically. There are two basic aspects to the newly applied methods of data handling. One relates to the practical problems of dealing with large quantities of data. When one is handling literally thousands of analytical results it is no longer practical to compare specimen to specimen manually in order to determine whether these specimens can be separated into statistically significant individual groups. Rather one can resort to computer based procedures such as clustering techniques which can rapidly generate matrices of similarity characteristics between all pairs of specimens being considered and then group the specimens together upon the basis of greatest mutual similarity. The second basic aspect relates to the need for the application of true multivariate statistical analysis to these multicomponent data. Such methods take into account correlations between elements in calculating probability contours for specimen groups in multi-component spaces. Both clustering techniques and a variety of multivariate statistical methods have been applied extensively in other fields involing multiparameter data, such as numerical taxometry in the biological sciences and intercomparison of specimens in classical archaeology itself. However,

*Research supported by the U.S. Energy Research and Development Administration.
CONTRACT NO. E(30-1)-16

DISCLAIMER

the application of

methods of these types in the field of archaeometry has been a relatively recent development. A number of different methods of both clustering of specimens into groups and multivariate evaluation of group membership are being applied in several different archaeometric laboratories. This paper, however, will deal exclusively with methods we have found to be practical and useful in the evaluation of our multicomponent neutron activation analyses and related studies of archaeological artifacts at Brookhaven National Laboratory. These methods have been developed and tested by a group of co-workers working under the general supervision of Dr. Garman Harbottle and myself. The methods have been applied most extensively and successfully to data on pottery and related clays.

The Use of Log Normal Distributions

We have fairly consistently throughout our treatments assumed that our data are log normally distributed and hence, as a rule, have worked with the logarithms of concentrations of elements in our specimens rather than the concentrations themselves. There are two reasons for this choice. First, when we have plotted for compositionally consistent groups of specimens histograms of frequency of occurrence versus logarithms of concentrations or versus concentrations themselves, we have indeed observed that the data more often are more normally distributed in the log concentration plots. This has been observed, particularly for trace elements, in a number of geological and forensic studies. The second reason is that working with logarithms of concentration in a sense standardizes one's data without having to make any arbitrary *a priori* assumptions as to how one's data is grouped. In logarithm of concentration space one observes for a compositionally matching group of specimens that the spread of data is approximately the same for each of the

individual components independent of the magnitude of the concentrations of these components. For example, in pottery the standard deviations in log concentrations for iron and potassium oxides which usually are present in concentrations of several percent are roughly the same as for europium and lutetium oxide which are usually present at only the parts per million level. This means that when the positions of specimens are plotted in log concentration space the compositionally similar groups of specimens will group together in clusters with roughly spherical symmetry, unless, of course, a high degree of correlation exists between certain of the elements present. Such spherically symmetric groups will not occur in ordinary concentration space unless all elements concerned are present at the same level of concentration or the data are standardized. However, the quasistandardization that occurs in log concentration space is achieved automatically without making any assumptions concerning the degree or nature of grouping within the assembly of specimens being studied. It is fairly obvious that this automatic near standardization and the approximately spherical groups it can produce much facilitates the application of clustering and other multivariate methods. Thus it can be seen that it is very convenient to work in log concentration space.

Clustering Methods

Our first step in determining whether a large assemblage of data, involving the analyses of many specimens for a sizable number of components, contains within it statistically meaningful compositional groups is to resort to computer programs that produce cluster analysis of the data. Such clustering, which is usually carried out in the log concentration space of most if not all of the elements determined, provides a preliminary indication of grouping within these data. For a variety of reasons we do not feel that clustering provides a fully reliable definition of the groups present. The final assignment

of individual specimens into groups is best based upon multivariate probability calculations. Clustering, however, is a rapid and usually quite reliable method of determining how a large set of specimens are likely to be subdivided.

For clustering we first resort to one of two computer programs written at Brookhaven by A. M. Bieber, Jr., NADIST and DISCMT. These programs calculate dissimilarity matrix elements for all pairs of specimens upon the basis of one of several different dissimilarity or distance parameters. These include six distance parameters which in different ways describe the proximity of specimens plotted as points either in elemental concentration or log concentration space. They are:

1) The normal Euclidian Distance between specimen points in this space

$$ED = \left[\sum_{i=1}^{n} (A_i - B_i)^2 \right]^{1/2}$$

where n is the number of variates and A_i and B_i the coordinates for the two specimens being compared;

2) The square of this distance, the Squared Euclidian Distance;

3) The Mean Squared Euclidian Distance, that is, the Squared Euclidian Distance divided by the number of variates. In instances where data are missing for individual specimens the Mean Squared Euclidian Distances involving these specimens, which are calculated for a lower than normal number of coordinates, will be close to what they would have been had no data been missing by virtue of the distance being averaged over the number of coordinates actually employed;

4) The Mean Euclidian Distance;

5) The "City Block" Distance, which is just the linear sum of the differences of the coordinate positions for a pair of specimens

$$CBD = \sum_{i=1}^{n} (A_i - B_i) \quad .$$

6) The Mean Character Difference, i.e., the Mean City Block Distance

$$MCD = \frac{1}{n} \sum_{i=1}^{n} (A_i - B_i)$$

There are also two correlation parameters available, 1) the Pearson Correlation Coefficient R, and 2) Cosine Theta, where theta is the angle positions between lines drawn from an arbitrary origin to the position points of a pair of specimens.

Clustering is carried out by the computer program AGCLUS which was written by D. C. Olivier at the Department of Psychology and Social Relations, Harvard University. With this program, clustering can be carried out in a number of ways. Each method starts with those specimens which are most similar, i.e., which have the least dissimilarity parameter relative to each other, clustered together. Additional specimens (or eventually or clustered together clusters) are then added to existing clusters upon the basis of one of several criteria, e.g., the clustering together of units with least remaining mutual dissimilarity parameters, or units which when clustered together will form the most compact new clusters. Eventually all specimens will be clustered together and the level of the dissimilarity parameter at which each joining together of subunits has occurred recorded.

Eventually a computer plotted dendrogram of the clustering is produced, which is a display of the order in which specimens were clustered together and the levels at which clustering occurred. This is done by a plotted series of horizontal and vertical lines which link together successive clusters. Such a dendrogram showing the cluster analyses of 63 Aegean pottery sherds

upon the basis of Mean Squared Euclidian Distance is shown in Figure 1. The distance toward the right of which the horizontal lines are terminated and joined is an indication of the level of the distance parameter at which grouping into clusters occurred.

Preliminary Univariate, Element by Element, Evaluation of the Groups Indicated by Clustering

Although univariate statistical techniques do not provide as definitive a basis for final assignment of specimens into groups as do multivariate techniques, they can be applied much more easily and rapidly to the data and they can confirm some significant divisions between groups of specimens. If two groups of specimens are found to be significantly separated on the basis of single variate they will remain significantly separated as more variates are taken into account. Hence significant separations into subdivisions established upon the bases of element by element probability consideration should persist when true multivariate procedures are applied to the sample of specimens. However, as shall be demonstrated later, subsequent multivariate analysis can determine that what, upon the basis of univariate analysis, appear to be single groups of specimens can sometimes be meaningfully divided into distinctly separate subgroups. Such additional subdivision can occur only when a significant degree of correlation occurs between some of the variates. Hence, univariate techniques can indicate and confirm valid subdivision within the statistical sample but these subdivisions may not be single normally distributed groups.

We have found it to be quite convenient and useful to resort at this stage to a computer program ADSTAT, written by me at Brookhaven, which converts concentration data to logarithms and calculates averages and standard deviations for each variate. Tables are printed in which the data itself is tabulated together with the mean values, the standard deviations

expressed, in effect, as a percentage of the mean, standard deviation ranges and the ranges defining 95 percent probability limits (or other limits if so requested) for group membership based upon Student's t statistic. All individual concentrations lying outside of these limits are flagged in the table. It is obvious that if several concentrations for an individual specimen are so flagged the specimen very probably does not fit the group in which it has been included.

A computer output printer plot of the statistical analysis is produced in which on a logarithmic scale the mean values for each component are plotted bracketted with a standard deviation range indicated by plus signs and the Student's t probability range by an extending set of minus signs. Figure 2 shows such a plot. Because the concentration ranges of different components can differ by as much as four or even five orders of magnitudes, it has been expedient in these plots to shift the points for some low concentration components over toward higher values by fixed numbers of decades so that high concentration and low concentration components appear in nearly overlapping positions. This permits the plots to be accommodated conveniently in only three decades of concentration along the abscissa. Since for an individual component the same shift is made in all plots which are to be compared, this shift is in no way confusing.

In practice it has been found to be very helpful to compare two or more of these plots by superimposing them over a light box. One can then see at a glance which elements differ in the groups of data being compared and whether the deviations are significant. One can also use ADSTAT to produce analogous plots for individual specimens. These plots can then be compared to each other by superposition or to group plots to ascertain, again at a glance, how closely the specimen resembles the group in all components.

Since these plots are on a logarithmic scale, a uniform shift of one plot relative to another in a horizontal direction is equivalent to altering all components of the specimen being compared by a constant multiplication factor. Hence if two specimens differ from each other by just a dilution factor, which might be approximately the case if two pieces of pottery made from the same clay contained substantially different amounts of a relatively pure temper, their plots could be brought into near superposition by such a horizontal shift.

The program ADSTAT allows for the same sort of shift to be made mathematically. This data treatment, which I have called "best relative fit", allows one set of data to be adjusted to another set through alteration of all items in it by a constant multiplication factor so chosen as to achieve a best least-squares matching between the log concentration values for the two sets. This is done by calculating the factor

$$f = \left(\prod_{i=1}^{n} \frac{A_i}{B_i} \right)^{1/n}$$

where n is the number of variates being matched, A_i are the actual concentrations for the specimen being adjusted and B_i the corresponding concentrations in the set chosen for comparison. If now all values A_i are divided by this factor the least squares fit in log concentrations will have been achieved. The process has some interesting properties such as the fact that if one adjusts all specimens within a group to the mean values of that group, then the mean values for the adjusted group will be the same as those of the original group. The computer program allows one to select arbitrary sets of variates for such fitting and to leave the remaining data either unchanged or altered along with the fitted variates. A somewhat tricky but quite useful application of this option is to make a best relative fit adjustment for a group of

specimens upon the basis of one element alone. The concentrations for all other adjusted elements will then be transformed into concentrations relative to concentrations of this element.

Multivariate Probability Calculations

When one is working with a single variable parameter and is attempting to determine 1) whether a given set of specimens represents a sampling of one or more than one population, or 2) whether a given specimen can be assigned with confidence to a given population one usually resorts to the concept of a normal distribution as characterizing the frequency of occurrence of specimens within a single population. One first either tests to determine whether the sample of specimens presumed to represent the population is reasonably normally distributed, or for small samplings assumes it to be so upon the basis of having experienced this when examining larger groups of specimens. One then calculates upon the basis of normal distribution the probabilities that the separation between proposed different groups are indeed significant or the probabilities that individual specimens would belong to specified groups.

It is of course quite possible to carry out the same calculations on a multivariate basis. Statisticians have for many years worked with multivariate normal distribution functions. There even exists a multivariate parameter, Hotelling's T^2, which is analogous to the univariate Student's t parameter in that it brings into account in a probability calculation the additional uncertainties arising from having a small number of specimens in a sample. In both the univariate and multivariate case the normal distribution function can be written in the same basic form

$$f = K e^{-u^2/2}$$

In the univariate case K is the constant $\pi/2$ and u is the standardized distance $(x - \bar{x})/\sigma$, where x is the coordinate of an individual data point, $\bar{x}$ is the coordinate of the centroid of a normally distributed group to which the data point is being compared and σ the standard deviation of that group. In the multivariate case K is a constant that depends upon the number of variates and u is a multidimensional standardized distance which is equal to the Euclidian Distance from the center of a group to the data point in question divided by the standard deviation for the group in that direction. In general in multivariate space the standard deviation for a population will be a continuously varying function of direction relative to the centroid of the population sample. The square of this standardized distance has been called by statisticians the Mahalanobis Distance, MD, and can be calculated as follows:

$$u^2 = MD = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \bar{x}_i) I_{ij} (x_j = \bar{x}_j)$$

where x_i and x_j are the coordinates for the variates i and j for a data point, $\bar{x}_i$ and $\bar{x}_j$ are the corresponding coordinates for the centroid of the population sample, I_{ij} is the ith-jth element of the inverse of the variance covariance matrix for the group, i.e., population sample, and n is the number of variates. I have written a computer program ADCORR which among other operations calculates the variance-covariance matrix and the centroid for any group of specimens and then determines from these the Mahalanobis Distances from the centroid of the group for all members of the group and for any additional specimens one might wish to compare to the group. From these the probabilities that the individual specimens could belong to the group and yet deviate from the centroid by as great a Mahalanobis Distance as they do are calculated. An option of

the program is that group member specimens showing a low probability of belonging to the group can be deleted from it, with this elimination process being reiterated after the statistics for the reduced group are calculated until a fully internally consistent group is reached.

There are two practical problems that can complicate the calculation of these probabilities. One is that one must have more specimens included within the group than the number of variates used in the calculation. The variance-covariance matrix calculated for a group whose number of members is less than or equal to the number of variates will necessarily be singular and hence cannot be inverted. The Mahalanobis Distance calculation requires the inverse of this matrix. What this means mathematically is that in an n dimensional variate space one needs at least n + 1 independent data points to define variance in all discussions. If one has only n + 1 specimens in the group the calculation will of necessity show all specimens to be equally probable members of the group regardless of their relative coordinates. As is true in all calculations of statistical probabilities, the results become more meaningful as the number of specimens increases and ideally one should have at least several times the number of specimens defining a group as the number of variates. However, in practice this is often not possible and fortunately a statistic, Hotelling's T^2, which is a multidimensional generalization of Student's t, has been developed which allows one properly to adjust probability calculations for the additional uncertainty introduced by relatively small samples. The probabilities by program ADCORR are all based upon the distribution of Hotelling's T^2. However, this does not solve all of one's problems for if one is working with a smaller than ideal number of specimens the inclusion of an additional specimen will tend to alter the variance-covariance matrix in such a way as to accommodate the specimen within the group.

Therefore, unless the number of group members is adequately high the calculated probability of an individual specimen belonging to a group will be significantly higher when that specimen is included within the group than when it is excluded from the group. When refining a group with this method one encounters some specimens which show low probability of belonging to a group when they are included within it, and therefore can be excluded from the group with confidence, and some specimens which show a high probability of belonging to the group even when excluded from it and hence can be included within the group with confidence. However, one also can encounter intermediate specimens for which one calculates a reasonably high probability of belonging to a group when the calculation is made including them within the group and a significantly low probability of belonging to the group when they are excluded from it. One, of course, feels uncertain about such specimens and tends to exclude them from the group one is defining. It is a wise procedure therefore to test all marginal group members to see if they fall in the ambiguous category.

In a multivariate space all points of equal probability will, of course, have equal Mahalanobis Distances relative to a population centroid and will define a hyperellipsoidal surface in that space. Figure 3 is a correlation diagram for the elements scandium and iron upon which is plotted the distributions of specimens from three regions of the Middle East. In each instance the specimens include clay samples from the three regions and pottery sherds which because they were found in the same regions as and were closely similar in composition to the clays could be presumed to have been fabricated from these local clays. The clays in question are the Red Field Clays which typify the deposits of the Palestinian coastal plain, the Limestone Hill Clays which are found in the central upland of Israel, and the geologically recently

deposited Alluvium from a number of sites along the Nile, ranging from Aswan to Cairo. Around each of the data sets is plotted the probability ellipsis which define, upon the basis of Hotelling's T^2 distribution, the ninety-five percent confidence limits of containment of each of the compositional types. Such correlation diagrams with probability ellipsis can be computer plotted through use either of the program ADCORR or a program RAPLOT, which is yet to be discussed.

The Need for Multivariate Data Handling

In Figure 3 one can see that the elements iron and scandium are highly correlated in each pottery group. This is to be expected because the elements iron and scandium tend to be highly correlated in nature, as many geological studies of rock and mineral compositions have shown. However, because the two Palestinian clay and pottery groups are primarily only offset from each other along nearly parallel correlation regression lines, the individual iron concentrations and scandium concentrations in the two groups of specimens largely overlap. Hence if one were to analyze this assembly of data only element by element upon the basis of iron and scandium, one would fail to observe that the Red Field Clays and the Limestone Hill Clays are significantly different in composition from one another. One must use a method of analysis that fully takes into account the correlation occurring between variates as well as the absolute values of the coordinates of the variates to achieve a full resolution of one's data. A Mahalanobis Distance calculation automatically takes into account all of the correlations existing between variates for a group being analyzed. In the case shown in Figure 3 a Mahalanobis Distance calculation showed that the Limestone Hill Clay specimens all had less than 0.001 percent probability of belonging to the Red Field Clay group. Thus it can be seen that when correlation between variates is involved, significant separations between groups of specimens can

occur which would be missed in univariate, element by element, analysis but observed and shown to be signficant in a true multivariate analysis.

Use of Characteristic Vectors of the Variance-Covariance Matrix

It can be shown that the axes of the probability hyperellipsoids generated in a Mahalanobis Distance calculation for a normally distributed group are the characteristic vectors, i.e., eigenvectors, of the variance-covariance matrix of that group. The calculation of characteristic vectors and their corresponding characteristic values are well described in almost all standard references dealing with matrix algebra, and I shall therefore not comment upon the methods of their calculation here. I shall simply describe the characteristic vectors as a set of n mutually orthogonal coordinate vectors in an n dimensional multivariate space with origins at the centroid of a data group which are oriented in the multivariate space along regression lines of correlation to the extent that these exist for the data points of the group. Hence they provide a basis for a description of the distribution of the group from which all correlation has been removed. That is to say, the characteristic vectors constitute a new set of variates between which no correlation exists for the data group being analyzed. The set of characteristic basis vectors are similar to the original orthonormal set of elemental basis vectors with the exceptions only that their origin have been transferred to the centroid of a group and their directions in multivariate have been rotated to conform to the directions of maximum and minimum variances within the group. Hence a transformation to characteristic vectors in no way alters the relative positions of data points in a multivariate space but only provides a new description of this distribution.

In Figure 3 the characteristic vectors for the group of the Palestinian Red Field Clay specimens are drawn through the group and marked A and B. It can be seen that the vector A[1] lies along the regression line for correlation between iron and scandium for the group, also[2] is in the direction of maximum variance within the group and also[3] is the major axis of the probability ellipse for the group. The vector B is perpendicular to A, is in the direction of least variance for the group and is the minor axis of the probability ellipse of the group. It is interesting to note that the elongations of the other two groups in Figure 3 are nearly parallel to each other so that the characteristic vectors of all three sets and indeed that of the composite of all of these data points would be nearly parallel. This is a state of affairs which we have come to expect to be quite common in clay and pottery analyses. The elemental correlations occurring in different clays tend to be similar even in instances in which the mean relative abundances of the correlating elements are significantly different. The parallel placement of groups with internal correlation is in part the result of working in log-concentration space because to the extent that the ratios of concentrations of a pair of correlating elements in data sets are constant within each set the correlation regression lines for each set when plotted in log concentration space will have a slope of one. This tendency for groups with pronounced correlation to be parallel to one another in log concentration space and hence have characteristic vectors that are parallel to one another is another advantage of using the logarithms of concentrations rather than the concentrations themselves.

Those familiar with factor analysis will recognize that working with the characteristic vectors of the variance-covariance matrix is similar to the method of principal component analysis. However, there is a significant difference between the applications normally made of the characteristic vectors

in principal component analysis and the applications that are useful in this instance which stems from the fact that one is usually using principal component analysis to clarify the relationships between variates while in our investigations we are more concerned with the relationships between specimens. In principal component analysis one obtains the characteristic vectors of the variance-covariance matrix or the correlation matrix and then concentrates one's attention upon those vectors that explain most of the variance or correlation. Usually the vectors with little variance are ignored. Figure 3, however, makes it clear that in our studies, where we are attempting to discriminate between significantly different groups of data, there is no a priori reason to expect any of the characteristic vectors to be more discriminating than others. It is apparent in Figure 3 that the characteristic vector of greatest variance, A, for the Red Field Clay group would discriminate between this group and the Nile Alluvium but fail to discriminate between it and the Limestone Hill Clays. In contrast to this the characteristic vector of least variance, B, discriminates well between the two Palestinian groups but would fail to discriminate between the Red Field Clays and Nile Alluvium. Accordingly in our investigation we look at all characteristic vectors equally carefully to ascertain which if any of them will discriminate between groups. As aids in determining whether individual characteristic vectors are discriminating, the program ADCORR will print out the coordinates for each specimen along the characteristic vectors and also histograms of the distributions of these coordinates for the total assembly of specimens along the characteristic vectors. Histograms showing the distributions of specimens of the groups shown in Figure 3 along the two characteristic vectors calculated from the iron and scandium data for the group formed by combining all of the specimens together are shown in Figures 4 and 5. The subdivision of the specimens into three separate groups is apparent in these histograms.

Once one has determined which of the characteristic vectors are discriminating it is easy to determine in turn which of the original component elements are most involved in this discrimination because one of the tables printed by ADCORR lists the components of the original basis compositional elements projected onto the characteristic vectors. From this table one can infer which elements are most involved in the constitution of an individual characteristic vector.

Standardized Multivariant Coordinates

When one calculates the characteristic vectors for a variance-covariance matrix one transforms the matrix into a simple diagonal one, that is, one for which all off diagonal covariant terms are zero. The transformed matrix is the variance-covariance matrix for the characteristic vectors. The diagonal matrix elements, which are referred to as the characteristic values of the original variance-covariance matrix, are the variances for the characteristic vectors. The fact that the covariance terms of the matrix are all zero results from the fact that there is no correlation between the characteristic vectors.

It is, of course, possible to define the position of all data points in the multivariate hyperspace by coordinates along the characteristic vectors. Such coordinates are the projections of data point positions upon the characteristic basis vectors. These coordinates will define vectors between the centroid of the group and the data points in question, and the sum of the squares of sets of these coordinates for each specimen will be equal to the squared Euclidian distances between the centroid and the data point for that specimen. If one now divides each of the characteristic vector coordinates for a data point by the square root of the corresponding characteristic value, i.e., by the square root of the characteristic variance, one obtains a set of standardized coordinates which will define a standardized vector from the centroid of the group to the data

point whose length is the Euclidian Distance between these points divided by the standard deviation of the group in that direction. The sum of the squares of these standardized characteristic vector coordinates for a data point will therefore be the Mahalanobis Distance for that point. The characteristic vectors for a group are a unique set of basis vectors for achieving the transformation to standardized dimension throughout a multivariate space in the manner just described.

When one does transform to multivariate coordinates which are standardized with respect to a particular group, the hyperellipsoids of equal probability for that group are transformed to hyperspheres. One can say that the group plotted in this transformed space will have "spherical" symmetry. All other groups which are similarly shaped and oriented parallel to the group which forms the basis of the transformation will be similarly transformed into spherically symmetric clusters.

Such a transformation of elongated, highly correlated groups into roughly spherical internally uncorrelated groups is shown in Figure 6. The three data groups plotted in Figure 6 are the same ones, Palestinian Red Field Clays, Limestone Hill Clays and Nile Alluvium, plotted for the elements iron and scandium in Figure 3. In this instance we first calculated the variance-covariance matrix for the Palestinian Red Field Clay group, determined the characteristic vectors of this matrix, and then calculated standardized characteristic vector coordinates for all specimens. In Figure 6 the specimens have then been plotted in terms of their standardized characteristic vector coordinates.

The clustering procedures based upon distance or size parameters as measures of similarity rarely take into account correlation. Accordingly these procedures do not serve to separate well adjacently situated highly correlated

groups such as the Red Field Clay and Limestone Hill Clay groups as plotted in Figure 3. However, after the data for such groups have been transformed as they have for Figure 6, they clearly can be separated into clusters much more definitely and unambiguously. We have found it to be very useful to transform an assemblage of specimen analyses into standardized characteristic vector coordinates by means of program ADCORR as a preparatory step for cluster analysis. In many instances when we have carried out such transformations on pottery data the subsequent clustering has been significantly improved.

The Handling of Missing Data

When one is analyzing many specimens for a sizable number of components, it is almost inevitable that for a variety of practical reasons some of the concentrations will be missing for some specimens. If one, however, has the determined values for, let us say, twenty-two or twenty-three out of twenty-four components of a specimen, one would quite sensibly try to make as much use as one could of the existing data. One would be loath to eliminate such a specimen from consideration in forming or calculating the properties of related groups simply because its data set is not fully complete.

There are several ways one can proceed in surmounting this difficulty. It has been mentioned that in calculating a similarity matrix for clustering one can use the Mean Euclidian Distance or Mean Character Difference where the distances are divided by the number of variates actually used in calculating a distance between two specimens and hence the effect of a missing set of coordinates is reasonably averaged out. One can use a somewhat analogous approach in calculating a variance-covariance matrix by calculating each matrix element for the total number of specimens within a group for which data exists for both of the two variates of that matrix element. In this approach the number (n-1), that occurs in the denominator of each variance or covariance

matrix element and is one less than the number of specimens, might be different for different matrix elements. Those matrix elements for pairs of variates for which a full set of data exists will be calculated upon the bases of the full set of specimens. Those matrix elements for pairs of variables for which some data is missing will be calculated just as they would be for that subset of the group of specimens which has complete data for this pair of variates. This would seem to be a reasonable approach and one of the options in ADCORR is to compensate for missing data in this way when calculating a variance-covariance matrix. However, we have found this approach to lead to some rather peculiar results in practice, including even the calculations of negative Mahalanobis Distances for some specimens. We have tended to avoid the use of this option.

Another approach would be to substitute for a missing datum a value that would have a minimal effect upon the calculation concerned. One might think that the average value for the missing variate for the group would be a satisfactory substitution. However if significant correlation exists between variates within the group the substitution of group average value for one or more of the correlating variates might be a very distorting choice. Consider again Figure 3 with the very elongated distribution of the Red Field Clay group in iron-scandium spaces. If only an iron value were available for a specimen and this value differed significantly from the mean of the iron values, then the substitution of a mean scandium value might well place the specimen completely outside of the group. In fact one can see in the case cited that the arbitrary replacement of scandium values for those Red Field Clay specimens lying within the ninety-five probability ellipse would transfer a good half of those specimens to positions outside of that ellipse.

A considerably more sensible value to substitute for a missing datum would be one chosen so that the specimen would conform to the group with greatest probability. This means choosing the missing datum or data to minimize the Mahalanobis Distance for the specimen. This can be done in a very straightforward way by setting the partial derivatives of the expanded Mahalanobis Distance function with respect to the missing variates equal to zero. If there are n variates for which data is missing for a specimen the n partial derivatives with respect to these variates will form a set of n linear equations, the simultaneous solution of which would be the substitution coordinates which would most closely fit the specimen to the group. To make this process more clear I shall develop the equations for the cases of a single missing datum and two missing data.

Let us write the function for the Mahalanobis Distance in the form

$$MD = \sum_i \sum_j X_i I_{ij} X_j$$

where $X_i = (x_i - \bar{x}_i)$ and $X_j = (x_j - \bar{x}_j)$ are the deviations of specimen coordinates for variates i and j from the centroid of a group, I_{ij} is the ith, jth matrix element of the inverse of the variance-covariance matrix for the group and both summations are made over all variates. If a missing datum is designated X_a the criterion that its selection lead to a minimum value of the Mahalanobis Distance lead to the equations

$$\frac{\partial}{\partial X_a}(MD) = \frac{\partial}{\partial X_a}\left(\sum_i \sum_j X_i I_{ij} X_j\right) = 0$$

$$2I_{aa}X_a + \sum_{i \neq a} I_{ai}X_i + \sum_{i \neq a} I_{ia}X_i = 0$$

Since $I_{ai} = I_{ia}$, the last equation reduces to

$$2I_{aa}X_a + 2\sum_{i\neq a} I_{ai}X_i = 0$$

$$X_a = \frac{\sum_{i\neq a} I_{ai}X_i}{I_{aa}}$$

in which the X_i for $i\neq a$ are all existing data for the specimen in question.

In the case of two missing data for a specimen, X_a and X_b, one has a pair of partial derivatives set equal to zero

$$\frac{\partial}{\partial X_a}(MD) = 2I_{aa}X_a + 2I_{ab}X_b + 2\sum_{i\neq a \text{ or } b} I_{ai}X_i = 0$$

$$\frac{\partial}{\partial X_b}(MD) = 2I_{ab}X_a + 2I_{bb}X_b + 2\sum_{i\neq a \text{ or } b} X_{bi}X_i = 0$$

resulting in the pair of linear equations

$$I_{aa}X_a + I_{ab}X_b = -\sum_{i\neq a \text{ or } b} I_{ai}X_i$$

$$I_{ba}X_a + I_{bb}X_b = -\sum_{i\neq a \text{ or } b} I_{bi}X_i$$

the simultaneous solution of which will provide the substitution values for the two missing data. The sets of linear equations one obtains for greater numbers of missing data are, of course, closely analogous to these.

Program ADCORR provides as one of its options the calculation of substitutions for missing data upon this basis of minimizing the Mahalanobis Distance for specimens. In doing this the program first calculates a variance-covariance matrix for which group average values are substituted for data missing for specimens of the group. Having calculated a new set of missing data values from the inverse of this matrix a new variance-covariance

matrix is calculated using the new missing data values, which in turn allows one to calculate an improved set of missing data values. This process is iterated until the missing data substitution values and the corresponding variance-covariance matrix have effectively become constant. Based upon our experience in using it we are of the opinion that this method is the most logical and effective one for compensating for missing data.

Auxiliary Programs

Alan Bieber, Jr. has written a series of computer programs which have supplemented our more basic programs well and have proven to be quite useful. The first of these is RAPLOT which plots correlation diagrams between pairs of variates using a normal printer for producing these plots, thus avoiding the need to use auxiliary equipment such as a Calcomp plotter. Probability ellipses may be plotted around groups of data points. The program also calculates and tabulates 1) the ratios between pairs of variates for each specimen, 2) the means and standard deviations of these ratios for separate groups as well as 3) the means and standard deviations for the separate variates themselves, and 4) the Pearson product-moment coefficient R for pairs of variates.

A second program HISTEL produces a printer output histogram of the distributions of specimens upon the basis of their concentrations or log-concentrations of selected compound. A third program SKWURT analyzes in one dimension the skewness or kurtosis of group of data which one is testing to determine whether it can be regarded as a sample of a normally distributed population.

Summary

Clearly this set of procedures do not begin to exhaust the methods which can be effectively applied to the classification of specimens into compositionally

consistent and significantly different groups. They have, however, provided an effective approach to this problem, and one that takes into account the interdependence between concentration levels of some components as well as the individual magnitudes of these concentrations. Because strong correlations do frequently exist between elements present in pottery, we believe that multivariate techniques which take such correlations into account ultimately must be employed in order to fully resolve a set of data. However, much can be accomplished by more simple element by element methods, and we continue to use monovariate techniques along with the multivariate ones. In general, it has been found to be both more convenient and accurate to work with logarithms of concentrations rather than concentrations themselves. We have usually found clustering techniques to be our most useful preliminary tool for grouping large amounts of data, and multivariate probability calculations to provide the most reliable final criteria for the assignment of specimens to groups.

Acknowledgements

I should like particularly to acknowledge the significant and long range contributions of Dr. Garman Harbottle to the development and testing of these data handling techniques. We have worked together closely at all stages of this investigation. Among a number of archaeologists who have collaborated with Dr. Harbottle and myself in the development and testing of these methods Drs. Alan Bieber, Jr. and Ronald Bishop have made outstanding contributions to the computer based statistical procedures that have been evolved.

Figure Captions

Figure 1

Dendrogram of the cluster analysis of 63 Aegean sherds generated by Program AGGLUS, using type 5 dusting with Mean Squared Euclidian Distance.

Figure 2

Plot generated by Program ADSTAT of mean concentrations, standard deviation ranges, and 95 percent confidence ranges for various components in 69 Palestinian Red Field Clay specimens.

Figure 3

Iron-scandium correlation plot generated by Program RAPLOT for three groups of Middle Eastern Pottery. 95 percent probability ellipses are plotted around each data group.

Figure 4

Distributions of Middle Eastern clay and pottery specimens along the characteristic vector of Greater Variance based upon iron and scandium concentrations in all specimens.

Figure 5

Distributions of Middle Eastern clay and pottery specimens along characteristic vector of Lesser Variance based upon iron and scandium concentrations in all specimens.

Figure 6

Distributions of the three Middle Eastern pottery groups of Figure 3 plotted in the two dimensional standardized characteristic vector space based upon iron and scandium concentrations of the Palestinian Red Field Clay specimens.

AYIOS STEPHANOS, BERBATI, AND RELATED MATERIALS -- TYPE 5 CLUSTERING ON SMED

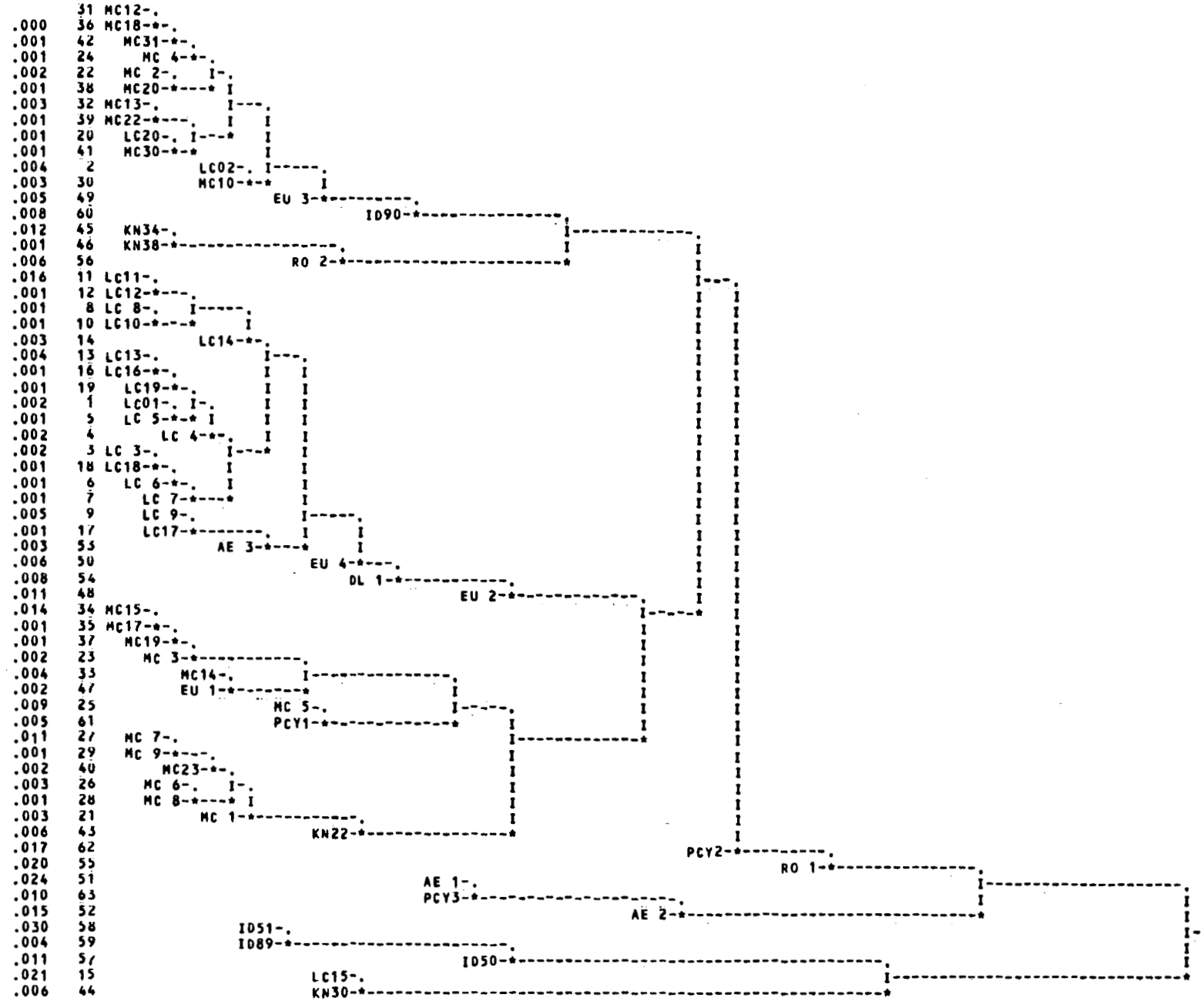

Fig. 1

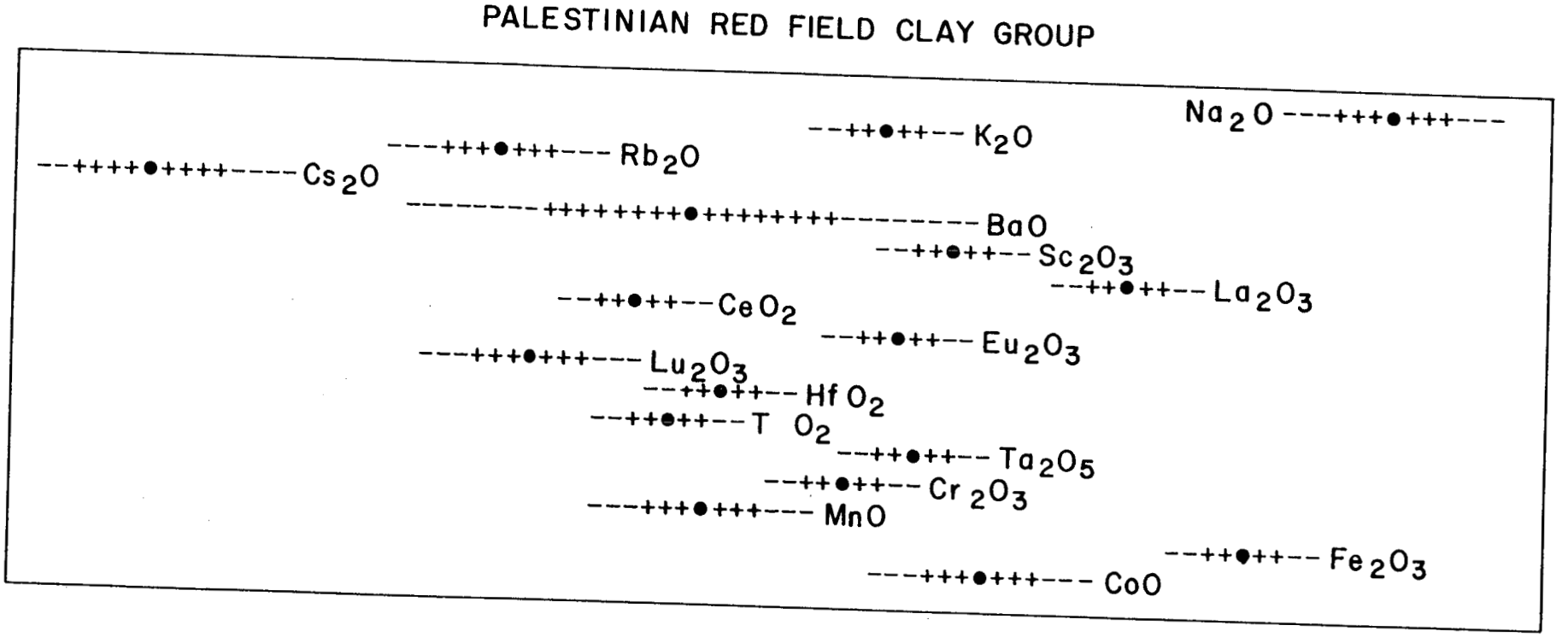
PALESTINIAN RED FIELD CLAY GROUP
Na2O
K2O
Rb2O
Cs2O
BaO
Sc2O3
La2O3
CeO2
Eu2O3
Lu2O3
HfO2
TiO2
Ta2O5
Cr2O3
MnO
Fe2O3
CoO

Fig. 2

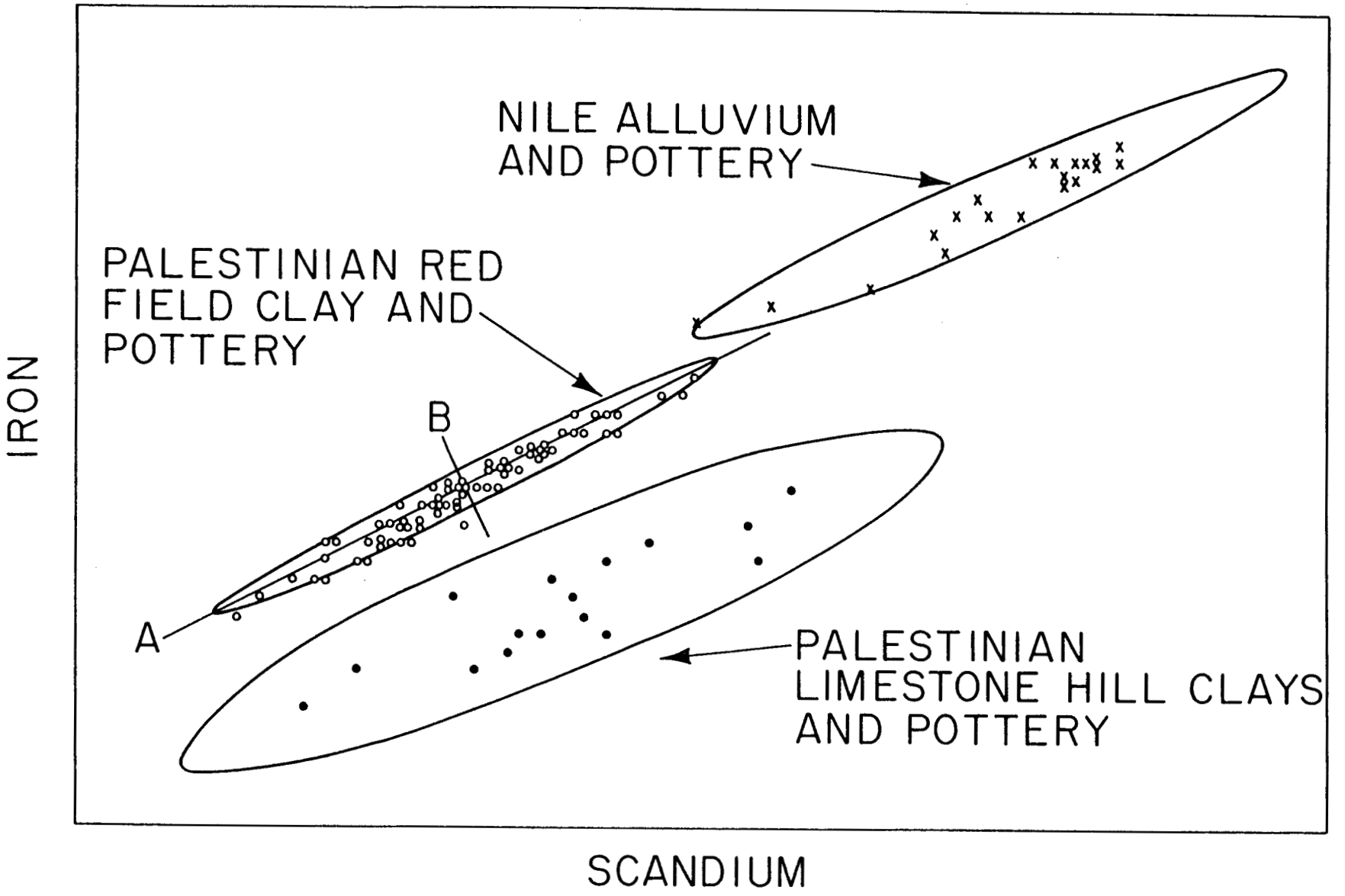
NILE ALLUVIUM
AND POTTERY
PALESTINIAN RED
FIELD CLAY AND
POTTERY
B
A
PALESTINIAN
LIMESTONE HILL CLAYS
AND POTTERY
IRON
SCANDIUM

Fig. 3

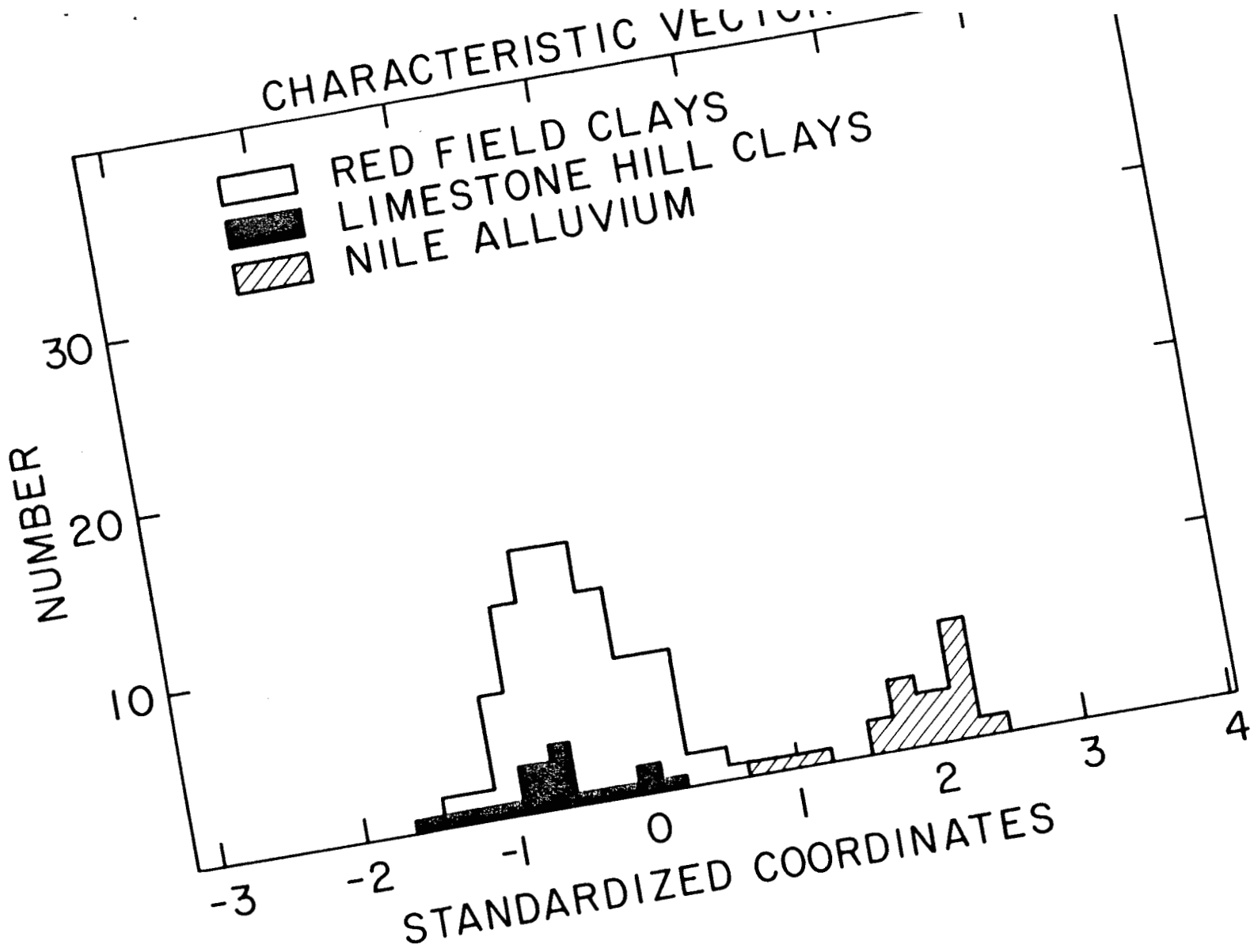
CHARACTERISTIC VECTOR
RED FIELD CLAYS
LIMESTONE HILL CLAYS
NILE ALLUVIUM
NUMBER
30
20
10
-3
-2
-1
0
1
2
3
4
STANDARDIZED COORDINATES

Fig. 4

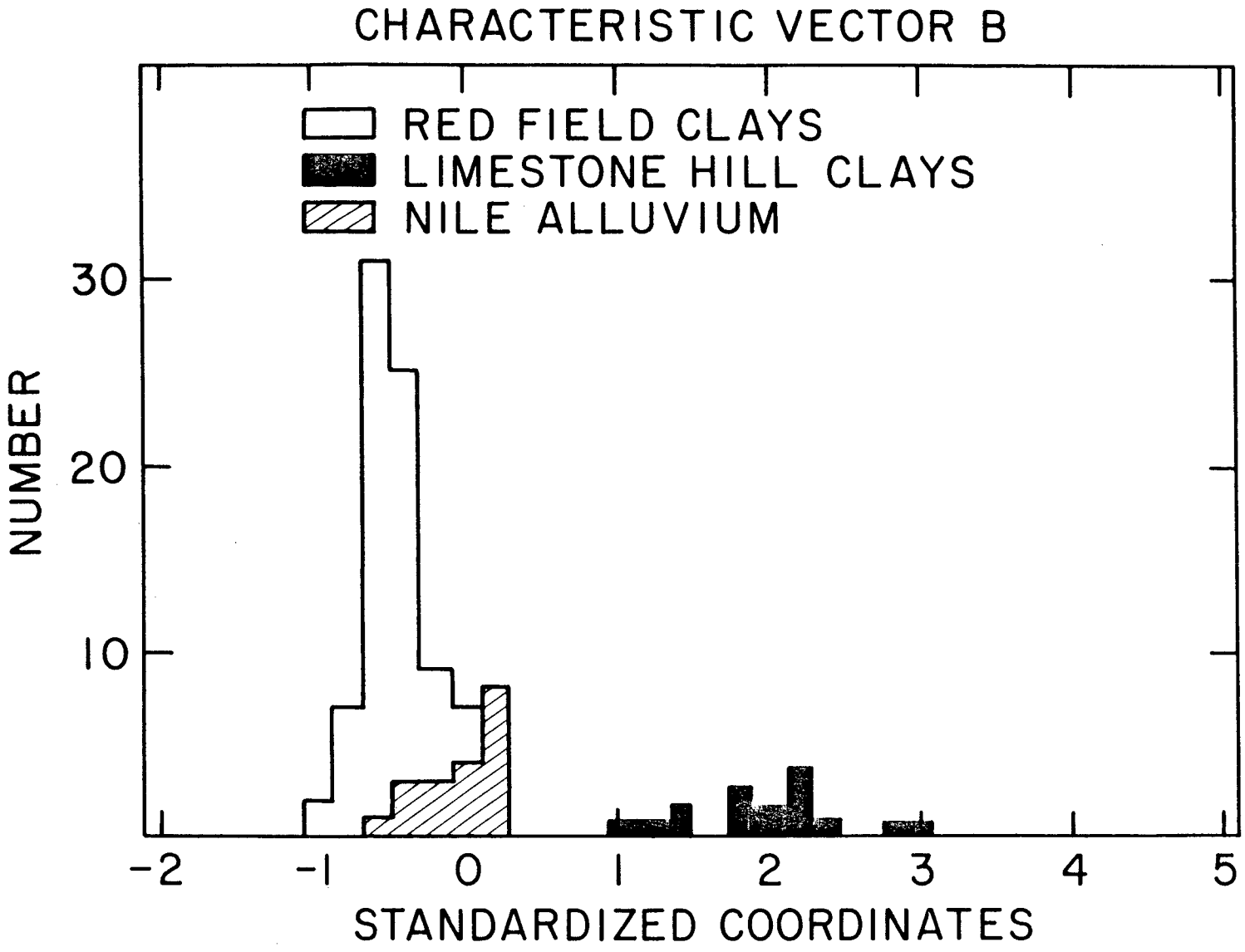
CHARACTERISTIC VECTOR B
RED FIELD CLAYS
LIMESTONE HILL CLAYS
NILE ALLUVIUM
NUMBER
30
20
10
-2
-1
0
1
2
3
4
5
STANDARDIZED COORDINATES

Fig. 5

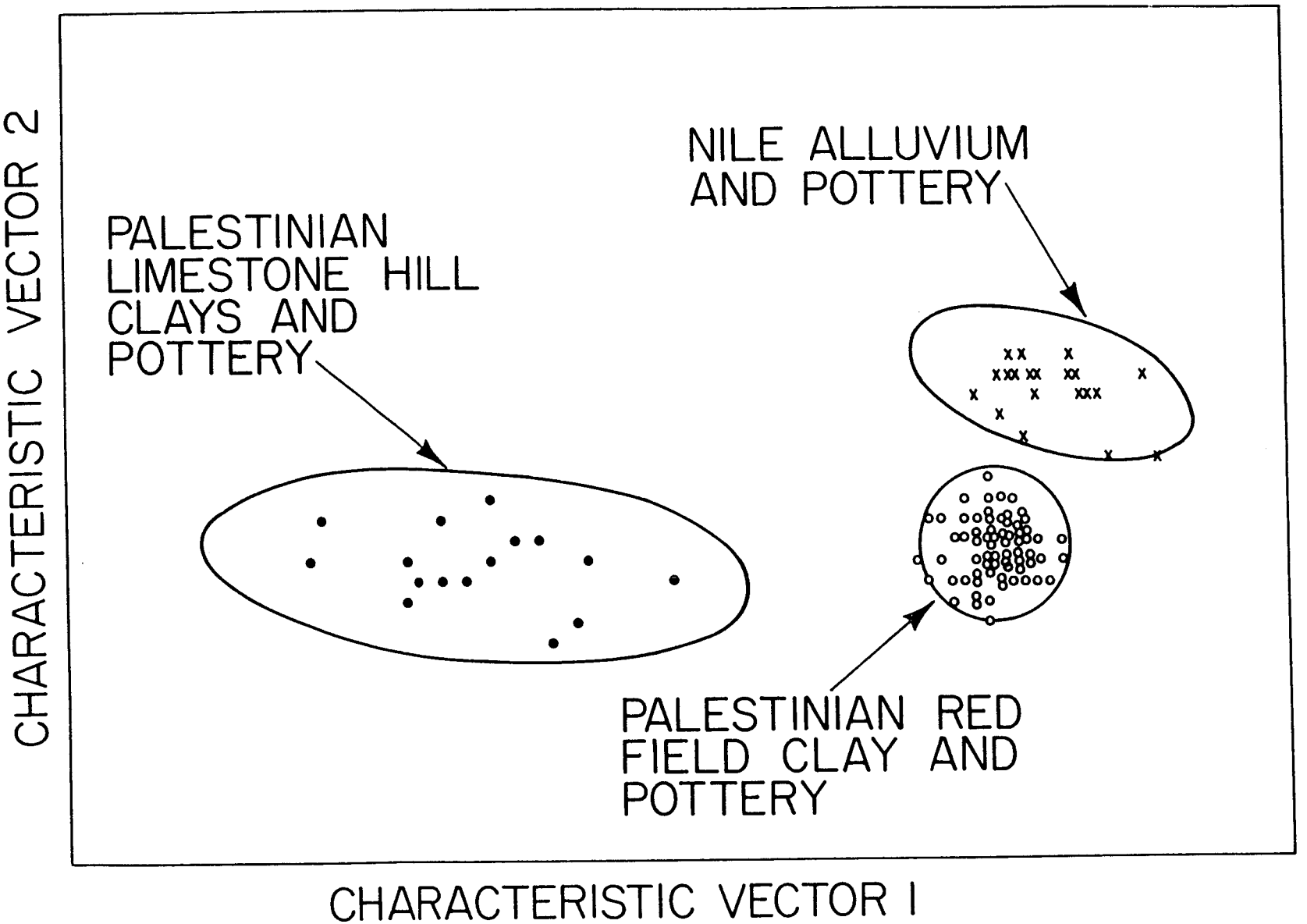
CHARACTERISTIC VECTOR 2
NILE ALLUVIUM AND POTTERY
PALESTINIAN LIMESTONE HILL CLAYS AND POTTERY
PALESTINIAN RED FIELD CLAY AND POTTERY
CHARACTERISTIC VECTOR I

Fig. 6