Mathematical Geology, Vol. 3, No. 1, 1971

Numeric Examination of Multivariate Soil Samples[1]

A. J. B. Anderson[2]

Numerical methods for the examination of multivariate soil samples are presented in geometric terms. Techniques of coordinate representation by principal components, by nonmetric scaling, and by a new method are discussed, as are techniques for agglomerative hierarchic cluster analysis. These are illustrated by two sets of previously published data. KEY WORDS: classification, cluster analysis, principal components analysis, numerical taxonomy.

INTRODUCTION

Rayner (1966) described numeric methods for investigating relationships between soil samples with many known properties or attributes. Using similar techniques, Muir and others (1970) compared the classification of four soil series by numerical and traditional methods. The purpose of this paper is to draw the attention of soil scientists to more refined methods of clustering and coordinate representation using the data presented in those papers to exemplify results.

A GEOMETRIC FORMULATION

Suppose there are n soil samples on each of which m attributes have been measured. All the procedures to be discussed require the formation of a table or matrix of the $\frac{1}{2}n(n-1)$ similarities between pairs of samples. Rayner (1966) showed how such a matrix can be constructed and how a similarity between any two profiles can be derived from similarities between their constituent horizons. This definition of intersample similarity is not in any sense unique but does present a reasonable means of determining the "closeness" of two samples.

The concept of closeness is a geometric one, and implies that the n samples can be thought of as being represented by points $P_1, P_2, \ldots, P_n$ in Euclidean space even if some or all of the attributes are not quantitative. The validity of such an approach can be proved by the following argument.

J. C. Gower (personal communication) showed that Rayner's matrix S is positive semidefinite. Hence, there exists an $n \times p$ matrix X of quantitative elements such that

$$XX' = S$$

[1] Manuscript received 18 May 1970.

[2] Rothamsted Experimental Station. Present address: Medical Research Council, Western General Hospital, Edinburgh (UK).

Then X gives the coordinates of a configuration of points $P_1, P_2, \ldots, P_n$ in p-dimensional Euclidean space. In fact, there is an infinite number of such matrices, each representing a translation or rotation of the same configuration of points. In practice, p is usually equal to $(n-1)$. The distance d_{ij} between any two points P_i and P_j is given by

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{p} (x_{ik}-x_{jk})^2 \\ &= \sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - 2\sum_{k=1}^{p} x_{ik}x_{jk} \\ &= s_{ii}+s_{jj}-2s_{ij} \\ &= 2(1-s_{ij}), \qquad \text{because } s_{ii} = s_{jj} = 1 \end{aligned}$$

Because the configuration of a set of points is defined, except for rigid translations and rotations, by the interpoint distances, the geometric approach is valid, and $\sqrt{2(1-s_{ij})}$ represents the "distance" between samples i and j.

Although such a formulation gives a familiar and well-defined meaning to the observed sample values, the amount of data has not been reduced. To examine the structure of the data, or to form hypotheses, a more parsimonious summarization of the inherent information is required. We can provide this in two ways and these are discussed in the following sections.

COORDINATE REPRESENTATION

Numeric results are frequently presented in the form of a diagram, and it is reasonable to seek a means whereby relationships among soil samples with respect to many attributes can be similarly represented by low dimensional spatial configurations of points. Sometimes a single two-way scatter plot proves adequate but, frequently, more dimensions are needed and diagrammatic representations are correspondingly more complex. In general, outlying samples, clusters of similar samples and other more complex patterns of association may be detected by visual inspection of the coordinate representation, and instead of plotting points on such diagrams the values of observed attributes or of subsidiary variables can be given to help identify causal factors visually.

We have seen how the data can be represented by n points in a p-dimensional Euclidean space. We now seek a means of compressing most of this information into q dimensions, where q is small. In precise terms, the samples are to be represented in q dimensions by points $Q_1, Q_2, \ldots, Q_n$, and $d_{ij}(q)$ is the Euclidean distance between Q_i and Q_j. The following paragraphs discuss three approaches to this problem.

PRINCIPAL COMPONENTS ANALYSIS

This is the method used in the earlier papers and is so well known that we shall note only two points:

(1) The points $P_1, P_2, \ldots, P_n$ are projected perpendicularly onto a q-dimensional hyperplane.
(2) The orientation of the hyperplane is such that it accounts for the greatest possible variation. This is equivalent to minimizing the quantity

$$\sum_{i,\,j} [d_{ij}^2 - d_{ij}^2(q)]$$

These concepts are most easily understood by considering projection from a plane ($p = 2$) onto a line ($q = 1$).

Essentially, principal components analysis is applied to the configuration in p space defined by X. However, the computation can be shortened. Gower (1966) has pointed out that when the elements of S are adjusted to give S^* by subtraction of the row and column means and addition of the general mean, the derived configuration is unaltered (because $s_{ii}^* + s_{jj}^* - 2s_{ij}^* = s_{ii} + s_{jj} - 2s_{ij}$). Hence, if we choose the columns of X to be the latent vectors of S^*, each being scaled so that its sum of squares equals the corresponding latent root, we have $XX' = S^*$ and therefore the configuration $P_1, P_2, \ldots, P_n$ must necessarily be referred to principal axes. This is the procedure reported by Rayner, and use of the axes corresponding to the q largest latent roots gives the representation which is "best" in the sense that the maximum variation in q dimensions is accounted for.

Of the techniques to be described, principal components analysis is the simplest. Its main weakness is the possibility of poor representation of some samples even though the overall proportion of variation is high. Anderson (1970*a*) has discussed the examination of the residuals P_iQ_i in this type of analysis to help detect such anomalies. These are easily computed, because if G is the centroid of $P_1, P_2, \ldots, P_n$, then $P_iG^2 = s_{ii}^*$ so that $P_iQ_i = \sqrt{s_{ii}^* - Q_iG^2}$. For comparative purposes, the percentage contribution of any residual to the total residual variation can be expressed as

$$R_i^2 = 100 \times \left(P_iQ_i^2 \Big/ \sum_1^n P_iQ_i^2 \right)$$

Notice also that the cosine of the angle $P_i\hat{G}Q_i$ is given by Q_iG/P_iG.

MINIMIZATION OF A QUADRATIC LOSS FUNCTION

The requirement of orthogonal projection, which is basic to the previous method, leads to simple mathematical equations, but places an arbitrary constraint on the solution. Anderson (1970*a*) suggested that the main criterion of the adequacy of a coordinate representation must be the closeness with which $d_{ij}(q)$ approximates d_{ij} for all pairs of samples i and j. It is reasonable, therefore, to seek the representation that minimizes the quantity

$$L = \sum_{i,\,j} [d_{ij}(q) - d_{ij}]^2$$

or, more generally,

$$L_w = \sum_{i,j} w_{ij}[d_{ij}(q) - d_{ij}]^2$$

where the w_{ij} provide differential weights for the comparisons. For example, if more importance attaches to the accurate representation of large distances, we can use $w_{ij} = d_{ij}$.

The coordinates of $Q_1, Q_2, \ldots, Q_n$ which minimize L or L_w can be found only by iterative approximation from an initial configuration. If a Newton-Raphson technique is used, a matrix of order nq must be inverted at each stage and it seems more reasonable to employ a steepest descent method. More research is required on this, particularly as regards avoidance of local minima. The best scheme so far devised is as follows:

(1) Start with the principal components solution in $t > q$ dimensions. (For $q = 2$, t should be 4 or 5).
(2) Find the best solution in t dimensions.
(3) Find the best of the t solutions in $t-1$ dimensions derived from starting configurations obtained by dropping each axis in turn ($t = t-1$). Repeat until $t = q$.

NONMETRIC MULTIDIMENSIONAL SCALING

There are circumstances in which the requirement that $d_{ij}(q)$ should approximate d_{ij} can be replaced by the constraint that $d(q)$ should be monotonically related to d, that is, $d_{hi}(q) \geqslant d_{jk}(q)$ if $d_{hi} \geqslant d_{jk}$ for all h, i, j, and k. This is similar to the previous method, but is nonmetric in the sense that the values of the similarity coefficients are ignored and only their ranking is used. Such an approach is more robust to sampling errors and can be used with similarity measures that are not Euclidean.

Clearly, it may not be possible to find a representation in q ($<p$) dimensions such that the condition of monotonicity is exactly satisfied. Kruskal (1964*a*, 1964*b*) has defined a measure of nonmonotonicity (called "stress") that can be minimized to provide a "best" configuration. Once again, the computation is iterative. However, the computer storage required is approximately four times greater than is needed for the previous method, so that, in practice, the value of n is severely limited.

It should be noted that the stress function is not continuous and experiment shows that, unless the number of samples is large ($n > 30$, say) points may be subjected to appreciable displacements without any alteration in the minimum value. Indeed, the information lost by considering only the rank order of the distances is sometimes most important. For example, if the samples form two well separated main clusters, the histogram of the intersample distances is bimodal, the peaks corresponding to within- and between-cluster comparisons. This is ignored, and the resulting representation necessarily gives the appearance of a uniform continuum. In the same manner, the presence of outliers can be obscured.

EXAMPLE

As an example of the application of these methods to some real data, we have re-analyzed the similarity matrix given in Table 2 of Rayner's paper. This relates to soil samples of brown earths (BE), some of which are described as acidic [BE(A)] or lessivated [BE(L)], together with some gley soils and a rendzina (Rz) sample.

Figure 1 reproduces the two-dimensional representation obtained by a principal components analysis (i.e., Rayner's fig. 5), and Table 1 gives the corresponding residuals analysis. It will be seen from this that samples 8, 18, and 20 have large residuals, suggesting that these are placed misleadingly near the center. Samples 3 and 14 have small cosine values, but these are near the centroid and therefore cannot be poorly fitted. The internal evidence of Figure 1 confirms the indication that sample 8 is badly placed and the brown earths are not compact; the representation is otherwise reasonably satisfactory.

Figure 2 gives the two dimensional configuration derived by minimizing the sum of squares of $d(2)$-d. Sample 8 has now been "pushed out" into the gley soils region. The brown earths form a more compact set, but the two lessivated soils are separated owing to the improved position of sample 20.

Figure 3 shows the similar representation derived by nonmetric scaling. In the absence of clustering, this similarity is typical.

Of the three methods, the quadratic minimization procedure is to be preferred.

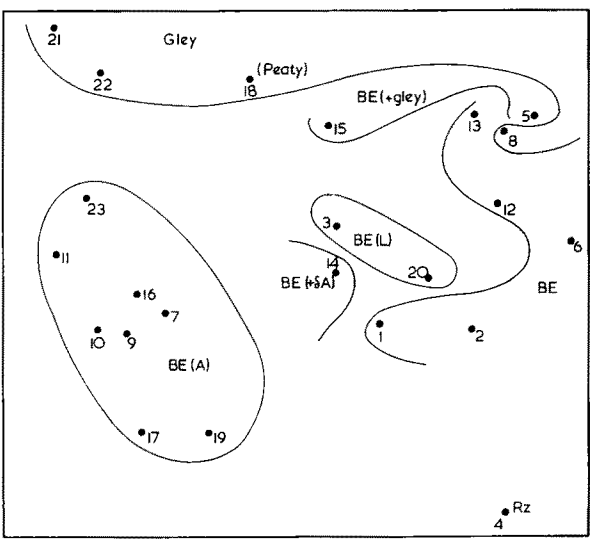

Figure 1. Principal components representation of Glamorganshire soils: BE—brown earth; BE(A)—acid brown earth; BE(L)—lessivated brown earth; Rz—rendzina.

Table 1. Analysis of Residuals from Two-Dimensional Principal Components Representation of Glamorganshire Soils

Sample(i)	P_iG^2	=	Q_iG^2	+	$P_iQ_i^2$	R_i^2	$\cos(P_i\hat{G}Q_i)$
1	0.2096		0.0252		0.1844	4.74	0.347
2	0.2199		0.0706		0.1494	3.84	0.566
3	0.2059		0.0038		0.2022	5.20	0.135
4	0.4114		0.2166		0.1948	5.01	0.726
5	0.2728		0.1370		0.1358	3.49	0.709
6	0.2607		0.1416		0.1191	3.06	0.737
7	0.2257		0.0396		0.1860	4.78	0.419
8	0.3309		0.1047		0.2262	5.81	0.563
9	0.2126		0.0674		0.1452	3.73	0.563
10	0.2696		0.0843		0.1853	4.76	0.559
11	0.2289		0.1047		0.1242	3.19	0.676
12	0.2319		0.0777		0.1542	3.96	0.579
13	0.2481		0.0901		0.1580	4.06	0.603
14	0.1881		0.0043		0.1838	4.73	0.151
15	0.1986		0.0293		0.1694	4.35	0.384
16	0.2238		0.0516		0.1722	4.43	0.480
17	0.2790		0.1077		0.1713	4.40	0.621
18	0.2716		0.0562		0.2154	5.54	0.455
19	0.2092		0.0782		0.1309	3.37	0.612
20	0.2572		0.0344		0.2228	5.73	0.366
21	0.3352		0.1974		0.1378	3.54	0.767
22	0.2879		0.1242		0.1638	4.21	0.657
23	0.2427		0.0843		0.1584	4.07	0.589
Total	5.8214		1.9308		3.8906	100.00	(mean)0.533

It is free from both the difficulties of interpretation inherent in principal components analysis, and the computational problems presented by nonmetric scaling. And, above all, the criterion to be satisfied seems closest to intuitive ideas about what a coordinate representation should provide.

HIERARCHIC CLUSTERING

The techniques so far discussed have been aimed at reducing the number of attributes required to describe the set of n samples. The following procedures, on the other hand, attempt to compress information on the intersample structure by finding groups of samples whose attributes are relatively similar. This does not imply that a causal mechanism for such divisions can then be isolated. In geometric terms, we must partition the p-dimensional configuration in such a manner as to form an unspecified number of clusters of relatively close points. There must be some quantitative indicator measuring whether one clustering scheme is better than another, but unless the number of samples is small it is impracticable to examine this for all possible divisions.

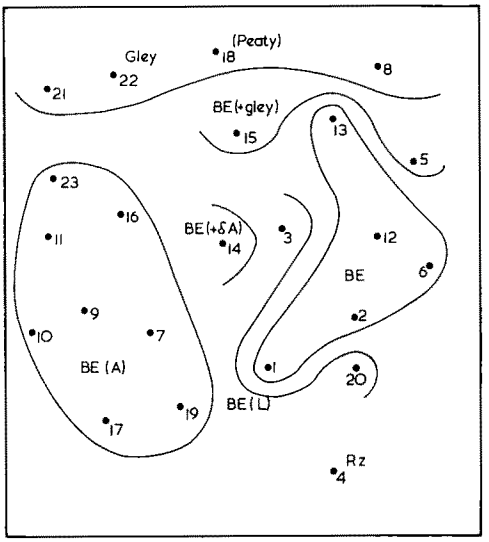

Figure 2. Quadratic loss function representation of Glamorganshire soils. Abbreviations same as Figure 1.

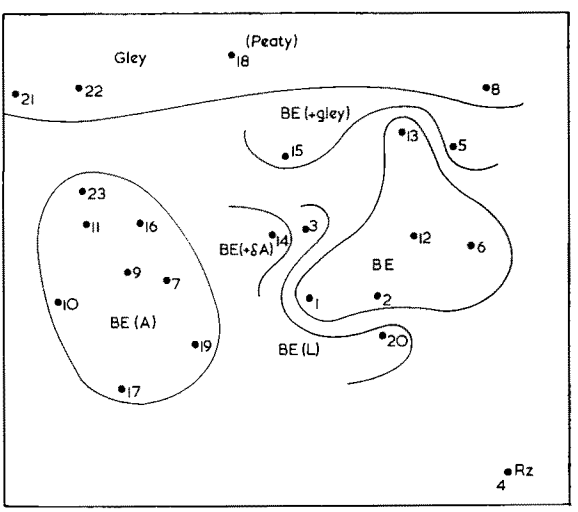

Figure 3. Nonmetric scaling representation of Glamorganshire soils. Abbreviations same as Figure 1.

One simplification is provided by hierarchic clustering, which may be divisive or agglomerative. Divisive methods such as those described by Anderson (1970*b*) are not entirely free from computational problems and we shall discuss here only agglomerative solutions, including that used by Rayner. For these, we require a measure D of the affinity of any two groups of samples. So that the discussion can be cast in geometric terms, we shall regard D_{ij} as the distance between groups I and J.

In agglomerative clustering procedures, n initial groups, each containing one sample, are successively amalgamated in $n-1$ steps, until all have been combined into one group. At each stage, the two groups chosen for fusion are those that are nearest according to the definition of D. These two groups are fused, and the distances of the new group from each of the remaining groups is calculated. To do this efficiently, we require some formula for $D_{k \cdot ij}$, the distance of group K from the union of groups I and J. Lance and Williams (1966) have pointed out that a general relation covering most distance metrics is given by

$$D_{k \cdot ij} = \alpha_i D_{ki} + \alpha_j D_{kj} + \beta D_{ij} + \gamma |D_{ki} - D_{kj}|$$

The result of the analysis can be represented by a dendrogram, using intercluster distance where Rayner uses similarity.

Various measures of intercluster distance have been suggested. We now review these, giving appropriate forms for α_i, α_j, β, and γ in the above relation. We shall suppose that clusters I, J, and K contain n_i, n_j, and n_k samples, respectively.

Nearest Neighbor Clustering

This is also called Single Linkage Clustering. Rayner's analysis can be regarded as an approximation. The distance between I and J is the least of the $n_i n_j$ distances between elements of I and elements of J. Clearly,

$$\alpha_i = \alpha_j = -\gamma = \tfrac{1}{2}, \qquad \beta = 0$$

that is,

$$D_{k \cdot ij} = \tfrac{1}{2}(D_{ki} + D_{kj} - |D_{ki} - D_{kj}|)$$

Farthest Neighbor Clustering

This is also called Complete Linkage Clustering. Here, D_{ij} is the greatest of the individual intersample distances, so that

$$\alpha_i = \alpha_j = \gamma = \tfrac{1}{2}, \qquad \beta = 0$$

that is,

$$D_{k \cdot ij} = \tfrac{1}{2}(D_{ki} + D_{kj} + |D_{ki} - D_{kj}|)$$

Group Average Clustering

In this situation, the intercluster distance is taken as the average of the $n_i n_j$ intersample distances, and

$$\alpha_i = \frac{n_i}{n_i + n_j}, \qquad \alpha_j = \frac{n_j}{n_i + n_j}, \qquad \beta = \gamma = 0$$

that is,

$$D_{k \cdot ij} = (n_i D_{ki} + n_j D_{kj})/(n_i + n_j)$$

A so-called unweighted analysis is possible using $\alpha_i = \alpha_j = \frac{1}{2}$, but D is not then explicitly defined.

Centroid Clustering

Gower (1967) suggested using the squared between-centroid distance as a measure of intercluster distance, and showed that

$$\alpha_i = \frac{n_i}{n_i + n_j}, \qquad \alpha_j = \frac{n_j}{n_i + n_j}, \qquad \beta = -\alpha_i \alpha_j, \qquad \gamma = 0$$

that is,

$$D_{k \cdot ij} = \left[n_i D_{ki} + n_j D_{kj} - (n_i n_j) D_{ij}/(n_i + n_j) \right]/(n_i + n_j)$$

An unweighted version with $\alpha_i = \alpha_j = \frac{1}{2}$ is termed Median Clustering. Again, D is not explicitly defined.

Minimum Variance Clustering

This method unites clusters so as to minimize within-cluster variance and hence D_{ij} is the sum of squares between clusters I and J. Anderson (1966) showed that

$$\alpha_i = \frac{n_i + n_k}{n_i + n_j + n_k}, \qquad \alpha_j = \frac{n_j + n_k}{n_i + n_j + n_k}, \qquad \beta = 1 - \alpha_i - \alpha_j, \qquad \gamma = 0$$

that is,

$$D_{k \cdot ij} = \left[(n_i + n_k) D_{ki} + (n_j + n_k) D_{kj} - n_k D_{ij} \right]/(n_i + n_j + n_k)$$

The results of this type of clustering can be presented as an analysis of variance, because the sum of squares being subdivided is that due to the samples + samples × attributes effects.

In general, nearest neighbor grouping produces elongated clusters within which pairs of dissimilar samples can occur, whereas centroid clustering, group average, and farthest neighbor methods lead to spherical clusters of high internal affinity. In taxonomic work, where plant or animal species can be related through evolutionary chain mechanisms, the former structure is not unreasonable, but the soil scientist is more usually interested in compact grouping. On the other hand, nearest neighbor and centroid analyses give good indication of the presence of single-element groups such as the isolated rendzina soil in Rayner's study. Outlying or erroneously measured samples can, of course, cause the same effect.

For all but the minimum variance clustering method the distance function is independent of the number of units in the clusters. This is not so for the minimum variance technique. Thus, if the samples in two groups are duplicated, the corresponding intergroup distance is doubled. Hence, there is a tendency for equal-sized clusters

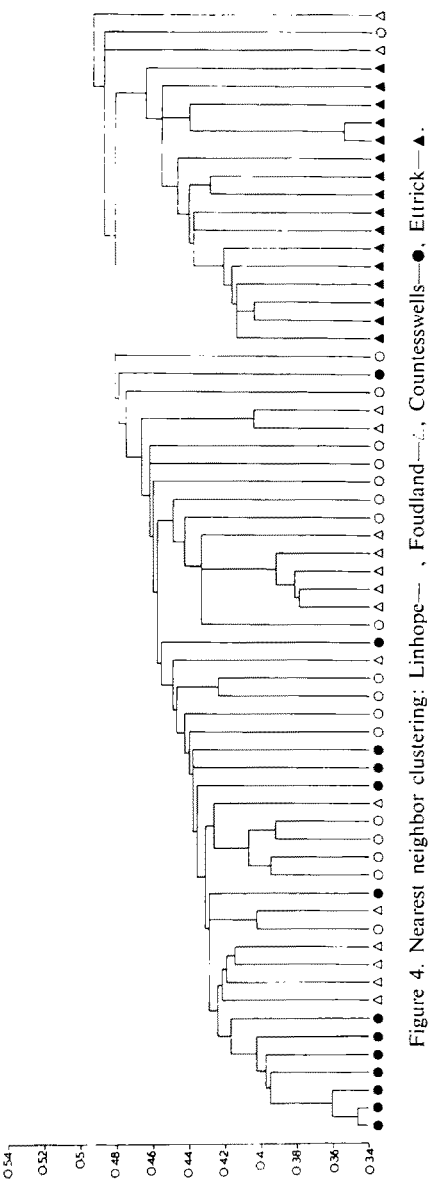

Figure 4. Nearest neighbor clustering: Linhope—, Foudland—, Countesswells—●, Ettrick—▲.

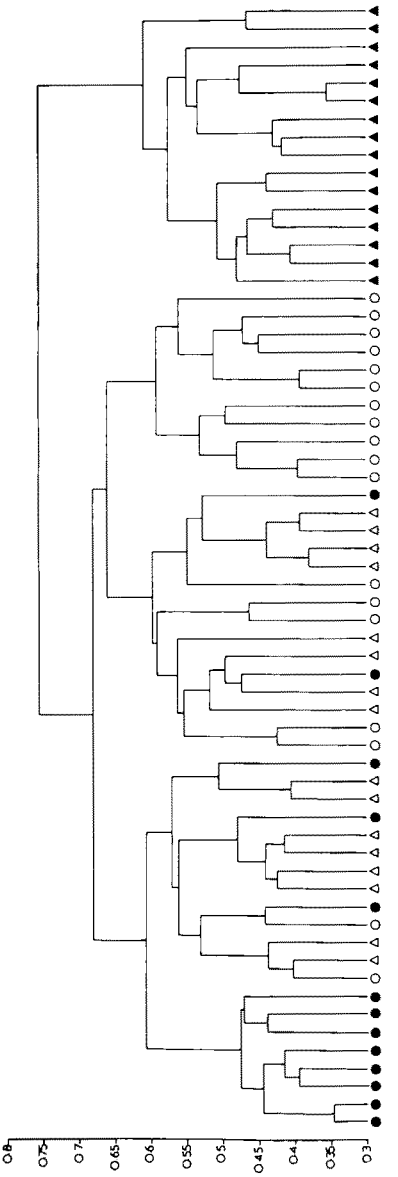

Figure 5. Farthest neighbor clustering. Symbols same as Figure 4.

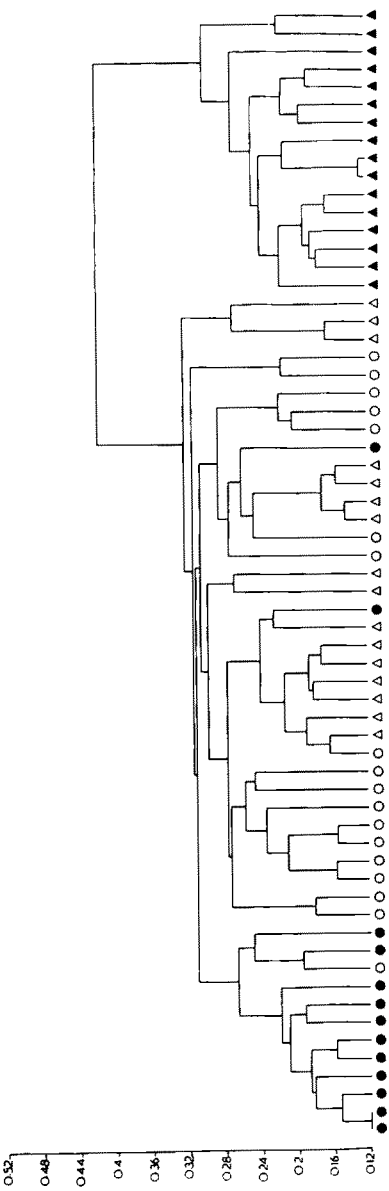

Figure 6. Group average clustering. Symbols same as Figure 4.

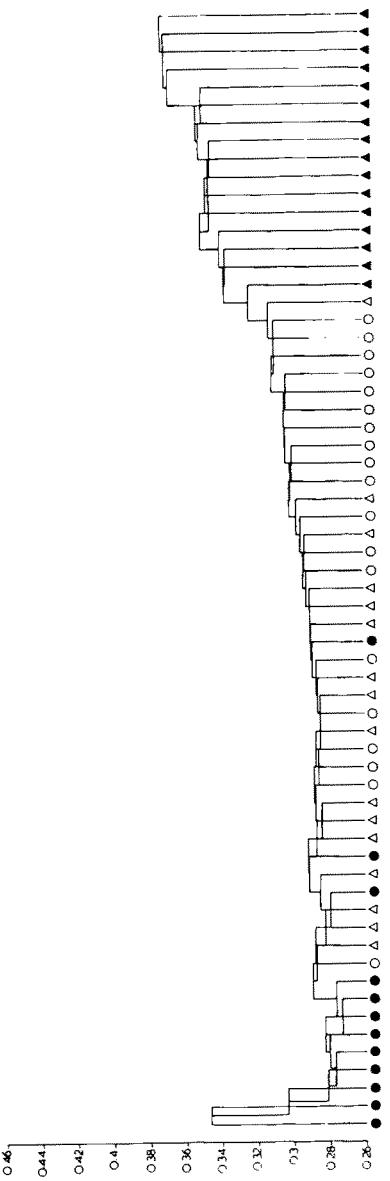

Figure 7. Centroid clustering. Symbols same as Figure 4.

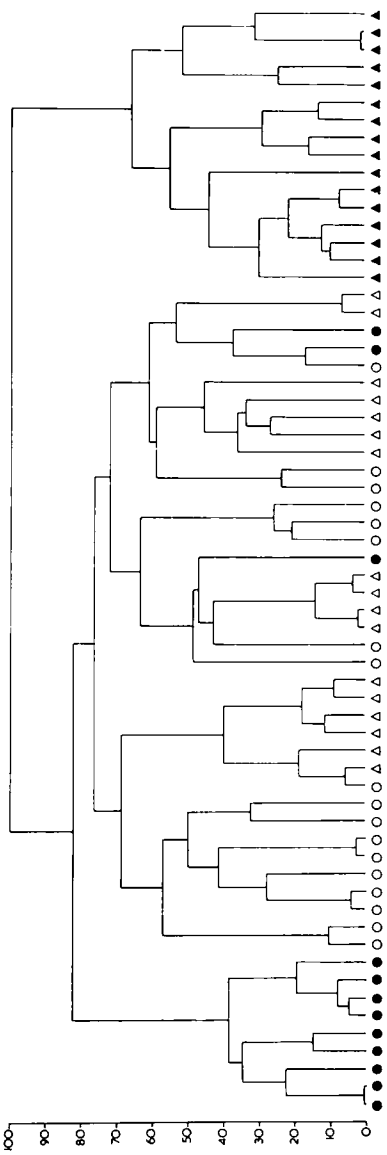

Figure 8. Minimum variance clustering. Symbols same as Figure 4.

to be formed, because the smaller clusters unite preferentially. This does not affect statements about the structure of the n samples. However, corresponding population inferences depend on the (unknown) sampling fractions of the parent clusters.

Example

A study of 63 profiles belonging to four well established soil series is reported by Muir and others (1970), the series having been selected to provide a reasonable comparison between traditional and numeric methods of classification. The soil surveyors' subjective description indicated that the Foudland series is intermediate between the Linhope and Countesswells series and that the Ettrick series is markedly different from the other three. The same attributes were available as in Rayner's study, and for numeric analysis the same measure of interprofile similarity was used.

These data now have been examined further, and Figures 4–8 show the dendrograms resulting from application of the five clustering methods described above.

(1) The tendency of nearest neighbor clustering to produce clusters containing single elements (profiles) is clearly demonstrated. Differences between the dendrogram shown in Figure 4 and that given by Muir and others (1970*a*) are due to the use of the approximate sorting strategy in the earlier work. These differences are not negligible. However, even the accurate analysis fails to indicate any separation of the Linhope, Foudland, and Countesswells series.

(2) Farthest neighbor clustering clearly segregates the Ettrick series. Clusters consisting entirely of Linhope and Countesswells profiles also are apparent, each associated with groups containing a jumble of profiles mostly from the intermediate Foudland series.

(3) Group average clustering likewise isolates the Ettrick series. Some partitioning of the remaining series can be detected but is obscured by the tendency to fragmentation.

(4) The nonmonotonic dendrogram of Figure 7 is typical of centroid clustering. A cluster is frequently nearer to the centroid of two clusters than to either of the individual clusters. This makes the segregation of even the Ettrick series undetectable.

(5) The minimum variance method isolates the Ettrick soils and then all but three of the Countesswells soils. There is some evidence of a Linhope group and a Foudland group, with a residue of jumbled samples.

These results exemplify the typical behavior of the various distance measures. Except for centroid clustering, all the methods indicate the segregation of the Ettrick series reasonably well. Only farthest neighbor and minimum variance clustering succeed in making any differentiation of the other series. The relationship of the minimum variance technique to the analysis of variance of the data suggests that further investigation of this method would be worthwhile.

ACKNOWLEDGMENTS

I thank Miss Bridget I. Lowe for assistance in the analyses presented in this paper, and Dr. J. H. Rayner and Mr. J. C. Gower for helpful discussions.

REFERENCES

Anderson, A. J. B., 1966, A review of some recent developments in numerical taxonomy: unpubl. master's thesis, Univ. Aberdeen, 50 p.

Anderson, A. J. B., 1970*a*, Ordination methods in ecology: Jour. Ecology, in press.

Anderson, A. J. B., 1970*b*, Hierarchical cluster analysis: Appl. Stat., in press.

Gower, J. C., 1966, Some distance properties of latent root and vector methods used in multivariate analysis: Biometrika, v. 53, no. 3–4, p. 325–338.

Gower, J. C., 1967, A comparison of some methods of cluster analysis: Biometrics, v. 23, no. 4, p. 623–637.

Kruskal, J. B., 1964*a*, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis: Psychometrika, v. 29, p. 1–27.

Kruskal, J. B., 1964*b*, Nonmetric multidimension scaling: A numerical method: Psychometrika, v. 29, p. 115–129.

Lance, G. N., and Williams, W. T., 1966, A generalised sorting strategy for computer classifications: Nature, v. 212, no. 5058, p. 218.

Muir, J. W., Hardie, H. G. M., Inkson, R. H. E., and Anderson, A. J. B., 1970, The classification of profiles by traditional and numerical methods: Geoderma, v. 4, no. 1, p. 81–90.

Rayner, J. H., 1966, Classification of soils by numerical methods: Jour. Soil Sci., v. 17, no. 1, p. 79–92.