J. Natn. Sci. Coun. Sri Lanka 1986 14(2) : 233–250

CHARACTERIZATION OF SOIL PROFILES FOR NUMERICAL CLASSIFICATION

P. WICKRAMAGAMAGE
Department of Geography, University of Peradeniya, Peradeniya, Sri Lanka

(Date of receipt : 9.4.86)
(Date of acceptance : 30.12.86)

Abstract : Soil individuals are three-dimensional natural bodies with their properties varying both in vertical and lateral directions. The tendency of certain soils to develop horizons, has influenced the collection of soil information and the soil profile is divided into a series of Master Horizons and each Master Horizon is sub-divided. Data on soil horizons are used to characterize soils for classification purposes. Numerical taxonomists have used various models of the soil profile in order to compute the relative similarity between soil individuals, but only a few attempts have been made to assess the effect of different soil profile models on the results of classification. In this study two soil profile models were compared and the effect of inter-attribute correlation on the inter-individual similarity was examined. It was demonstrated that it was not necessary to use data for a large number of depth levels or soil horizons because of the correlation between depth levels. It was also shown that the use of rigorous mathematical methods to characterize soil individuals only increases the computational load while similar results can be obtained from much simpler methods of soil profile description.

1. Introduction

Soils are anisotropic entities, the properties of which vary both in lateral and vertical dimensions. The lateral variation can be represented by a point sampling procedure whereas the vertical variation is represented by sampling from a series of depth levels. Although soils are three-dimensional bodies, the study of which has been based upon a two-dimensional entity known as the soil profile (a vertical cross-section through soil). The use of the soil profile as the basic unit of soil classification was first introduced by Dukuchaev in Russia in the 19th Century. This concept was introduced to America by Marbut[8] who claimed that soils at maturity developed a soil profile, the features of which could be used to characterize them. However, with the development of numerical taxonomic methods, it became necessary to find a suitable model of the soil profile that could be used as the basic taxonomic unit.

Conventional soil taxonomists have considered that the soil profile consists of a set of genetic horizons. Grouping of soil profiles according to

the nature and arrangement of soil horizons has gained a wide acceptance among soil taxonomists in almost all parts of the world. This is the basis of the definition of the Soil Series both in USA[9] and British[1] classifications. This method, despite being the most practical one, leads to subjectivity; and the Soil Series defined this way show a high degree of heterogeneity.[11]

Lance and Williams[3] have recognized four soil profile models that have been used in soil classification to characterize soils.

1. Depth levels as arrays of independent attributes,
2. Mean values of soil properties averaged over all depth levels (horizons),
3. A 'linked level system'; the similarity between soil profiles is defined as the average similarity between their corresponding horizons,
4. Parameters of a depth dependent function fitted to the soil properties.

Lance and Williams[3] reckon that the theoretically most sound model is the fourth one, despite the heavy computational load involved.

A fifth model, that can be used in numerical classification is the one which uses mean values of the attributes for the Master Horizons. This model does not discard too much information as the second model does and is not as complex as the fourth. However, it must be emphasized that the information loss due to averaging over all depth levels depends on the level of inter-attribute correlation. It has been demonstrated that depth levels are correlated[11] and elimination of the correlated attributes has no adverse effect on the classifications.[7] Therefore, the use of the first model seems to have been based on a faulty assumption. This investigation attempts to compare the fourth model and the three-horizon model.

2. Data and Methods

Data for 32 soil profiles were obtained from the published data of the United States Department of Agriculture.[9] The soil profiles used in this study were selected in such a way that they all have data at least for seven depth levels in order to fit a fifth degree polynomial. Moreover, the soil profiles chosen had data for all or most of the ten soil properties used in this study (Table 1). This list does not form an exhaustive set of soil properties, but it was chosen mainly because of the availability of quantitative data.

The two models that were used to characterize soils for numerical classification are:

(a) the orthogonal polynomial model,
(b) the three-horizon model.

Table 1. Soil properties used to characterize soil profiles

1. Percentage Silt
2. Percentage Clay
3. Percentage Organic Carbon
4. Percentage Dithionite Extractable Iron as Fe
5. pH (1:1 soil/water suspension)
6. Exchangeable Ca me/100g Soil
7. Exchangeable Mg me/100g Soil
8. Exchangeable Na me/100g Soil
9. Exchangeable K me/100g Soil
10. Cation Exchange Capacity (CEC) me/100g Soil

A fifth degree polynomial function of the following form was fitted to all ten soil properties.

$$Y_i = b_{0i} + b_{1i}X + b_{2i}X^2 + \ldots + b_{ki}X^k \ldots\ldots\ldots \quad (1)$$

where, Y_i — value of property i at depth X.

This model is flexible enough to fit a wide range of trends if a sufficiently large value is chosen for k. For soil properties k = 5 has been suggested by Colwell.[2] For statistical analysis a much more convenient form of this function can be obtained.

$$Y_{xi} = c_{0i}\,\ell_{\gamma_{0x}} + c_{1i}\,\ell_{\gamma_{1x}} + \ldots\ldots + c_{ki}\,\ell_{\gamma_{kx}} \ldots\ldots\ldots \quad (2)$$

where, $\ell_{\gamma_{kx}}$ is the value of orthogonal polynomial of degree k (k = 1, 2, . . . 5) at depth x. c_{ki} is the polynomial coefficient for the kth power of X.

The main advantage of this model is that each term of the polynomial function can be computed independent of others, and therefore it is not necessary to compute the whole function whenever the power of the polynomial function is changed. The original function was proposed for equally space X values (independent variable), but a modification has been described by Robson[6] and a computer programme for it has been written by Mather.[4] The coefficients c_{ki} can be used in numerical analysis in the place of original observations.

The second model was obtained by taking the mean values of ten soil properties for three Master Horizons, A, B and C. In cases where all three horizons were not present, only the available horizons (Master Horizons)

were used to compute the inter-individual similarity matrix. By this method, the number of attributes was reduced from 70 (10 soil properties for 7 depth levels) to 30. Product-moment correlation between the 30 attributes was computed and classification of all thirty attributes was done by average linkage method.

Two similarity measures (distance type) were used to generate inter-individual similarity matrices;

$$(a)\ d_{ij} = (1 - r_{ij})^{/2}$$

where, d_{ij} — similarity between ith and jth individuals
r_{ij} — product-moment correlation between ith and jth individuals.

$$(b)\ d_{ij} = 1/p \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$$

where, x — attribute values standardized to unit variance and zero mean.
p — number of attributes.

Euclidean distance requires the attribute vectors to be mutually independent (orthogonal). However, attributes of both models may not meet this requirement and therefore the possibility of masking certain attributes and its effect on the classifications should be examined. This is one of the objectives of this study. The relationship between the two soil profile models was also examined by product-moment correlation between inter-individual similarity matrices generated using the two models.

Classification of soil profiles was done by Ward's Error Sum-of-Squares (ESS) method.[10] This method was preferred to other agglomerative strategies because it tends to produce well defined clusters. Possible misclassifications by this method may be corrected by a suitable reallocation procedure if the hierarchy is not of interest. However, average linkage method was used to classify soil attributes since they very often show well defined clusters when depth levels are treated as arrays of independent attributes.[11]

3. Results and Discussion

Orthogonal polynomial coefficients were calculated for all ten soil properties; only in a few cases were poor fit recorded. A fifth degree polynomial seemed to be adequate to represent the vertical variation of soil properties considered in this study.

A series of inter-individual similarity matrices were calculated using similarity measure (a) and (b) described in Section 2.

The similarity matrices calculated for the two soil profile models using similarity measure (a) were classified by Ward's ESS method and two dendrograms were drawn to represent the classifications (Figure 1, a & b). Both dendrograms show well defined clusters with somewhat similar composition. The Group consisting of 1, 26, 28, 29 is common to both classifications; all soil profiles in this Group belong to Mollisols Order.[9] Again the Group consisting of 6, 7, 9, 10, 13, 14, 17, 30 can be identified from both dendrograms as a single group. Most soils in this Group belong to Alfisols Order. The other soil profiles have produced clusters which can only be described as broadly similar in the two dendrograms. However, it is worth noting that all soil profiles of any given order have not clustered together to form a single group.

Similarity between the classifications obtained for the two soil profile models confirm the findings of Moore, Russel and Ward[5] who concluded that no additional information could be gained by fitting mathematical functions to soil profile data.

Similarity between classifications produced for the two soil profile models can be traced back to the inter-individual similarity matrices. A sample of 30 similarity values was chosen randomly from one similarity matrix and plotted against 30 corresponding values taken from the other similarity matrix (Figure 2) and product-moment correlation between the two matrices was calculated. The scattergram (Figure 2) and product-moment correlation coefficient show a strong linear relationship ($r = 0.9$, $n = 30$). There is a slight scatter of the higher values, but it affects only those individuals which have a very low similarity to other individuals.

Further two classifications were obtained by Ward's ESS method using squared Euclidean distance (similarity measure b) as the similarity measure. The classifications are represented by two dendrograms (Figure 3, a & b). Although some groups (e.g. 2, 26, 28, 29 and 6, 9, 30) can still be identified from the two dendrograms, they are not very similar. This may well be due to the effect of inter-attribute correlation on the similarity measure. Since both classifications were obtained from the same clustering strategy, the difference between the classifications should be related to the nature of relative similarity between individuals. There are two possible explanations for this :

(a) difference between the soil profile models,
(b) the effect of inter-attribute correlation on the similarity measure.

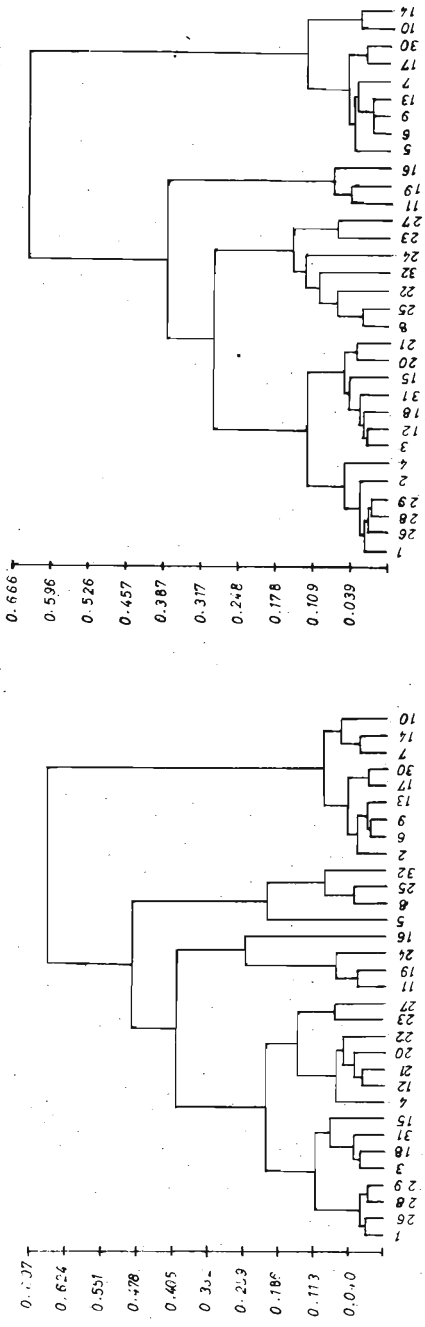

Figure 1. Classification of 32 soil profiles by Ward's method with $(1 - r_{ij})/2$ as the similarity measure, (a) three-horizon model (b) orthogonal polynomial model.

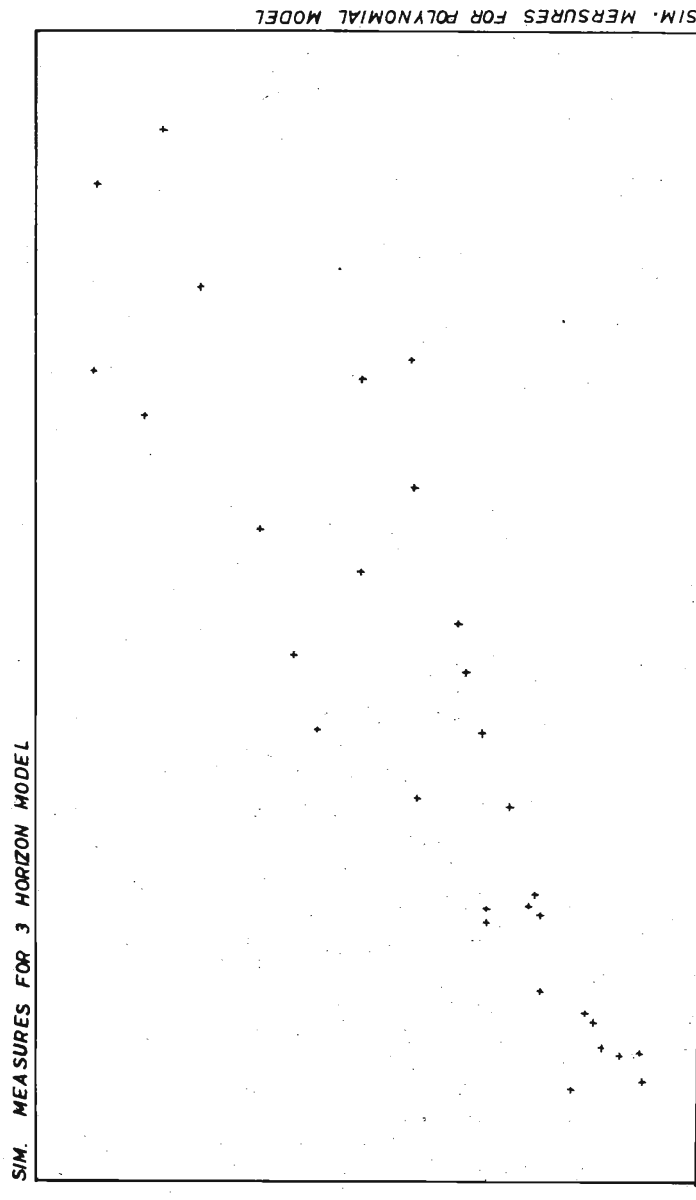

Figure 2. Relationship between inter-individual similarity $(1 - r_{ij})/2$ matrices calculated from the two soil profile models.

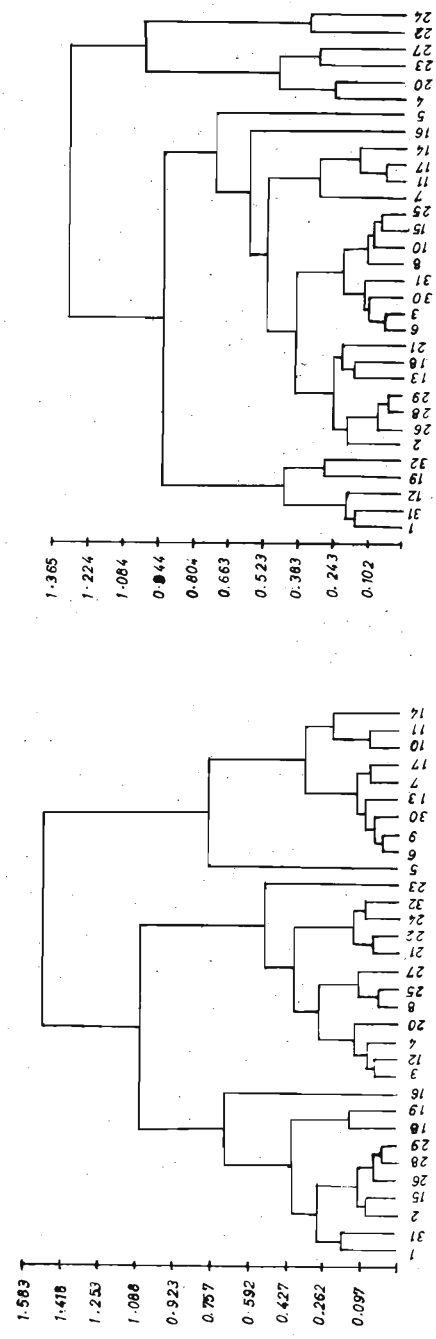

Figure 3. Classification of 32 soil profiles by Ward's method with sq. Euclidean distance as the similarity measure, (a) three-horizon model (b) orthogonal polynomial model.

It was shown earlier that when similarity measure (a) was used both models produced similar classifications. Therefore it was felt necessary to examine the relationship between the two similarity matrices calculated using similarity measure (a). If there is a strong linear relationship between the similarity matrices, the classifications obtained from any clustering strategy should be similar. The relationship between two similarity matrices computed using the same data can be shown by a scatter plot and product-moment correlation. A scatter plot was produced for a sample of 30 similarity values randomly drawn from the two similarity matrices (Figure 4). There is a considerable scatter of points at all levels of similarity with a low correlation of 0.58. Therefore, it can be concluded that the difference between the two classifications (Figures 3, a & b) is due to the difference in the relative similarity between individuals. This may well be due to unequal effect of inter-attribute-correlation on the similarity metric. The attribute vectors of the three-horizon model are highly correlated as can be seen from Figure 5, which was obtained from average linkage method with product-moment correlation as the similarity measure. It can be seen from Figure 5 most soil properties measured at different depth levels are correlated. In other words, the assumption that depth levels can be used as arrays of independent attributes has no basis. A number of fairly well defined groups of attributes can be identified from the dendrogram (Figure 5).

Except for a few, most soil properties measured at different depth levels are correlated. This indicates that it is not necessary to measure soil properties at a large number of depth levels for numerical classification of soils, because most of such observations do not contain additional information about the soil under study. However, multiple sampling from soil profiles may have other important uses.

Inter-attribute correlation seems to affect the relative similarity between individuals when squared Euclidean distance is used as the similarity measure. To minimize this effect, two sub-sets of attributes were chosen from the two soil profile models; eight attributes of the three-horizon model were masked so that inter-attribute correlation would not exceed 0.90. This level of correlation was arbitrarily chosen but would give some idea about the effect of strong correlation between attributes on the classification. The number of attributes in the orthogonal polynomial model was also reduced by taking the first three coefficients (C_{0i}, C_{1i} & C_{2i}). Two inter-individual similarity matrices produced using the revised data. This improved the similarity between the two classifications to some extent (Figure 6, a & b) due to the improvement in the relationship between the two similarity matrices as is demonstrated by the scatter plot (Figure 7). Scatter of points has reduced and correlation has increased from 0.58 to 0.79. This suggests that the relative similarity between individuals is similar for both soil profile models when inter-attribute correlation is eliminated.

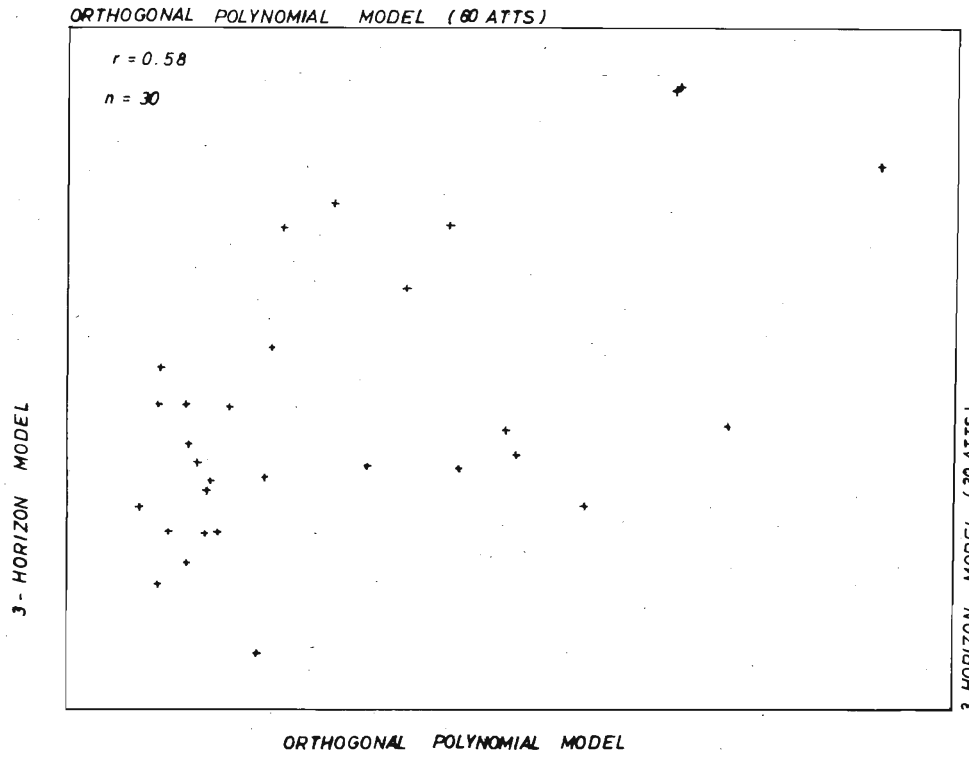

Figure 4. Relationship between the two inter-individual similarity (Euclidean distance) computed from the two soil profile models.

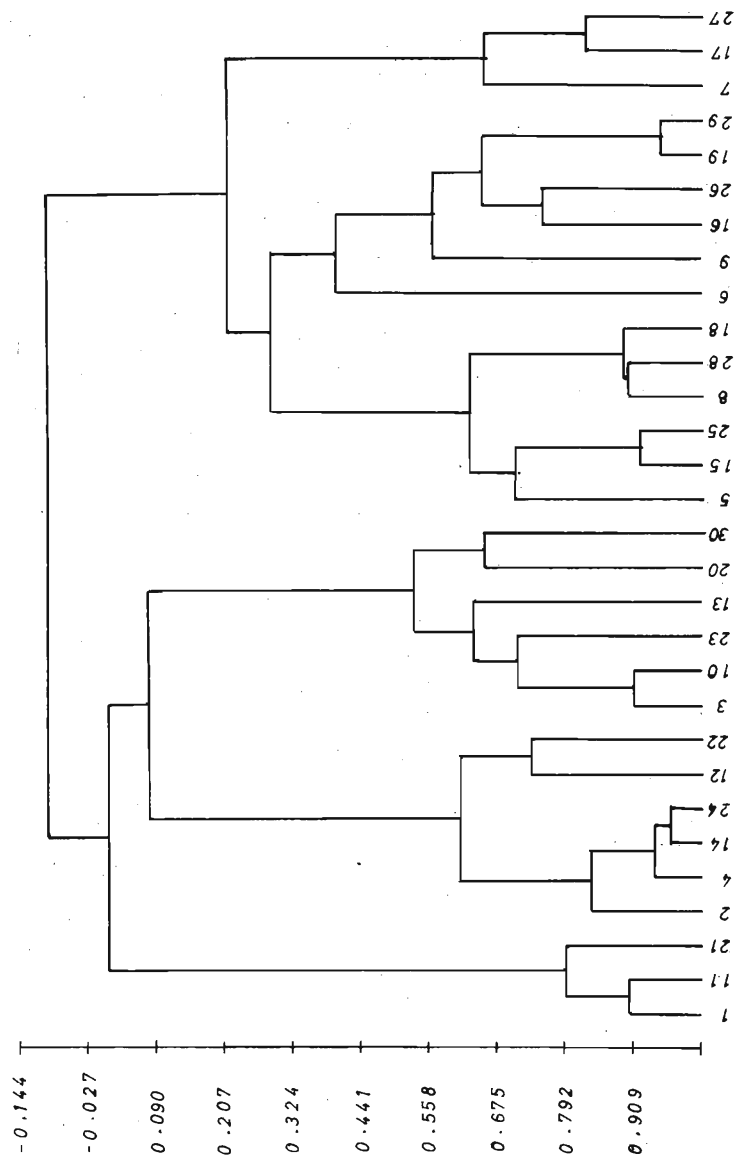

Figure 5. Classification of 30 soil attributes (10 soil properties for 3 horizons) by Average Linkage Method with product-moment correlation as the similarity measure.

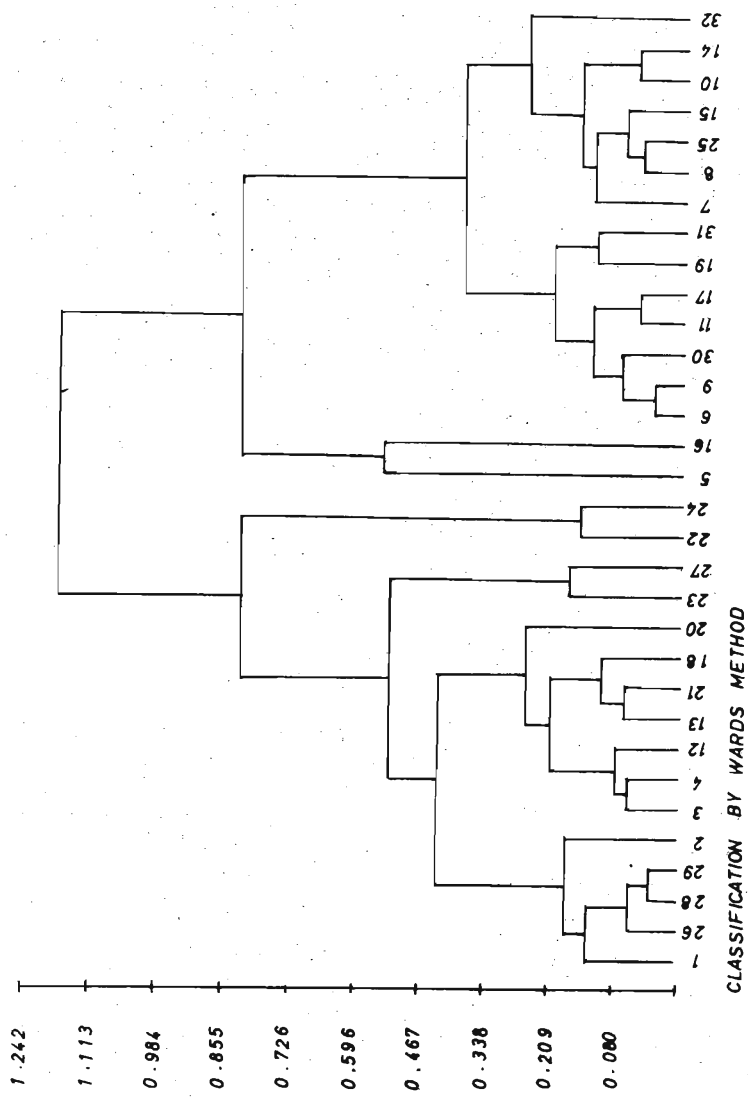

Figure 6a. Dendrogram produced by Ward's ESS method with Euclidean distance as the similarity measure after masking 30 attributes (orthogonal polynomial model)

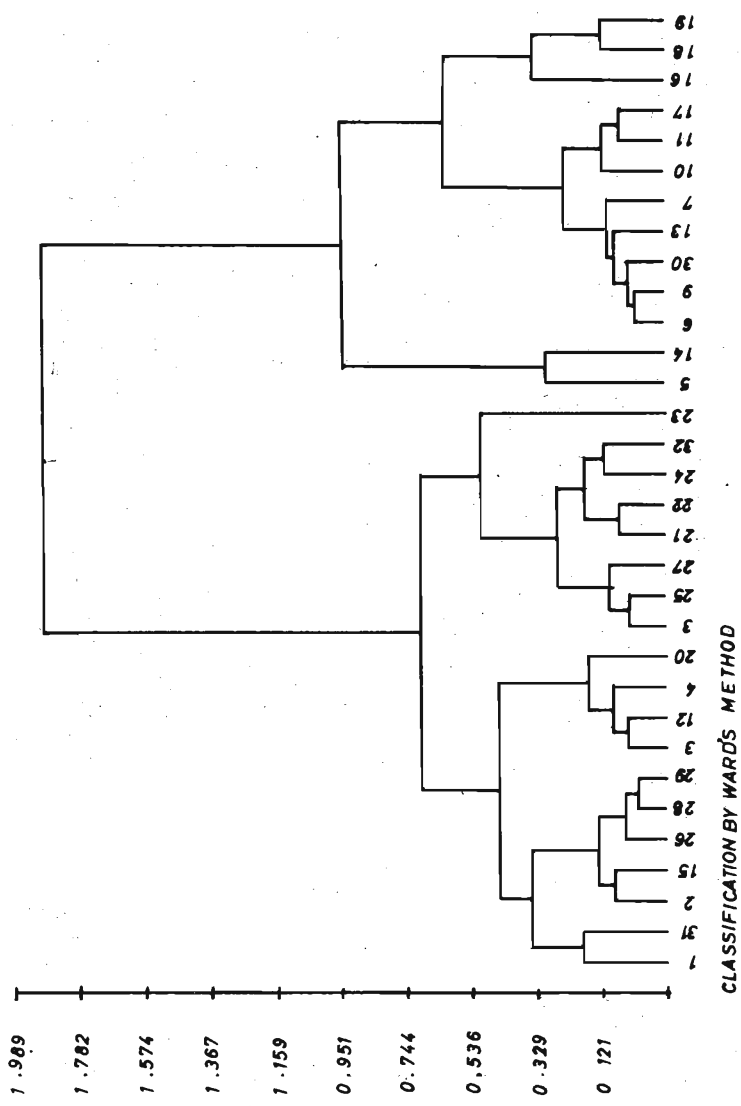

Figure 6b. Dendrogram produced by Ward's ESS method with Euclidean distance as the similarity measure after masking eight attributes (3–horizon model).

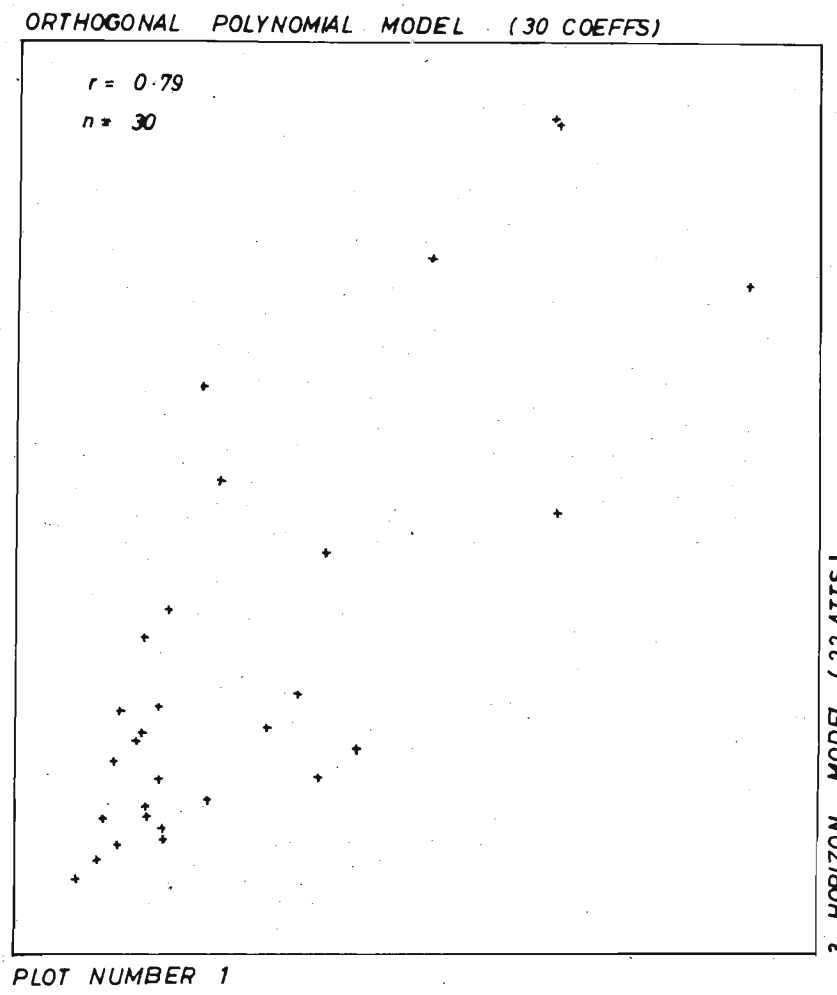

Figure 7. Relationship between two Euclidean matrices calculated for two soil profile models after masking 30 attributes of the polynomial model (Coefficient $C_3 \leftrightarrow C_5$) and 8 attributes of the 3—horizon model.

The relationship between the similarity matrices calculated using all attributes and a sub-set of attributes was examined for both soil profile models (Figure 8, a & b) assuming that it might reveal some information on the effect of inter-attribute correlation on the relative similarity between individuals. The scatter plots (Figure 8, a & b) show a curvi-linear relationship between the similarity matrices. Since the two matrices are not linearly related, the relative similarity calculated using correlated attributes tends to produce different classifications. Therefore, the selection of attributes for numerical classification is of fundamental importance. Equally, it has been shown that although goodness-of-fit of orthogonal polynomial functions depends on the value chosen for k, it is not necessary to use higher degree polynomials for numerical classification of soils. Such polynomials may be used to explain the variation of soil properties with depth.

4. Conclusions

The results reported above suggest that soils may be characterized for numerical classification by observations made at a few depth levels or horizons. Since soils are multi-level bodies, it has been the normal practice to use all observations on intrinsic soil properties made of all identified soil horizons. This has led to the collection of and the analysis of a large quantity of information despite the fact that such information may include redundant attributes. It was demonstrated in this analysis that it was necessary to eliminate redundant attributes prior to numerical classification for two reasons; (a) redundant attributes contain no additional information about the population under study and (b) they can distort the measurement space which would produce meaningless results.

The effect of inter-attribute correlation on the similarity measure seems to vary from similarity measure to similarity measure. The similarity measure (a) used in this analysis was not affected by inter-attribute correlation. Euclidean distance requires the attribute vectors chosen to be mutually orthogonal. Therefore it is required either to eliminate the correlated attributes or to transform data in order to achieve a mutually orthogonal set of vectors.

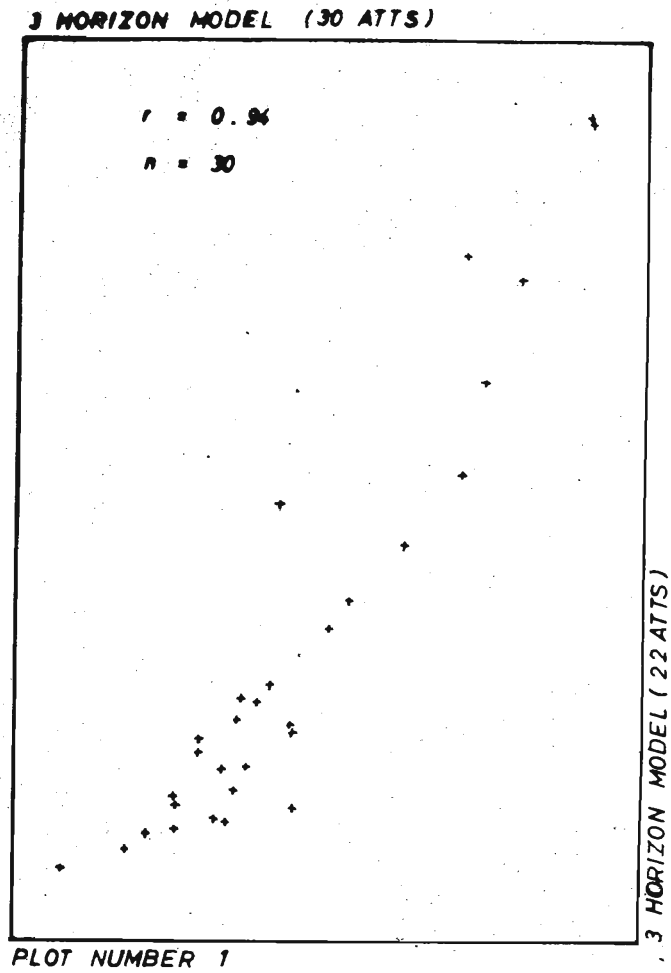

Figure 8a. Relationship between two Euclidean distance matrices calculated for the 3—horizon model before and after masking attributes.

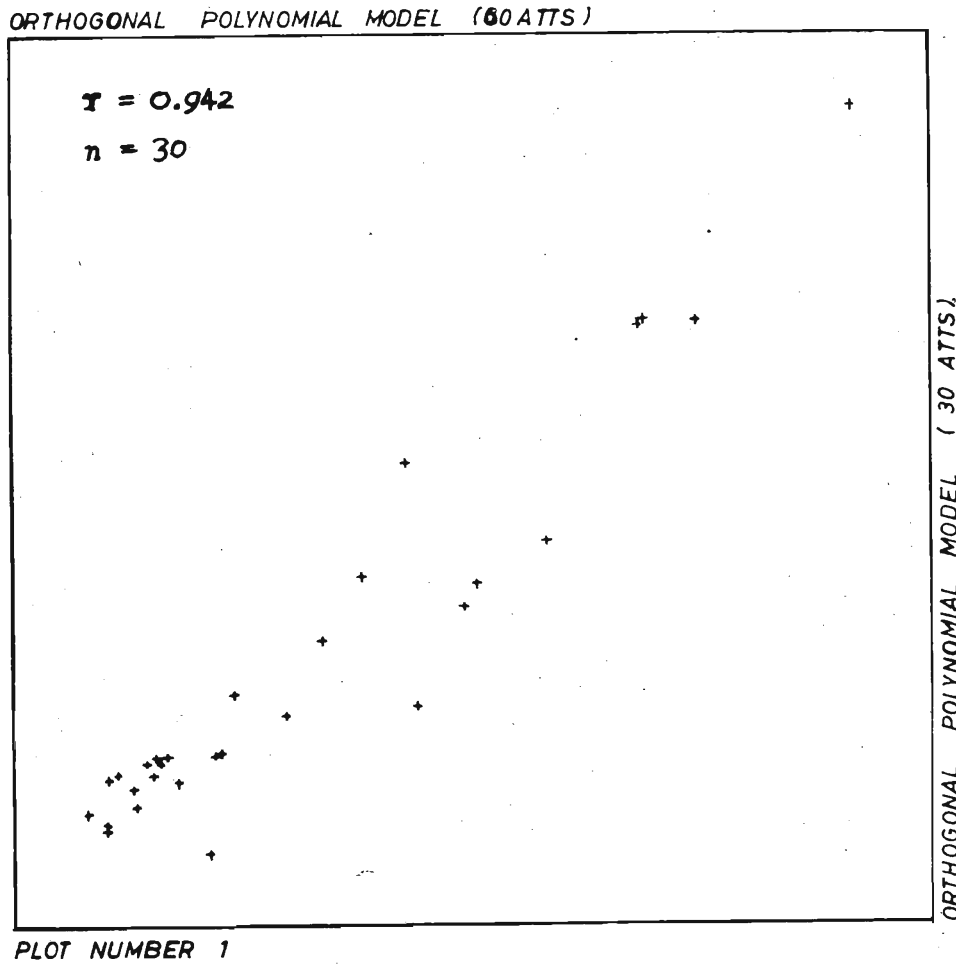

Figure 8b. Relationship between similarity matrices calculated using 80 and 30 polynomial coefficients.

References

1. AVERY, R. J. (1980) Technical Monograph No. 14, Soil Survey of Great Britain, *Harpenden,* 67pp.

2. COLWELL, J. D. (1970) *Aust. J. Soil Res.,* **20**: 221–238.

3. LANCE, G. N. & WILLIAMS, W. T. (1967) "A note on the classification of multi-level data"., *Comput. J.,* **9**: 381–383.

4. MATHER, P. M. (1976) *Computational Methods of Multivariate Analysis in Physical Geography,* Wiley & Sons, London, 532pp.

5. MOORE, A. W., RUSSELL, J. S. & WARD, W. T. (1972) *J. Soil Sci.*, **23**: 193–209.

6. ROBSON, D. S. (1959) *Biometrics,* p. 187–191.

7. SARKAR, P. K., BIDWELL, O. W. & MARCUS, L. F. (1966) *Soil Sci. Soc Am. Proc.,* **30**: 269–272.

8. SIMONSON, R. W. (1952) *Soil Sci.*, **74**: 249–257.

9. USDA (1975) *Soil* Taxonomy, *A basic system of soil classification for making and interpreting soil survey.* Government Printing Office, Agricultural Hand Book, 436pp.

10. WARD, J. H. (1963) *J. Am. Stat.* Ass. **58**: 236–244.

11. WICKRAMAGAMAGE, P. (1982) *Studies in Numerical Taxonomy of Soil,* (Unpublished Ph.D thesis, University of London).