ADVANCES IN

AGRONOMY

Prepared under the Auspices of the

AMERICAN SOCIETY OF AGRONOMY

VOLUME 28

Edited by N. C. BRADY

International Rice Research Institute
Manila, Philippines

1976

ACADEMIC PRESS • New York San Francisco London

A Subsidiary of Harcourt Brace Jovanovich, Publishers

STATISTICAL METHODS IN SOIL CLASSIFICATION RESEARCH

Rodney J. Arkley

Department of Soils and Plant Nutrition, College of Natural Resources, University of California, Berkeley, California

I. Introduction: Objectives and Problems of Soil Classification

The purpose of classification is to organize the members of a large popu
of objects into groups or classes of objects so that the nature of, and rel
ships between, the objects can be more easily understood. This purpose n
limited to understanding related to a specific purpose such as the irrigabi
soils and based upon only those attributes relevant to the purpose;
purpose may be to develop a more general classification based upon as

attributes as possible and useful for a wide range of purposes. Most of the research reported herein is of the latter kind with a few exceptions.

Prior to 1955 when Hughes and Lindley first used a statistical procedure to reclassify members of six soil series, soil classification was based primarily on subjective judgment. That is not to say that the judgment applied has not been good; some of the best minds in soil science have been focused on soil classification. Nevertheless, both the selection of criterion variables for classification, their effective weighting by application at different categorical levels in a hierarchical classification, and of boundary values for separations have all been made primarily on the basis of fallible human judgment.

Although some statistical analysis was carried out on soil data prior to 1955, the amount was limited primarily by the tedium of statistical analysis and the hand sorting of data. However, with the development of the electronic computer this tedium has been removed, and the application of numerical and statistical methods to soil classification has developed rapidly as indicated by the number of references cited in this paper, most of which deal directly with soil classification.

To be both comprehensible and most effective, the differentiating characteristics or criterion variables used to form classes should contain the maximum possible information. That is, they should be those which have the most predictive value for the nature and behavior of the soil when subject to external influences. These criterion variables then should be those which are covariant with other properties (accessory characteristics) not used as criterion variables. The number of soil characteristics that might be used for soil classification is very large, and the selection of criterion variables from this list is very likely to be suboptimal by subjective methods. However, the covariance among all variables can readily be examined with the use of the computer by calculating the product-moment correlation coefficient between all pairs of variables and with direct examination or analysis of the resulting correlation matrix. On this basis an optimal set of differentiating characteristics can be chosen. Gibbons (1968) effectively argued the importance of covariance to the usefulness of soil classification. The extent to which soil properties are covariant will be discussed later in this paper.

Soil classification in the past has been primarily hierarchical in nature, with one or more criteria used at each categorical level to divide soils into mutually exclusive classes. Such a classification is helpful in the understanding of relationships among soils, but some relationships may be seriously distorted. This occurs when a group of soils which is relatively similar in all other respects, is subdivided into two groups at all lower categorical levels by small differences in a particular differentiating characteristic. This is a general problem of hierarchical classification, in that it is a divisive procedure and the divisions are dichotomous; that is, they either have or have not a certain property or the value of a certain measured property is above or below a specified level. Avery (1968)

points out that a successful hierarchical classification can be made only if the differentiating criteria can be ordered in accordance with the number of other attributes (accessory characteristics) associated with them. He points out further that soil variation is not of this character, presumably because soil characteristics result from the interaction of several factors.

This is the crux of a major problem in soil classification. Are soils to be considered as made up of discrete natural individuals, or do the individual soil profiles represent points in a multivariate continuum, which considered as a whole contains no distinct boundaries for purposes of classification? Experience has shown that soils may occur as discrete, relatively homogeneous bodies when considered only within a local area. But as investigations extend to broader areas more and more soils of intermediate character are encountered which bridge the gaps between the original discrete soil bodies. This leads to problems in soil correlation. The recognition and identification of soil classes, such as soil series in the field, in the face of this kind of soil variation is the bane of the soil surveyor's existnece. His decisions must be made primarily on the basis of field observations of a number of soil characteristics, and the soils classified on the basis of variations in one or many of these simultaneously.

Avery (1968) argues forcefully that a coordinate system of classification is more appropriate for soils than a hierarchical system. The advantage to a coordinate system being that each differentiating criterion is given equal weight, at least a priori. It should be pointed out also that a coordinate system can be used to construct a number of hierarchical systems, by arranging the differentiating criteria in different categorical orders.

In the following discussion of the various statistical methods applied to soil classification, it will be observed that much of the work has assumed that soils (mainly soil profiles) fall into natural clusters or groups which can then be ordered into a classification. For the limited sets of soils used, this appears to be a correct assumption. Although little work has been done toward the development of a coordinate classification scheme, it is made clear that some of the statistical methods described can be used effectively for this purpose assuming that soil characteristics vary in such a way as to form a continuum through the whole population of soils. A method is suggested by which soils can be classified using a set of well-separated centroids or conceptual modal soils and classes formed on the basis of the general affinity of real soils to the centroids.

II. Numerical Taxonomy or Cluster Analysis of Soils

A. GENERAL THEORY

Numerical taxonomy is defined by Sneath and Sokal (1973) as the grouping by numerical methods of taxonomic units into taxa on the basis of their

character states. Groupings are formed using the following general procedures: First, data for a number of units, such as soil profiles, are assembled including a sizable number of selected variable properties for each unit. Because this discussion is confined to the classification of soil units, they will be simply referred to as soils. The data are commonly arranged into a matrix consisting of soils by columns and soil properties by rows. Next an over-all estimate of resemblance is obtained between pairs of soils by some mathematical function of all differences between the values for each property of the two soils. After numerical values for the estimate of resemblance (either estimates of similarity or of difference can be used) between all pairs of soils included in the study are obtained, the matrix of $n(n-1)/2$ values is subjected to a sorting strategy which forms groups of similar soils. The nature of the groups formed and their relationships or taxonomic structure can be presented in various ways; these may include dendrograms, reordered matrices, ordination, or simply tables of coordinates.

B. DATA SELECTION

The choice of soils to be included in the data should be such that the number of soils is large and the general kinds of soils included are well represented. For example if the soils included are mainly well drained and without evidence of wetness, then the inclusion of a very few poorly drained soils may interfere with the analysis because those soil properties associated with wetness may not be representative of the range of variation in those properties. This is particularly important in cluster analysis as the order of cluster formation is affected by the number of soils in the clusters or groups formed in some clustering techniques.

The selection of soil properties is even more important to a successful analysis than the selection of soils. Although all kinds of both field and laboratory data can be used, there are certain kinds that should be excluded or that need special treatment. Some soil properties, such as the field moisture content, are generally irrelevant to soil classification and should be excluded. Logically correlated properties, such as dry and moist colors, are generally so highly covariant that one or the other should be included. Particle size distribution values for sand, silt, and clay always add up to 100% and so one of the three should be eliminated from the data. The inclusion of large numbers of logically related properties should be avoided, as they tend to create an inadvertent extra weight to such a group of properties in the classification. For example, in the initial list of properties used by Sarker *et al.* (1966) were 6 particle size ratios, 5 of which were intercorrelated above the 0.90 level. This kind of redundancy among properties should be avoided or dealt with by analysis of variables as discussed later in this paper.

Soil data obtained from laboratory analysis are almost always continuous

variables as are a number of field measurements such as thickness and depths of recognizable characters. Hue, value, and chroma (soil color) are continuous variables even though the color chart commonly used is made up with discrete steps. However, field observations such as soil texture, structure, and consistency are usually discrete or multistate variables and need to be treated with special care. Data of this kind should be coded in such a way as to reflect their proper rank order reflecting their importance to soil behavior or development. For example, soil texture classes might be coded in order of their relative clay content or water retention characteristics.

Structure is a particularly difficult soil property to code as a ranked multistate variable. For surface soils an appropriate order of structure type might be single grained, massive, platy, crumb, granular, subangular blocky and angular blocky; for subsoil layers the order might be single grained, massive, platy, granular, subangular blocky, angular blocky, prismatic, and columnar. Structure grade or distinctness such as weak, moderate, and strong coded 1, 2, 3 might well be multiplied by the code for structure type coded 0 to n and a value for size of peds added to give a single value for type-grade-size of structure. The system suggested is only one of many possible ways that soil structure might be treated. Arkley (1971) and Cipra *et al.* (1970) have used type and grade omitting structure size with some success. Barkham and Norris (1970) treated soil structure type, grade, and size as separate characters.

Soil color mottling is troublesome to code. Cipra *et al.* (1970) used a combination of abundance, size, and contrast of mottles scaled from 0 to 8. Cuanalo and Webster (1970) used abundance percent and position on peds separately. Rayner (1966) used abundance, size, and contrast as separate variables.

Dichotomies such as the presence or absence of earthworms, concretions, carbonates, iron pans, manganese stains are sometimes used (Rayner, 1969; Muir *et al.*, 1970). Dichotomies require special treatment which will be discussed in relation to the standardization of variables. The same is true for unranked multistate variables.

C. WEIGHTING AND STANDARDIZATION OF VARIABLES

1. *The Problem of Weighting of Variables*

Sneath and Sokal (1973) present cogent arguments in favor of weighting all variables equally, especially where a classification is intended to be a "natural" or basic classification for general use rather than one for a specific objective. These arguments against "a priori" weighting appear to be on sound rational grounds. This is in direct opposition to the methods of orthodox or conventional

hierarchical classification wherein the differentiating characteristics used at higher categorical levels take precedence over those at lower levels, and therefore have greater effective "weight." For example, in the Soil Classification of the Soil Survey Staff, U.S. Department of Agriculture (1960), certain diagnostic horizons are considered more important than others; a case in point is the use of the mollic epipedon at the "Order" level to separate the Mollisol soil order from other orders, irrespective of the nature of the subsoil horizons to a large degree, whereas most of the other orders are separated on the basis of the nature of subsoil horizons. Decision such as the one cited are based on intuition or human judgment, both of which are fallible.

Sneath and Sokal (1973) also argue in favor of the use of a large number of variables (i.e., soil properties) in numerical taxonomy, on the grounds that the use of variables greatly evens out the effective weight which each one contributes. This argument presupposes that all pertinent groups of covariant properties are about equally represented in the data. In the data used for numerical classification research on soils, this is clearly not true in many cases. Some kinds of measurement on soil properties are more easily obtained than others, or have been of more interest to the investigator, and so are overrepresented and thus unduly weighted. Also the use of a large number of variables involves a great deal of time and expense, especially if the classification is intended to encompass a large number of individuals. However, in the first stages of analysis, the use of a large number of variables standardized so as to give equal weight to each is certainly a sound approach. For the final classification it may be possible to reduce the number of variables to a manageable but still effective size by analysis of the covariance among them.

2. *Covariant Soil Variables*

In the past, covariance among soil variables was rarely analyzed, but with the advent of electronic computers the tedium of the calculation of correlation coefficients has been removed. In several papers involving numerical taxonomy of soils, correlation matrices have been published, revealing how much covariance exist among soil variables. Moore and Russell (1967) analyzed 10 trace elements in 28 soil profiles and found that the correlation matrix of 45 r-values contained 27 which were significant ($P < 0.01$) and ranging from 0.49 to 0.90; Moore *et al.* (1972) show 50 of 91 r-values significant ($P < 0.01$) ranging from 0.18 to 0.80 for 14 variables in 4 layers of 40 soils. Sarker *et al.* (1966) found that 39 of 61 soil properties were correlated with at least one other at the level of $r > 0.50$; Russell and Moore (1967) show 57 of 136 r-values significant ($P > 0.01$) ranging from 0.40 to 0.80 for 17 variables and 43 soils.

Reexamination of my own data sets used for analysis of variables also revealed similar levels of communality among variables as shown in Table I (Arkley, 1971).

TABLE I
Number of Variables Significantly Correlated with other Variables in Analyses Reported by Arkley (1971)

Number of		Variables with r-values significant at $P < 0.01$			
Soils	Variables	0	1+	>2	>4
59	21	4	17	8	1
621	23	0	23	21	20
220	34	0	34	30	26
87	44	0	44	42	40
87	53	2	51	46	38

The extensive covariance among soil variables in the widely differing data sets described in Table I is strong evidence that a long list of soil variables is not necessary to classify soils effectively by either conventional or numerical methods. Also, it is evident that analysis of variables should be among the first steps in the development of a classification system. There should be no objection to weighting variables according to their predictive values as revealed by the analysis of variables as this would not be considered a priori weighting.

3. *Standardization of Variables*

Most procedures for obtaining an estimate of resemblance require that the variables be standardized to a common range of values. It is clearly inappropriate to compare differences in a variable with a range of 0.0 to 1.0 with those in a variable with a range of 100 to 1000. For continuous variables standardization may be by range, i.e.,

$$X' = (X - X_{min}) / X_{max} - X_{min})$$

or by variance, i.e.,

$$X' = (X - \bar{X}) / \mathrm{SD}_X$$

The former gives each variable a range of 0.0 to 1.0, the latter a mean of 0.0 and a standard deviation of ±1.0. These methods can also be applied to ranked multistate variables, but with more risk of injecting spurious information.

For data containing both continuous and discrete variables, either dichotomous or multistate, Grigal and Arneman (1969) applied a method proposed by Talkington (1967): For a variable that can assume a number of discrete and mutually exclusive states (i.e., soil structure types) which is coded as 1.0 for no

more than one of these states and 0.0 for all others, the maximum possible contribution to the differences or the sum of the square of the differences between two individuals is 2.0 (Table II). Continuous variables are therefore standardized so that the maximum contribution to the squared difference is also 2.0; such variables are standardized by range and then multiplied by $2^{1/2}$ or 1.414.

Another method for equalizing the contribution of discrete and continuous variables based upon information theory has been developed by Burr (1968) but so far has not been applied to soils. Continuous variables are standardized to a mean of 0 and a standard deviation of $\pm 2^{1/2}$ by the formula

$$X' = (X - \overline{X}) / (1.414 \times \mathrm{SD}_X)$$

And dichotomous and multistate variables by the formula

$$M'^2 = M(t-1) / [2tp_s (S_M - 1)]$$

where M' is the standardized variate, M is an unstandardized variate (as coded in Table II), t is the total number of individuals (soils) with nonmissing data, p_s = the proportion of t in state $s(s_n/t)$, s_n is the number of individuals in state s, and S_M is the number of possible states of variate M.

Burr proposes to call this procedure standardizing by reciprocal proportions since the weight of each state is weighted inversely to its frequency of occurrence (p_s). The formula given above weights a multistate variate M equally with a continuous variable. If one considers that each state s should be weighted equally with a continuous variable, then the parameter $(S_M - 1)$ can be omitted from the formula.

The problem of highly skewed data should be considered in the standardization of variables. In some cases it would be appropriate to use a logarithmic or square root transform for known skewed distributions as was done by Moore and Russell (1967). Talkington (1967) advocates a slight truncation of the range for extreme values which occur very rarely as was done by Grigal and Arneman

TABLE II

Variable states (s)	Individuals A	Individuals B	Difference (d)	d^2
s_1	0	0	0	0
s_2	1	0	1	1
s_3	0	0	0	0
s_4	0	1	1	1
			2	2

(1969). In this connection it should be pointed out that the use of ratios as data variables may well lead to highly skewed distributions because of the hyperbolic nature of ratios; in general they should be avoided. In any case a careful examination of the data for errors or aberrant data should precede the analysis, simply as a good analytical practice. Where highly skewed variables are suspected, they should be examined in a frequency distribution and perhaps plotted against related variables in a scatter diagram before a decision is made as to their treatment.

D. MEASURES OF SIMILARITY OR DIFFERENCE

Various procedures are available for calculating an over-all estimate of resemblance, which can be based upon measures of either similarities or differences. Sneath and Sokal (1973) use the term "similarity coefficient" to cover coefficients both of similarity and of dissimilarity; the one being the complement of the other. They describe four somewhat fuzzy classes of similarity coefficients as (1) distance coefficients, (2) association coefficients, (3) correlation coefficients, and (4) probabilistic coefficients, of which the first and third have been used in soil studies most commonly.

1. *Distance Coefficients*

The simplest practical form of "distance" measure called mean character difference (MCD) is:

$$\mathrm{MCD}_{jk} = \frac{1}{n} \sum_{i=1}^{n} |X_{ij} - X_{ik}|$$

where $X_i \ldots n$ are standardized variates and j and k are two individuals such as soil profiles. However, this coefficient is rarely used and suffers from the fact that a large difference in a single variable is inadequately represented in the coefficient. MCD has been applied by Moore and Russell (1967) and Webster and Burrough (1972).

A much more commonly used distance coefficient is the familiar Euclidean distance (d) in the form:

$$d_{jk} = \left[\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - X_{ik})^2 \right]^{1/2}$$

The expression $1/n$ is introduced into the equation in order to equalize differences introduced by missing data, or differing numbers of variates used. The average Euclidean distance coefficient has the advantage of being more readily visualized and can be plotted in two or three dimensions although not in n-space where n is greater than 3. It also has other statistical properties which are advantageous. This coefficient has been used in soil studies by Cipra *et al.* (1970), Crichton (1975), Grigal and Arneman (1969), Moore *et al.* (1972),

Lamp (1972), and Webster and Burrough (1972). Cuanalo and Webster (1970) used d^2 rather than d.

Another distance that has been used several times is referred to as the Canberra metric (d_c). It was developed by Lance and Williams (1967b) and has the advantage of needing no prior standardization of variables. It is in the form of

$$d_c = \sum_{i=1}^{n} (|X_{ij} - X_{ik}|)/(X_{ij} + X_{ik})$$

However, it has the unfortunate characteristic that a difference in the upper part of the range of a variate (i.e., when the denominator is large) is minimized as compared to an equal absolute difference in the lower part of the range and thus is sensitive to proportional rather than absolute differences between individuals (soils).

Another coefficient of similarity [of Bray and Curtis (1958)] was applied to soils by Hole and Hironaka (1960), the first to use numerical taxonomy in soil classification, and later by Bidwell and Hole (1964), Bidwell *et al.* (1964), Sarkar *et al.* (1966), and Moore and Russell (1967). The Similarity Index (SI) mathematically restated is:

$$SI_{jk} = \sum_{i=1}^{n} |X_{ij} - X_{ik}| \Big/ \sum_{i=1}^{n} (X_{ij} + X_{ik})$$

All variables must be standardized to a common range and positive in sign. It is particularly useful where the data are all in percentages as in species composition of plant communities.

Another distance coefficient, Mahlanobis D^2, is suitable only for use with predetermined groups of individuals. It was applied to soils by Hughes and Lindley (1955) and by Van den Driessche and Maignien (1965). It is based on comparisons of variances within and between groups and is closely allied to discriminant function analysis. However, it is computationally involved and heavily dependent upon multivariate normality. The reader should consult Sneath and Sokal (1973) and Rao (1948) for details of the method.

2. *Simple Matching Coefficient*

For data coded as two state characters (0 and 1), a simple matching coefficient (S_m) can be used. S_m is simply the number of matches 2 vs. 1 or 0 vs. 0) divided by the total number of matches and mismatches. A related coefficient called the coefficient of Jaccard (S_J) omits zero matches from both the numerator and denominator. The coefficient S_m was examined by Moore and Russell (1967) but was applied to continuous variables by coding data in rank intervals. This involves considerable loss of information content, and so is not recommended. It was also used for classifying soil bacteria by Brisbane and Rovira (1961). The coefficient of Jaccard was used by Varty and White (1964) to classify montmorillonite clay.

3. *Correlation Coefficients*

The product moment correlation coefficient (r) has also been used as a similarity coefficient in soil studies (Cipra *et al.*, 1970; Cuanalo and Webster, 1970; Dryden, 1935; Moore and Russell, 1967; Moore *et al.*, 1972; Russell and Moore, 1967). However, the use of r in this way has several disadvantages in soil classification. It is a measure of pattern rather than of magnitude of difference; for example if the values of X_{ij} are 1, 2, and 3, and of X_{ik} are 7, 8, and 9, then $r_{jk} = 1.0$ indicating the individuals j and k are identical, while they evidently differ in terms of magnitude. Another problem arises in ranked multistate variables, in that the direction of scaling may influence the sign of the correlation coefficient as pointed out by Eades (1965).

4. *Comparison of Similarity Coefficients*

Moore and Russell (1967) compared the results of using five different similarity coefficients, namely, (1) simple matching coefficient (S_m), (2) mean character difference (MCD), (3) Euclidean distance (d), (4) the Canberra metric (d_c), and (5) correlation coefficient (r). They concluded that the Euclidean distance is probably most appropriate for soils because it is sensitive to magnitude, is metric, and provides a model that can be readily visualized. Webster (1975) criticizes the Euclidean distance measure because of its sensitivity to magnitude, in that a single large difference makes a disproportionate contribution to the calculated dissimilarity between pairs of individuals. This might be true in the case where only a very few soil properties are used; but where a larger number is used, say 15 or more, then the contribution of a single large difference is less and appears to the author to be appropriate for soil classification.

Euclidean distance is especially useful for data containing all continuous variables, and can be used effectively for mixed data containing both continuous and multistate variables. For data consisting mainly of 2-state variables, the simple matching coefficient (S_m) or the coefficient of Jaccard (S_J) would be more appropriate.

E. SORTING STRATEGIES

Generally, the matrix of pair-wise similarity coefficients produced from the analysis of the data matrix is very large as the number of pairs is equal to $n(n-1)/2$, where n is the number of individuals. Thus the similarity matrix usually cannot be adequately interpreted by simple visual inspection. A large and confusing array of sorting strategies have been developed for describing the

pattern of relationships within the matrix. At least ten different methods have been applied to soil data of which only the more commonly used will be described in any detail.

Sneath and Sokal (1973) describe the various methods and combinations of methods for sorting a similarity matrix in "Taxonomic Structure." The most commonly used procedures in soil studies are included under systems called "sequential, agglomerative, hierarchic, nonoverlapping clustering methods" by Sneath and Sokal. In these procedures, the first step is a search of the matrix for the most similar pair (or pairs) of individuals (maximum similarity or minimum difference) which are joined to form the initial cluster(s). Then the matrix is reexamined for maximum similarity among remaining pairs or individuals or between individuals and prior clusters or between pairs of prior clusters, and the most similar pairs are joined to form new or enlarged clusters. This process is repeated until all individuals are accounted for, and the degree of similarity at each combination recorded. The results are generally presented in the form of a dendrogram or a phenogram, as it is called by Sneath and Sokal.

This seems to be a perfectly straightforward method, but there are a variety of algorithms available for the definition of maximum similarity between clusters of individuals and clusters.

1. *Single Linkage or Nearest Neighbor Clustering*

This method uses the criterion for joining based upon the two most similar individuals between the two clusters. This procedure has the undesirable characteristic of frequently leading to long, wandering clusters; a result that is referred to as "chaining." Nevertheless, single linkage clustering has been used by a few soil scientists, namely Anderson (1971), Rayner (1966), and Muir *et al.* (1970), and it has been compared with other methods by Moore *et al.* (1972).

2. *Complete Linkage or Farthest Neighbor Clustering*

This method forms clusters directly opposite in character to those formed by nearest neighbor sorting. It is based upon the similarity of the least similar pairs of individuals in the two clusters. It forms tight, hyperspherical clusters which join others generally at low levels of similarity and often leave a number of isolated individuals.

3. *Average Linkage Clustering*

This most commonly used clustering methods in soil studies are intermediate between the extremes of the two methods described previously and are forms of arithmetic average linkage methods. The simplest form is the unweighted pair-group method using arithmetic averages (UPGMA) and was first used by Rohlf (1963). Individuals or clusters are joined on the basis of the average similarity between all pairs of individuals in one cluster and those in another cluster, after the initial pair or pairs are joined. This method is thus based upon average between-group differences and produces results very similar to the more exact centroid sorting described below. It has been used in soil studies by Anderson (1971), Barkham and Norris (1970), Bidwell and Hole (1964), Bidwell *et al.* (1964), Cipra *et al.* (1970), Cuanalo and Webster (1970), and Sarkar *et al.* (1966).

Unweighted pair-group centroid method (UPGMC) is similar to the UPGMA method except that it is based upon the centroid of each cluster, which is calculated from the original data considered as coordinates in n-space for n variables or from the scores of individuals in n-space for n-dimensions derived from analysis of variables. The centroid is defined as the average coordinates of the members of a cluster. For soil studies this is an attractice method because the centroid of a group of soils is conceptually akin to the modal soil concept of the soil series, in which the modal soil is considered to be one near the centroid of the members of that soil series. It also can be represented accurately in two or three dimensions using 2 or 3 standardized variables or factor score dimensions. UPMGA usually gives similar results but it lacks the latter advantage. Unweighted centroid sorting has been examined by Anderson (1971), Campbell *et al.* (1970), Cuanalo and Webster (1970), and Moore and Russell (1967).

Weighting can be applied to the calculation of either the arithmetic average or the centroid method. These are called "weighted pair-group method using arithmetic averages" (WPGMA) and "weighted pair-group centroid method" (WPGMC), respectively, by Sneath and Sokal (1973). In these procedures the individual most recently added to a cluster is weighted equally with all previous members of the cluster. The WPGMA method was used by Grigal and Arneman (1969) but they gave no reason for its selection. There seems to be no rational purpose in using this kind of weighting for soils.

4. *Variable Group Clustering*

Rather than permitting clusters to form by pairs of individuals and/or clusters, it is possible to allow several individuals and/or clusters to join at a single step in the procedure. This requires that an arbitrary criterion level of similarity be specified at each step in the clustering for joining individuals or clusters. The criterion may be based on the change of average within-group similarity introduced by the merger after the first step in which initial clusters are formed by joining individuals above another specified criterion level of similarity.

Variable-group centroid sorting is also amenable to an iterative procedure which converges on a stable configuration. As individuals are added to initial

clusters formed with specified within-group similarity, the centroids are shifted, possibly in the direction of individuals on the outer fringe of other clusters. So after the first clustering, the centroids are established and all individuals are reallocated to the centroid which has the most similar coordinates (closest in n-space). This process is reiterated until the centroids remain stable. Usually, only a few iterations are required. Unweighted variable-group centroid sorting with iteration was used by Arkley (1968, 1971); criteria for joining were based upon change in within-group variance. Cuanalo and Webster (1970) compared a weighted variable-group method with weighted and unweighted pair-group methods and found the results to be very similar for the three sorting strategies. Sokal and Sneath (1963) describe the variable-group methods, but in their later book (Sneath and Sokal, 1973) they indicate that there is little to choose between the methods and so omit the variable-group procedures, mainly because they are more difficult to program for the computer than pair-group methods.

5. *Flexible Sort Clustering*

Lance and Williams (1967a) developed a general formula for SAHN clustering procedures in the form

$$D_{(i,j),k} = a_i D_{i,k} + a_j D_{j,k} + bD_{i,j} + c|D_{i,k} - D_{j,k}|$$

where D is a measure of difference or dissimilarity, i and j are a joined pair of individuals or groups, and k is a candidate for joining the group, a (alpha) is a parameter that may be set at $\frac{1}{2}$ or a function of the numbers of individuals (t) included in i,j, or k such as

$$a_i = t_i/t_{i,k} \quad \text{and} \quad a_j = t_j/t_{j,k}$$

as in UPGMA and UPGMC; c is generally zero except that it is equal to $-\frac{1}{2}$ for nearest neighbor and $+\frac{1}{2}$ for farthest neighbor sorting; b is usually zero except in unweighted centroid sorting (UPGMC) where b is equal to $-a_i a_j$ and D is squared Euclidean distance.

Lance and Williams (1967a) proposed a method which they called Median sorting in which $a_i = a_j = \frac{1}{2}$, $b = -\frac{1}{4}$, and $c = 0$. This places the centroid of the combined group midway between the centroids of i and j irrespective of the number of members in i and j rather than nearer the larger of group i or j as in true centroid sorting.

Flexible sorting was applied as above with $b = -\frac{1}{4}$ by Campbell *et al.* (1970), Crichton (1975), Moore *et al.* (1972), Russell and Moore (1967, 1968). Moore and Russell (1967) used another form of flexible sorting in which $a_i = a_j$, $b = 1 - (a_i + a_j)$, and $c = 0$. This procedure was intended to produce especially tight clusters of high within-group similarities. Anderson (1971) also used a "minimum variance clustering" which minimizes the within-group variance in which

$$a_i = (t_i + t_k)/(t_i + t_j + t_k), \quad a_j = (t_j + t_k)/(t_i + t_j + t_k), \quad b = 1 - (a_i + a_j)$$

and $c = 0$. The clustering algorithm between cluster k and the combined cluster (i,j) is thus

$$D_{k,ij} = [(t_i + t_k)D_{ki} + (t_j + t_k)D_{ki} - t_k D_{ij}]/(t_i + t_j + t_k)$$

where t is the number of individuals in clusters i, j, and k. This author points out that the results can be presented as an analysis of variance, because the sum of squares being subdivided is that due to the samples + samples × attributes effects.

6. *Information Content Clustering*

A clustering technique based upon information theory has been used by Norris and Dale (1971) and by Moore *et al.* (1972) applied to all two-state variables (transition matrix). Clustering is based upon the minimum information gain upon the fusion of two pairs or groups of individuals. The method is explained in detail in Sneath and Sokal (1973); however, they indicate serious reservations with regard to the assumptions on which the method is based. Since most soil data include continuous and ranked multistate variables, the method is not generally applicable in soil classification (see also Section IV,B,3).

7. *Sorting Strategies: Divisive*

Divisive sorting systems have been applied to soil classification in only a few cases. Divisive systems begin with the whole population, and, progressively divide it into smaller and smaller groups using the similarity matrix. Traditional categorical soil classification proceeds in a similar way, generally making separations using a single dichotomous criterion for separation (monothetic) or perhaps two or more criteria simultaneously (polythetic).

Norris (1972) used the technique of "Association analysis." The variables are analyzed for covariance and the first is selected which has highest communality with other variables which can be computed from the maximum sum of r^2 for a variable compared to each other variable ($\Sigma_{j=1}^{n} r_{i,j}^2$). Thus this is the variable with the highest predictive value. The population is divided into two classes using this variable as a criterion. Thereafter each subset is further subdivided by the same procedure. The process is terminated when a specified number of classes have been formed or when the maximum sum of r^2 falls below a specified level for all remaining variables.

Karmeli *et al.* (1968) devised a procedure for specific design purposes similar to "Association analysis," which was called "Maximum Uniformity Classifica-

tion." The method suggested uses separation on a primary variable with high communality together with minimum variance and *t*-tests and discriminant functions to increase the separation of groups.

8. *Comparison of Sorting Strategies*

Using the same similarity matrix as Muir *et al.* (1970), Anderson (1971) compared five sorting strategies: nearest neighbor, farthest neighbor, group average (UPGMA), centroid (UPGMC), and minimum variance clustering. He concluded that the 63 soil profiles sampled from 4 soil series were classified most effectively by farthest neighbor and minimum variance clustering.

Campbell *et al.* (1970) compared centroid, flexible ($b = -0.25$), and median sorting and concluding that the flexible sort ($b = -0.25$) appeared to be the most effective in forming groups, but the appearance of the dendrogram resulting from centroid sorting appeared to be equally effective to this writer.

Cuanalo and Webster (1970) compared unweighted pair-group (UPGMA), weighted pair-group (WPGMA), and weighted variable-group methods. They concluded that all gave similar results and printed only the results from UPGMA.

Moore and Russell (1967) compared nearest neighbor, farthest neighbor, centroid sort and flexible sort (see above). They found the typical chaining effect of nearest neighbor sorting, but the other three produced dendrograms in which 7 groups could be consistently detected in the 28 soils used.

Examination of these comparisons suggests clearly that in soil studies, the single-linkage (nearest neighbor) sorting with its strong tendency to form long chains rather than tight spherical clusters is to be avoided. Unweighted pair-group or group average clustering (UPGMA), unweighted centroid sorting (UPGMC), and minimum variance clustering are appropriate for soil studies as they produce moderately "tight" initial clusters of reasonable homogeneity and the central "core" of the cluster can be identified. However, as pointed out by Moore and Russell (1967), clustering and its representation in a dendrogram provides a two-dimensional representation of a multidimensional system (*n*-space) and clearly must contain distortion of the true configuration. They also point out that distortion is least at the lower levels of the hierarchy produced; thus the initial groups formed have the most validity, larger groups formed by merging smaller groups the least. Thus groupings formed at the higher levels should be treated with caution.

F. PRESENTATION OF RESULTS OF SORTING PROCEDURES

A similarity matrix can be shown effectively in a shaded similarity matrix as in Fig. 1 provided the indiviudals are ordered in such a way that the highest

similarities are at or near the diagonal. However, if the number of individuals is large, manual sorting to produce the optimum order may be very difficult. In the example shown in Fig. 1, the order was according to the results of a clustering procedure (Fig. 2).

By far the most common visual presentation of the results of the similarity matrix sorting procedures is in the form of a dendrogram such as that shown in Fig. 2. In a dendrogram, the individuals are indicated by symbols at the end of each branch, and the length of the branches are proportional to the degree of dissimilarity; the scale of dissimilarity used is often shown, although in this case it was not. Dendrograms have the advantage of being readily interpreted, but it should be remembered that they are two-dimensional representations of a

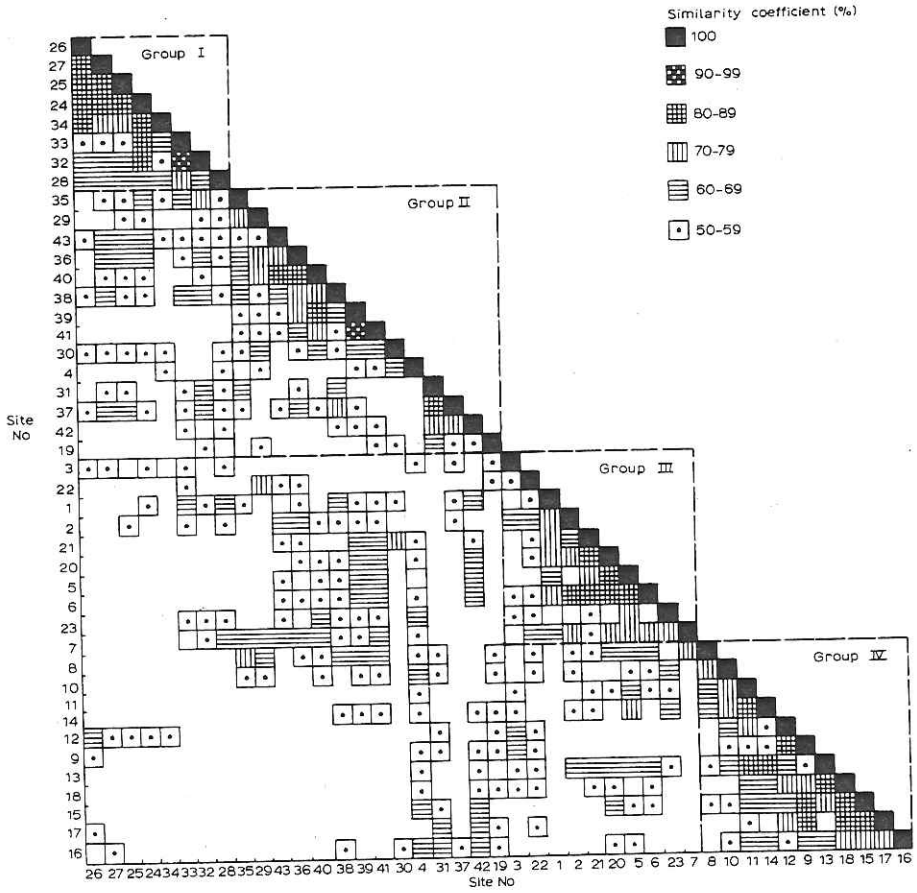

FIG. 1. Shaded similarity matrix for 43 soil bodies. Reprinted from Russell and Moore (1967) by permission of the publisher, Geoderma.

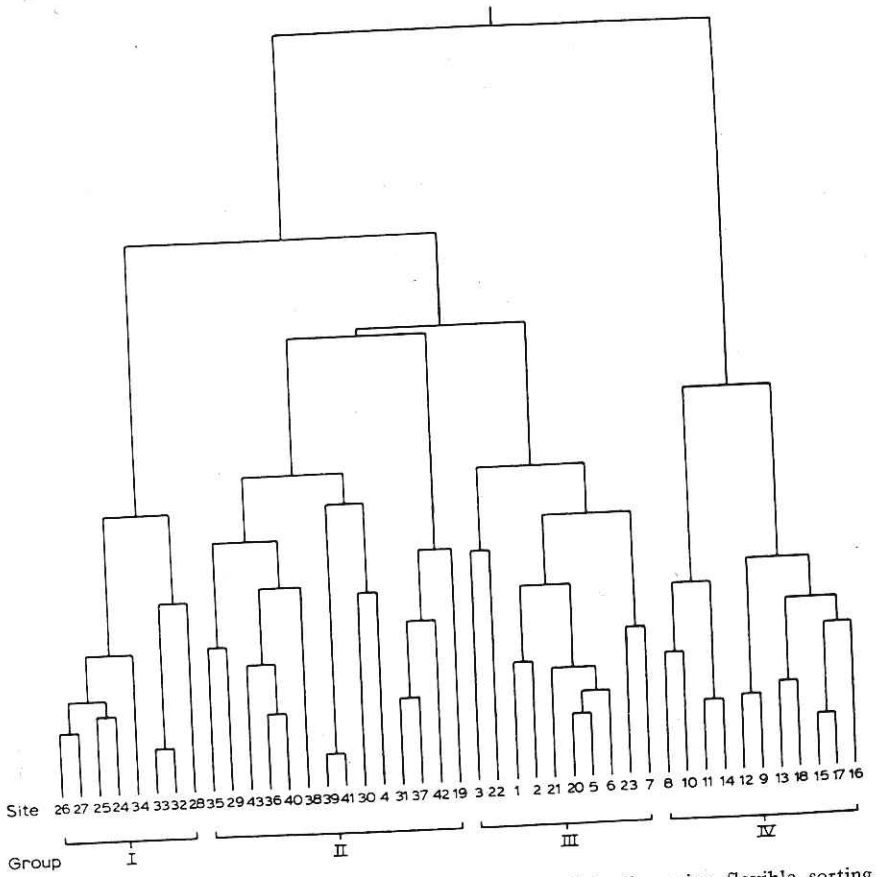

FIG. 2. Dendrogram showing relationships between soil bodies using flexible sorting procedure ($b = -0.25$). Reprinted from Russell and Moore (1967) by permission of the publisher, Geoderma.

multidimensional system and so suffer some distortion. Sometimes some of the distortion can be relieved by changing the order of the initial clusters so as to bring the nearest members of two clusters adjacent to each other.

III. Ordination of Soils

Another approach to the examination of taxonomic structure is through ordination, which is normally used when the distribution of indiviudals in n-space tends to be continuous rather than in distinct clusters. Thus some form of ordination seems to be appropriate for soils when large numbers of individuals

are included, if one considers the universe of soil individuals to be represented by one large hyperspherical swarm of points in n-space with n corresponding to the numbers of independent factors or processes involved.

Two kinds of ordination have been used in soil studies. Q-type ordination or analysis of objects operates on a similarity matrix of objects and examines the distribution of objects in object-space to find dimensions along which the objects are distributed. R-type ordination or analysis of variables operates on the covariance matrix variables to find dimensions in variable-/space. These dimensions are formed by clusters of covariant variables. Objects then can be located in varible-space by their coordinates along the axis of these dimensions.

A. Q-TYPE ORDINATION

Ordination of this type was used in one of the first attempts at numerical classification of soils by Hole and Hironaka (1960) and Bidwell and Hole (1964); however, the method is suboptimal as shown by Rayner (1969).

The method of principal coordinates analysis (PCO) developed by Gower (1966) operates best on a Euclidean distance matrix (dissimilarity coefficients) but may be used with other similarity coefficients. The procedure is discussed in detail by Webster (1975) who points out that PCO gives a good view of the general structure of a population but that distances between indiviudals are imperfectly represented. This is just the opposite of cluster analysis; thus PCO is complementary to cluster analysis in exploring relationships in $\hat{n}$-space. A problem of representing the results of PCO comes from the fact that for soils a number of principal axes are often produced, whereas only two, or at the most three, can be shown in a single diagram. Thus it may take a number of diagrams, each showing a pair of principal coordinates to represent the results. The interpretation of such two-dimensional scatter diagrams must be done with care, because two individuals which appear to be closely similar in one diagram, may actually be at an considerable distance along another axis in another diagram. Another problem is that variation along a principal coordinate is not always interpretable as to its source; i.e., variation due to differences in reaction, clay content, etc.

The first use of this approach was by Rayner (1966) who utilized the method of Gower (1966); and it was again used in a later publication by Rayner (1969). Anderson (1971) used Rayner's similarity matrix and applied a Q-type technique called "Nonmetric Multidimensional Scaling," which uses the rank order of the elements of a similarity matrix rather than the raw matrix. Muir *et al.* (1970) also used PCO on 63 soil profiles of four soil series in Scotland. They plotted the individual soils on the first two coordinates but found that three of the four soil series could not be well separated. This is not surprising since the first two

coordinates accounted for only about 28% of the total variance. Norris and Dale (1971) used a slightly modified form of PCO as part of a procedure for comparing the classifications derived from two sets of data (field and laboratory) on the same soils and found a high degree of correlation between the two; canonical correlations between the two sets for the first three principal coorinates were 0.89, 0. 62, and 0.58, respectively.

B. R-TYPE ORDINATION

1. *Principal Components Analysis*

The most commonly used R-type procedure is the Principal Components Analysis (PCA) which operates on the matrix of correlation coefficients (r) between all pairs of variables. It extracts orthogonal, independent dimensions from the data which are called principal axes. The coordinates of individuals on these axes are linear combinations of the original variables. Commonly, as few as three principal axes will account for a large portion, as about 75%, of the original total variance of the data matrix.

Like Principal Coordinates Analysis (PCO), it provides an accurate representation of the relationships between major groups and clusters, but is less accurate in reproducing differences between closely similar individuals. Also like PCO, it is often difficult to interpret the significance of an axis (dimension) in that it is made up of a combination of all variables. Sometimes an axis may be interpreted rather generally by examination of the so-called factor loadings of the variables on the dimension. However, a three-dimensional representation of the individuals plotted in relation to the three axes gives a general picture of the taxonomic structure. More importantly, the coordinates for individuals in each dimension (factor scores) can be used as the data for one of the clustering methods, rather than the raw data matrix. This greatly reduces the number of variables used in clustering.

Principal components analysis has been used in a number of soil studies: Anderson (1971) applied this method to the data of Rayner (1966) and then refined the results by examination of the residuals (residual variance) on each individual after extracting the variance represented by two dimensions; he also applied another procedure called "minimization of a quatradic loss function," which is a procedure for extracting the best minimum number of dimensions from the total number.

Barkham and Norris (1970) used principal components analysis on both the vegetation and soils and examined the realtionships between the two by comparing the first two vegetative components with four soil components and various soil variables and by canonical correlations between components of the two

systems. They considered this procedure to provide a workable strategy to investigate a complex ecosystem.

Cipra *et al.* (1970) used PCA together with cluster analysis to study relationships among 59 world soils. A three-dimensional diagram based on the first three principal components revealed no very distinct clusters, but a number of remotely related individual soils. The three axes could be interpreted as composites of various soil properties but these did not appear to be very logically related. The first component contained high factor loadings on chroma and clay content. However, Arkley (1971) using factor analysis on these same data found chroma and clay content to fall into separate, independent dimensions. Still, Cipra *et al.* (1970) considered this a meaningful method of visualizing relationships between soils.

Cuanalo and Webster (1970) also used both principal components with projection of individuals on the first two dimensions together with clustering methods. They concluded that soil data be first examined by ordination by principal componenets before attempting a classification of a set of soils.

Norris (1971a) applied PCA to two groups of soils using two sets of date: laboratory and field for each group. From the examination of two-dimensional diagrams, they concluded that the variation of soils can be characterized by the variation of a relatively few properties, and that there is a considerable correspondence between groupings based upon field and laboratory data separately. They suggest that where this is generally the case, soil classification might well be based on a relatively few dimensions defined by a few soil variables, but rarely by one singly. Norris (1972) applied the results derived above (Norris, 1971a) to applied problems in soil mapping and classification and as an aid in understanding the causes for soil variation.

Norris (1971b) also used PCA to assist in the solution of a statistical problem which arises in the use of soil data resulting from the fact that many variables are included. This problem is called matrix singularity and may result from a variable being completely determined by one or more other variables (i.e., there is a high degree of correlation among groups of variables) or when the matrix is overdefined (i.e., there are more variables than individuals). Thus the use of PCA reduces the large number of variables to a few components or dimensions, and the overdefinition is relieved with a minimum loss of information.

2. *Principal Factor Analysis*

Principal factor analysis (PFA) has been used much less frequently than principal components analysis in soil studies. PFA differs from PCA in that in the latter the diagonals of the correlation matrix of variables are filled with unities, whereas in PFA the unities are replaced by so-called communalities, the percentage of variation due to the common factors. In PCA the axes are

orthogonal, whereas in factor analysis they can be rotated and can be permitted to become obliquely related, i.e., correlated. This is particularly suited to soil data, which are often highly intercorrelated, as discussed earlier in this paper.

Arkley (1968, 1971) used a form of factor analysis described by Tryon and Bailey (1970) which they call Cumulative Communality Cluster Analysis of variables (CC5), wherein each dimension is defined by a minimal number of definer variables which are both intercorrelated and have closely similar patterns of correlations with other variables in the analysis; at the same time the dimensions are held to be as independent of each other as possible. This has the great advantage that the dimensions produced in this way are defined by one a few variables, rather than by factor loadings on all variables, and thus are readily identifiable. Arkley (1971) analyzed the data for 59 soils, used also by Cipra *et al.* (1970), by both PCA and PFA and obtained five dimensions which were identical except for the number of definer variables. The exact comparison was made possible by a program by Tryon and Bailey (1970) called cluster summary analysis (CSA2); this program extracted definer variables from PCA which could be compared with those from CC5. The computer program package developed by Tryon and Bailey is a tremendously powerful tool, as it contains a wide variety of options including various methods of principal component and factor analyses which can be applied with great simplicity for purposes of comparison. Also, the output contains a thorough set of statistical parameters for analytic purposes. The package also contains an iterative centroid clustering program for indiviudals, which can be modified at will.

Arkley (1971) used the programs described to analyze six different sets of soil data and found that five dimensions accounted for 85% of the squared raw correlations in 220 California soils, 87% in 620 California soils, and 72.1% in 59 world soils; seven dimensions accounted for 75.5% of the squared raw correlations in 148 Ohio soils, and 80.9% in 86 world soils. In every case only about three or four definer variables were required for each dimension. The evidence is clear that relatively few, say 20 to 25, variables are sufficient for soil classification.

C. PRESENTATION OF RESULTS OF ORDINATION

Ordination can be presented as two- or three-dimensional scatter diagrams of the component or factor scores of the individuals along each axis. A typical two-dimensional scatter diagram is shown in Fig. 3 in which the two axes are the first and second components of a principal components analysis (PCA). However, in interpretation of such a diagram it must be kept in mind that individuals that appear to be in close proximity on the diagram, may actually be separated at considerable distance along the axis of a third component.

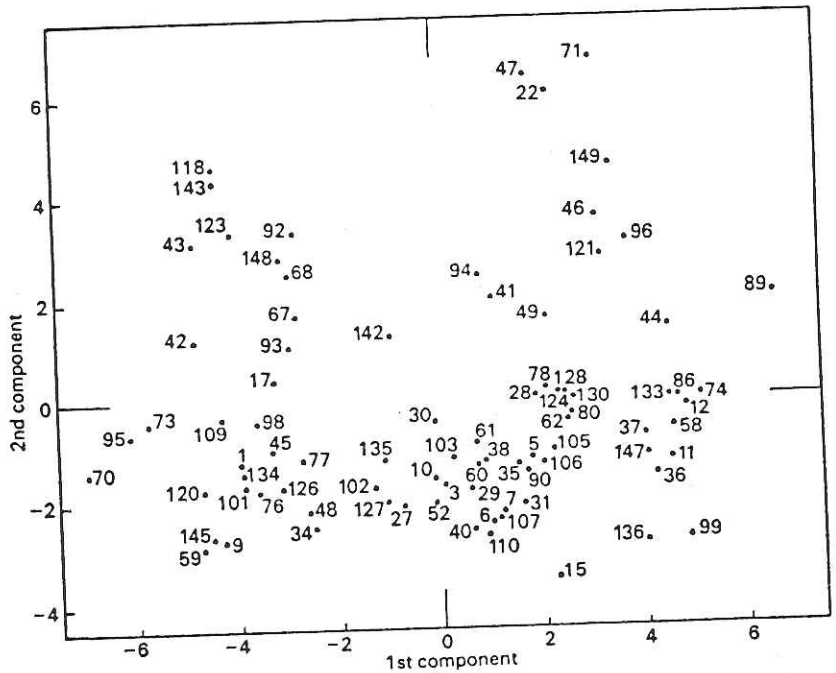

FIG. 3. Scatter diagram in which the first principal component is plotted against the second for 85 soil profiles. Reprinted from Cuanalo and Webster (1970), by permission of the publisher, Oxford University Press.

Three-dimensional diagrams are more difficult to prepare than two-dimensional ones, but can be prepared using a computer; and these can represent relationships more completely. The pin and ball diagram, a commonly used type, is shown in Fig. 4 and a rotated view of the same is depicted in Fig. 5. Again it should be kept in mind that a fourth component may well separate individuals which appear in close proximity. Where a number of independent factors or components are needed to represent the taxonomic structure, a number of diagrams may be required to present a complete picture of the structure.

IV. Soil as an Anisotropic Entity

Soil descriptions and analytical data have been treated in several different ways in an attempt to solve the problem of homology, i.e., the comparisons of profiles, layers, or horizons. It is sometimes difficult to decide which horizons should be compared between soils with different sets of horizons designated in the conventional way by soil scientists. Comparisons based upon identical depths

Context : soil profile as group

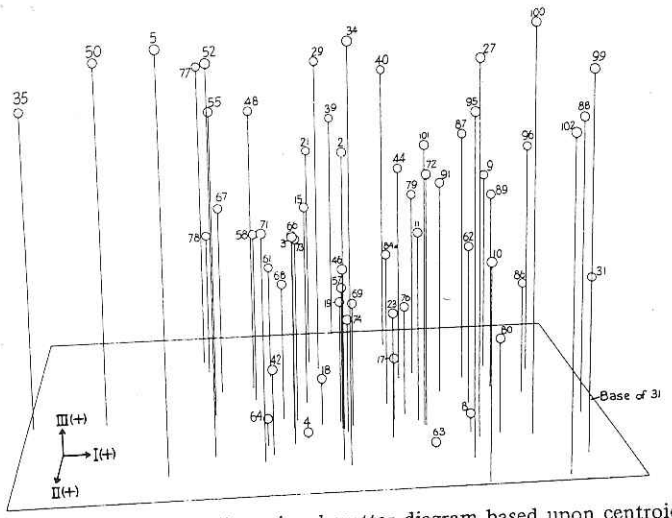

FIG. 4. Front view of a three-dimensional scatter diagram based upon centroid component analysis projections. Reprinted from Cipra *et al.* (1970), by permission of the publisher, American Society of Agronomy.

are often inappropriate due both to inherent differences in thickness in the soil or to depths altered by erosion.

A. SOIL PROFILE AS AN ARRAY OF SOIL PROPERTIES

One approach that has been used considers the soil profile as an entity, and the conventional designation of master soil horizons A, B, C, and R in the soil description have been used as a basis of comparison. Soil properties such as the color and structure of the A1 and B2 horizons, texture of the A1, B2, and C, and difference between maximum clay in the B and minimum clay in the A horizon have been used. This treatment has been applied by Bidwell and Hole (1964), Bidwell *et al.* (1964), Cipra *et al.* (1970), Hole and Hironaka (1960), Moore and Russell (1967), Rozkov (1974), Sarkar *et al.* (1966), Webster (1973), and Arkley (1968, 1971).

The use of soil "profile" data or soil horizon data has been criticized on the grounds that (a) the designation of horizons by the field scientist is subjective and thus violates one objective of numerical classification which is maximum objectivity, and (b) that it is subject to human error, even when the designation of horizons follows defined rules. This author is inclined to the view that the designation of horizons according to well-established rules such as those published in soil survey guides and handbooks or in "Soil Classification" (Soil

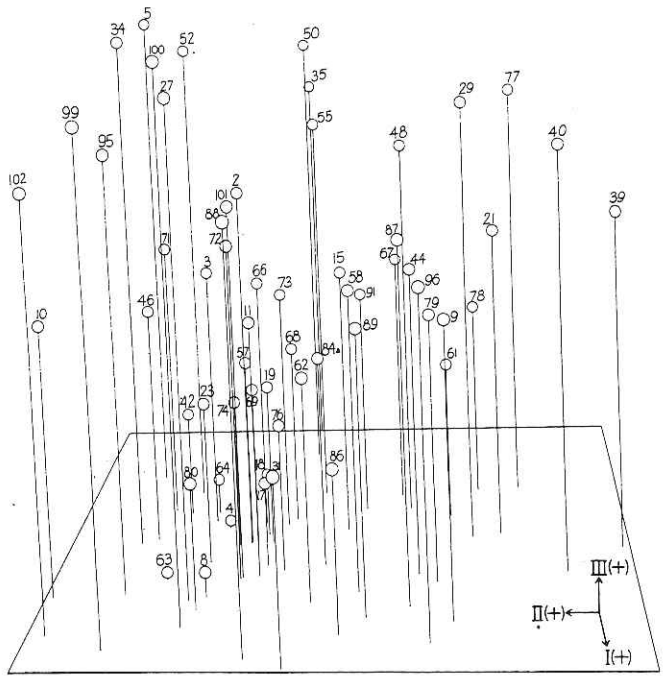

FIG. 5. Same as Fig. 4 viewed from the reader's right.

Survey Staff, U.S. Department of Agriculture, 1960) is sufficiently objective and accurate for the purposes of soil classification when carried out by a trained soil scientist. However, this is a personal view biased by my own long expeirence as a soil surveyor, and I have no wish to dispute the opposite view. The use of selected "profile" data or data by specified pedogenic horizons does have certain advantages, in that the mathematical treatment does not involve the analysis of a three-dimensional data matrix of variables X layers (or horizons) X soil individuals.

B. SOIL DATA BY LAYERS OR HORIZONS

1. *Uniform Soil Depth Layers–Direct Comparison*

Soil samples taken at specified depths and compared directly have been used by Barkham and Norris (1970), Cuanalo and Webster (1970), Hughes and Lindley (1955), Horris (1971a, 1972), Norris and Loveday (1971), and Webster and Burrough (1972). This method is appropriate only over small areas where

the arbitrary depths chosen can reasonably be expected to sample comparable portions of the soil profiles.

2. *Horizon or Layer Matching Methods*

Rayner (1966) devised a method of calculating similarity between soils by average of maximum similarity between all pairs of horizons; matching first the horizons of soil A with soil B and then soil B with A. There are two difficulties with this procedure. First, the amount of computer time is multipled by twice the number of horizons, and second, maximum similarity may well occur between surface horizons and deep horizons as in the C horizons, which seem inappropriate. This method was also used by Lamp (1972), and the similarity matrix of Rayner (1966) was also used by Anderson (1971) and Muir *et al.* (1970).

In order to remedy both objections to Rayner's method, Grigal and Arneman (1969) devised a similar method except that for a given horizon of soil A, comparisons are limited to three of soil B. The three are the horizons of soil B at the equivalent depth to the horizon of soil A plus the one just above and just below. The horizons of soil B are then compared to A in the same way, and the average of the maximum similarity between pairs used for comparison of the two profiles.

3. *Soil Profiles as a Sequence of Layers*

Norris and Dale (1971) developed a method based on a statistical device called a "transition matrix." In their procedure, all layers or horizons are grouped by one of the clustering methods into classes which are assigned an arbitrary number as a designation. Each profile then is described by a sequence of horizon-type numbers. Thus for a 10-layer profile in which the first 3 layers are of type 2, and the remaining layers type 5, the soil profile is defined as 2,2,2,5,5,5,5,5,5,5. Each soil profile then is redefined as a transition matrix which records the number of times each type-number follows every other type-number down the sequence. In the example given above, the transition matrix of $n \times n$ horizons types for that soil would contain all zeros except for matrix position (2,2) would contain a 2, position (2,5) would contain a 1, and position (5,5) would contain a 6. Soil profiles defined by transition matrices can be sorted on the basis of minimum information gain statistic of Dale *et al.* (1970). Norris and Dale (1971) claim several advantages to the method: (1) horizons or layers are conveniently and objectively classified, (2) the transition matrix includes information about relative position, and (3) the number and thickness of layers need not be identical. A major difficulty is the amount of computer time and computer storage space involved. This author agrees that the first advantage mentioned is a real one. However, I examined their matrices of profile × horizon types and found that a simple matching coefficient followed by UPGMA gave almost identical soil profile groupings as their much more complicated method. Also it appears to me that there is considerable loss of information in converting the horizon data into a transition matrix.

The transition matrix method was also used by Moore *et al.* (1972). They point out that grouping based upon the transition matrix is dependent entirely on the relation between each layer and the one preceding it.

C. SOIL PROFILE AS AN ARRAY OF DEPTH FUNCTIONS

Colwell (1970) introduced a procedure applicable to continuous variables such as chemical data. In this method, the depth function for each variable is characterized by fitting the data to an orthogonal polynomial. Colwell found that a polynomial of the fifth degree gives a reasonably accurate fit. The coefficients of the polynomial were used to characterize mean depth trends of four Great Soil Groups for 9 variables, and their confidence intervals at various depths. However, no classification of individual soils was attempted.

Campbell *et al.* (1970) used the coefficients of orthogonal polynomial depth functions as a means of characterizing soil profiles, but concluded that such a smooth depth function will not adequately describe the profile, and did not use the technique further. Moore *et al.* (1972) used orthogonal polynomial coefficients to represent shape and profile means to represent magnitude as variables for comparing soil profiles. They found wide variation in goodness-of-fit between profiles and between soil properties. However, when the profile mean was weighted equally (5 ×) to the 5 polynomial coefficients, they found profile groups similar to those produced by layer by layer comparison weighted by a negative exponential depth function ce^{-cy} where y is depth in centimeters and c is a constant. They adopted $c = 0.02$ as most appropriate.

V. Statistical Methods for Comparing Classification

As pointed out by Sneath and Sokal (1973), there is no general agreement on the optimal classification, except in cladistics where the optimal classification is one best representing the branching pattern of organisms through evolutionary history. For soils there is clearly no such criterion. However, it is possible to make some statistical comparisons which may be useful.

A. COPHENETIC CORRELATION

This procedure is described in detail by Sneath and Sokal (1973); it analyzes the level at which the individuals are joined to each other in a dendrogram, and compares that matrix of similarity with the original one by correlation coefficient between the two arrays of values. The higher the correlation coefficient, the more accurately the dendrogram represents the original pattern of similarities. This method was applied by Lamp (1972), and by Grigal and Arneman (1969). The latter compared different sets of data on the same soils and found the cophenetic correlation between field properties alone, and other data sets were 0.817 for comparison with all properties, and greater than 0.74 for three other data sets. Cipra *et al.* (1970) also applied cophenetic correlation to ordination by three principal components, and to dendrograms based on Euclidean distance and correlation coefficient; the correlations obtained were 0.78, 0.83, and 0.62, respectively.

B. COEFFICIENT OF ASSOCIATION

Grigal and Arneman (1969) used a method of comparing classifications developed by Goodman and Kruskal (1954) to compare numerical classifications with the Seventh Approximation Classification of the U.S. Department of Agriculture and found a low order of correspondence attributed to the differences in criteria used.

C. WILK'S CRITERION

Webster (1971) describes this method based on within group and between group variance extended to multivariate data. He illustrated its use comparing soil map classes at three different map scales and by comparing classes formed according to profile appearnace and by numerical methods. He concluded that the criterion can be used to compare classifications effectively. Webster and Burrough (1972) used Wilk's criterion to help determine the number of classes most appropriate to use in soil mapping over a small area and to determine the similarity between clusters and previously established soil series.

VI. Conclusions and Evaluation

From this welter of different methods employed, including different ways of treating soil profile data, different coefficients of similarity, clustering, and divisive methods, is it possible to decide which are the best methods to apply to soils? Although there is no criterion by which a classification can be judged to be the best, it does appear that some methods can be applied more logically to soils than others.

A. THE CHOICE OF METHODS

1. Data Selection and Standardization

First let us consider the kinds of data that have to be dealt with in soil classification. Field descriptions contain much information that is probably best treated as either ranked or unranked multistate variables. Laboratory data are primarily continuous variables. Thus most soil data sets contain mixed variable types. The data scaling method of Talkington (1967) with standardization of variables by range, and that of Burr (1968) with standardization by variance can accommodate mixed data. If Euclidean distance is used as the similarity coefficient, the results from either method will be about the same. Very rare extreme values on a variable should probably be truncated as advocated by Talkington (1967). Variables with highly skewed distributions should be transformed by logarithmic or square root transformation. The choice of variables to be used will depend, of course, upon the nature of the data available or the purpose of the classification. However, for a general classification my investigations (Arkley, 1968, 1971) suggest that a minimal set of soil properties should include at least one or several measurements representing the following dimensions:

1. Soil reaction such as surface and subsoil pH, carbonate depth, exchangeable Na or S.A.R.
2. Hue and chroma.
3. Texture or contents of clay and sand, or clay and silt, and gravel or stone content.
4. Soil color value such as thickness of surface layer with value of 3 or less, or color value per se.
5. Depth to and degree of mottling and/or other evidence of wetness or poor drainage.
6. Degree of profile differentiation such as difference in clay content between the surface and B-horizon or subsoil, clay films, and structure of the subsoil.
7. Solum thickness.

Most of these, as can be seen, are properties measurable in the field. Since the purpose of soil classification is related mainly to land use or plant growth in the field, it is relevant only to soil distributions that are mapped by field methods primarily, supplemented by laboratory analysis only to a limited degree. As indicated in the section (II,C,2) on data selection, soil variables in a data set are

often highly correlated. Norris (1971a), Grigal and Arneman (1969), Hole and Hironaka (1960), and Norris and Dale (1971) all found a high degree of similarity between the soils classified by numerical methods when the results of laboratory versus field data were compared.

2. *Similarity Coefficients*

Comparisons of two or more similarity coefficients are included in several papers including those of Moore and Russell (1967), Cipra *et al.* (1970), and Cuanalo and Webster (1970). These comparisons provide no basis for unequivocal recommendations, but it appears that the Euclidean distance is most suitable to soil data on theoretical grounds as well as its ease in conceptual representation. The correlation coefficient and the Canberra metric seem to be less suitable.

3. *Sorting Strategies*

Of the numerous sorting strategies available, nearest neighbor or single linkage sorting seems to be the least useful for soil classification because of its characteristic "chaining." Furthest neighbor or complete linkage sorting produces the least chaining but is conceptually less satisfactory for soils than centroid sorting (UPGMC) or average linkage (UPGMA). Centroid sorting requires a larger computer storage capacity, but it has the advantages or producing initial clusters of specified degree of within-group homogeneity and also can be applied iteratively to find a stable taxonomic structure. The minimum within-group variance clustering of Anderson (1971) is also an attractive form of centroid clustering as it can be presented as an analysis of variance.

4. *Ordination Methods*

Ordination per se does not provide a classification but does reveal relationships between individuals and groups when presented in one or several two- or three-dimensional scatter diagrams. Ordination by principal coordinate analysis (PCO), a Q-type analysis of the similarity matrix shows relationships between individuals but the axes are often difficult to interpret. Principal components analysis (PCA), an R-type analysis of the variables is more easily interpretable, but still the components are made up of variable weights on all variables, and further examination or analysis is required to identify the meaning of each component. Factor analysis (PFA), another form of R-analysis, can be used for ordination as well as a means of reducing the number of variables in the analysis, and the axis can be identified as to the nature of the dimensions they represent. In applying either PCO or PFA, the correlation matrix should be examined for

redundant variables (i.e., with very high correlation coefficients as > 0.95) and one of the pair eliminated before proceeding further.

B. A SUGGESTED PROCEDURE FOR A GENERAL SOIL CLASSIFICATION

The cluster analysis of soils described herein is based on the assumption that soils do indeed fall naturally into discrete clusters or classes. For large numbers of soils, as in an entire nation, a whole continent, or the world, this assumption may not be valid and soils may well form a continuum in multidimensional variable space. For the latter case, a procedure suggests itself which is based upon the idea of arbitrarily defined, well-separated centroids, each centroid thus representing an hypothetical or perhaps an actual modal soil. In order to define the centroids, each variable or dimension is divided into appropriate segments of its total range and the midpoint of each segment established as an arbitrary centroid. For example if we accept the meaningful range of soil clay content to be from 0 to 50, the segments would be 0–10, 10–20, 20–30, 30–40, 40–50+ and the mid-point centroids would be located at 5, 15, 25, 35, and 45% clay. With this approach, soil individuals would be allocated to the centroid to which it is nearest in n-space on the basis of Euclidean distance. For soils considered as a continuum, this has the greatest advantage over conventional heirarchical classification in that no dichotomous separation is made on the basis of a single variable separated at a single point along its range. For example the soils classified as Mollisols are separated from other soils at a high category on the basis of a specified thickness of a dark, organic rich surface soil. With this kind of dichotomy a small variation in this single property for two soils separates them regardless of the degree of similarity of all other properties. The allocation of soils by over-all similarity to definite centroids avoids this problem entirely, so that each class is distinct from every other in at least one dimension, and individuals within a class are all closely similar to the defined centroid.

It is interesting to consider the number of classes formed by this kind of procedure. The number depends upon the number of segments into which each variable is divided raised to the power of the number of variables as follows:

Number of Segments	Number of dimensions or variables					
	4	5	6	7	8	9
3	81	243	729	2,187	6,561	19,683
4	256	1,024	4,096	16,384	65,536	262,144
5	625	3,125	15,625	78,125	390,624	1.95×10^6

On the basis of intuition, it appears that seven dimensions divided into five segments each producing 78,125 classes of which perhaps 20% might remain unoccupied would be adequate for a classification of the soils of the world. However, it might be necessary to first stratify the soils such as separating organic soils from mineral soils, and perhaps separating soils of arid regions from those of humid regions as different soil properties might be used as criterion variables for classification of the two major kinds of soils. Soils of subhumid regions then might be classified according to both systems, thus avoiding a dichotomous separation between arid and humid regions. Factor analysis of the kind used by Arkley (1968, 1971) seems to be an effective way of finding a suitable number of independent dimensions for such a classification procedure.

This kind of system could be used for setting up a number of hierarchical arrangements depending upon the categorical order in which the dimensions are used, or it could be used as a coordinate system of classification. The latter would be particularly useful to show relationships among soils. For example, if we have five segments labeled VL, L, M, H, VH and seven dimensions, then a sequence of soils varying only in one dimension would be found in classes labeled such as:

VL, H, H, H, H, H, H
L, H, H, H, H, H, H
M, H, H, H, H, H, H
H, H, H, H, H, H, H
VH, H, H, H, H, H, H

Assuming all classes to be occupied, there would be 15,625 possible such sequences to be examined. These should very well keep the pedologist of the world occupied for some time.

Finally, it appears that cluster analysis of soils is a most effective means of classifying soils when the number of distinct soils or soil groups is limited as within a relatively small land area. For large areas including large numbers of soils, clusters of similar soils are likely to be either nonexistent or existing only with diffuse boundaries. In which case a coordinate system based upon predefined centroids is more likely to produce an effective classification system.

REFERENCES

Anderson, A. J. B. 1971. *J. Int. Assoc. Math. Geol.* **3**, 1–14.
Arkley, R. J. 1968. *Trans. Int. Congr. Soil Sci., 9th, 1968* Vol. IV, pp. 187–192.
Arkley, R. J. 1971. *Soil Sci. Soc. Am., Proc.* **35**, 312–315.
Avery, B. W. 1968. *Trans. Int. Congr. Soil Sci., 9th, 1968* Vol. IV, pp. 169–176.
Barkham, J. P., and Norris, J. M. 1970. *Ecology* **51**, 630–639.
Bidwell, O. W., and Hole, F. D. 1964. *Soil Sci. Soc. Am., Proc.* **28**, 263–268.

Bidwell, O. W., Markus, L. F., and Sarkar, P. K. 1964. *Trans Int. Congr. Soil Sci., 8th, 1964* Vol. V, pp. 933–941.
Bray, J. R., and Curtis, J. T. 1958. *Ecol. Monogr.* **27**, 325–349.
Brisbane, P. G., and Rovira, A. D. 19611. *J. Gen. Microbiol.* **26**, 379–392.
Burr, E. J. 1968. *Aust. Comput. J.* **1**, 97–99.
Campbell, N. A., Mulcahy, M. J., and McArthur, W. M. 1970. *Aust. J. Soil Res.* **8**, 42–58.
Cipra, J. E., Bidwell, O. W., and Rohlf, F. J. 1970. *Soil Sci. Soc. Am. Proc.* **34**, 281–287.
Colwell, J. D. 1970. *Aust. J. Soil Res.* **20**, 221–238.
Critchton, J. E. 1975. Ph.D. Thesis, Dep. Soil Sci., University of Sydney, New South Wales, Australia.
Cuanalo, H. E. de la C., and Webster R. 1970. *J. Soil Sci.* **21**, 340–352.
Dale, M. B., McNaughton-Smith, P., Williams, W. T., and Lance, G. N. 1970. *Aust. Comput. J.* **2**, 9–13.
Dryden, L. 1935. *Am. J. Sci.* **29**, 393–408.
Eades, D. C. 1965. *Syst. Zool.* **14**, 98–100.
Gibbons, F. R. 1968. *Trans. Int. Congr. Soil Sci., 9th, 1968* Vol. IV, pp. 159–168.
Goodman, L. A., and Kruskal, W. H. 1954. *J. Am. Stat. Assoc.* **49**, 123–163.
Gower, J. C. 1966. *Biometrika* **53**, 325–338.
Grigal, D. F., and Arneman, H. F. 1969. *Soil Sci. Soc. Am., Proc.* **33**, 433–438.
Hole, F. D., and Hironaka, M. 1960. *Soil Sci. Soc. Am., Proc.* **24**, 309–312.
Hughes, R. E., and Lindley, D. V. 1955. *Nature (London)* **175**, 806–807.
Karmeli, D., Pitkovski, G., and Regev, J. 1968. *Technion-Isr. Inst., Tech. Fac. Agric. Eng., Publ.* **50**, 1–10.
Lamp, J. 1972. D. Agric. Dissertation, Fac. Agric., Christian-Albrechts University, Kiel.
Lance, G. N., and Williams, W. T. 1967a. *Comput. J.* **9**, 373–380.
Lance, G. N., and Williams, W. T. 1967b. *Aust. Comput. J.* **1**, 15–20.
Moore, A. W., and Russell, J. S. 1967. *Geoderma* **1**, 139–158.
Moore, A. W., Russell, J. S., and Ward, W. T. 1972. *J. Soil Sci.* **23**, 193–209.
Muir, J. W., Hardie, H. G. M., Inkson, R. H. E., and Anderson, A. J. B. 1970. *Geoderma* **4**, 81–90.
Norris, J. M. 1971a. *J. Soil Sci.* **22**, 69–89.
Norris, J. M. 1971b. *Pedobiologia* **11**, 410–416.
Norris, J. M. 1972. *J. Soil Sci.* **23**, 62–75.
Norris, J. M., and Dale, M. B. 1971. *Soil Sci. Soc. Am., Proc.* **35**, 487–491.
Norris, H. M., and Loveday, J. 1971. *J. Soil Sci.* **22**, 395–400.
Rao, C. R. 1948. *J. R. Stat. Soc., Ser. B* **10**, 159–193.
Rayner, J. H. 1966. *J. Soil Sci.* **17**, 79–92.
Rayner, J. H. 1969. *Syst. Assoc.* **8**, 31–39.
Rohlf, F. J. 1963. *Ann. Entomol. Soc. Am.* **56**, 798–804.
Rozhkov, V. A. 1974. *Geoderma* **12**, 175–182.
Russell, J. S., and Moore, A. W. 1967. *Geoderma* **1**, 47–68.
Russell, J. S., and Moore, A. W. 1968. *Trans Int. Congr. Soil Sci., 9th, 1968* Vol. IV, pp. 205–213.
Sarkar, P. K., Bidwell, O. W., and Marcus, L. F. 1966. *Soil Sci. Soc. Am., Proc.* **30**, 269–272.
Sneath, P.H.A., and Sokal, R. R. 1973. "Numerical Taxonomy." Freeman, San Francisco, California.
Soil Survey Staff, U.S. Department of Agriculture. 1960. "Soil Classification." US Govt. Printing Office, Washington, D.C.
Sokal, R. R., and Sneath, P.H.A. 1963. "Principles of Numerical Taxonomy." Freeman, San Francisco, California.