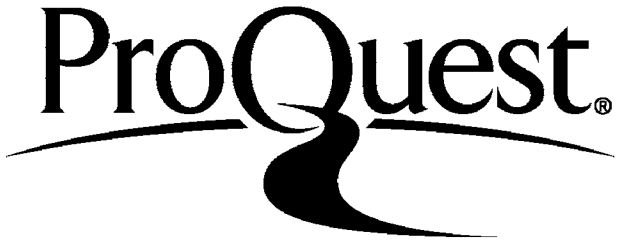
STUDIES IN NUMERICAL TAXONOMY OF SOILS

by

Piyasena Wickramagamage

Thesis submitted for the Degree of Doctor of Philosophy
in the University of London, Geography Department,
Bedford College,
December 1982



ABSTRACT

A series of established numerical taxonomic strategies was applied to soil data from three sources: USDA (1975), De Alwis (1971) and the Soil Survey of England and Wales. The first two sources provided data for 41 soil profiles, which were classified without reference to their geographical location. The data obtained from the Soil Survey of England and Wales related to a particular geographical area (West Sussex Coastal Plain) and the geographical relationship between soil individuals was also examined.

Two methods of soil characterization (soil profile models) were compared with respect to their effect on the results produced by two hierarchical agglomerative strategies based on two measures of inter-individual similarity. Comparison of results, obtained from the agglomerative strategies for the two soil profile models, was made. The nature of inter-attribute correlation for depth levels modelled as arrays of independent attributes was examined, and all attributes were classified on the basis of inter-attribute correlation.

Seven hierarchical agglomerative strategies were examined with respect to their goodness-of-fit in the original space and also the relationship between goodness-of-fit and clarity of clusters was examined. From these comparisons, two agglomerative strategies were chosen to represent two classes of strategy : (a) strategies

with minimum of distortion, (b) strategies with a greater distortion but clear clusters. The average linkage method from the first category and the Ward's error sum of squares (ESS) method from the second category were selected.

These two strategies were applied to the data sets described above using two measures of similarity namely (a) squared Euclidean distance and (b) Mahalanobis D^2, and a divisive strategy, REMUL, was also applied to classify the soil populations. The classifications obtained from these strategies were compared by Wilk's Criterion Λ and the classification which had the lowest Λ was treated as the best initial partition. The best two partitions of the two populations obtained from the agglomerative strategy, Ward's ESS method, were further analysed. The optimum number of groups (G) in each population was decided by the relationship between ΛG^2 and G. The soil profile groups produced by these methods were further examined and improved by a reallocation strategy based on the Mahalanobis distance between individuals and the group centroids. Reallocation was done using 30 attributes from the uppermost soil horizons.

Canonical analysis was performed on the populations both before and after the classification. Canonical plots were produced and a comparison was made with the dendrograms obtained for the best partitions.

The classifications obtained were examined in relation to parent material classes. The spatial

relationship of the soil groups of the West Sussex Coastal Plain was also investigated.

As shown by this study, it is possible to produce a better classification of soils by numerical taxonomic methods compared with traditional methods. For this end, it is not necessary to use all attributes of soils, but a sufficiently large number of properties, which can be empirically determined, is adequate for the purpose of producing a natural classification. The soil groups produced by numerical methods showed a closer association with parent materials.

ACKNOWLEDGEMENTS

This thesis is based on the research carried out at the Department of Geography, Bedford College, University of London, under the supervision of Dr. G. C. Fisher, whose constant help and encouragement is acknowledged gratefully, and I am also grateful to the College for providing facilities for my research.

Special mention should be made here of Dr. G. N. Lance, Director of the Avon Universities Computer Centre, Bristol, for lending the cluster analysis computer package, TAXON, and his help in the implementation of it at the University of London Computer Centre (ULCC). The staff of Bedford College Computer Centre also helped me in the course of this project.

Some of the diagrams of this thesis were drawn by Ms. C. S. Westie of the Department of Geography, and the typing was done by Lloyd Typing Bureau.

Finally I am grateful to all those who helped me in many ways during the four years at Bedford College.

TABLE OF CONTENTS

LIST OF FIGURES

page

page

page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.0 The definition of soil

The study of soils as an independent science is mainly due to the works of the Russian School of Pedology led by Dukuchaev. The concept that soils were independent natural bodies was first introduced by Dukuchaev and they were conceived as products of a combination of processes and factors namely climate, living matter, parent material, relief and age. This was a revolutionary idea (Soil Survey Staff, 1960;p.1), which made it possible to investigate soils themselves rather than infering from other factors. This Russian concept of soil,according to Marbut "established the study of soils firmly as an independent science with criteria, point of view, method of approach, process of development applicable to soils alone and inapplicable to any other series of natural bodies" (quoted by Basinski, 1959). These developments were not known to the rest of the world until the Glinka's famous text on world soils was translated into English (Soil Survey Staff, 1951, p.3). The early American and W. European pedologists did not examine the soil in depth, often confined their investigations to the plough layer and soils were studied as a part of related sciences (e.g. geology). Soils were sometimes treated as static

storage bins for plant nutrients (Soil Survey Staff, 1951, p.1).

Having identified soils as independent natural bodies, Dukuchaev tried to present a definition of soil. As quoted by Glinka (1928, p.2), the soil was defined by Dukuchaev as "the layers of materials lying on the surface of the earth or near it which have been changed by natural processes under the influence of water, air and living and dead organic matter". In his definition of soil the presence of genetic horizons with properties reflecting the effects of local and zonal soil forming processes led to the exclusion of those soils which have no genetic horizons or not thick enough (USDA, 1975). On the other hand it is not possible to distinguish between soil and parent materials. The definition of soil adopted by the U.S. Soil Survey can be treated as the logical development of the Russian concepts of soils and the contribution of Marbut and others in the 1920s and 1930s. In presenting the 7th Approximation System, the Soil Survey Staff (1960, p.1) defined the soil as a "collection of natural bodies on the earth's surface containing living matter and supporting or capable of supporting plants. At its upper limit is air or water. At its lateral margins it grades to deep water or barren areas of rock, ice, salt or shifting desert sand dunes. Its lower limit is perhaps the most difficult to define... The lower limit of soil, therefore, is the lower limit of the common rooting of the perennial plants". In this definition the genetic horizons are not included and therefore soils of

recent origin are also included.

1.1 Principles of soil taxonomy

According to Gilmour (1936) "classification is primarily utilitarian. It is a tool by the aid of which the human mind can deal effectively with the almost infinity variety of the universe. It is not something inherent in the universe but it is a conceptual order imposed on it by man for his own purposes". Classification therefore, is essentially utilitarian in that its uses could be preconceived and limited or could have an undetermined number of uses. As has been pointed out by Gilmour (1937) it is stated in logic that more propositions could be made regarding the constituent members of the natural classification than about the population as a whole. Broadly speaking there are two types of classifications, namely,

(1) classification for a pre-defined purpose or purposes.

(2) classification for a large number of purposes usually on the basis of many properties of the soils.

The former classification is described by various terms such as extrinsic, artificial or special purpose. It best serves the purpose or the purposes defined and is based on the characteristics (attributes) specially relevant to the proposed need. The classification of soils for agricultural use or for engineering purposes are examples of special purpose classifications.

The second type is generally known as natural, general purpose, intrinsic or taxonomic classifications. The soil taxonomist who is interested in soil classification needs such a system to serve as a frame of reference for soil mapping and other soil studies. However, there are conflicting views regarding the nature of a natural classification and the natural group or the taxon. Early ideas of a natural system were based on Aristotelian logic. As has been pointed out by Sneath and Sokal (1973, p.19) the purpose of the Aristotelian system as applied to taxonomy is to discover the essence of a taxonomic group (natural taxon) in such a way that essence is expressed in axioms that give rise to properties which are inevitable consequences. They have illustrated this logical system by the example of a triangle on a plane surface. The essence of a triangle is expressed by its definition as a figure bounded by three straight sides, and an inevitable consequence is that any two sides together are longer than the third. This logical system is described as a "system of analysed entities" which is not suitable to classify natural entities which represent a "system of unanalysed entities" (Sneath and Sokal, 1973, p. 19-20). It is not possible to define natural groups in such a way that many consequences follow from the definition without exception (Sneath, 1964). Although these ideas have been expressed in relation to biological taxonomy, they also have relevance to soil classification. Soil groups cannot be defined using "essential characteristics"

since they are not known to the soil taxonomist prior to classification. The 'a priori' weighting of characteristics of soils would lead to a special purpose classification (Gilmour, 1936; Kubiena, 1958; Basinski, 1959). Therefore, the Aristotelian system cannot be applied to soil taxonomy.

Another approach to natural classification is one which is based on the phylogenetic relationships and can only be applied to the products of organic evolution (Crowson, 1970,pp.95-114). The biological taxonomists who have adopted this approach stress that the natural taxa should be constructed to reflect evolutionary relationships. This concept has no parallel in pedology, and cannot be treated as an adequate basis for all natural taxa.

Adanson rejected ideas of 'a priori'assumptions and pointed out that natural taxa should be based on the similarity measured taking all characteristics into consideration (Sneath and Sokal, 1973, p.5). Therefore, a natural taxonomic classification should ideally be based on the intrinsic properties of the objects to be classified. According to Gilmour (1936, 1960) the terms 'artificial' and 'natural' are relative and no classification can be based on all attributes for reasons given below. Therefore a classification can be more natural than another depending on the range of attributes used in the definition of groups. A classification based on a large number of attributes is more useful as a general reference system.

Soil taxonomists are not in agreement as to what a true 'taxonomic' classification is. Soils have developed in diverse parent materials and environments, and therefore are not related in a biological sense. However, concepts of soil genesis have entered into soil classification directly or indirectly (Avery, 1968). A natural soil classification was conceived by many as one which would reflect genetic relationships. This concept of natural classification was derived from biology (Gilmour, 1961). The most important characteristics, which were considered as the basis of classification, were derived by circular argument by inspection of natural groups already recognized empirically (Cain, 1958). However, the Russian pedologists always recognized the 'genetic soil type' as the basic unit of soil classification (Basinski, 1959). Even the most recent soil classifications in the USSR have paid more attention to pedogenetic factors and the geographical environment than to soil properties. This emphasis on pedogenesis can be found in soil classification systems elsewhere as well. For example, the current USDA system has chosen some differentia which reflect soil genetic processes. The use of genetic homogeneity as an objective of soil classification runs into troubles for three reasons.

(1) The genesis of most soils is not known at the time of the classification or is controversial.

(2) The apparently genetically homogeneous soils may not necessarily be homogeneous in their intrinsic properties. Therefore, such a classification is no more than a special purpose classification of soil genetic factors and processes.

(3) The pedogenetic environments have changed over time and consequently some soils have developed under more than one genetic environment.

It may not be possible to determine the genesis of soils before classification but a successful soil classification could lead to better understanding of their genesis.

The natural soil classification defined by Kubiena (1953) was one which, it was claimed, was based on all characteristics of soils. The use of all characteristics of soil, however, is not possible (Gibbons, 1968) for three reasons:

(1) some attributes may not be known at the time of classification.

(2) the attributes that are known but not evaluated cannot be used. This is a particular problem when soil survey data are used to classify soils.

(3) the constraints of data manipulation. This constraint has been remarkably reduced by the development of electronic computers. But certain mathematical techniques do require the number of attributes to be less than the number of individuals.

Apart from these constraints some attributes could be excluded from the classification strategy when there is a strong inter-attribute correlation (Sarkar et al, 1966). Thus those attributes which have the maximum variance and the largest number of accessory properties should be used to classify soils.

Kubiena's concept of natural classification is not accepted by Leeper (1956), who suggests that a classification could be simply good or bad but not natural or artificial as claimed by the former. This is contrary to the general principles of taxonomy and cannot be accepted because this attitude would only lead to a special purpose classification. Muir (1962) has used the periodic classification of elements to illustrate the concept of an ideal natural classification and suggests that the soil taxonomists should try to produce a comparable classification system. But the task of the soil taxonomist is much more difficult. There was a unifying theory regarding the nature of the elements of the periodic table prior to the classification and the laws of chemistry were well established unlike those of soil science. As far as soil is concerned there is no such theory that could help define and interpret soil groups. Also soil individuals are not discrete entities like elements in the periodic table. The soil universe is made up of an infinite number of individuals which are not mutually exclusive. Kubiena's approach to soil classification is based on Adansonian principles of natural taxonomy.

Although it is not possible to give equal weights to all attributes for the reasons given above, equal weights can be given to all those attributes which are used in classification. When data are available on a large number of soil properties a representative set of attributes could be chosen to characterize soil individuals. In this study all attributes used were given equal weights.

1.2 Soil universe

Classification of soils involves the definition of the soil universe. The soil universe is made up of all soil individuals and classes, and therefore is a super class (Knox, 1965). The nature of the soil universe is fundamental to the understanding of the soil classification. Gibbons (1968) has suggested that there are three models of the nature of the soil universe.

(1) The soil universe is essentially particulate (made up of discrete natural individuals). The classes are identified by peaks in the distribution curves of attributes (Kubiena, 1958); individuals and classes are found but not constructed. Soil individuals do not have clearly defined boundaries (section 1.3).

(2) The soil universe is essentially continuous, both classes and individuals are constructed not found (Knox, 1965). It is important to recognize the continuous nature of the soil universe but it is possible, for practical purposes, to describe soil individuals in the field. The continuous nature of soils may hide the differences between soils and the task of the taxonomist is to uncover the hidden differences.

(3) The soil universe is continuous, the individuals are real and found but the classes are abstract (Soil Survey Staff, 1960, p.1-11).

The third view was adopted here and the classification of soils is considered as the identification of those soil groups whose members are similar in their intrinsic properties. Therefore, soil groups are not truths to be discovered by man but they are constructed for practical purposes and such a classification is essentially utilitarian. All soil groups are associated with a certain degree of variation in their properties. A successful classification would minimize this variation.

1.3 Soil individual

The main difficulty in the definition of the soil individual is due to the fact that it cannot be identified as an exclusive entity like biological organisms or the elements in the periodic table. But a soil body is a natural entity, the properties of which can be observed in the field. Knox (1965) suggests that only a particulate universe has natural individuals and the individuals in a continuous universe are artificial individuals created in the absence of natural individuals for the purpose of classification. This distinction between the artificial and natural individuals may be considered as an over simplification. As has been suggested by Soil Survey Staff (1960) "a soil individual is not found as a distinct entity clearly separated from all others, but grades on its margins to other soil individuals with unlike properties".

Therefore, the soil individual is a natural individual which can be found in the field.

Knox (1965) has described eight possible soil bodies which can be considered as soil individuals, from which six soil bodies can be identified as forming the concept of soil individuals for various purposes.

(1) Primary particles, such as crystals or crystal fragments. These particles do not form a soil and therefore cannot be used as the soil individuals in the classification of soils.

(2) Hand specimen, these are samples of soil materials used for laboratory determinations. A given hand specimen is treated as a homogeneous sample and is only a part of a soil body and does not represent the soil body as a whole.

(3) The soil horizon. The soil horizon has all the characteristics of the hand specimen plus the thickness. A soil horizon can be observed in the field (where present) and is a layer of relatively homogeneous soil material. The soil horizon has been used in soil classification as the soil individual by Rayner (1966). Fitzpatrick (1967) has proposed a system of soil classification using the soil horizon sequences to determine the similarity between soils. However, the soil horizon cannot be used as a soil individual, as it is only a part of the soil (section 1.0).

(4) The soil profile. This is a vertical cross-section of soil without the lateral dimension. It represents the vertical variation of soil properties and the

horizonation can also be observed where present. The importance of the soil profile as a unit of classification was introduced to the U.S. Soil Survey by Marbut (Simonson, 1952). The concept had been described by Dukuchaev much earlier than Marbut but was not known to those outside Russia. Simonson (1952) claims that the introduction of the soil profile to soil science was comparable to the introduction of anatomy to biology some centuries ago. Marbut believed that all soils at maturity developed a soil profile and the features of the soil were expressed as the features of the soil profile. This concept has influenced the methods of the modern soil surveys in the world. Both in the USA and the UK representative soil profiles are used to characterize the Soil Series. Therefore, collection of soil data has been done using the soil profile as the basic unit of sampling. For practical reasons the soil profile has an area of about $1m^2$ and the depth is left undetermined. However in practice the depth of soils is considered to be the rooting depth of the perennial plants (Soil Survey Staff, 1960, p.1). The soil profile is the most effective unit of soil which could be used as the soil individual. The main criticism of the use of the soil profile as the basic unit of classification has been that it does not represent soil in three dimensions but it is not possible to collect data on larger soil bodies efficiently.

(5) Pedon. The pedon is the smallest three dimensional unit that could be called a soil (Soil Survey Staff, 1960, p.2-4). It includes all the characteristics of the soil profile plus lateral variation. The area of a pedon varies from $1m^2$ to $10m^2$ depending on the variability of the soil horizons.

(6) The soil landscape unit. This is a geographical body of soil which is a mappable unit unlike the previous five soil bodies. The soil series, which is the lowest category of the soil survey of England and Wales classification belongs to this. The soil landscape unit cannot be considered as the soil individual because of the high degree of heterogeneity involved in the definition of such landscape units. For detailed soil surveys they are useful as mapping units,the identification of which should be done after the classification of soils.

The other two soil bodies, the delineated soil body and the soil type, decsribed by Knox (1965) are also soil landscape units which have been used for mapping purposes rather than for soil classification.

Among the soil bodies considered, the soil profile is the most convenient unit which could be used as a soil individual for the collection of data and classification. The other units described above are either not representative soil bodies or they are too large to be homogeneous enough to be used in soil classification. Therefore, the classification of soils by numerical methods, as used in this work, will involve the classification of soil profiles. Identification of soil mapping units similar to soil profile groups is the task of the soil

surveyor.

1.4 Taxa

The meaning of the taxonomic group or the taxon was given earlier but it is necessary to define it in relation to soil classification. Classification of soil individuals (profiles) involves the determination of affinity (taxonomic similarity defined by Sneath and Sokal, 1973, p.31-40) between individuals with respect to all, or an adequate number of soil properties. This has been done by the traditional taxonomists by defining a set of diagnostic features. In the USDA system, the Soil orders are defined by using a small number of diagnostic features. In the Soil Survey of England and Wales Classification System (Avery, 1973, 1980), the higher categories are identified using diagnostic features termed 'keys'.

The taxonomic class (taxon) is necessarily a polythetic class based on all characteristics or an adequate number of characteristics. The objective of numerical taxonomy is to define such groups. The USDA (1975, p.9-10) considers that the choice of the attributes (characteristics) should be made in such a way that the chosen attributes should have the greatest number of accessory attributes. Such attributes have been generally considered as related to the genesis of soils. It has been pointed out by Webster (1977) that the soil properties do not covary as had been expected earlier. This may be due to the fact that,such correlations may not

exist between soil properties observed for different soil populations. The nature of correlation between soil properties may vary with soil groups. But Sarkar et al (1966) demonstrated that the sixty one soil attributes used in their classification of Kansas soils were correlated and when the number of attributes was reduced to twenty two, the same results could be obtained. Because of such disagreements, it is necessary to classify soil attributes prior to the classification of soils.

1.5 Historical development of soil classification

Prior to the works of Dukuchaev and his colleagues, soil was studied as a part of other sciences such as geology and agriculture. There was no clear definition of soil although the importance of it had been recognised. Some early attempts to classify soils were made in Western Europe. Thaer in 1853 proposed a soil classification based on the textural properties at the primary level and agricultural properties at the lower categorical levels. In 1886 Richthofen proposed a classification with the emphasis on the geological properties and nomenclature. These early classifications reflect the state of contemporary pedological knowledge. Buol et al (1973,p.174) describe such classifications as technical classifications.

Dukuchaev (1846-1903) recognized soils as independent natural bodies and attempted to classify them, in part, with respect to their properties. Dukuchaev's

work on the Russian Chernozem states that "soils must be classified according to their properties" (quoted by Buol et al, 1973, p.175). However, in practice he used soil properties only at the lower categorical level and the highest categories were separated using environmental factors on the assumption that they were related to the broad climatic and vegetational zones. Dukuchaev and his colleagues (specially Sibirtsev and Glinka) gave a great importance to soil genesis and the soil properties were chosen in such a way that they would reflect the genetic environment and the factors of soil formation.

These ideas are comparable with the views of the early biological taxonomists that a taxonomic classification should be based on Aristotelian logic. This trend continued in the USSR. The soil classification proposed by Kovda et al (Fitzpatrick, 1980, p.174) is claimed to be an historical genetic classification using properties which reflect the evolution of soil in time. This system breaks with the old Russian tradition by using soil properties rather than environmental factors as a basis for soil classification but it is identical with the old system in that it was the interpretation of facts (data) rather than facts themselves which were used. Another recent Soviet soil classification proposed by Rozov and Ivanova is described by Avery (1968) as a coordinate system. In this system the categories below 'Types' are based on their relationship with three groups of soil properties (Coordinate axes).

Axis 1 The properties of soil and environment that can be little changed by man.

Axis 2 The moisture characteristics of soil

Axis 3 Bio-physico-chemical soil ranges

i) peculiarities of organic matter decomposition

ii) saturation and absorption complex and cation exchange complex

iii) general structure of the soil profile and the presence or absence of carbonates, gypsum and soluble salts.

This system is comparable with that of Avery (1968) in principle but the categories above Types have not been worked out. There has not been a substantial change of emphasis in the tradition of soil classification in the USSR.

The concepts of soil classification worked out by Dukuchaev and subsequently developed by Sibirtsev and Glinka were introduced to the West by Marbut by translating the German edition of Glinka's text on the world soils. The US soil classifications used during the period 1899-1922 have been described by Buol et al (1973, p.176-177) as single factor soil classifications with a bias towards geological techniques and the nomenclature. In Western Europe and in America the concept of soil geology (geological derivation of soils) prevailed during the 19th century (Cruickshank, 1972, p.13-31).

Hilgard (1833-1906) was the first in the USA to recognize soils as independent natural bodies (Buol et al, 1973,p.176), though his ideas were not applied in operational soil surveys in the USA. After him Coffey suggested in 1912 that soils were natural bodies, the classification of which should be based on soil properties. These ideas did not have an impact on the pedological thinking of the USA until Marbut introduced the Russian concepts.

Marbut can be considered as the founder of modern soil taxonomy in the USA. He not only introduced the ideas of Dukuchaev, Sibirtsev and Glinka, but developed his own ideas of soil classification and survey. The contribution made by Marbut has been summarized by Buol et al (1973, p.171-181) under three headings.

(1) The establishment of the soil profile as the unit of study and the emphasis on soil properties (section 1.3).

(2) Establishment of criteria for soil series.

(3) The preparation of the first hierarchical multi-categoric system. This concept has been developed by the Soil Survey Staff (1960) to produce the 7th Approximation soil classification system.

Marbut put more emphasis on the soil properties than genesis in devising his system of soil classification. He divided soils into two major classes:

i) Pedalfers

ii) Pedocals

Pedalfers are those soils that accumulate sesquioxides while Pedocals have a horizon of carbonate accumulation. This system encountered problems when attempts were made to include Brownearths most of which accumulate neither iron nor carbonates (Fitzpatrick, 1980, p.126). However, this system can be treated as the true beginning of modern US soil classification. Along these lines a comprehensive multi-categoric system was proposed in 1938 by Baldwin, Kellog and Thorp (Buol et al, 1973, p.179) but they included zonal concepts of Sibirtsev. The highest categories of this system, unlike those of Marbut, were defined in genetic terms (Bidwell and Hole 1964a).

All the soil classifications proposed in the USA prior to the 7th Approximation were qualitative to varying degrees. The classes were not defined using quantitative measurements and as a result the decisions on criteria were made subjectively. However, it can be seen over time there was more incorporation of quantitative data at various levels.

The 7th Approximation system (Soil Survey Staff, 1960) was proposed as a general purpose classification based on a large number of observed soil properties. The general purpose classification was conceived as a multi-categoric system along the lines of the previous soil classifications of the USA, and such a system was conceived as hierarchical in organisation. The system could be put to a multitude of uses at its various categorical levels. The 7th approximation is an attempt

to rationalize the criteria used to define various classes although the decisions on the choice of them have been made subjectively. Ragg and Clayden(1973, p.12-13) have summarized the criticisms of the USDA system coming from various sources. These criticisms reflect the conflicting views held by soil taxonomists of different countries. For example, while some (Webster, 1968) object to the system for using genetically important properties, others have pointed out that no adequate consideration has been given to such properties (Duchaufour,1963; Gerasimov et al, 1964). However, the choice of the differentia has been subjectively made and therefore a large amount of information gathered on soils has not been used consistently and objectively. Webster (1968) claims that the fundamental fault of the system is its hierarchical organisation of the categories.

In the last century and in the early part of this century, important contributions were made towards the recognition of the study of soils as an independent science in Western Europe. In this respect, works of Müller (Sweden) and Ramann (Germany)are of great importance. In Great Britain the study of soils started formally with the establishment of a research institute at Rothamsted, England in 1843. Soil survey in Great Britain dates back to 1911 (Cruickshank, 1972, p.23) when the first of special surveys were published. Robinson (1932) and Avery (1956) have been responsible for the development of the Soil Survey of England and Wales. A separate soil survey for Scotland was established in 1930. Three soil

* The three classifications are those of Soil Survey of England and Wales, Scottish Soil Survey and Irish National Soil Survey. Both England and Wales and Scottish classifications are multicategoric although the categories, their position in the hierarchy in particular, are somewhat different. For example, there are ten Major Groups at the highest level of the England and Wales classification, whereas the Scottish Soil Survey has defined three categories called Soil Divisions. The Irish classification has derived from the old European classification system.

classification systems have been used in modern times in different parts of the British Isles. Since the beginning of modern British soil classification in the 1930s, morphological properties have been widely used. The soil series were defined by taking the nature and the sequence of soil horizons into consideration. The most recent soil classification proposed by Avery (1980) for England and Wales is a somewhat different approach, which originated as a coordinate system. This is also a multi-categoric system which was originally described as non-hierarchical by Avery (1968). The class differentia have been subjectively determined, as a result this system also has failed to eliminate the basic problem of subjectivity of all traditional classifications. The similarity between soil profiles has been determined on the basis of the presence or the absence of pre-selected diagnostic characteristics. In addition the existence of three soil classification systems for the British Isles may lead to inconsistencies and creates difficulties in the communication of soil information among different authorities.*

The majority of the traditional soil classifications have been devised to serve the needs of the country concerned and therefore they have taken local conditions into consideration. The USDA system was proposed as a comprehensive classification drawing samples from a wide range of geographical environments, in order to represent as many diverse soils as possible. But tropical

soils were under represented. The soil classification systems devised for individual national soil surveys cannot be applied to other areas successfully. Even the USDA system has been proved to require modification when it is tried in other countries (Ragg and Clayden, 1973; Kesseba et al, 1972). The soil classes defined in different countries according to different classification systems cannot be compared easily. Therefore, a great need is felt for an objective system of soil classification using as many properties as necessary.

Several common features of the traditional soil classifications can be identified.

(1) Each of them has been devised to serve a national need and therefore the taxa of the system are limited to those which occur in the country for which the classification system was proposed. Although the USDA system is supposed to be a comprehensive system applicable to other countries, the attempts to use it elsewhere revealed the need for modifications (Ragg and Clayden, 1973; Kesseba et al, 1972).

(2) The use of 'a priori' assumptions on the diagnostic criteria. The diagnostic features have been defined prior to the classification even though they should have been discovered after the classification was devised.

(3) The affinity between soil profiles has been determined subjectively. The decision on the similarity between soil profiles and between groups has been judged by the surveyor in the field, or a soil taxonomist, purely on the basis of experience and intuition.

(4) The properties chosen were weighted without any empirical justification.

(5) Almost all the properties used were related to the soil morphology (Muir et al, 1970).

Despite the drawbacks of the soil classification systems hitherto produced, a general improvement can be detected. This development from early single factor classifications to more complex systems such as USDA system using a wide range of information on the soils themselves can be viewed as a considerable progress. Over the years the knowledge of the nature of soils has increased tremendously, so that a considerable amount of information about soils is available. The main problem of the classification of soils today is how best this information could be used to devise a taxonomic classification. Crowther (1953) conceived the problems arising from the multi-dimensional nature of soils and suggested that a coordinate system would be suitable to classify soils but the objective use of such a model was not then possible. The advent of the electronic computer made numerical taxonomy possible, consequently a whole range of techniques is available to handle a large amount of data. These methods could be used to replace old subjective methods of soil classification.

1.6 Numerical taxonomy of soils - a review of previous work

The term numerical taxonomy has been defined by

Sneath and Sokal (1973, p.4) as "the grouping by numerical methods of taxonomic units into taxa on the basis of their character states". The taxonomic units are the soil individuals in the form of soil profiles. The character states are the attributes which are presented in a numerical form. Therefore the phenetic similarity between individuals could be determined using a metric. Numerical taxonomy is a further development of Adansonian taxonomy.

Sneath and Sokal (1973, p.11) claim that the principal aims of numerical taxonomy are repeatability and objectivity, which most soil taxonomies proposed earlier are lacking. Modern data processing systems are capable of handling a large quantity of data faster and as a result, the use of numerical methods in the soil classification has become very much easier.

Prior to 1955 the use of numerical taxonomic methods was limited due to the fact that a large quantity of information could not be handled without the aid of a sufficiently powerful computer which was not available at the time. The selection of a set of attributes for classification and identification was done without appreciating the inter-attribute correlations and such methods are described by Arkley (1968) as suboptimal.

The traditional soil classifications have usually been hierarchical with a small number of differentia at each categorical level. This could distort the relative

relationships between soil individuals, and groups can be constructed in such a way that soils similar in a few characteristics but dissimilar in all other characteristics could be grouped together.

The application of numerical methods to soil classification is to achieve objectivity and repeatability but it must be noted that the early use of numerical methods in the classification of soils encountered several problems, for example the selection of a suitable similarity measure. A large number of similarity measures (similarity here refers to both similarity and dissimilarity measures) have been used to generate an inter-individual similarity matrix between soil individuals. At the early stages of numerical taxonomy product-moment correlation coefficient was a popular measure of similarity. As has been pointed out by Sokal and Sneath (1973, p.117), the choice of the similarity measure has been made without adequate theoretical justification. Moore and Russell (1967) demonstrated that five different similarity measures produced different classifications. There are some theoretical objections to the use of some similarity measures. Eades (1965) has objected to the use of product-moment correlation when the attributes are measured on different scales. The Euclidean distance metric has been used ignoring the correlation between attributes. This metric can be used only if the attribute vectors are mutually orthogonal. Kawaguchi and Kyuma

(1977) used the Euclidean distance after orthogonalizing the attribute vectors by means of principal component analysis (PCA).

Several sorting strategies of agglomerative cluster analysis are available but the outcome of the strategy depends very much on the similarity measure. Also these strategies may be different from each other in terms of the clarity of the clusters in the dendrogram produced.

Both principal component analysis and principal coordinate analysis (PCO) have been (Rayner, 1966; Cuanalo and Webster, 1970; Webster and Burrough, 1972; Campbell et al 1970; Norris, 1971) applied to soil classification but these methods have often failed to isolate clusters (Webster, 1976, 1979).

1.7 Aims and procedures of the present study

It has been found that the classifications obtained by numerical methods disagree not only with the traditional classsifications but also among themselves (section 1.6). When the numerical taxonomic methods were first introduced to soil taxonomy no attempt was made to examine their properties or their behaviour in relation to soil data. This problem still retards the progress of numerical soil taxonomy. Therefore, it has become necessary to evaluate at least the widely available procedures in relation to soil classification and the current work attempts this. The data used in this study was obtained from three sources (chapter 3).

It was shown earlier that there is no common agreement among the soil taxonomists on the nature of the natural soil classification. However, a growing number of soil taxonomists tends to agree that such a classification should be based on the measured properties of soils. But this decision alone would not produce a unique classification as the methods used dictate the nature of a classification, It is necessary to determine the best method of soil description. A given soil property may be measured at a series of depth levels and there are several ways of presenting such information (Lance and Williams, 1967a).Moore, Russell and Ward (1972) compared three soil profile models and concluded that both original data,weighted by an exponential function and the coefficients of depth functions fitted to all the attributes used in the study produced similar classifications. The use of all measurements of a given attribute may not be necessary,but the effect of the elimination of some depth levels and the nature and effect of inter-attribute correlations are examined (chapter 4). The use of depth functions is re-examined.

Although the hierarchical agglomerative strategies tend to produce similar classifications when the same inter-individual similarity matrix is used, the degree of distortion introduced by the clustering strategies varies. The dendrograms produced by them are different in terms of the clarity of the clusters. For example,

the single linkage sort method is known to suffer from the chaining of individuals rather than producing clusters, when there are intermediate types of individuals in the sample. In this study seven classificatory strategies were compared using the cophenetic correlation which was considered as a measure of distortion (chapter 5).

Although the ultimate objective of the taxonomic classification is to discover 'natural' groups (as defined earlier) it is difficult to assess such classifications in mathematical terms. The artificial classifications can be evaluated in relation to the utility of the classification defined, the taxonomic classifications cannot be evaluated so easily. But it may be possible to define a statistical criterion to compare classifications for numerical optimality (chapter 2).

It is possible to reallocate individuals until a measure defined achieves its optimum value. But in practice this is not feasible as the limit of computer time could impose a constraint. A possible way round this problem is to use other classificatory strategies and then find the best classification as determined by a suitable criterion and finally perform reallocation by an appropriate strategy. In this process both similarity measures and the classificatory strategies have been compared (chapters 6 and 7). The final classifications are compared with the traditional classifications and the relationship between the new soil groups and the parent material is also examined.

CHAPTER 2

A REVIEW OF NUMERICAL CLASSIFICATORY METHODS AND TEST CRITERIA

2.0 Introduction

Numerical classification is based on the phenetic similarity between individuals. Therefore, the measures to determine the phenetic similarity are of as fundamental importance as the classificatory strategies. In this chapter, methods which have been used to measure the similarity between individuals, classificatory strategies, methods to improve classifications and some possible tests for the optimality of classifications, are discussed. As outlined earlier, the main problem of traditional taxonomy is the lack of objectivity in classifying, and subsequently identifying, natural groups. The numerical strategies are aimed at eliminating the inherent subjective decision making process from traditional taxonomy and producing stable and 'natural' classifications. However, experience suggests that various numerical classificatory strategies produce incompatible results when applied to the same data. Therefore, it was felt necessary to examine the properties of some widely available numerical strategies and assess the relative merits of different methods in relation to soil classification.

2.1 Similarity measures

Similarity is defined here as a measure of closeness between pairs of individuals or groups, in a multi-dimensional space.

The data matrix X_{np} (n = individuals and p = attributes) can be presented as follows:

$$
\begin{array}{llll}
x_{11} & x_{12} & \ldots\ldots\ldots & x_{1p} \\
x_{21} & x_{22} & \ldots\ldots\ldots & x_{2p} \\
\ldots\ldots\ldots & & & \\
\ldots\ldots\ldots & & & \\
x_{n1} & x_{n2} & & x_{np}
\end{array}
$$

The space defined by more than two attribute vectors cannot be represented graphically but there are algebraic means to manipulate such data or to reduce the dimensionality. The matrix X can be examined in two ways, as has been pointed out by Cattel (1952):

i) Q analysis. Association between individuals (row vectors).

ii) R analysis. Association between attributes (column vectors).

The distinction between the two methods is not always clear because R analysis can be a preliminary step prior to the Q analysis as principal component analysis (PCA) is performed to obtain a sub-space with mutually orthogonal vectors before calculating the inter-individual similarity matrix.

2.1.1 Similarity measures for quantitative data

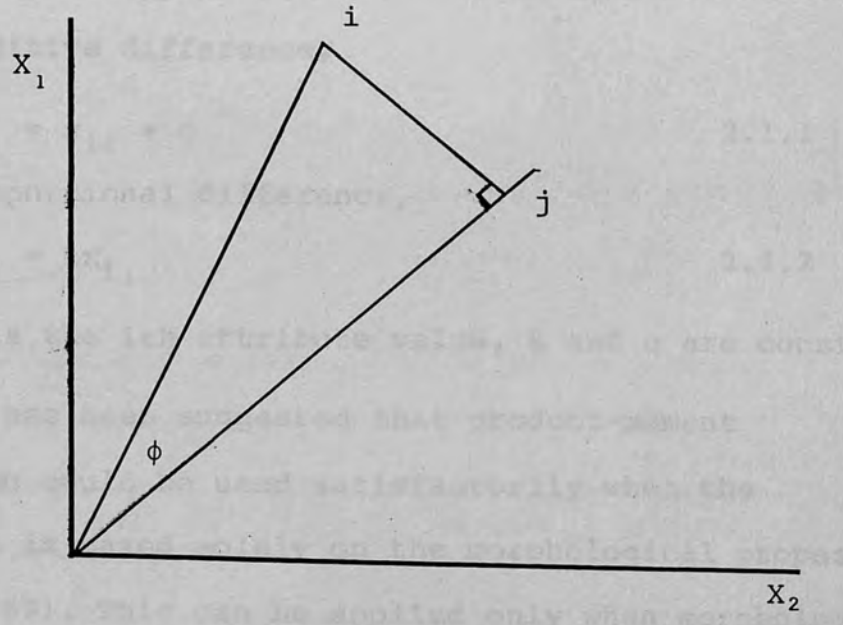
The inter-individual similarity matrix is the basis of agglomerative classificatory strategies, and also can be used to compare a set of groups of a given population. The measures generally used in determining the resemblance between individuals or groups can be considered in two ways.

(a) Coefficients of angular separation

(b) Distance coefficients

For simplicity, both kinds of measures are described here as similarity measures.

The former involves the determination of the angle between the row vectors of the matrix X.



This angle can be expressed as its cosine as used by Battacharyya or a distance $d^2_{ij} = 2 - 2\cos\phi$ as did Edwards and Cavalli-Sforza (Boyce, 1969). The most

widely used angular coefficient is the product-moment correlation which has been used in soil classification by Russell and Moore (1967. The use of product-moment correlation as a similarity measure has been criticised by Mincoff (1965) and Eades (1966) because it requires both directional and dimensional properties of the attributes to be the same. But soil attributes do not meet these requirements because different properties are measured using different scales and units. This problem can be solved to some extent by standardizing the data matrix to zero mean and unit variance. Product-moment correlation ignores both additive and proportional differences between individuals. The additive and the proportional differences can be explained as follows:

Additive difference,

$$x_{i1} = x_{i2} + c \qquad 2.1.1$$

Proportional difference,

$$x_{i1} = kx_{i_2} \qquad 2.1.2$$

where x_i is the ith attribute value, k and c are constants.

It has been suggested that product-moment correlation could be used satisfactorily when the comparison is based solely on the morphological properties (Boyce, 1969). This can be applied only when morphologically similar groups are required. This, therefore, is not a suitable measure of similarity for soil classification since it could produce soil groups with a high degree of heterogeneity.

The second group of similarity measures involves the determination of distance between individuals in a suitable space. The Euclidean space is preferred by Williams and Dale (1965) for three reasons.

(1) Many simple, robust and powerful methods are available in the Euclidean space.

(2) They have an advantage in hierarchical classification. The Euclidean functions possess the property associated with each level of division is a measure which falls as the hierarchy descends.

(3) It is easy to have an intuitive perception of the Euclidean systems.

A given distance measure is a metric only if it satisfies the following axioms.

i) $d_{ij} = d_{ji}$

ii) $d_{ij} \leq d_{ik} + d_{jk}$

iii) $d_{ij} = 0$ only if $i=j$

iv) If $i \neq j$ then $d_{ij} > 0$

All distance measures in use are not metric. Sneath and Sokal (1973,p.120-121) have identified several categories of distance measures in relation to their metric properties.

a) metric. Those which satisfy all four axioms above.

b) semi-metric or pseudometric. Those measures which satisfy the axioms I, II and III but not IV.

c) ultra-metric. The axiom II can be relaxed in the following way.

$$d_{ij} \leq \max[d_{ik}, d_{jk}]$$

Those measures which satisfy the axioms I,III, IV and the revised II are called ultrametrics.

d) non-metric, the rest of distance measures.

The concept of distance can only be illustrated graphically in a two dimensional space. But the theorems of three dimensional geometry can be extended to the pth dimension (Sneath and Sokal, 1973, p.122). A series of distance measures in the Euclidean space can be generalized in the following way known as Minkowski metrics.

$$d_{(ij)r} = \Sigma(|x_{ik} - x_{jk}|^r)^{1/r} \qquad 2.1.3$$

When r=1 the Minkowski metric becomes the City Block or the Manhattan metric. Three standardization methods have been used with this metric.

(1) Standardization by $(X_{ik} + X_{jk})$ is known as Canberra metric (Lance and Williams, 1967a).

(2) Standardization by the range r_k was suggested by Gower (1971). The range can either be the sample range or the population range if known. The Gower metric has not been used in soil classification to the author's knowledge.

(3) Standardization by $|X_{ik} + X_{jk}|$, which is known as Bray-Curtis measure (Clifford and Williams, 1976) because it was first used by them.

All three forms of the Manhattan metric can only be used when all data are non-negative and non-zero because in both occasions the metrics take the highest value 1, irrespective of the absolute values of X_{ik} and X_{jk}. They are not invariant under the rotation of the vectors and also cannot be used with coefficients of depth dependent functions. The main advantage is the self-standardization built into the metric. The devisive program REMUL (Lance and Williams, 1975) employs the Canberra metric for reallocation of individuals after a monothetic split and also for the global reallocation after the final split. Webster and Burrough (1972) compared it with the Euclidean metric and concluded that the results were comparable. This metric has been popular for its simplicity.

The Coefficient of Divergence defined by Clark (1952) to measure the similarity between different populations of snakes is also based on the Manhattan metric.

$$CD_{ij} = \frac{1}{n} \sum_{k=1}^{p} \left(\frac{X_{ik} - X_{jk}}{X_{ik} + X_{jk}} \right)^2 \qquad \text{.....2.1.4}$$

Where CD_{ij} is the similarity between ith and jth individuals. p is the number of common properties of the two populations or individuals.

When r=2 the Mincowski metric becomes the Euclidean metric of the following form.

$$d_{ij} = \left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}} \qquad \text{..... 2.1.5}$$

The most widely used form of this metric is its squared form divided by the number of attributes p. All the metrics so far considered can be divided by the number of attributes common to the two individuals to solve the problem of missing attribute values. The Euclidean metric is sensitive to aberrant character values and therefore it is affected by linear transformation of data but it is invariant under rotation of the vectors. The Euclidean metric requires the data to be standardized by a suitable method and the most widely used method is the transformation to zero mean and unit variance as follows:

$$x_{ij} = \frac{X_{ij} - \overline{X}_j}{\sigma_{xj}} \qquad \ldots\ldots\ldots\ldots \quad 2.1.6$$

where σ_x is the standard deviation of x

The Euclidean distance was first introduced to numerical taxonomy by Sokal (1961) and subsequently used by other workers. It has been used in soil classification by Russell and Moore (1967), Moore, Russell and Ward (1972) and Webster and Burrough (1972). This can be a reliable measure of relative similarity between individuals only if the attribute vectors define an Euclidean space, in which all vectors are mutually orthogonal. But this is not the case in reality since soil attributes are correlated to varying degrees. Therefore the space defined by the original attributes is distorted. This problem could be solved in two ways.

(1) By choosing a sub-set of uncorrelated attributes.

(2) By orthogonalizing the attributes by a suitable method.

The former method was used by Sarkar et al (1966) to classify a sample of soil profiles. Alternatively a set of mutually orthogonal vectors can be obtained from originally correlated attributes by performing principal component analysis (PCA) on the data. The new vectors obtained by this method contain the same information about the population, as they are linearly related to the original data, and it is also possible to obtain a space of a reduced dimension as some attributes are redundant. Kawaguchi and Kyuma (1977) used this method to transform data prior to the calculation of the Euclidean distance similarity matrix to classify a population of paddy soils sampled in South East Asia. But the Euclidean distance has been used as a similarity measure in soil classification in the past without taking this requirement into consideration (Russell and Moore, 1967).

A similarity measure was proposed by Pearson (1926) known as Coefficient of Racial Likeness (CRL) to classify anthropological populations.

$$CRL = \frac{1}{n} \left[\frac{(X_{ik} - X_{jk})^2}{(S^2_{ik/n_i}) + (S^2_{jk/n_j})} \right] - \frac{2}{n} \quad ....2.1.7$$

Where S^2_{ik} and S^2_{jk} are variances for the groups i and j for the k'th attribute. n_i and n_j are the number of individuals of the groups i and j.

Mahalanobis (1936) proposed a measure based on the CRL as a discriminant function for a two group situation. This measure is known as the Mahalanobis D^2, which has been generalized to more than two groups by Rao (1952). The Mahalanobis D^2 between two mean vectors can be represented in the following form.

$$D^2 = (\mu_i - \mu_j)V^{-1} (\mu_i - \mu_j)^T \quad \ldots\ldots \quad 2.1.8$$

Where μ_i and μ_j are mean vectors for the groups i and j. V is the pooled within groups variance-covariance matrix.

When the properties of the individuals are measured at a single depth level, V can be replaced by A (Total Variance-Covariance matrix) as has been pointed out by Webster (1975) but it is not necessary for soil individuals as most of them have data for several depth levels.

The Mahalanobis D^2 is a discriminant function based on several assumptions, which are:

(1) The population under consideration has a multivariate normal distribution.

(2) Homoscedasticity of the dispersion matrices V_k ($V_1=V_2= \ldots=V_k$)

(3) The samples are randomly drawn.

Originally Sokal and Sneath quoted by Sneath and Sokal (1973, p.127-128) objected to this measure for four reasons.

(1) Individuals or taxonomic units were not drawn randomly from different populations.

(2) The computational load involved was too high for the computers available at that time.

(3) D^2 required quantitative variables.

(4) The validity of the underlying assumptions was questionable.

According to the same authors (Sneath and Sokal, 1973, p.127) the first three objections have now become irrelevant but the fourth can still be important. The use of D^2 can be justified on the grounds that it is robust enough to give reasonable results under the violation of the normality and the homoscedasticity of the dispersion matrices V_k.

The homoscedasticity of the dispersion matrices can be tested using the following test.

$$C = - \sum_{k=1}^{g} n_k - 1 \log_e \frac{|V_k|}{|V|} \quad \ldots\ldots\ldots \quad 2.1.9$$

V_k is the within-group variance-co-variance (var-covar) matrix for the kth group; g is the number of groups.

$C \sim \chi^2$ with $p(p+1)(g-1)/2$ degrees of freedom.

C, however, is more sensitive to departures from the normality than the tests based on the dispersion matrices.

The Mahalanobis distance can be used to determine the similarity between soil individuals because they are heterogeneous bodies consisting of several layers (horizons) and therefore can be treated as groups for the purpose of calculating the inter-individual similarity matrix. The scatter matrix V_k(k=1, 2,...,n) can be calculated for the soil individuals from which the pooled within-group var-covar matrix V can be calculated. A typical soil individual can be represented as a data matrix of the following form.

$$x_i = \begin{pmatrix} x_{11} & x_{12} & \cdots & \cdots & x_{ip} \\ x_{21} & x_{22} & \cdots & \cdots & x_{2p} \\ \cdots & \cdots & & & \cdots \\ \cdots & \cdots & & & \cdots \\ x_{d_1} & x_{d_2} & \cdots & \cdots & x_{dp} \end{pmatrix}$$

Where d is the number of depth levels and p is the number of attributes. The Mahalanobis distance between two soil individuals is the distance between their centroids. The two soil individuals occupy a space similar to two groups rather than two points, when they are represented by single vectors. Mahalanobis D^2 was used in this study (chapters 7 and 8) to solve the problems associated with the Euclidean distance. There are several attractive features in the Mahalanobis distance. It is invariant under the linear transformation

of data and therefore, it is not necessary to standardize data prior to the computation of D^2. When the attributes are correlated, which is the case with soil attributes, D^2 measures the relative similarity more accurately than the distance measures based on the Euclidean space. To compute the D^2 values the problem of missing values was solved in this study by using attribute means. This is possible only when the number of missing values is few.

Blackith and Rayment (1971) compared D^2 with the Euclidean distance in the original space and found that the Euclidean distance values were exagerated unequally over all points of the sample because the vectors were not orthogonal. Therefore, the Euclidean distance in a distorted space is not suitable for the purpose of classification.

As has been demonstrated by Webster (1977, p.194-195), the Mahalanobis distance can be displayed geometrically by a canonical vector analysis. The groups produced by a classification or the individuals themselves as 'groups' can be plotted in a two dimensional space whose vectors are mutually orthogonal. The best fit two dimensional space is the one which represents the greatest proportion of the variation of the population. The total number of canonical vectors which can be obtained equals one less than the number of groups or attributes, which ever is smaller. The percentage variance represented by the first two canonical vectors is dependent on the number of vectors in the new space. When both the number of groups (g) and the number of attributes (p) are more than three,

it is possible to find the best fit two dimensional space which represents the greatest proportion of the variance.

Canonical vector analysis is similar to principal component analysis except for the fact that the matrix $(W^{-1}B)$ is used in the place of the total variance-covariance matrix (A). The latent roots λ_i of the matrix $(W^{-1}B)$ can be found by solving the determinantal equation:

$$|W^{-1}B - \lambda I| = 0 \quad \ldots\ldots\ldots \quad 2.1.10$$

The ith canonical vector c_i is given by;

$$|W^{-1}B - \lambda I| c_i = 0 \quad \ldots\ldots\ldots \quad 2.1.11$$

The centroids of the groups or soil individuals can be obtained by the following relationship.

$$z_i = \mu_t c_i \quad \ldots\ldots\ldots\ldots \quad 2.1.12$$

where z_i mean vector in the transformed space.

μ_i mean vector in the original space

c_i canonical loadings.

The multi-level observations of soil individuals can be represented by their centroids (chapters 6 and 7). Only the soil properties are treated as independent attributes which helps reduce the number of attributes in the original space. A number of methods have been used to solve the problem of soil heterogeneity. Canonical analysis was performed on the classifications obtained by different methods (chapters 6 and 7). Webster (1976,1977, 187-200)

has shown that when the soil groups are homogeneous, they occupy distinct parts of the canonical space. Therefore, canonical vector analysis was preferred to principal component analysis on the unclassified population to identify groups in a two dimensional space. The canonical plots of centroids of soil individuals can be compared with the dendrograms produced by the classificatory strategies to identify possible misclassifications.

The probablity contours of the canonical space are circular as canonical vectors are mutually orthogonal. The radius of the probability circles for the groups is $\sqrt{\chi^2}$ and for the group centroids it is $\sqrt{\chi^2/n_k}$ (Webster, 1977, p.196) with 2 degrees of freedom.

2.1.2 Similarity measures for qualitative data

Like quantitative attributes, qualitative attributes can also be used to compute the similarity between individuals or groups. Some such measures have been developed independently for the processing of qualitative data whereas some of them are extensions of their quantitative counterparts. Both product-moment correlation and the Euclidean distance, for example, could be applied to qualitative data. Cheetham and Hazel (1969) have listed twenty three similarity measures for binary data

The angular separation between the row vectors of the data matrix X can be measured as the cosine of the

angle between them or in the form of the product-moment correlation coefficient. However, it is necessary first to produce a contingency table for pairs of individuals in the following form.

Ind. i	Ind. J: 1	Ind. J: 0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	

a - number of characters coded 1 in both ith and jth individuals

d - number of characters coded 0 in both ith and jth individuals

The other two quadrats b and c represent the number of mismatches.

Therefore, the following information can be derived from the contingency table.

Total number of attributes $n = a + b + c + d$

Total number of matches $m = a + d$

Total number of mismatches $u = c + c$

There are methods to transform qualitative data into quantitative data and vice versa. The way in which it may be done has been demonstrated by Wishart (1969a) as follows:

Binary to Numeric

1	1.0
0	0.0

Ordered m. state to Numeric

1	1.0
2	2.0
3	3.0

But unordered (disordered) multi-states have to be transformed first to binary if they are to be transformed to numerical (quantitative) data.

Unordered m. state to Binary

white	1
brown	2
red	3
yellow	4

1	2	3	4
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Quantitative data can be transformed to binary or ordered multi-states by dividing the attribute range into a several parts and assigning binary codes. The most convenient binary codes in use are 1 for the presence and 0 for the absence of a particular state. However, the codes to be assigned to binary attributes depend on the computer program available, some of them use alphabetic codes to separate quantitative attributes from qualitative attributes. Transformation of quantitative attributes to

binary form involves discarding of a certain amount of information and is therefore conceptually unattractive. On the other hand, loss of information due to transformation of binary attributes to numerical mode may be less. The ordered multi-state attributes can be treated as quantitative attributes as the magnitude of the ranks is meaningful. Most classificatory programs treat this type of data as quantitative for computational simplicity, but unordered multistate and binary data are usually treated as a separate category. Unordered multistate attributes are the most difficult type and these are processed after transforming to binary mode. It is more convenient to have all attributes in a single mode for the purpose of numerical analysis. The cluster analysis package CLUSTAN 1C (Wishart, 1969a) can handle only one data mode at a time but there are mixed data programs now available, the most important of which is the Canberra computer package TAXON (CSIRO).

2.1.3 Similarity measures for mixed mode data

Calculation of the similarity matrix without first transforming the mixed mode data to a single mode has been discussed by Talkington (1967), Burr (1968) and Lance and Williams (1968, 1975). This is particularly important for the soil taxonomist whose information about soils usually contains mixed attribute types (chapter 3). Therefore, the use of qualitative attributes, where possible, should be encouraged. But it must be noted that certain

similarity measures may require all quantitative data (e.g. Mahalanobis D^2), in which case transformation of qualitative attributes to quantitative form may be attempted on an experimental basis. In this section, the use of mixed mode data to compute similarity measures is reviewed.

Lance and Williams (1967a)pointed out that one important desirable property of a similarity measure was that it should be additive over attributes. In the case of the Euclidean metric, this is achieved by taking the squared form. This property of a similarity measure can be used to devise a system of standardization which would ensure that all attributes have equal weight. This system can be applied to mixed mode data as has been demonstrated by Burr (1968). According to Lance and Williams (1967a)the first known Euclidean system for mixed mode data was that of computer program TAXAN of Burr. Burr (1968) has discussed eight methods of standardization for both quantitative and qualitative data, but preference has been given to two methods.

(1) Standardization by mean squared distance between pairs (MSDBP) $2\sigma_m^2$.

(2) Half the MSDBP is the sample variance for a given attribute as was demonstrated by Burr (1968) σ_m^2 .

σ_m^2 for quantitative attributes can be calculated as follows:

$$\sigma_m^2 = \frac{n_m \Sigma X_{im}^2 - (\Sigma X_{im})^2}{n_m(n_m - 1)} \quad \ldots\ldots \quad 2.1.13$$

where n_m is the number of individuals available for attribute m

The variance for a qualitative attribute can be obtained by the following relationship.

$$2\sigma_m^2 = \frac{n_m - \Sigma f_{ms}^2}{n_m(n_m - 1)} \quad \ldots\ldots \quad 2.1.14$$

Where f_{ms} is the number of individuals in state s of the mth attributes.

For the multistate attributes the variance can be calculated separately and added together. Thus the Euclidean distance between a pair of individuals i and j can be calculated by;

$$d_{ij}^2 = \frac{k_m^2 \Sigma (X_{im} - X_{jm})^2}{p_{ij}} \quad \ldots\ldots \quad 2.1.15$$

Where k_m^2 is the standardization factor.

p_{ij} is the number of attributes common to ith and jth individuals. For binary attributes $(X_{im} - X_{jm})=0$ when both individuals share a particular character state.

The Canberra metric is self standardizing and it is suitable for mixed mode data as has been demonstrated by Lance and Williams (1968, 1975). Similar methods can be applied to other similarity measures too. Since the

present work was based mainly on quantitative data more attention was paid to methods related to such attributes. Some of the similarity measures described above for quantitative data can be used for mixed mode data by choosing the corresponding metric for qualitative data as listed below (Table 2.1.1).

TABLE 2.1.1 Some Similarity Measures for Mixed Mode Data

Metric for Quantitative Data	Metric for Qualitative Data
Squared Euclidean distance standardized by the range	Simple matching coefficient $(a+b)/(a+b+c+d)$
Gower metric	"
Bray-Curtis measure	Czechanowski measure $(b+c)/(2a+b+c)$
Canberra metric	Jacard measure $(b+c)/(a+b+c)$

The similarity measures described above can be applied to mixed mode data ensuring equal weights to all attributes. But the properties of these measures should be examined empirically.

2.2 Classificatory strategies

Cluster analysis strategies have been used to obtain a hierarchical structure of a population or to partition a population of n individuals into k groups ($k < n$), which should be more useful and informative than the population as a whole. At this point, however,

a formal definition of a cluster is necessary. Several attempts in this direction have produced vague definitions as has been pointed out by Sneath and Sokal (1973,p.194/195). Rao (1952) recognized the fact that a formal definition of a cluster is not easy and as a result it has been kept vague and according to Bonner (1964) the ultimate criterion of evaluating a cluster should be the value judgement of the taxonomist. Everitt (1974,pp.43-48), however, considers that clusters are continuous high density regions separated by low density regions in a multi-dimensional space. This concept has an intuitive appeal but in the case of soils it has been demonstrated that an even distribution of points in a two dimensional space could be expected (Webster, 1977). This may have resulted from the fact that soils grade from one group to another, and also that soil boundaries can be obscured when a multivariant population is reduced in two dimensions. The notion of density in a multi-dimensional space can be expressed as the number of points per hypervolume (Sneath and Sokal, 1973). The main advantage of this definition of the cluster is that it does not need to have a particular shape. The clusters defined by high density regions in a multi-dimensional space should have the greatest homogeneity and therefore minimisation of within group variance can be considered as an objective of the classificatory strategies.

There is a whole range of cluster seeking strategies available to the numerical taxonomist and therefore the

properties of such strategies should be examined before applying any to the data. The numerical strategies used to identify groups can be divided into two categories.

(1) Classificatory methods (strict sense)

(2) Ordination

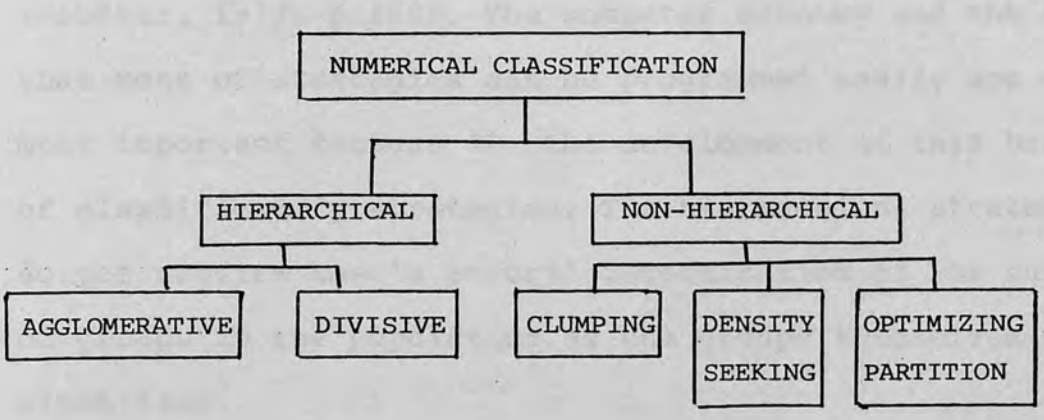
These two types of strategies are different mathematically and may have different purposes. Ordination is not necessarily for identifying groups in multivariate populations, it can also be used as a method of data reduction or as a method of data transformation for further statistical analysis. The ordination methods have been used by numerical taxonomists to identify clusters by projecting the population on to a two dimensional sub-space. In this context ordination is a non-hierarchical cluster seeking procedure, but here it is discussed as a separate group of strategies since ordination has a wider use than classification.

What are here called the classificatory strategies can be divided into two broad classes and in turn each class of strategies can be subdivided:

(1) hierarchical strategies

(2) non-hierarchical strategies

This distinction is important because of the computational implications. As will be demonstrated later a large number of strategies in each category have common computational procedures. The taxonomy of classificatory methods can be illustrated by the following diagram.



Sneath and Sokal (1973, p.202-214) have considered eight aspects of clustering of which the hierarchical/non-hierarchical dichotomy seems to be appropriate for this discussion as these two forms of clustering have distinct computational procedures and also conceptual relationships with the nature of a given population. It may be fair to suggest that the most widely used clustering strategies are hierarchical for two major reasons.

(1) The classificatory strategies developed by biological taxonomists placed a considerable emphasis on hierarchical methods on the assumption that biological organisms were hierarchically related and the task of the taxonomist was to uncover that hierarchy. This is important as the development of numerical taxonomic methods was due to the works of the biological taxonomists. Therefore the computer software is available for a wide range of hierarchical classificatory strategies.

(2) Mathematical and computational considerations (Webster, 1977, p.160). The computer economy and the fact that most of strategies can be programmed easily are the most important factors in the development of this branch of classificatory strategies. The hierarchical strategies do not require the 'a priori' determination of the number of groups in the population as the groups themselves are classified.

Because of these reasons, the majority of studies on the use of numerical methods for soil classification are confined to the hierarchical strategies. This may also have resulted from the fact that the major traditional soil classification systems, such as the USDA system are hierarchical despite the fact that soil populations are not hierarchically structured like biological populations.

2.2.1 Hierarchical classificatory strategies

The cluster analysis strategies described as hierarchical can again be divided into two categories as agglomerative and divisive methods. The distinction between the two types of strategies is in the way in which a population is divided into groups. The agglomerative methods are essentially based on an inter-individual similarity matrix and the strategy proceeds by fusing the most similar pairs of individuals or groups working up the hierarchy, whereas the divisive strategies begin at the population level by dividing the whole population into two groups on the basis of a selected criterion. Lance and Williams (1975) suggest that the

hierarchical divisive methods are preferable to the agglomerative methods for three reasons.

(1) The divisive strategies begin at the highest level of information by considering the whole population.

(2) The divisive strategies are computationally economic. They do not require an inter-individual similarity matrix and the divisive process does not have to proceed to the level of individuals and therefore large populations can be handled conveniently.

(3) The divisive strategies allow the use of reallocation procedures if the hierarchy is not the main interest. Therefore correction of misclassifications after each split and after the final split is possible. Lance and Williams (1975) used reallocation to obtain a polythetic classification in the divisive program REMUL (CSIRO).

The main disadvantage of the divisive methods is that most of them are monothetic, whereas agglomerative methods are polythetic as the similarity matrix can be calculated using any number of attributes. The cluster analysis package CLUSTAN (Wishart, 1969a), has eight agglomerative strategies and CSIRO computer package TAXON has one mixed data and one quantitative data divisive program.

2.2.1 (a) Agglomerative Methods

Lance and Williams (1967a)have demonstrated that all agglomerative strategies so far proposed obey the following linear relationship and a large number of new

strategies can be obtained simply by changing the parameters of the equation. The distance $d_{k(ij)}$ between the individual k and the group formed by fusing ith and jth individuals can be obtained by the equation.

$$d_{k(ij)} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} - \gamma |d_{ki} - d_{kj}| \quad \ldots\ldots\ldots 2.2.1$$

Where the parameters α, β and γ are arbitrarily determined. They control the structure of the dendrogram drawn for the classification. The values of those parameters in the equation for eight strategies are listed in the Table 2.2.

The agglomerative strategies which obey the equation above are described as combinatorial (Lance and Williams, 1967b) in that after each fusion the new similarity matrix can be computed from the previously calculated similarities, saving a considerable amount of computer time and storage. Therefore, it is not necessary to save the original data matrix.

These strategies are compatible in that the similarity measures calculated later are exactly the same kind as the initial similarity measures. An incompatible strategy is one which has lost this property and which Lance and Williams (1967b) consider to be inappropriate for cluster analysis.

The agglomerative strategies can be considered in terms of their effect on the original space. There

are two types of such effects. They can either be space conserving or space distorting. The strategies which do not cause a distortion of the space defined by the original similarity matrix due to the group formation are considered as space conserving strategies (Lance and Williams, 1967b).But certain of the agglomerative strategies behave as if the space in the immediate vicinity of a group is contracted or dilated, such strategies are known as space distorting strategies. The space distorting strategies tend to move groups as they are formed nearer to another individual rather than allowing another individual to act as a group centroid. This tendency is known as chaining (Lance and Williams, 1966).

TABLE 2.2 Values for the Parameters of the Equation 4.2.1 For 8 Agglomerative Strategies

		α_i	α_j	β	γ
1.	Single Linkage Method	1/2	1/2	0	-1/2
2.	Complete Linkage Method	1/2	1/2	0	1/2
3.	Average Linkage Method	n_i/n_i+n_j	n_j/n_i+n_j	0	0
4.	Centroid Sort	n_i/n_i+n_j	n_j/n_i+n_j		0
5.	Median Sort	1/2	1/2	-1/4	0
6.	Ward's Error-Sum-of-Squares	$(n_k+n_i)/N$	$(r_k+n_j)/N$	$-n_k/N$	0
7.	Macquitty's Sim. Analysis	1/2	1/2	0	0
8.	Lance and Williams Flexible Sort Method	$1-(\beta+\alpha_j)$	α_i	$1-\alpha_i+\alpha_j$	0

Where $N = n_i + n_j + n_k$

Single Linkage (nearest neighbour) method

This method is compatible with any kind of similarity measure. The distance between two groups or an individual and a group is defined as the distance between the two nearest individuals. This method is known to suffer from chaining when there are intermediate types of individuals. The single linkage sort, however, has been used by Rayner (1966), and Moore and Russell (1967) in soil classification. Sneath (1957) has used it in ecological classification. This method of cluster analysis has become known through the works of Sneath (1957). As groups grow they become closer to the unclassified individuals giving rise to chaining. This is the main weakness of this strategy (Wishart, 1969b; Webster, 1977, pp.101-165). The chaining effect is more common in soil classification as soil groups tend to overlap on most dimensions. It was found by Russell and Moore (1967) that soil groups cannot be identified easily by this method.

Complete Linkage (furthest neighbour) method

This is the compliment of the single linkage method in that the nearest distance between two groups or a group and an individual is defined as the distance between their furthest members. This method suffers from the same problems as the single linkage method. The process of fusion could lose the monotonicity due to a process called 'reversing' (Webster, 1968) which is caused by a temporary increase of similarity after a fusion step.

Average Linkage method

The average linkage method was developed by Sokal and Mitchener (1958). The similarity between two groups or an individual and a group is defined as the average distance between all pairs of individuals of the two groups to be fused. This strategy is somewhat similar to the centroid sort but unlike it, the average linkage sort proceeds monotonically. This strategy also has the tendency to chaining (Moore and Russell, 1967; Cuanalo and Webster, 1970).

Centroid Sort

Webster (1977, p.166-167) suggests that from a geometrical point of view this is the most attractive strategy. But the main problem of the method is that it does not proceed monotonically as reversing occurs.

Median Sort

This is also known as weighted centroid method, which was proposed by Gower (1967) to eliminate the influence of group size on the fusion of groups and individuals. Anderberg (1973) has pointed out that the median sort strategy can be used only for distance measures, since similarity measures such as product-moment correlation were geometrically meaningless.

Ward's Error-Sum-of-Squares (ESS) method (Ward,1963)

This strategy is somewhat different from the rest and produces tight clusters. The fusion is between those

groups or individuals and groups which have the minimum increase of average distance between the centroid of the new group and its members. This strategy tends to produce spherical clusters with minimum variance. The method favours the fusion of small groups and individuals and is also strongly space dilating. The main problem of the strategy is that the minimization of error-sum-of-squares may require a reallocation of individuals at coarser levels of cluster formation and that is not possible because of the hierarchical nature of the strategy. Webster (1977, p.174-175) suggests that this strategy may look like a non-hierarchical strategy of minimizing within groups sum of squares, but it is strictly hierarchical. On the other hand, it is possible that this method imposes an artificial structure on an unstructured population. Therefore it may be advisable to use this method with caution.

Lance and Williams (1966) Flexible Sort

This method is based on the following quadruple constraint.

1. $\alpha_i + \alpha_j + \beta = 0$
2. $\alpha_i = \alpha_j$
3. $\beta < 1$
4. $\gamma = 0$

The user can change the strategy by varying the value of β between -1 and 1 but the use of -0.25 is suggested

by Lance and Williams (1967b) as most appropriate on the basis of the past experience with the strategy. The clarity of clusters depends on the choice of the value for β rather than the nature of the population under study. This is the main disadvantage of this strategy. Moore and Russell (1967), Campbell et al (1970) used this method in soil classification. Webster (1977, pp.172-174) suggests three advantages of the flexible strategy.

(1) All individuals are fused early

(2) Chaining could be avoided by choosing a suitable value for β

(3) The structures obtained by this method closely relate to the PCA plots of the same data.

Macquitty's Similarity Analysis

This method is the same as the flexible sort of Lance and Williams with $\beta = 0$. Chaining of large populations could occur.

Computer Time

The advent of fast computers has made it possible to handle large problems fast and efficiently and the constraint of computer time has been reduced to some extent. But computer economy is an important factor in data processing. The number of computations associated with an agglomerative strategy is very high compared to hierarchical divisive methods.

The number of claculations for the generation of an inter-individual similarity matrix is $p(n-1)^2$ according to Williams (1976a). The programs first compute the similarity matrix and then proceed to cluster analysis combinatorially, require the following number of calculations:

$$1/2\ p\ n(n-1) + 1/2(n-1)(n-2)$$

where p is number of attributes

n is number of individuals

If the original similarity matrix is calculated outside the program, then the number of calculations required is $1/2(n-1)(n-2)$. All the above mentioned strategies are available in the CLUSTAN 1C package on the CDC 7600 computer at the University of London Computer Centre (ULCC), and five agglomerative strategies are available in the TAXON cluster analysis package (CSIRO, Canberra, Australia) which was implemented by the author on the CDC 7600 at ULCC. Certain properties of these strategies are examined in Chapter 5.

2.2.1 (b) Hierarchical Divisive Strategies

These cluster analysis methods are the complement of the agglomerative methods. The clustering process begins at the population level by splitting the population into two groups on the basis of a selected criterion. Theoretically the subdivision of the population can proceed down to the individual level as suggested by Edwards and Cavalli-Sforza (1965) but in practice it can be done only if the population is very small (Everitt, 1974, pp.18-21) even with a large computer.

The main advantages of hierarchical divisive strategies over the hierarchical agglomerative strategies were discussed in the section 2.2.1. Mcnaughton-Smith et al.(1964) suggest that the divisive methods are preferable as 'false' decisions made at the early stages of fusion in agglomerative strategies will distort its subsequent course. However, there are $(2^{n-1} - 1)$ ways of dividing n individuals into two groups and it can be computationally impossible when the population involved is very large.

Williams (1976b) reviewed the divisive algorithms available then and recognized three types of strategies in respect of the type of data they could be applied to.

(1) All binary data strategies.

(2) Quantitative strategies.

(3) Mixed data strategies

These programs can either be monothetic (split on a single attribute) or polythetic (all attributes are considered).

The first monothetic strategy for binary data was that of Williams and Lambert (1959). It has no provision for missing attribute values. "Association analysis" of Williams and Lambert (1960) is the most prominent monothetic divisive strategy according to Sneath and Sokal (1973, p.203). It has been widely used in ecology and is suitable only for binary data; quantitative data can be transformed to binary mode to be used with this strategy but the loss of information in the transformation process makes it unattractive.

The early method of polythetic division was by identifying the most dissimilar pairs and allocating the rest to them. This method is faced with two problems:

(1) It requires (1/2n(n - 1)) number of computations which can be extremely expensive and impractical for large populations.

(2) The strategy is unduly dependent on data.Some data sets have shown that once a deviant individual is set aside nothing else would join it (Williams, 1976b).

The TAXON (CSIRO, Canberra) program POLYDIV is for all numerical data and is a polythetic strategy with provision for missing values. This program proceeds by performing PCA to extract the first eigen-vector which subsequently is used to obtain two groups by splitting the population at the mean. This process is repeated a specified number of times without reallocating individuals. Since the division is based on the first eigen vector this strategy is suitable only when the number of attributes is small and the first eigen vector accounts for a great proportion of the variance. The absence of facilities for the reallocation of individuals is a disadvantage of this strategy. Lambert et al (1973) have devised a similar strategy for all quantitative data.

The first programmed mixed data divisive strategy was that of Boulton and Wallace (1973). This computer program splits the population randomly into two groups and then reallocates individuals until

stability is achieved. It also allows missing attribute values. Williams (1976b) had previously pointed out two disadvantages of this method.

(1) Division begins by splitting the population into two equal groups at random. When n is large and the division is nowhere near the optimum point as defined by the criterion chosen, the number of reallocations will be so great that the strategy can become very slow.

(2) The use of a stopping rule causes the subdivision less fine than usually required by the user. The use of stopping rules has been criticized by many workers (Bottomley 1971).

Another approach to division is based on discriminant analysis, a linear discriminant function is computed after the initial split and individuals are iteratively reassigned and the discriminant functions are computed again. The process is repeated until most mutually separated groups are found. This method has been used by Casetti (1964) and Dubes (1970) as has been illustrated by Anderberg (1973, p.155).

The TAXON program REMUL has been devised to eliminate some problems encountered by the previous strategies. This program aims to satisfy five distinct needs (Lance and Williams, 1975).

(1) The use of a wide range of data types. All four types of data described in the chapter 3 can be used, and the presence of missing values is allowed.

(2) The initial split is attempted somewhere near the optimum to minimize the number of reallocations.

(3) The reallocation must not be group size dependent.

(4) There is provision for the terminal reallocation among all groups.

(5) Provision is made for the allocation of new individuals to existing groups in order to process large populations.

This program performs three main functions:

(1) Primary division. This involves the splitting of the population or the sub-groups with the lowest homogeneity into two groups. The primary split is monothetic, based on the attribute which has the highest number of correlated attributes.

(2) Reallocation after each split. The transfer of individuals can take place between the two groups just formed. The distance between individuals and group centroids is calculated after notionally removing the individual under consideration from its parent group and those individuals which are closer to the group other than their parent groups are transferred. The distance metric used here is the Canberra metric, which computes the distance using quantitative and multistate attributes separately and finally adds them together as has been described by Lance and Williams (1975). Since only one individual is considered at a time the missing

attributes can be excluded and therefore it is not necessary to divide the metric by the number of attributes. The homogeneity of a given group is the average distance of its members from the centroid and this measure is used to decide which group should be split next.

(3) Reallocation after the final split. Until the final group structure is obtained the reallocation is performed only between the two groups just formed. But the global reallocation is allowed after the final split which is decided by the user. The distance between all group centroids and individuals is calculated as before and the transfer occurs when an individual is closer to any other group than its parent group.

This program has the advantage that it can handle large populations by first classifying a sub-sample and then allocating the rest to existing groups or initiating new groups if necessary.

The computer time required by the strategy is proportional to the square of the number of individuals and attributes.

2.3 Non-hierarchical methods of cluster analysis

In the previous section a series of hierarchical clustering methods, developed mainly to discover hierarchical relationships in biological populations was discussed. The application of such techniques to essentially non-hierarchical populations such as soils gives rise to both philosophical and practical problems.

Soils are not expected to be hierarchically related like biological organisms as they have developed in diverse parent materials under diverse environmental conditions. Because of the relative ease of information handling, hierarchical soil classifications were preferred by traditional soil taxonomists. The USDA Comprehensive System (USDA, 1975) is the best example of a hierarchical 'traditional' soil classification. As has been pointed out by Webster (1968) ordering soil groups in a hierarchical manner could seriously distort the actual relationships. Numerical soil taxonomists have, however, used hierarchical strategies to classify soils becuase of the ready availability of hierarchical computer programs. The main problem of the hierarchical strategies is that they do not allow one to correct misclassifications at the beginning of the strategy.

A whole range of non-hierarchical strategies have been developed for non-hierarchical populations and in this section the most widely mentioned strategies are reviewed, and the merits and the demerits of such strategies are examined. The non-hierarchical strategies involve the partition of the whole population into k ($k<n$) groups simultaneously and the subgroups are not ordered in a hierarchical manner. On the other hand the relationship between individuals in a given group is not examined.

Everitt (1974, p.7) identifies three major kinds of non-hierarchical clustering techniques.

(1) Optimization-partition techniques.

(2) Density or mode seeking techniques.

(3) Clumping techniques.

The first two methods produce mutually exclusive clusters, whereas the third method produces overlapping clusters. The clumping techniques have been used in linguistic classifications (Everitt, 1974, p.35-37).

The optimization partition techniques involve defining of a set of k seeding points in the multidimensional space and subsequent allocation of the rest of the population to the seeding points, and the reallocation of individuals so as to optimize a suitable function.

A wide range of methods to obtain the initial partition has been discussed by Anderberg (1973,pp.156-159) and Everitt (1974, p.24-25). There are two decisions to be made on the choice of seeding points.

(a) The value of k (number of seeding points). This may be done either by using the users knowledge of the nature of the population or by choosing an arbitrary value for k. The main disadvantage of this method, however, is that when little is known about the population the chances of making an error are very high.

(b) A method of choosing the k points in the multidimensional space. This is also an arbitrary process and different workers have used different methods. Since the nature of the initial classification determines the required number of reallocations, the convergence can be

very slow if the initial classification is nowehere near the optimum point. Some of the widely used methods to achieve an initial partition can be listed as follows:

(1) The first k individuals (Mcqueen, 1967).

(2) Systematic sampling of k individuals from the population.

(3) Subjectively choosing of any k number of seeding points.

(4) Random sampling from the data set (McRae, 1971).

(5) The most mutually distant k points (Thorndyke, 1953).

(6) The population is partitioned subjectively and then centroids of the groups are used as seeding points (Forgy, 1965).

All these methods may not produce a classification anywhere near the optimum point. The random sampling method may be theoretically more attractive but for small samples its validity is questionable. On the other hand the order of the individuals can affect the sampling process.

An initial partition of a given population can be obtained by a suitable agglomerative strategy. Such a partition will be much closer to an optimum partition if the similarity measure chosen measures the relative similarity between individuals accurately. When the population is highly fragmented, there is a high degree of probability of discovering groups and therefore the

number of groups in the population can easily be determined, But certain populations are not fragmented and therefore the number of groups should be determined either by the user's knowledge of the population or using a suitable criterion. The use of probabilistic measures for this purpose has been criticised (Lance and Williams, 1975) but a non-probabilistic measure has been used by Webster (1977,p.212, 1979) to determine the number of groups in soil populations. It has been demonstrated that when the number of groups (G) is plotted against ΛG^2 (section 2.7) a rapid drop of the line of the graph followed by a rise indicates the optimum number of groups. Although this method needs to be tested using known populations the past experience with soil data seems to be encouraging. The hierarchical tree can be used to determine the group membership in a given number of groups. This procedure may produce a better partition as it involves optimizing a function unlike random or arbitrary partitions.

The initial partition is improved by reallocating individuals until a suitable function reaches its optimum value. The optimization functions generally used are based on the following matrix equation

$$T = W + B \quad \ldots\ldots\ldots \quad 2.3.1$$

T = Total Sum-of-Squares and products (SSP) matrix.

W = Pooled within groups Sum-of-Squares and products (SSP) matrix

B = Between groups Sum-of-Squares and products (SSP) matrix.

Demirmen (1969) has suggested several optimizing criteria derived from the above relationship.

(1) Tr (W)

(2) Tr ($W^{-1}B$)

(3) $|W|$

(4) Wilk's Criterion Λ (section 2.7)

In addition to these functions, Rubin (1967) used the generalized distance of Rao (1948).

Although these methods can be used to obtain a better classification, there is no guarantee that all these functions will be optimized simultaneously. Demirmen (1969) claims that Tr ($W^{-1}B$) is not a reliable measure. Both the Wilk's Criterion and the generalized distance are statistical tests which require standard assumptions to be met. The most important consideration is the fact that these tests cannot be applied to classified data in a probabilistic context and therefore they can only be used to compare different partitions of a population. Webster (1977, pp.187-218) has used the Wilk's Criterion to compare classifications of the same soil population by various methods. Rubin (1967) has pointed out that the optimization by reallocation can produce suboptimal classifications by terminating the process when a local optimum is encountered. He has considered several procedures to overcome the problem of local optima but they can be extremely time consuming when the population under consideration is very large. It may be worth noting the fact that there is no way of

knowing whether the classification is suboptimal or if in fact it has reached the global optimality. Everitt (1974, p.24-30) notes that clustering techniques which seeks to optimize some criterion function usually find suboptimal solutions. The obvious way to achieve the global optimum classifications is to consider all possible partitions. This will involve the number of partitions given by the recurrence formula proposed by Fortier and Solomon (1966).

$$P(N,G) = (g^n - \sum_{|i|}^{g-1} g_{(g-i)} P(N,1))/g! \qquad 2.3.2$$

g = number of groups ≥ 2

N = number of individuals ≥ 2

P(N,G) is the number of partitions of N individuals into g groups.

$g_{(g-i)} = g(g-1)(g-2) \ldots (g-i+1)$

This is an impossible task even when the number of individuals is small. For example Gower (1967) has estimated that all possible partitions of 41 individuals into two groups would require $2^{40} - 1$ number of partitions which would require 540 years of computer time using even the fastest computer available.

If the initial partition is somewhere near the optimal partition, the number of reallocation steps involved will be very much reduced. Therefore an agglomerative clustering strategy with a suitable similarity measure can be used to produce an initial partition. Since the

properties of the similarity measures are of fundamental importance to agglomerative cluster analysis it is necessary to compare such measures. The optimality of the classification obtained by various methods can be examined by performing canonical analysis on the classification and plotting the individuals in a two dimensional space. Webster (1977, pp.187-218) has demonstrated that when the groups are projected onto the two dimensional canonical vector space, the relative position of the groups depends on the optimality of the classification. He found that unsatisfactory soil classifications failed to produce well separated groups in the canonical space.

The density seeking methods assume that there should be dense parts of the metric space which are separated by parts of low density areas. The most prominant of density seeking methods is the mode analysis of Wishart (1969b).Like other density seeking methods mode analysis is based on the single linkage method. The density search methods are usually based on a similarity matrix and therefore the properties of the similarity measure used initially is of great importance. Some soil populations are known to be evenly distributed (Webster, 1979) in the multidimensional space, and therefore the efficiency of these methods in finding soil groups may be restricted.

Some of the density seeking methods involve distributional assumptions, which are very often not

met and the effects of the violation of such assumptions have not been examined. They also suffer from the difficulty of identifying the global maxima like other classificatory methods.

2.4 Ordination

In the previous section, various taxonomic strategies to identify groups in a given population were discussed. However, the general nature of the population under consideration can be examined using another group of methods known as ordination which helps reduce the dimensionality of multivariate populations, while preserving the maximum amount of information about the population. A given multivariate population can be projected onto the best fit two dimensional sub-space, or a sub-space of a reduced dimensionality can be obtained by eliminating redundant attribute vectors for further statistical analysis. Since soil attributes are mutually correlated certain attributes can be eliminated from the analysis by means of principal component analysis (PCA). Sarkar et al (1966) have shown that only a few attributes are required to represent the information of the original population. Hole and Hironaka (1960) first used ordination to classify soils from Kansas. The two dimensional projections of multivariate populations are used to identify clusters which are dense areas separated by relatively low density areas. But such clustering according to Webster (1979), is the exception rather than

the rule. In ordination finer divisions of the population may not be evident as noted by Webb et al (1967). Rohlf (1968) also found that the distance between close neighbours was not well represented by PCA ordination.

There are four major ordination techniques, which have been used to study multivariate populations.

(1) Principal component analysis (PCA)

(2) Principal coordinate analysis (PCO)

(3) Multidimensional scaling

(4) Factor analysis

The first two methods have been used in soil classification but the use of other techniques is limited. Multidimensional scaling is for qualitative data but it can be applied to quantitative data after transforming to qualitative form. Factor analysis involves the rotation of vectors which can give rise to interpretational problems.

2.4.1 Principal component analysis (PCA)

This is a technique used to reduce the dimensionality of the multivariate populations. When the attributes are correlated the dispersion of the population takes an elipsoidal shape in more than two dimensions. The geometrical properties of this technique can be demonstrated when there are only two attributes and the laws of two dimensional geometry can be generalized to higher dimensions. The PCA transformation of a multivariate population involves obtaining the best

fit sub-space with mutually orthogonal new vectors which are linear combiantions of the old. The new vector Y is given by the following relation.

$$Y = X\,C^T \qquad \ldots\ldots\ldots \qquad 2.4.1$$

where X is the standardized data matrix.

C is the matrix of angles between the new coordinate axes and the old.

The eigen values of the matrix A (var-covar matrix) can be derived from the equation.

$$|A - \lambda I| = 0 \qquad \ldots\ldots\ldots \qquad 2.4.2$$

The number of eigen roots of the matrix is equal to the number of attributes but all the roots smaller than 1.00 may be ignored (Harman, 1976, p.185).

The new coordinates (PCA scores) are mutually orthogonal and therefore the var-covar matrix of Y is a diagonal matrix consisting of the eigen values λ_i

$$1/(n-1)Y^T Y = \Lambda \qquad \ldots\ldots\ldots \qquad 2.4.3$$

Since $Y = XC^T$

$$1/(n-1)\;\; Y^T Y = 1/(n-1)CX^T XC^T \qquad \ldots\ldots \qquad 2.4.4$$

and since $1/(n-1)X^T X = A$

$$\Lambda = CAC^T \qquad \ldots\ldots\ldots \qquad 2.4.5$$

On post multiplying by C the equation 2.4.5. becomes

$$\Lambda C = CA \qquad \ldots\ldots\ldots \qquad 2.4.6$$

By solving this equation the values of the matrix C can be obtained. The contribution of each eigen value to the total variance is obtained by dividing the eigen

values by the tr (Λ) since

$$\sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} S_i^2 \quad \ldots\ldots\ldots \quad 2.4.7$$

PCA has been used in soil classification by Hole and Bidwell and Hole (1964b), Hironaka (1960), Russell and Moore (1967), Cuanalo and Webster (1970) but very often it is difficult to identify clear clusters from the two dimensional plots as soil individuals are evenly distributed over the space.

The other use of PCA is to transform the original space to a space with mutually orthogonal vectors and also to eliminate redundant vectors. The new coordinate vectors can be used in the computation of Euclidean distance between individuals. Kawaguchi and Kyuma (1977) used this technique in the classification of paddy soils sampled from a wide geographical area. They computed the Euclidean distance matrix after transforming the original attributes to PCA scores.

Divisive strategies, POLYDIV for example, have used PCA to reduce the dimensionality of multivariate populations in order to maximize between group Sum-of-Squares more easily. As the first PCA vector or the first few vectors account for the greatest percentage of the variance of the population they can be used to identify the discontinuities of the population.

The interpretability of PCA vectors can be increased by computing the correlation between the original attribute

vectors and the new vectors as done by Kawaguchi and Kyuma (1977).

Webster (1979) argues that PCA plots are not suitable for identifying soil groups because of the fact that attempts in the past have shown the lack of clustering. This may have resulted either from lack of abrupt discontinuities or the fact the model is not capable of identifying such discontinuities.

2.4.2 Principal coordinate analysis (PCO)

This method was proposed by Gower (1966) based on the Q matrix (inter-individual distance matrix) to overcome certain problems encountered by PCA. Unlike PCA, PCO can be applied to mixed data which is very useful to soil taxonomists. The Euclidean distance in the original space is not a true measure of the relative similarity when the attributes used are not mutually orthogonal. This problem can be solved by transforming this matrix by means of PCO which also reduces the dimensionality of the population. PCO has been applied to soil classification by Rayner (1966, 1969), Campbell et al. (1970) and Webster and Butler (1976)

Gower (1966) has demonstrated that the distance a_{ij} of the original space is related to the distance in the new space d^2_{ij} in the following way.

$$d^2_{ij} = a_{ii} + a_{jj} - 2a_{ij} \quad \ldots\ldots\ldots \quad 2.4.8$$

The elements of the diagonal of the matrix A (inter-individual distance matrix in the original space) can be replaced by unities.

$$d_{ij}^2 = 2(1 - a_{ij}) \quad \ldots\ldots\ldots \quad 2.4.9$$

The most widely used transformation for the side entries is by replacing them with $(1 - a_{ij})$.

Then,

$$d_{ij}^2 = 2a_{ij} \quad \ldots\ldots\ldots \quad 2.4.10$$

In this way the square roots of the original values can be preserved. If the a_{ij}'s are replaced by $(1 - a_{ij}^2/2)$ the original values themselves can be preserved.

The matrix A should fulfil two conditions to be used in principal coordinate analysis.

(1) The matrix A must be symmetrical, $a_{ij} = a_{ji}$

(2) The matrix A must be positive semidefinite.

Thus the new coordinates define an Euclidean space provided that no significant negative roots are found for the matrix A. The new space is also called Gower space (Williams, 1976c).

The matrix A was reduced by principal component analysis originally, but Gower (1966) later used a simpler method. According to the new method all a_{ij}'s are replaced by $(a_{ij} - r_i - c_j + g)$ (where r_i is the row mean of the ith row, c_j is the mean of the jth column and g is the grand mean). The roots and vectors of the new matrix are computed. The new coordinate vectors can be used to project the population onto two dimensions. The first few vectors usually represent the greatest proportion of the variance

of the population. The main problem of this method is that in certain situations roots of the matrix A can take negative values and the space defined by such vectors is imaginary.

Both factor analysis and multidimensional scaling have not been widely used in soil classification. Although factor analysis can be applied to soil data, the interpretation of results is much more difficult than PCA (Webster, 1977, pp.153-156). Multidimensional scaling is for qualitative data and probably much more appropriate for ecological data.

2.5 The choice of cluster analysis strategies

It is necessary to assess the relative merits of the clustering strategies so far discussed and their suitability for soil classification. Even in soil taxonomy, hierarchical methods have been widely used because of the fact that they could be viewed as useful ways of summarizing soil data. But these methods have inherent deficiencies associated with them. Several of such problems of hierarchical strategies can be identified.

(1) The suitability of hierarchical strategies to classify soils is questionable (section 1.5). The hierarchical assumptions, however, can be ignored in the interpretation of results by emphasising the final group constellation and using it as an intermediate step of the classificatory strategy. In this study the hierarchical numerical strategies were used only to obtain a partition of the multivariate population with the intention of

subsequently improving it by optimization - reallocation methods where necessary.

(2) The hierarchical strategies do not allow reallocation at each fusion or division step and therefore the course of the strategy could be determined by the early steps, which can lead to unsuitable classifications. This cannot be avoided if the hierarchical nature of group structure is to be preserved.

(3) Although there are several agglomerative strategies, the resultant classifications, to a great extent, depend on the nature of the similarity measure used. For example, angular measures tend to produce groups with a high degree of heterogeneity, whereas distance measures are affected by inter-attribute correlation. The way in which similarity is defined is of fundamental importance since the soil taxonomist's objective is to produce homogeneous soil groups. The effect of the inter-attribute correlation can be eliminated by transforming the original space to an Euclidean space with mutually orthogonal vectors before calculating the distance matrix. It is also possible to use a distance measure which is insensitive to inter-attribute correlations (i.e. Mahalanobis D^2).

(4) Certain of the agglomerative strategies are subject to a phenomenon known as 'chaining' and also to 'reversing' due to the fall of the distance after a fusion step. On the other hand, the methods which produce clear clusters are associated with varying degrees of distortion which can lead to artificial partition of a population.

The problems associated with agglomerative strategies can be eliminated by using a non-hierarchical strategy. As has been pointed out earlier, these strategies depend on the nature of the initial partition of the population for their efficiency. Although there are various methods for obtaining an initial partition, most of these methods are arbitrary. Among the methods which have been used in the past, the use of a suitable agglomerative strategy to obtain an initial partition seems to be most attractive as they are generally polythetic and the chance of making errors is very much less than with monothetic methods or arbitrary methods.

In using an agglomerative strategy to obtain an initial partition two important decisions have to be made.

(1) A suitable similarity measure must be chosen. The properties of the widely available similarity measures were examined in theoretical terms of the previous section but empirical studies on different similarity measures are required.

(2) The choice of an agglomerative strategy. The properties of certain agglomerative strategies were discussed earlier and again their ability to produce homogeneous clusters can be examined using real data. The relative homogeneity of clusters can be examined by a suitable test criterion; in this study Wilk's Criterion was used.

2.6 Methods of reallocation

It was shown earlier that a population of n individuals with p attributes could be partitioned in a large number of ways in order to optimize a given function. This is however, not achievable because of the computational load involved. One, therefore, has to be content with a less than optimum partition and ways to improve that partition should be found.

Reallocation of individuals can be done in several ways. Friedman and Rubin (1967) did this by moving a single individual to every group other than the one it was found in and computing the criterion function to be optimized. If a given move improves the classification that move is made permanent. This process is repeated until such moves do not lead to the improvement of the classification. They state that after several moves a local optimum can be found. This process is able to find the local optimum after half a dozen moves and to move above the local optimum Friedman and Rubin (1967) proposed two methods.

(1) 'Forcing passes'. The process begins with the currently best known partition. Taking a single group at a time, all members are transferred in sequence to the nearest group and the criterion function is calculated. After treating all members of that group the best partition so far found is retained. The process is repeated for all groups, and forcing passes are repeated until no improvement is achieved. This process is

fast since every possible move of an individual is not considered.

(2) 'Reassignment pass'. This method involves reassigning each individual to the nearest group centre of gravity measured in the following way. The distance between an individual i and a group centroid C_k is given by the function,

$$d(i,C_k) = (i - C_k)\ V^{-1}(i - C_k)^T \qquad 2.6.1$$

where V is the pooled within groups var-covar matrix.

This measure is the Mahalanobis D^2 described earlier (section 2.1.1). The scatter matrix V and the group centroid C_k are retained until all n individuals are reassigned. This process is repeated until the reassignment does not improve the classification.

Another method of reallocation is that of Lance and Williams (1975) which is based on the distance between individuals and group centroids determined by the Canberra metric. The Canberra (TAXON) divisive program REMUL employs this method to improve the classification obtained by a monothetic split. The Canberra metric has the advantage that it can be applied to mixed data sets and is computationally simple.

Many other reallocation procedures are achieved by moving individuals and recalculating the optimization function. A method very similar to the Friedman and Rubin (1967) method based on the Mahalanobis distance has been proposed by Webster (1977, p.187-200). The distance

measured by the Mahalanobis D^2 between the individuals and the group centroids is claculated and those individuals which are closer to any other group than their parent group are transferred.

Rao (1948) has pointed out that the discriminant space is divided into regions, each of which is associated with one of the groups. The individuals closer to that part of the discriminant space belong to the associated group. Thus the Mahalanobis distance can be employed to reallocate individuals to improve a given classification as used by Burrough and Webster (1976). In this study the number of groups in the population was determined by the relationship between G and ΛG^2 discussed by Webster (1977, p.212).

2.7 Test criteria

In the univariate case the tests of significance are generally used to test the significance of the difference between population means. These tests have been derived from the normal distributional curves and associated with several assumptions.

(1) The normal distribution of the observations.

(2) The samples are independently and randomly selected.

(3) In the case of analysis of variance, it is assumed that the variances are equal.

These theoretical assumptions are not strictly true for real data but the tests available are known to

perform well under the violation of the standard assumptions. Departures from the normality do not affect the results if the number of observations (n) is large enough as the Central Limit Theorem indicates that most natural populations approach normal distribution as n increases. The univariate statistical tests have proven that even small samples are unaffected by the departures from the normality as the tests are robust and insensitive to such departures.

The standard assumptions and their implications in multivariate tests are more complex and it has been more difficult to test the departures from the assumptions. Most univariate tests of significance have been generalized in multivariate data. Departure from normality has little effect on large samples as univariate Central Limit Theorem can be generalized to multivariate data (Ito, 1969). The most important of all assumptions for multivariate tests is the assumption of homoscedasticity (homogeneity) of within group dispersion matrices. The multivariate significance tests are more sensitive to heteroscedasticity of dispersion matrices than to departures from the multinormality. But provided that the differences between dispersion matrices are not large the tests can be applied (Webster, 1977). The homoscedasticity of dispersion matrices V_i can be tested as described earlier but the test is more sensitive to departures from the normality than the tests based on dispersion matrices. If the sample is large, it is possible

to proceed with the significance tests even when homoscedasticity of V_i is not true.

The significance test to be used depends on the number of groups in the population. When there are two groups the multivariate generalization of the Student's t-test, Hotelling's T^2 can be used. This test is related to the Mahalanobis D^2 for two groups in the following way.

$$T^2 = (n_1 n_2/(n_1 + n_2)\ D^2 \qquad \ldots\ldots \qquad 2.7.1$$

where n_1 is the number if individuals in the group 1

n_2 is the number of individuals in the group 2

The significance of the difference between two mean vectors can be tested. $T^2(n_1 + n_2 - p - 1)/p(n_1 + n_2 - 2)$ is distributed as the F ratio with p and $(n_1 + n_2 - p - 1)$ degrees of freedom.

The null hypothesis testing here is,

$H_0 : \mu_1 = \mu_2\ V_1 = V_2$

Although there is a possibility that the rejection of H_0 could be due to $V_1 \neq V_2$ rather than $\mu_1 \neq \mu_2$, it has been demonstrated that this test like its univariate counterpart is more sensitive to difference between the means than to the differences in V_i (Harris, 1975). Ito and Schull (quoted by Harris, 1975) have shown that the true significance level of T^2 is not affected by discrepancies between V_1 and V_2 provided $N_1 = N_2$ is large. But when $N_1 \neq N_2$ is small it is worth testing the assumption $V_1 = V_2$.

When there are more than two groups this test cannot be applied to test the significance simultaneously but a pairwise testing is possible. This can lead to commiting the type 1 error and therefore the power of the test can be reduced. Tests for more than two groups are available. Rao (1948) generalized the Mahalanobis D^2 to more than two groups in the following way.

$$D^2 = \sum_i \sum_j V^{-1} N_k (\bar{X}_{ik} - \mu_i)(\bar{X}_{jk} - \mu_j) \qquad 2.7.2$$

N_k is number of individuals in group k.

This measure has been used by Rubin (1967) to find the best partition of a population. When it was applied to a population with known V and G (number of groups) the best partition was the one which recovered all the members of sub-population.

Another test can be derived from the matrix equation

$$T = W + B \qquad \ldots\ldots\ldots \qquad 2.7.3.$$

and is known as Wilk's Criterion Λ.

$$\Lambda = \frac{|W|}{|T|} \qquad \ldots\ldots\ldots \qquad 2.7.4$$

W pooled within group SSP matrix

T total SSP matrix

$n \log_e \Lambda \sim \chi^2$ with $(n - 1 - (p + g)/2)$ degrees of freedom.

Wilk's criterion Λ was used in soil classification by Webster (1971; 1977, pp.187-216), and he concluded that it was a useful measure to choose the best classification among many.

Although the tests described above are powerful and robust under the violation of standard assumptions, they cannot be applied to classified data in a probabilistic context. The groups obtained by numerical classificatory methods are not samples drawn from wider independent populations and therefore the tests can only be used as criteria to compare classifications obtained by different methods. Both the generalized D^2 (Rao, 1948) and the Wilk's Λ are invariant under linear transformation and consequently independent of the scale of measurement. In this study the Wilk's Λ was used as a criterion to choose the best classification with respect to the minimum within group variance.

It may be possible to use two different sets of attributes for classification and testing to preserve the independence of groups. But it is not possible to judge the validity of a given classification purely in terms of a statistical criterion. It may be necessary to evaluate the classification in respect of the pedological meaning of the groups.

2.8 Computer programs used

This study was based on two classificatory computer packages: (a) CLUSTAN 1C 2nd Release, and (b) TAXON (CSIRO, Canberra). The main advantage of the

TAXON package over CLUSTAN is that it accepts mixed mode data for most classificatory strategies and also missing attribute values are allowed. CLUSTAN programs do not allow missing attribute values at the moment and also only one data mode is accepted. However it has more agglomerative clustering options, including all those available in TAXON.

The computer program MULVAN (Webster, 1971) was used to compute D^2, Hotelling's T^2, Canonical vectors and Wilk's Criterion Λ. It was also used to reallocate individuals of the classifications obtained by other methods. This program accepts only quantitative attributes and missing values are replaced by attribute means. Therefore, it is necessary to choose data in such a way that the number of missing attribute values is minimal.

CHAPTER 3

NATURE AND SOURCES OF DATA

3.0 Introduction

Computer based soil information systems are becoming a common feature in the modern soil surveys, especially in the developed countries. This is a main consequence of the development of high speed computers and the substantial increase in the amount and the variety of soil information. Automation of soil information storage, retrieval and processing has made it possible to apply mathematical methods to soil classification in the place of traditional subjective procedures of soil taxonomy. At the time when the knowledge of soil was limited, it was considerably easier to group soils with respect to a single property or a few properties. But the expansion of the knowledge on the nature of soils has produced data on a large number of properties which are used in soil classification. Thus the concept of similarity between soils can be defined taking as many characteristics as possible and new methods are required to determine the similarity, as the traditional method of defining similarity in terms of diagnostic features becomes inadequate. The traditional soil taxonomists use mainly morphological features to identify similar soils by the personal judgement of the soil surveyor. This strategy has led to inconsistencies of soil classifications. The 7th Approximation soil classification system of the Soil Survey Staff (1960) and its subsequent modifications used substantially increased amounts of soil data to define

various categories of the system. But the higher categories (Orders, Suborders, etc.,) have been defined using a few characteristics. The use of the readily observable features has been emphasized by most soil surveys in order to simplify classification. However, the recent development of methods of data processing has made it possible to use all types of soil data (section 3.3) in classifying soils into homogeneous groups.

The use of a large number of properties simultaneously to compare soils sampled from different areas is beyond the ability of the human mind. The methods of numerical taxonomy are of great use in this respect. The success of new methods depends not only on the suitability of the methods themselves, but also the quality of soil data, to which numerical methods are to be applied. Therefore, the nature of soil data in general, and the data used in this study in particular, need to be examined.

Soils are heterogeneous natural bodies, the properties of which vary in both lateral and vertical directions. The lateral variation can be represented by spatial sampling of soil profiles, whereas the vertical variation is represented by multiple measurements of soil properties. Observation of characteristics of the soil horizons is the standard procedure of the traditional soil survey. This involves making laboratory determinations of chemical, physical and mineralogical properties and recoding visible features of soil horizons.

3.1 Sampling Methods

The basic unit of soil sampling is the soil profile in both US Soil Survey and the Soil Survey of England and Wales. The USDA (1975, pp.2-5) considers that the three dimensional soil body, 'pedon', as the basic unit of soil sampling but when the pedon is larger than $1m^2$ it is not practical to use as a sampling unit. Soil sampling is a two stage procedure.

(1) site sampling

(2) sampling from the soil profile face for laboratory determinations.

The choice of the site is important in both the study of spatial variability of soil and for other pedological studies. For the purpose of classification, the soils sampled should represent the soils in a given area. There are two methods of site sampling.

(1) grid survey

(2) free survey

The former is important for statistical studies and it is a more objective method than the free survey method. The grid sampling, which may be systematic or random as desired, involves taking large numbers of samples per unit area. However, soil surveyors prefer the second method on the ground that it involves less sampling to obtain coverage and also to demarcate soil boundaries and as a result in less costly. On the other hand the first method does not require a great deal of expertise on the part of the soil surveyor,

whereas the second method is very much dependent on the soil surveyor's knowledge of the area and his field experience. The free survey method gives greater freedom to the field surveyor to choose representative profiles. Although the main emphasis is on the representative soil profiles, some intermediate soils are also sampled. Soil profiles which are sampled should characterize the soil mapping units of the area. The basic unit of soil mapping in detailed soil surveys of the Soil Survey of England and Wales is the soil series. In the USA soil type used to be the mapping unit used in detailed soil surveys,but now the soil series is used as the mapping unit (USDA, 1975,p.80-81). It is unavoidable that the free survey involves a considerable degree of subjective decision making, although it is planned using good field sheets (aerial photographs or detailed topographic maps) and geological information if available.

Once the site is selected a pit is dug to expose the vertical face of the soil. Samples from horizons are taken as described by Hodgson (1978,p.119) "to confirm, quantify or supplement information recoded in the field and to help the identification"of the body of soil according to a reference classification system.

The soil profile as a whole should be representative of the mapping unit identified and the samples from the horizons should be representative of individual horizons.For a small sample of soil to be representative, the soil body that the sample comes from, should be homogeneous.
If samples are collected ignoring horizons, at specified

depth intervals, it may be possible to fit mathematical models to describe the vertical variation of soil properties more easily, but it might not reveal the genesis of soil horizons. Therefore, soil surveyors place great importance on the sampling from genetic horizons. Three kinds of samples are taken from soil horizons:

(1) Disturbed bag samples for particle size, chemical and physical analysis.

(2) Undisturbed core samples for physical measurements.

(3) Box samples for micromorphological studies.

Most laboratory determinations are made on the first category. The Soil Survey of England and Wales takes samples from all horizons in a column one above the other. The column is about 20-50cm wide and extends deep enough to get 1-2kg of sample from the thinnest horizon. Before the samples are taken, the profile face is carefully marked to separate horizons. Samples are normally taken from the entire thickness of horizons, except where horizon boundaries are gradual or diffuse. Well distinguished bodies within heterogeneous horizons are separately sampled.

3.2 Types of soil information

The information about soils available from the soil survey is of two major kinds.

(1) field descriptions

(2) laboratory data

The field descriptions are made by the soil surveyor to characterize a given soil to be compared with the soils of other areas in the survey region or elsewhere. Since identification and grouping of soils are based primarily on the visible features of soils, the field description is an important task of the soil surveyor. As has been described by Webster (1976), three basic kinds of information are available from a field description sheet.

(1) General reference information that can be used to index both the site and the soil profile.

(2) Description of the site and its environment. i.e. topography,climate etc.

(3) Descriptions of the soil profile, horizon by horizon.

The third category of field description is directly relevant to the purpose of soil classification but other two types are important for the interpretation of results and geographical studies of soil. The field description of soils is done by the soil surveyor using his field experience and professional judgement and therefore, the resulting personal bias is unavoidable. The identification of horizons and their description can vary from one surveyor to another and the chances of making errors is very high. The soil profile is divided into three major horizons, designated as A, B and C by the Soil Survey Staff (1951, p .173-188) and each major horizon is subdivided as may be necessary. The Soil Survey of England and Wales (1967, p.39)

recognizes an eluvial horizon E as a separate major horizon. Both soil surveys attach a considerable importance to the identification of genetic horizons correctly as they have been used to group soil profiles. But for the numerical taxonomist, the horizon designation has no great importance since observed properties are used to group soil profiles ignoring the presence of genetic horizons, because of the subjective nature of the horizon nomenclature and identification on the basis of visible features in the field. It may be easy to divide a given soil profile into homogeneous layers, but the diagnosis of genetic horizons is not that easy. Therefore, grouping based on such information may lead to erroneous results. In this study the multiple observations of a given property are treated as representing the vertical variation of that property.

The field description of soils is greatly influenced by the requirement of the reference classification system in use and it may lead to giving preference to certain properties. This may be suitable for the traditional taxonomist but when numerical strategies are used it is necessary to have an unbiased set of soil properties to characterize the soil individuals.

The terminology of field description has been standardized to achieve a greater uniformity of soil description to be able to correlate soils found in different areas. The Soil Survey Handbook (Soil Survey of England and Wales, 1976) sets out numerical codes to be assigned to various soil character states (for qualitative information).

Nevertheless still a certain amount of subjectivity is involved in identifying the character states. It is very important to collect field data referring to carefully prepared tables such as Munsell colour charts to describe the soil colour. One may argue that the application of rational mathematical models to subjective data would result in false classifications, and therefore, the elimination or at least minimization of subjectivity involved with field data is a fundamental requirement. The US Soil Survey Manual (Soil Survey Staff, 1951) has outlined a comprehensive system of soil field description, which has been adopted elsewhere in the world. The system used by the Soil Survey of England and Wales (1976) is based on the US system but field sheets have been designed to record information in the field in a computer compatible format, which not only improves the quality of information but also makes it easier to input to computers (Webster, 1976; McDonald, 1981). The bulk of field descriptions is qualitative or semi-quantitative.

Laboratory determinations involve a wide range of chemical, physical and mineralogical properties of soils. As there are a large number of known soil properties, it is the convention of soil surveys that the most relevant properties to soil classification are determined. For example, when the soil is known to be noncalcareous, $CaCO_3$ content may not be determined. This is very much so in the British Soil Survey data, and it could be a constraint in the use of such data in numerical analysis.

As has been noted by Hesse(1971, p.4) there are three possible ways of introducing errors to laboratory determinations.

(1) personal errors

(2) sampling errors

(3) errors of method

The personal errors are due to personal characteristics which can influence results in a standard manner. They may not be eliminated entirely by doing duplicate determinations involving the same person.

Sampling errors are the most common due to the heterogeneity of soil bodies (Hesse, 1971, p.4). These errors can occur both in the field and in the laboratory, but the latter can be eliminated by using a standard procedure. Hesse(1971) reckons that although there are ways to reduce field errors, interpretation of results must always be done keeping in mind the probabilities of field errors.

The errors of method are difficult to detect. Most soil survey laboratory procedures have been standardized to preserve uniformity of data, but more than one method may be applied to different soils to determine the same property. It is well known that different techniques produce different results (Soil Survey of England and Wales, 1974), but it cannot be avoided because certain soils may need special methods. The data from USDA (1975) have been obtained using different analytical methods.

3.3 Attribute types

The term attribute is used here to denote any property state used to characterize soils. A property determined at two depth levels can be regarded as two attributes. The statistical terms variable or variate can be used only for continuous measurements, and soil information generally includes qualitative data as well as quantitative data.

Those attributes, which are used in numerical analysis can be divided into four types as suggested by Webster (1977, p.220-221).

(1) Binary or two state attributes

(2) Unordered or disordered multistates (nominal)

(3) Ranked or ordered multistates (ordinal data)

(4) Quantitative or numerical attributes (interval data).

This classification of soil attributes is compatible with the computer strategies used in numerical taxonomy. The first three types are also known as qualitative attributes.

The binary attributes are those which have only two possible states, i.e. presence or absence of a particular feature such as stones, gleying, etc. Each attribute state is given a numerical code, the usual practice is to denote presence by one and absence by zero. The magnitude of numerical codes has no meaning and therefore, one could use any two codes as may be desired.

The second and third types have more than two states, which may be as many as desired. Each state is given a specific numerical code. The magnitude of numerical codes given to unordered multistate attributes have no meaning although it is meaningful for ordered multistate attributes. Each code given to an ordered multistate attribute is a rank and therefore, can be treated as a quantitative attribute in a limited sense for the purpose of numerical classification. The ordered multistates can be influenced by observer errors more than the other three types.

Quantitative attributes are the most important type of soil information, not only because of rationality of such data but also a wide variety of statistical and mathematical models can be applied to them. But in the case of soils this type of information is the most difficult to come by because of the considerable costs involved. Therefore, laboratory determinations are usually done to supplement or confirm decisions made by field surveyors.

3.4 Sources of data

This study is based on data from three sources:

(1) Data published by the USDA (1975, p.485-743) for 32 soil profiles

(2) Data from De Alwis (1971) for 9 soil profiles

(3) Data from the Soil Survey of England and Wales for 65 soil profiles (West Sussex Coastal plain).

TABLE 3.2 Attributes used in the Classification of Soils of the West Sussex Coastal Plain

(a) Quantitative (numerical) attributes.

1. Percentage silt
2. Percentage clay
3. Percentage loss on ignition
4. pH (1:2.5 H_2O)
5. Percentage $CaCO_3$
6. Cation exchange capacity (C.E.C.)
7. Percentage base saturation
8. Exchangeable Ca (me/100g soil)
9. Exchangeable Mg "
10. Exchangeable K "
11. Exchangeable Na "
12. Percentage moisture content
13. Thickness (cm) of horizons
14. Colour value
15. Colour chroma

(b) Binary attributes

1.	Presence/absence of mottling	depth (horizon)	1
2.	"	"	2
3.	"	"	3
4.	"	"	4

(c) Multi-state attributes

i) Ordered multi-states

1.	Ped size in	depth (horizon)	1
2.	"		2
3.	"		3
4.	"		4

ii) Disordered multi-states

1.	Ped shape in	depth (horizon)	1
2.	"		2
3.	"		3
4.	"		4
5.	Colour (hue) in	depth (horizon)	1

6.	Colour (hue)	in depth (horizon)	2
7.	"		3
8.	"		4

The first two sources have used the field and laboratory procedures recommended by Soil Survey Staff (1972) and therefore a certain degree of uniformity could be expected. When a large number of people are involved in producing the soil information, it can be expected that a certain degree of variation of data can occur. At this stage, it is not possible to assess the accuracy of such data apart from observing that methods recommended by the Soil Survey Staff (1972) have been used in both field description and laboratory determinations. The data obtained from the third source is not comparable for two reasons.

(1) Laboratory determinations have not been done using the same methods as those of Soil Survey Staff (1972).

(2) Most soil profiles sampled by the Soil Survey of England and Wales do not have the same number of depth levels as those soil profiles chosen from the USDA (1975).

The data from the first two sources (Appendix 1) combined are for forty one soil profiles, thirty two of which are from the USDA (1975)and the rest from De Alwis (1971). The thirty two soil profiles belong to seven Orders, whereas the other nine soil profiles belong to red latosol of Oxisol Order of the USDA system. The USDA system has somewhat under represented tropical soils and therefore, the soil profiles sampled by De Alwis (1971) in Sri Lanka were included.

No attempt was made to represent the seven soil orders equally as the objective of the study was to evaluate

the merits of different numerical methods in producing homogeneous soil groups. Two main considerations were made in selecting the soil profiles from the published data of the USDA (1975). Firstly the availability of data for at least seven horizons (depth levels) as it was intended to fit mathematical curves to the soil properties chosen, and secondly the availability of data for most if not all properties. Therefore, the choice of soil profiles from the first two sources is to a certain degree subjective. The numerical strategies applied to the data make no assumptions on sampling and therefore, this has no effect on the results. The data published by the USDA (1975) is for representative soils in that the chosen soil profiles are considered by the surveyors as modal profiles. A modal profile is the best expression of a given soil and comparisons are made using such profiles sampled from different parts of a survey area or elsewhere.

Ten soil properties were chosen for this study to characterize the forty one soil profiles. All the properties are quantitative determinations.

TABLE 3.1 Soil Properties used to Characterize the Soil Profiles from USDA and De Alwis

1.	Percentage silt
2.	Percentage clay
3.	Percentage organic carbon
4.	Extractable iron as Fe (Pct)
5.	pH (1:1 soil/water)
6.	Exchangeable Ca me/100g soil
7.	Exchangeable Mg " "
8.	Exchangeable Na " "
9.	Exchangeable K " "
10.	Cation exchange capacity (C.E.C)

Any given property has not been determined by the same method for all soils but this fact was ignored in order to choose an adequate sample of soil profiles for this study. Although it is known that different methods may produce different results, it is difficult to predict their effect on the results obtained from numerical classificatory strategies. However, it is practically impossible to choose a sample of soil profiles whose properties have been determined by the same methods.

The majority of forty one soil profiles have at least seven horizons and missing data is not a considerable problem. When a particular soil profile does not have seven horizons it was considered that some horizons were missing.

The soil profile data obtained from the Soil Survey of England and Wales are from the West Sussex Coastal Plain survey area. A certain number of soil profiles were excluded from the analysis for the lack of data for the properties used to characterize soils and sixty five profiles were eventually selected. All four types of soil attributes were available but more emphasis was given to quantitative attributes. The complete data matrix is listed in the Appendix III. A part of the data set is listed in the soil survey memoir for the West Sussex Coastal Plain (Soil Survey of England and Wales, 1967, p.131-142), and the rest was obtained from the records at the Rothamsted Agricultural Experimental Station, Harpenden, England. The published data are for modal soil profiles which represent the soil series

of the area and the rest were originally treated as largely intergrade soils. This distinction, however, was not taken into consideration when the data were used for classification.

Although data are available on a large number of soil properties, a considerable number of them has to be excluded for lack of data for all chosen soil profiles. This was felt necessary for minimizing the number of missing cells in the data matrix.

For the purposes of numerical classification the quality of British data is poor compared to the data obtained from the USDA (1975). Most soil chemical properties have not been determined for all horizons, for example in many cases the organic carbon content has been determined for only the uppermost few horizons. This may not be a problem for the traditional taxonomist but the numerical classification is possible only if data on a given property has been determined for all horizons and all soil profiles. It is possible to treat them as missing data but the effect on the classification can be very great when there are a large number of such properties.

3.5 Missing data

As noted above missing data is a problem frequently encountered, when soil data collected by soil surveys is used in numerical classification. There are several reasons for this problem.

(1) Certain properties for certain soils have not been determined on the assumption that no appreciable amounts can be detected. For example, when the soil is regarded as noncalcareous $CaCO_3$ is not determined.

(2) Certain of soil properties are not determined for all horizons of a given soil profile.

(3) When there are soil profiles with an unequal number of horizons, those which have fewer than others have to be treated as having missing data for one or more lower horizons.

These problems are very common in the British data, consequently the data matrix has a large number of empty cells.

CHAPTER 4

THE SOIL PROFILE AS A BASIC UNIT OF CLASSIFICATION AND METHODS OF CHARACTERIZATION OF SOILS FOR NUMERICAL CLASSIFICATION

4.0 Introduction

The soil profile is the smallest unit, that can be used effectively in numerical classification of soils, since no smaller unit can be regarded as representing the total soil. Homogeneous entities can be described in terms of their properties conveniently, but the soil profile is an anisotropic entity (individual) which requires special treatment. As soils are three-dimensional bodies, any objective classification should take the vertical variation into account. A given soil profile property can, therefore, be considered as continuously varying characteristic which may not be represented by a single value.

The traditional soil classifications have treated the soil profile as a vertical cross-section of the soil consisting of horizons, usually recognized by the soil surveyor in the field. The concept that soils are layered natural bodies guided the sampling scheme and the soil description. Allocation of a given soil to a group was based on a few morphological characteristics. The American system of soil classification, over the years, improved the precision and objectivity of the definition of soil classes at lower categorical levels by using more and more

quantitative measurements of soil properties. The soil profile, however, remained as the unit of soil description and classification.

Characterization of the soil profile in terms of its properties is one of the basic problems the numerical taxonomist has to solve. Lance and Williams (1967a) suggested four soil profile models with varying degrees of generalization to meet the requirements of numerical taxonomic methods.

(1) to use a multi-level model of which all levels are treated as independent. Thus all observations for a given property are treated as independent attributes.

(2) to average over all depth levels (horizons) and use the average values of each property.

(3) to compute the similarity between corresponding depth levels and take the mean similarity.

(4) to use the parameters of depth dependent functions computed for all properties in the place of the original observations.

The existing soil survey data for profiles have usually been taken by the horizon, and as a result, most numerical methods cannot be applied to measure the similarity between soil profiles, having different sequences or different numbers of horizons. The only method, which is known to the author, takes horizons separately is Rayner's (1966) transition matrix approach. Depth levels should, therefore, be arranged in such a way that all the soil profiles to be compared have the same number of depth

levels and the horizon terminology is ignored. The second model above is the simplest of all but discards too much information according to Lance and Williams (1967a).The third method, known as the 'linked level system', was later abandoned by Williams and Lance (1967a) due to its similarity to the first method.

The last is the theoretically most attractive model but the amount of computation and the poor fit in some cases (Moore, Russell and Ward, 1972) may be disadvantageous compared to other simpler methods. Colwell (1969) used an orthogonal polynomial model of quintic form for the chemical characterization of three soil groups sampled in New South Wales, Australia, and concluded that a polynomial function of a sufficiently high degree can be used to represent soil properties for the purpose of numerical analysis. This approach was adopted by Moore, Russell and Ward (1972) in their classification of some Australian soils and they compared the results with two other soil profile models. They concluded that the classification from the orthogonal polynomial model was similar to that of the model in which depth levels were arrays of attributes (i.e. model 1).

4.1 Data and methods

The data used in this study were taken from USDA (1975). Thirty two soil profiles were chosen on the basis of the availability of an adequate number of observations per soil property per profile to allow fitting of a

polynomial function of the 5th degree and also the availability of data for most if not all properties considered. The properties used are listed in table 3. .

The soil profile was divided into three major horizons (section 3.3) in the convention of Soil Survey Staff (1951, p.173-188) to obtain a simpler model for the soil profile model 1 described earlier by averaging over all sub-horizons, and the resulting mean values for all properties chosen were used in the second analysis. Moore, Russell and Ward (1972) compared the polynomial model and the soil profile model 1 and concluded that both methods produced similar results when observations were weighted by an exponential function. This suggests that although the method of fitting mathematical curves to soil properties may be a theoretically sound idea, its contribution to soil classification is not very great. However, the use of depth levels as independent attributes may not necessarily mean all depth levels for which data are available should be used in characterizing soil individuals. The use of mean values of soil properties for three major horizons was intended as a simpler method and the results from both soil profile models were compared in order to determine the minimum number of depth levels required to represent the vertical variation of soil properties.

Orthogonal polynomial functions of the 5th degree were fitted for all soil properties. The general form of

the polynomial model is given by the equation 4.1.1.

$$Y_i = b_{0i} + b_{1i}X + b_{2i}X^2 + \ldots + b_{ki}X^k \quad \ldots \quad 4.1.1$$

where Y_i is the value for the property i of depth level X. This model is flexible enough to fit a wide range of trends if a sufficiently large value is chosen for k. For statistical analysis a more convenient form of this model can be obtained (Kendall and Stuart, 1961 and Colwell, 1969).

$$Y_{xi} = c_{0i}\,\xi_{0x} + c_{1i}\xi_{1x} + c_{2i}\xi_{2x} + \ldots + c_{ki}\xi_{kx} \quad 4.1.2$$

where ξ_{jx} (j=1,2,..k) is the value of the orthogonal polynomial ξ_j of degree j at depth x. The main advantage of this form of the polynomial model is that it is not necessary to compute all coefficients again whenever the power of the function is changed. Therefore, each term of the equation 4.1.1 can be calculated separately independent of the others. The original form of the function is for equally spaced x values (independent variables), and a modification of the computational procedure has been discussed by Robson (1959) and Mather (1976). The computer program used in this analysis is for unequally spaced independent variable (Mather, 1976, p.110-116). The coefficients of the orthogonal polynomial function c_k are used in numerical analysis.

The similarity between soil individuals was measured using two similarity measures which represent angular separation and distance in the Euclidean space.

(1) A similarity measure based on product-moment correlation

$$D_{ij} = (1 - r_{ij})/2$$

(2) Squared Euclidean distance

where D_{ij} is the distance between the ith and jth individuals and r_{ij} is the correlation between them.

All the data used were standardized to zero mean and unit variance. The average linkage method was used as a cluster seeking strategy and the structure of the population is illustrated by a dendrogram. The relationship between the two soil profile models was examined by means of correlation analysis between the similarity matrices and also the two classifications were compared. The number of coefficients was reduced to determine whether it is necessary to use all coefficients of a 5th degree polynomial function. Finally the nature of inter-attribute correlation was examined for the original data from USDA (1975) and also British data.

4.2 Results and discussion

Orthogonal polynomial coefficients were calculated for ten properties (table 3.1) of thirty two profiles. The goodness of fit of the model varied but only a few soil profiles showed poor fit. The curves fitted to most of the properties of the podzol profile (soil individual 14) showed a poor fit which has resulted from the presence of well developed horizons. Sharp variations of properties

over small distances make it difficult to fit smooth mathematical curves. The degree of horizonation of soils varies but podzols are noted for well developed horizons. Although it is possible to improve the fit for certain properties by increasing the power of the function 4.1.2, it may give an undue emphasis to random variations as suggested by Colwell (1969). On the other hand, addition of another attribute by increasing the power of the polynomial function could cause difficulties when the similarity matrix is calculated. Therefore, it was decided to proceed with the analysis assuming the effect of poor fit for some properties would not greatly affect the results of numerical classification. However, only in a small number of cases was poor fit found.

The three-horizon model was obtained taking the mean values of ten soil properties for three major horizons and whenever a major horizon was missing all the properties of the missing horizon were coded as missing.

A series of similarity matrices were generated using the similarity measures described in 4.1 on the basis of the attributes of the two soil profile models. A series of similarity matrices were also calculated after masking attributes of both models when the Euclidean distance was used as the similarity measure.

A series of classifications was obtained by the average linkage method and Ward's method and dendrograms were drawn for all classifications. The classifications obtained using the similarity measure (1) show a remarkable

similarity for both soil profile models (Fig. 4.1). The average linkage method does not show well defined clusters (Fig. 4.1 a and c) but the clusters produced by the Ward's ESS method can be clearly identified (Fig. 4.1. b and d). The relative position of individuals in both classifications is similar. This is because of the similarity between the two similarity matrices. The similarity measure (1) is based on product-moment correlation which does not take the additive and propertional differences between individuals into consideration. Although this measure may not be the best similarity measure to determine the affinity between soil individuals, it can be considered as a valid measure for comparing soil profile models. The similarity between the two soil profile models can be demonstrated by plotting the corresponding coefficients of the two matrices. A sample of thirty coefficients were randomly drawn from the two matrices and a scattergram was drawn (Fig. 4.1 e) and product moment correlation between the two was calculated. There is a strong correlation (r=0.9) between the matrices. It is clear from this analysis that little additional information can be obtained from the polynomial model despite its theoretical soundness. Since this comparison was made on the basis of product moment correlation, a second analysis was performed using the Euclidean distance as the similarity measure. The classifications obtained using this measure of similarity show a considerable difference between the two soil profile models (Fig. 4.2 a-d). Moore, Russell and Ward (1972)

Fig. 4.1a Dendrogram produced by average linkage method with $(1 - r_{ij})/2$ as the similarity measure (3-horizon model)

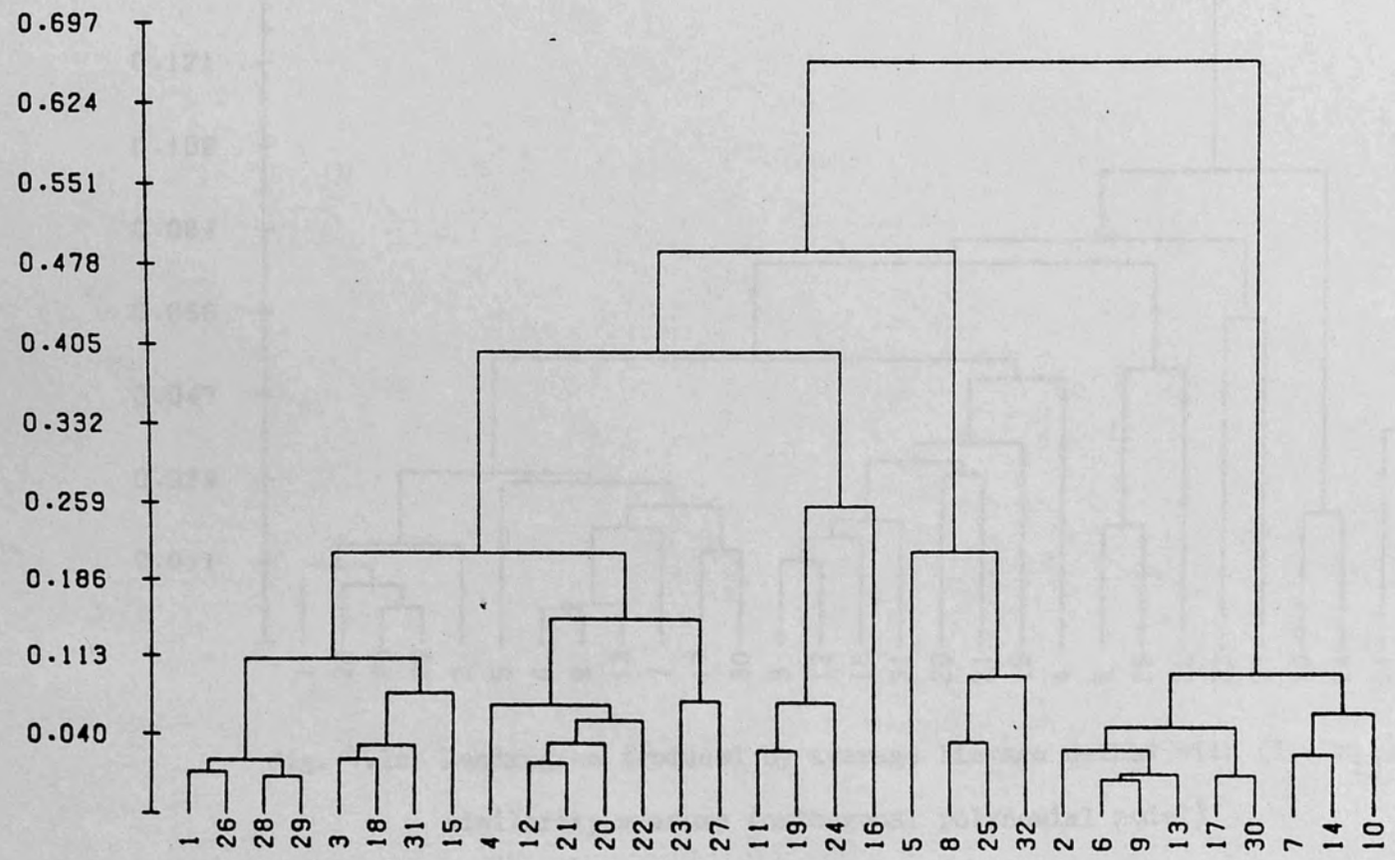


Fig. 4.1b Dendrogram produced by Ward's ESS method with $(1 - r_{ij})/2$ as the similarity measure (3-horizon model)

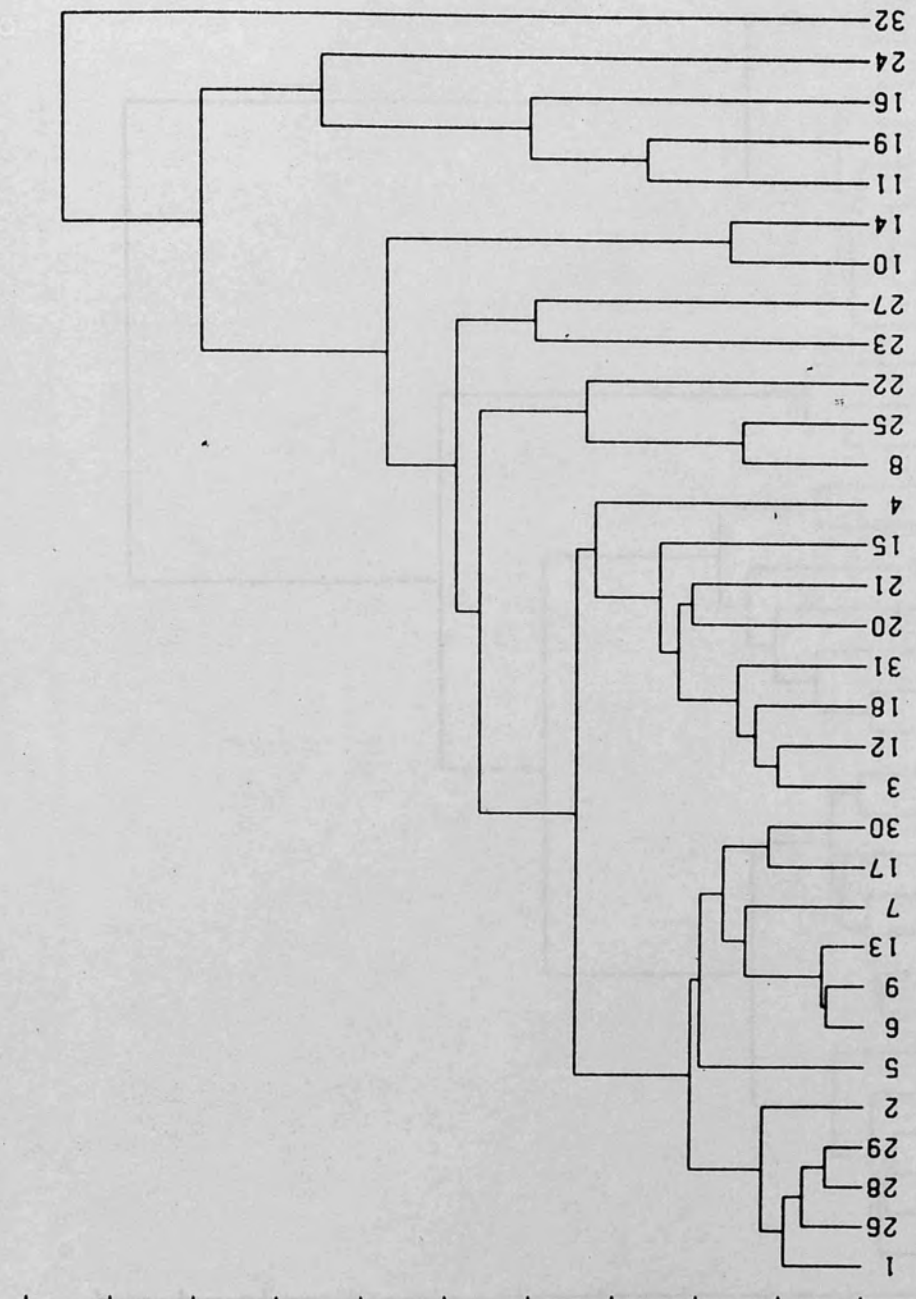


Fig. 4.1c Dendrogram produced by average linkage method with $(1 - r_{ij})/2$ as the similarity measure (orthogonal polynomial model)

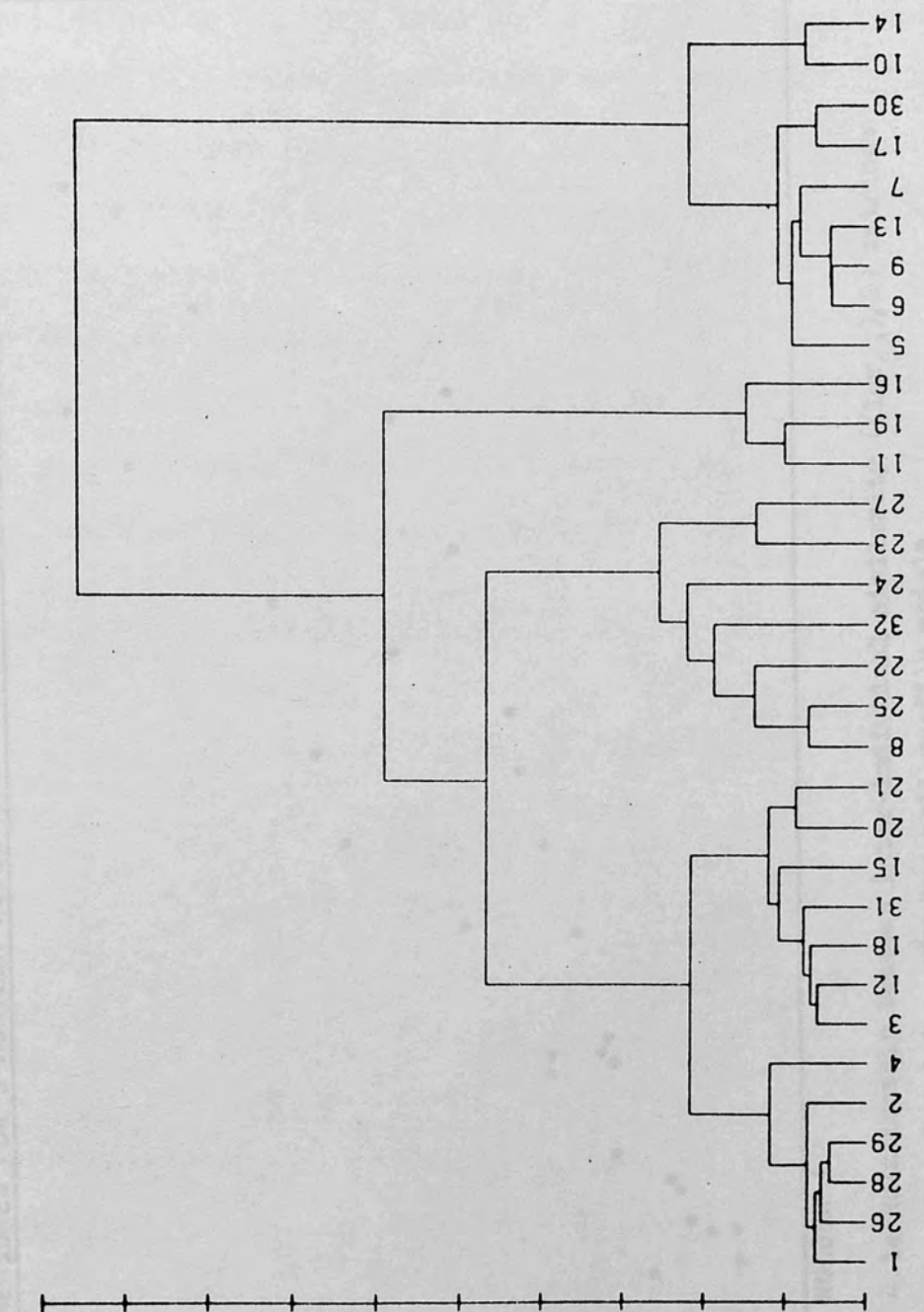


Fig. 4.1d Dendrogram peoduced by Ward's ESS method with $(1 - r_{ij})/2$ as the similarity measure (orthogonal polynomial model)

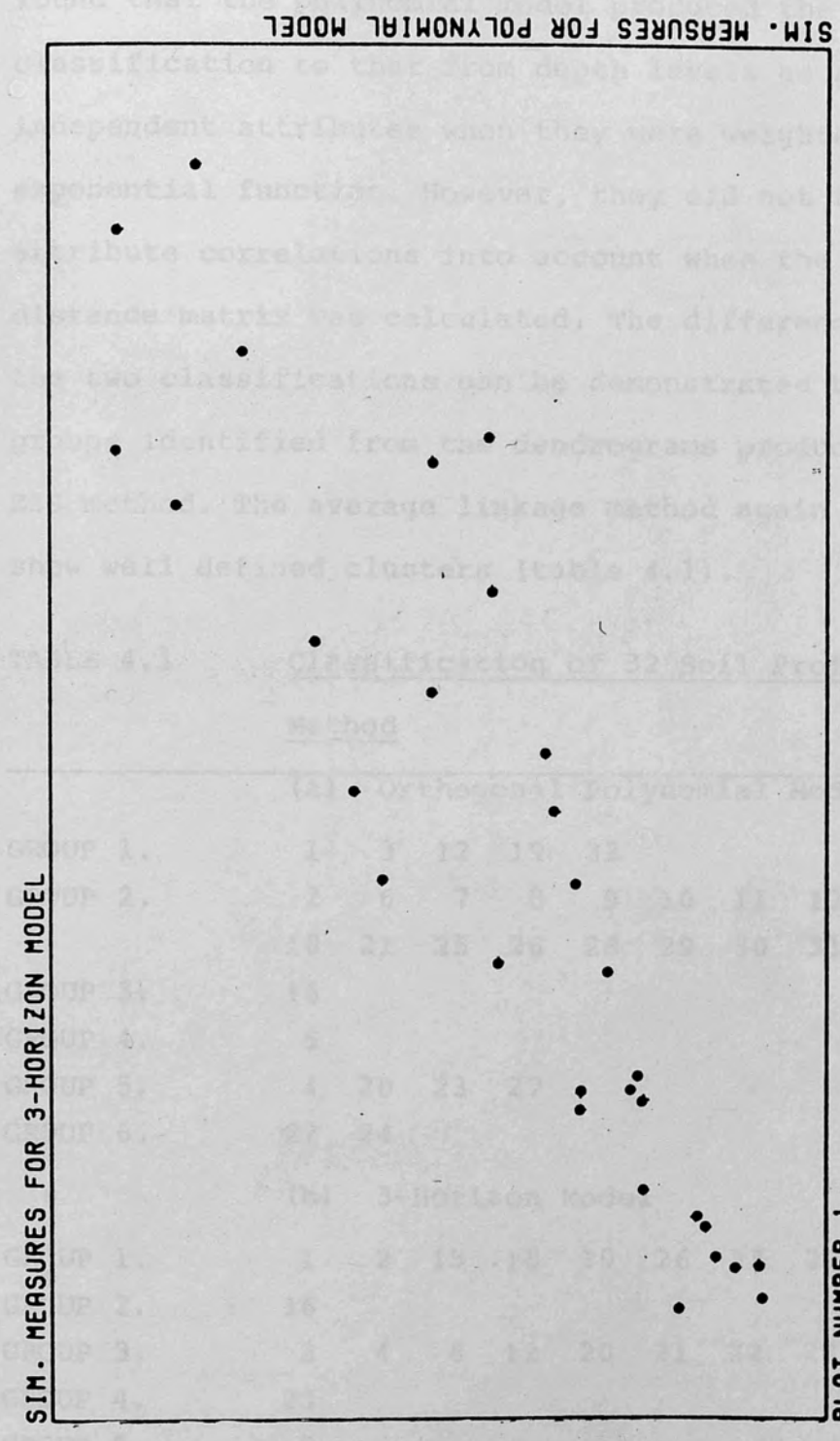


Fig. 4.1e Relationship between inter-individual similarity ($(1-r_{ij})/2$) matrices calculated from the two soil profile models

found that the polynomial model produced the same classification to that from depth levels as arrays of independent attributes when they were weighted with an exponential function. However, they did not take inter-attribute correlations into account when the Euclidean distance matrix was calculated. The difference between the two classifications can be demonstrated by the six groups identified from the dendrograms produced by Ward's ESS method. The average linkage method again failed to show well defined clusters (table 4.1).

TABLE 4.1 Classification of 32 Soil Profiles by Ward's Method

	(a) Orthogonal Polynomial Model										
GROUP 1.	1	3	12	19	32						
GROUP 2.	2	6	7	8	9	10	11	13	14	15	17
	18	21	25	26	28	29	30	31			
GROUP 3.	16										
GROUP 4.	5										
GROUP 5.	4	20	23	27							
GROUP 6.	22	24									
	(b) 3-Horizon Model										
GROUP 1.	1	2	15	18	19	26	28	29	31		
GROUP 2.	16										
GROUP 3.	3	4	8	12	20	21	22	24	25	27	32
GROUP 4.	23										
GROUP 5.	5										
GROUP 6.	6	7	9	10	11	14	17	30			

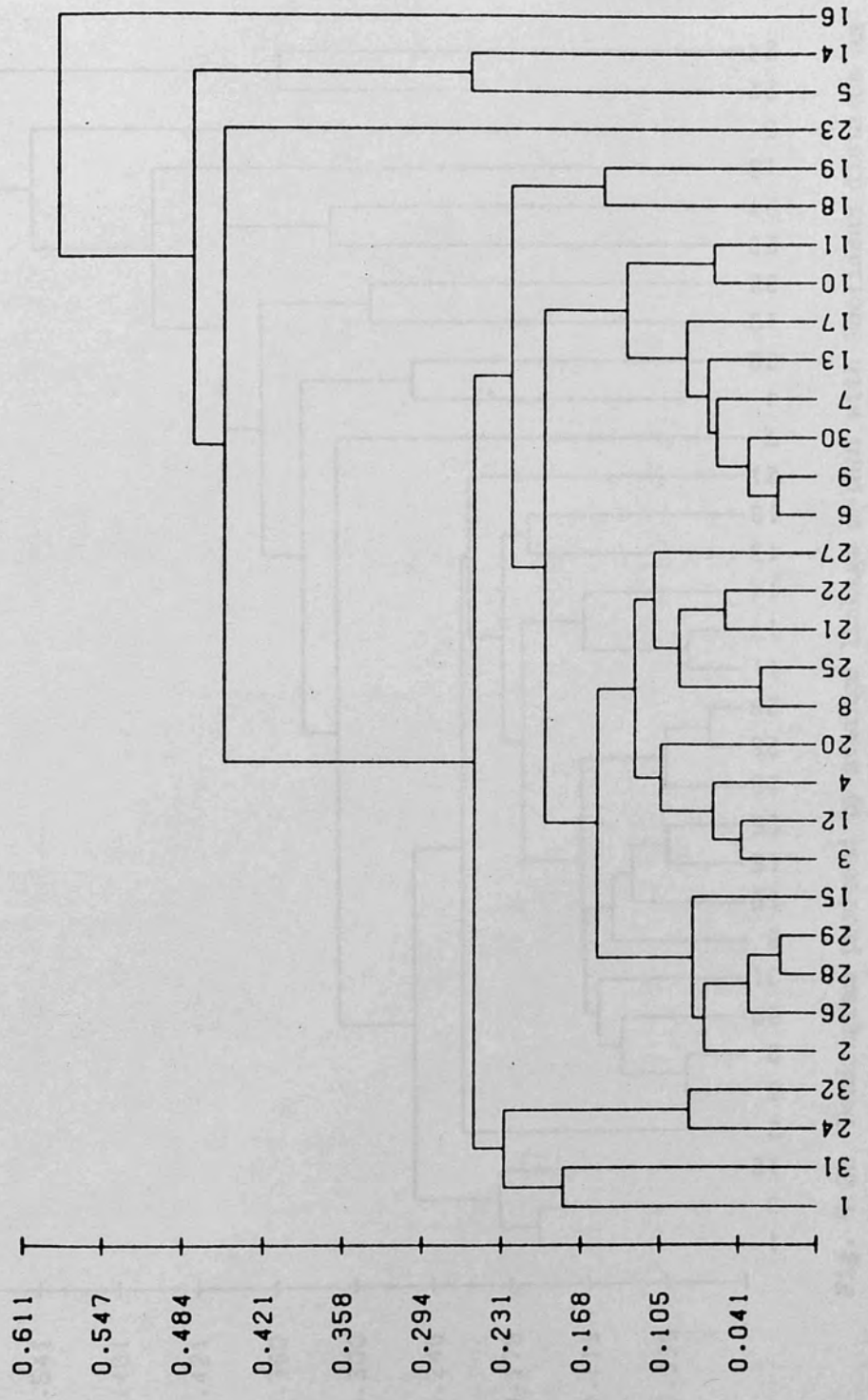


Fig. 4.2a Dendrogram produced by average linkage method with Euclidean distance as the similarity measure (3-horizon model)

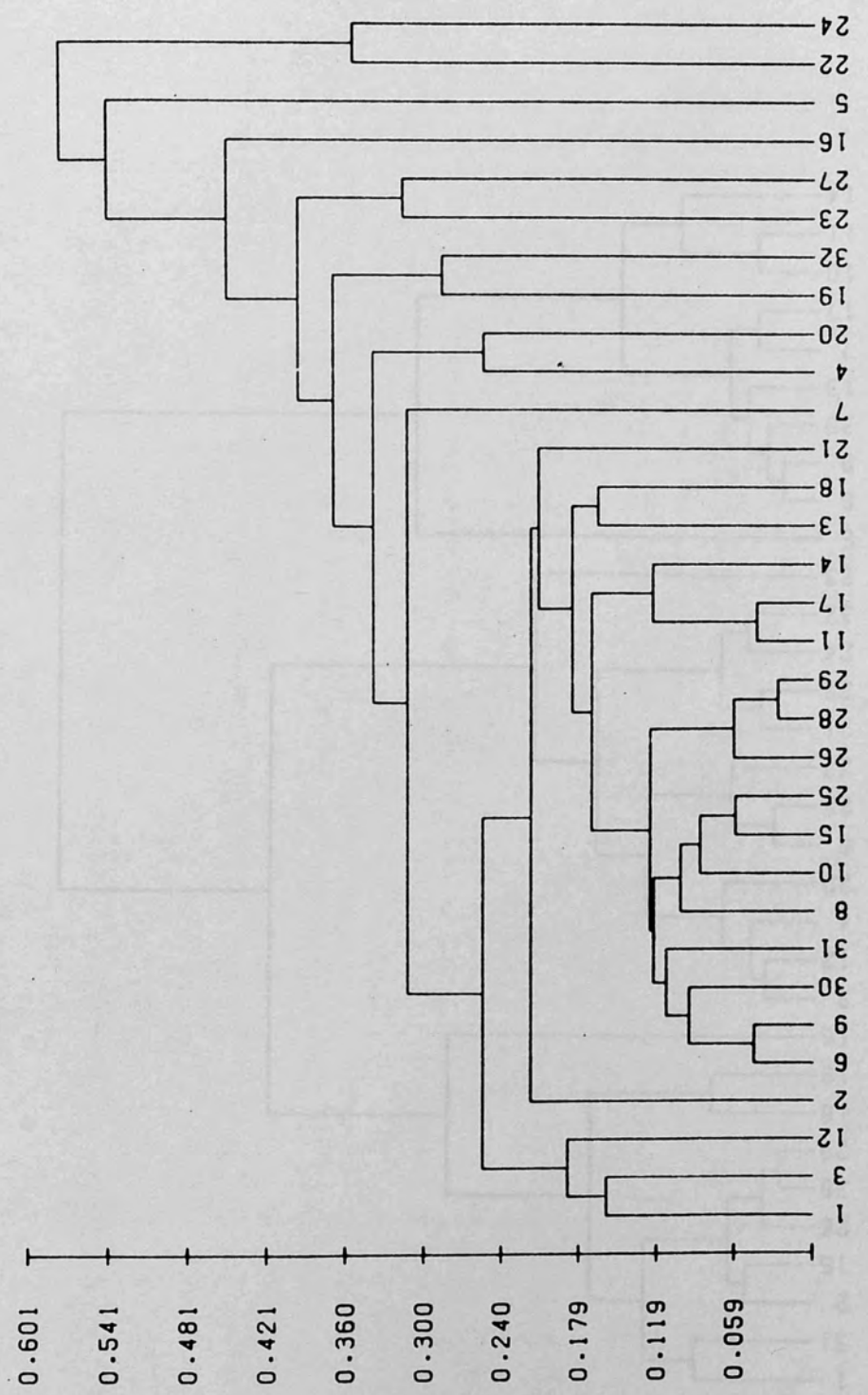


Fig. 4.2b Dendrogram produced by average linkage method with Euclidean distance as the similarity measure (orthogonal polynomial model)

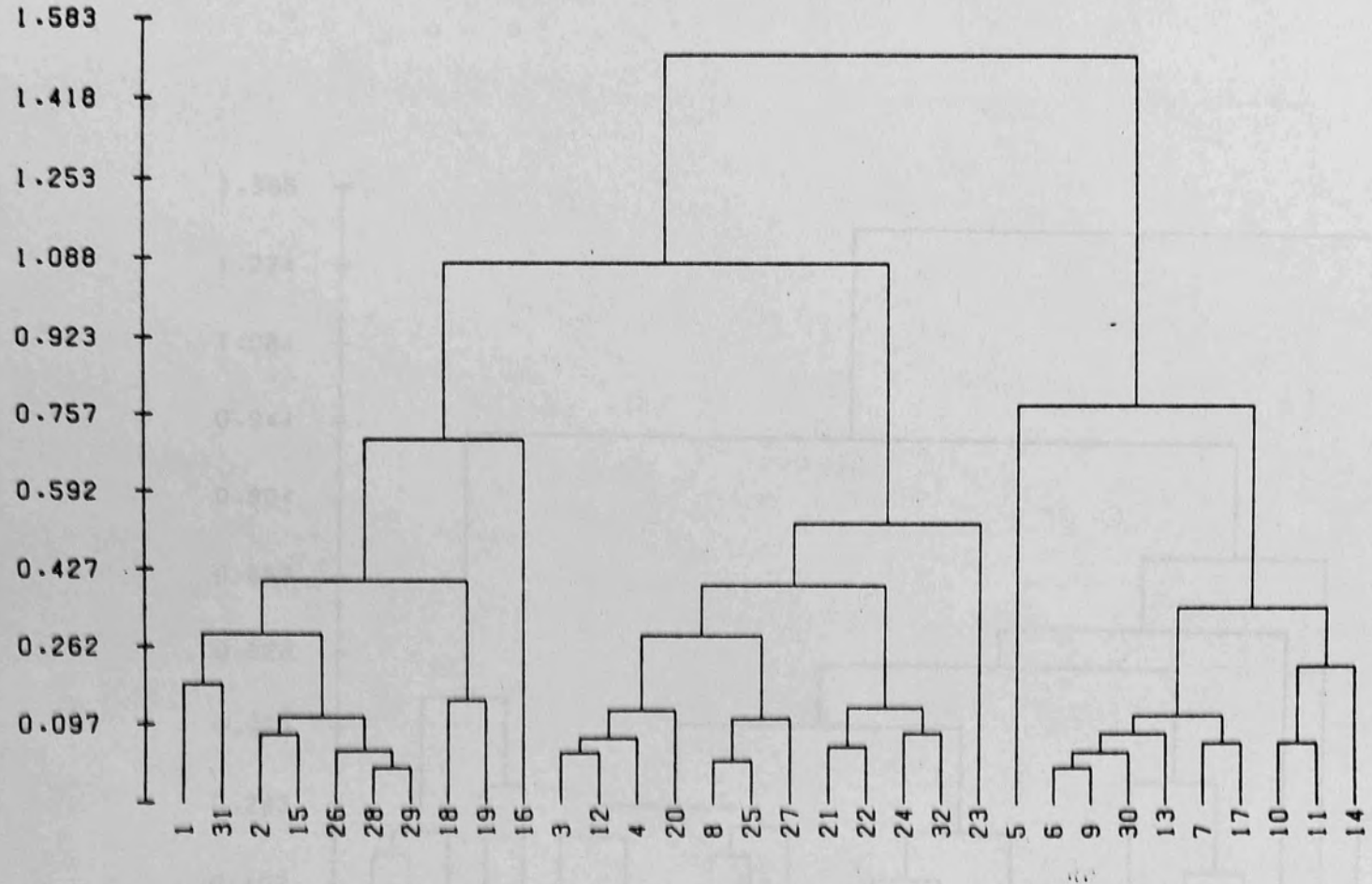


Fig. 4.2c Dendrogram produced by Ward's ESS method with Euclidean distance as the similarity measure (3-horizon model)

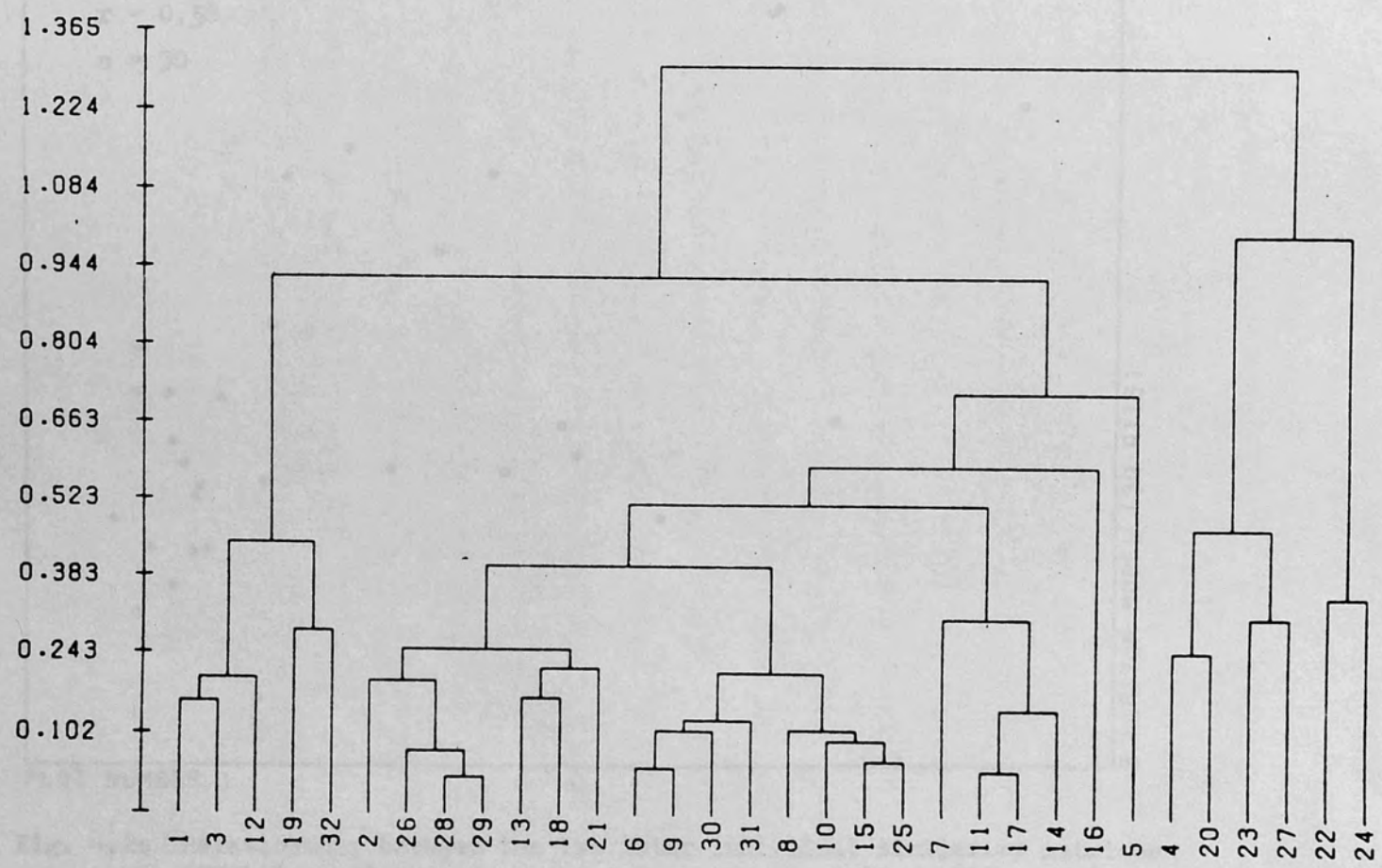


Fig. 4.2d Dendrogram produced by Ward's ESS method with Euclidean distance as the similarity measure (orthogonal polynomial model)

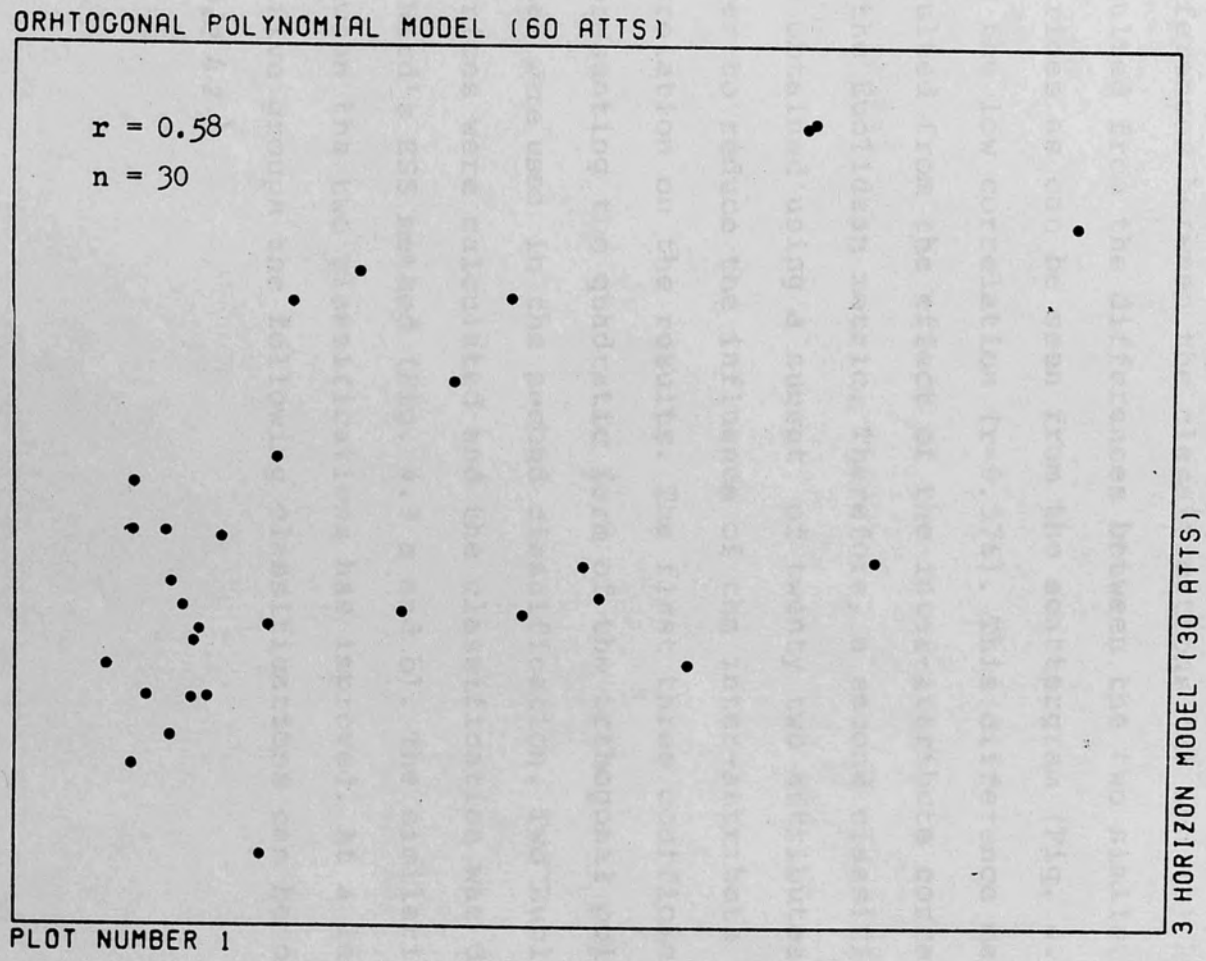


Fig. 4.2e Relationship between the two inter-individual similarity matrices (Euclidean distance) computed from the two soil profile models

There are only two groups in the two classifications which are similar, the rest are different from each other. Group 2 of the classification (a) is divided into two groups in the classification (b). This division can also be identified in the dendrogram (Fig. 4.2. b) of the classification (a), but not at this level of the hierarchy. Differences between the classifications (a) and (b) have resulted from the differences between the two similarity matrices as can be seen from the scattergram (Fig. 4.2. e) and the low correlation (r=0.576). This difference may have resulted from the effect of the inter-attribute correlation on the Euclidean metric. Therefore, a second classification was obtained using a subset of twenty two attributes in order to reduce the influence of the inter-attribute correlation on the results. The first three coefficients representing the quadratic form of the orthogonal polynomial model were used in the second classification. Two Euclidean matrices were calculated and the classification was done by Ward's ESS method (Fig. 4.3 a and b). The similarity between the two classifications has improved. At a level of five groups the following classifications can be obtained (Table 4.2).

TABLE 4.2 Classifications Obtained by Ward's Method from the two Soil Profile Methods

(a) 3-Horizon Model

GROUP 1.	1	2	3	4	12	15	20	26	28	29	31
GROUP 2.	8	21	22	23	24	25	27	32			
GROUP 3.	5	14									
GROUP 4.	6	7	9	10	11	13	17	30			
GROUP 5.	16	18	19								

(b) Orthogonal Polynomial Model

GROUP 1.	1	2	3	4	12	13	18	20	21	23	26	27
	28	29										
GROUP 2.	22	24										
GROUP 3.	5											
GROUP 4.	16											
GROUP 5.	6	7	8	9	10	11	14	15	17	19	25	30
	31	32										

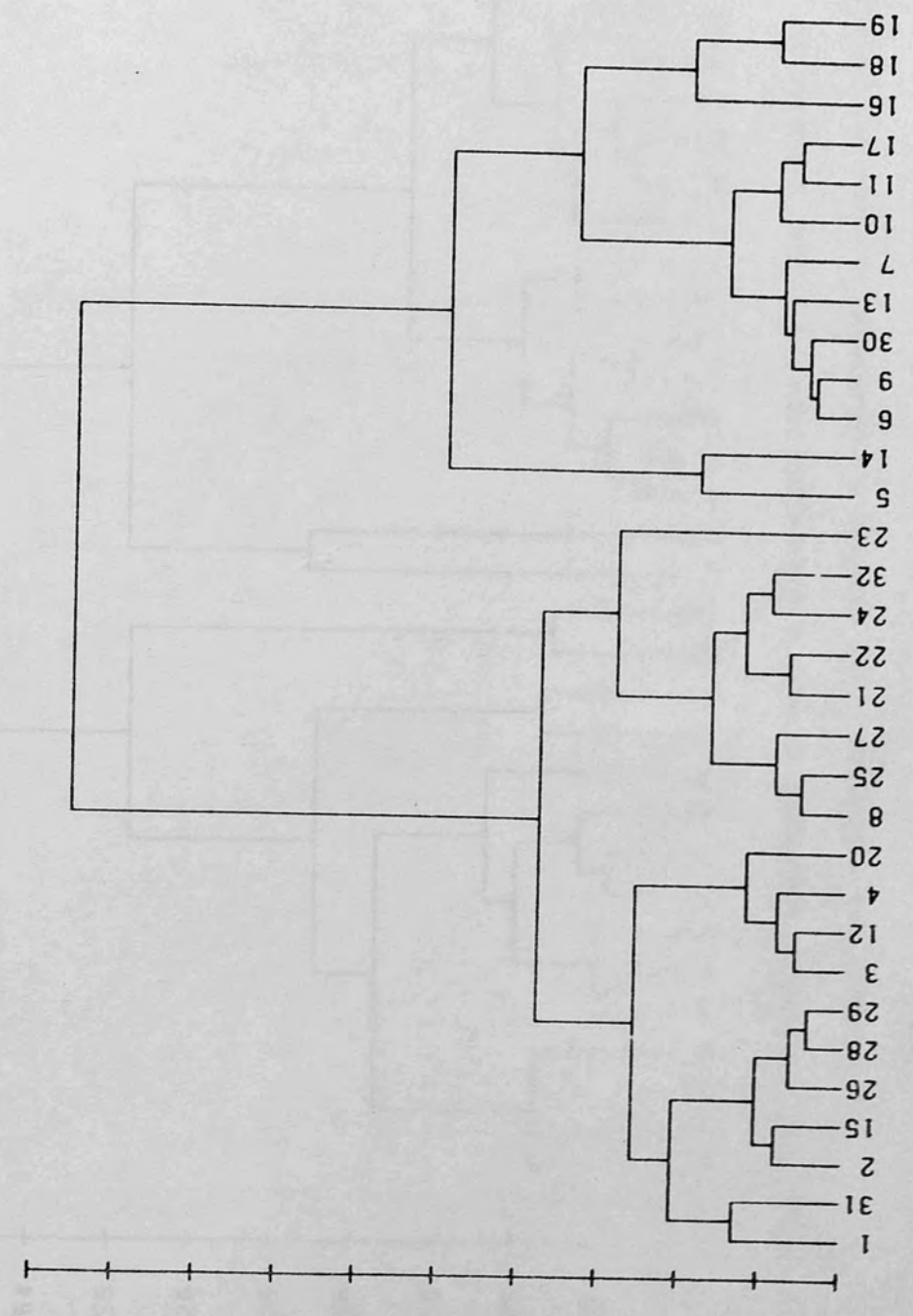


Fig. 4.3a Dendrogram produced by Ward's ESS method with Euclidean distance as the similarity measure after masking 8 attributes (3-horizon model)

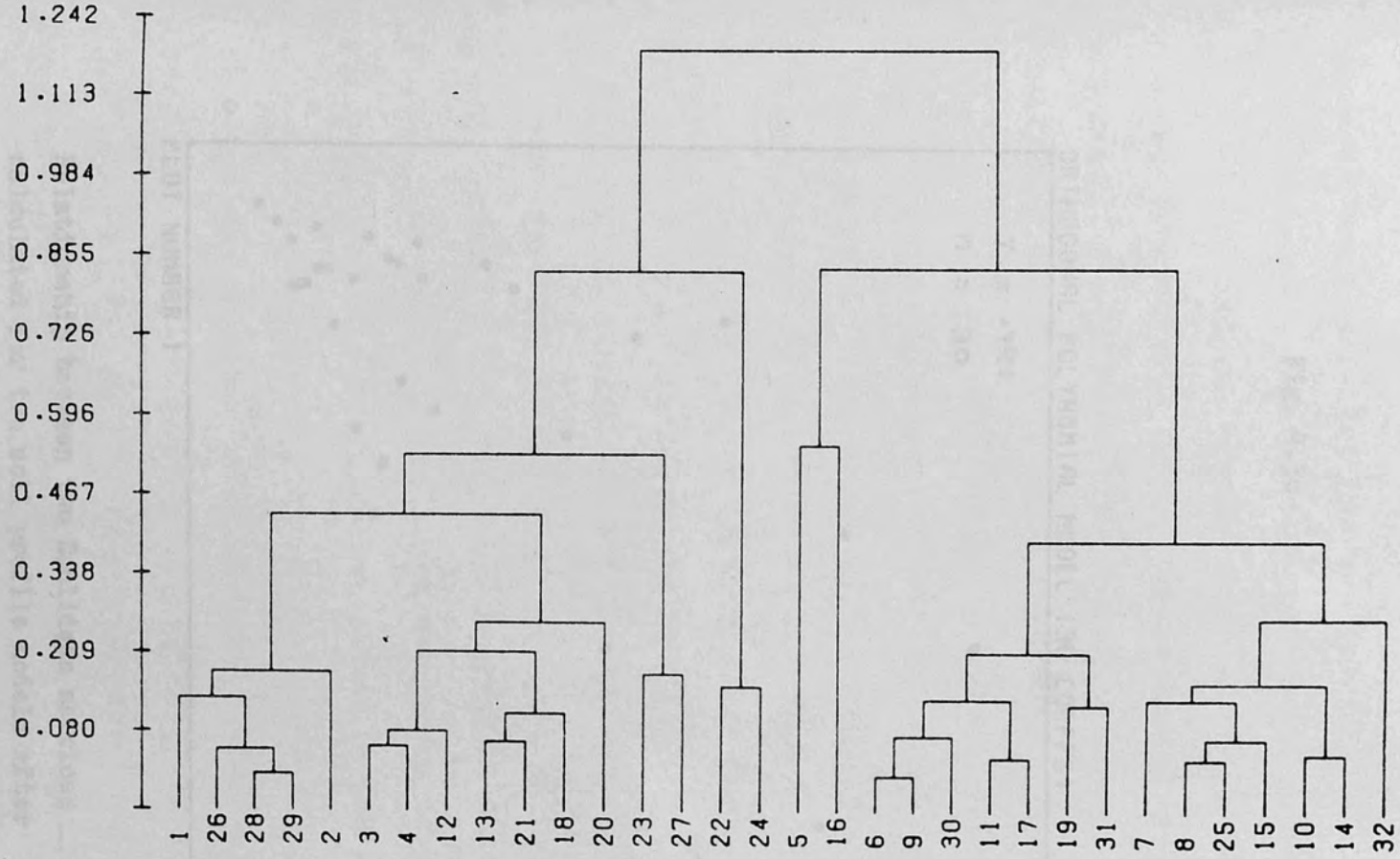
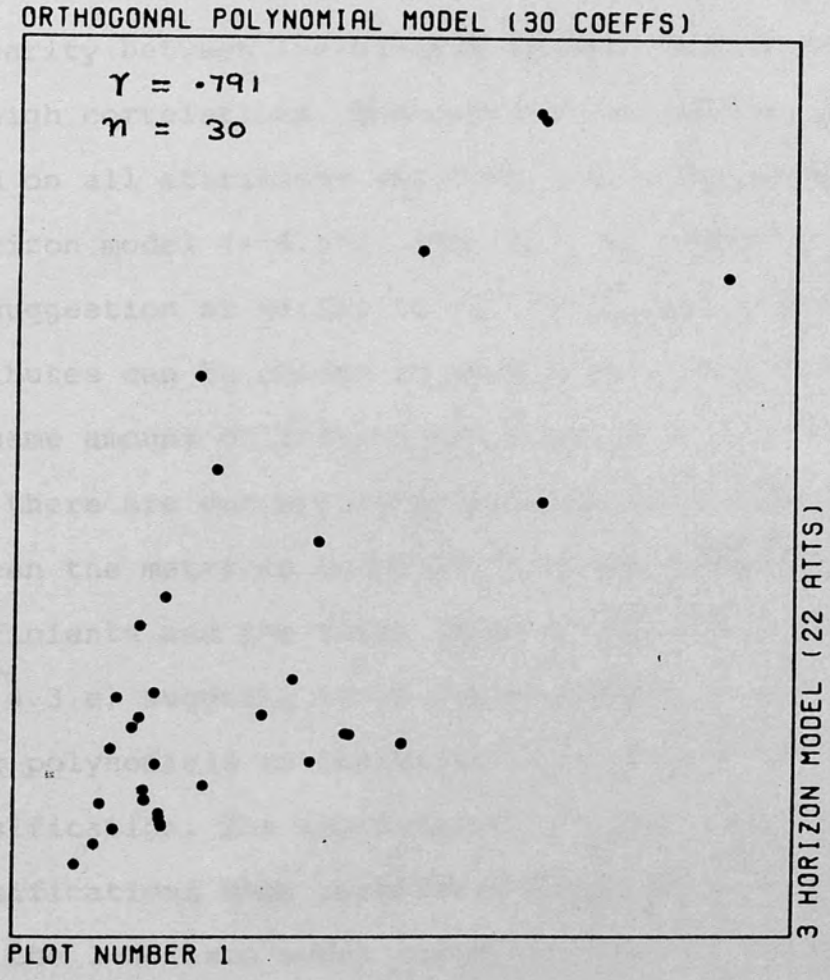


Fig. 4.3b Dendrogram produced by Ward's ESS method with Euclidean distance as the similarity measure after masking 30 attributes (orthogonal polynomial model)

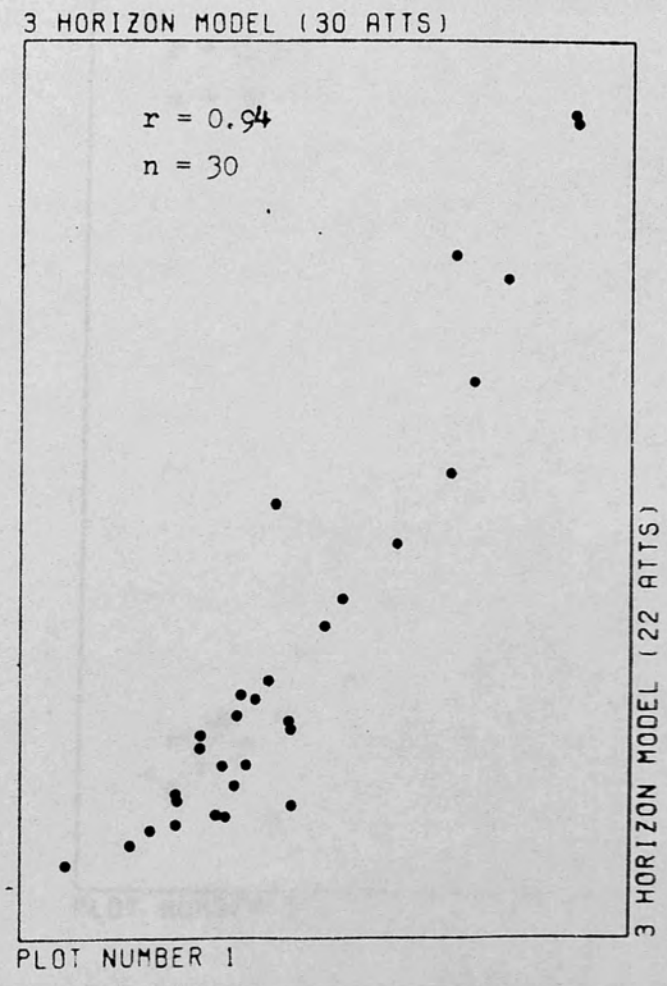
Fig. 4.3c



Relationship between two Euclidean matrices calculated for two soil profile models after masking 30 attributes of the polynomial model (coefficients $c_3 - c_5$) and 8 attributes of the 3-horizon model

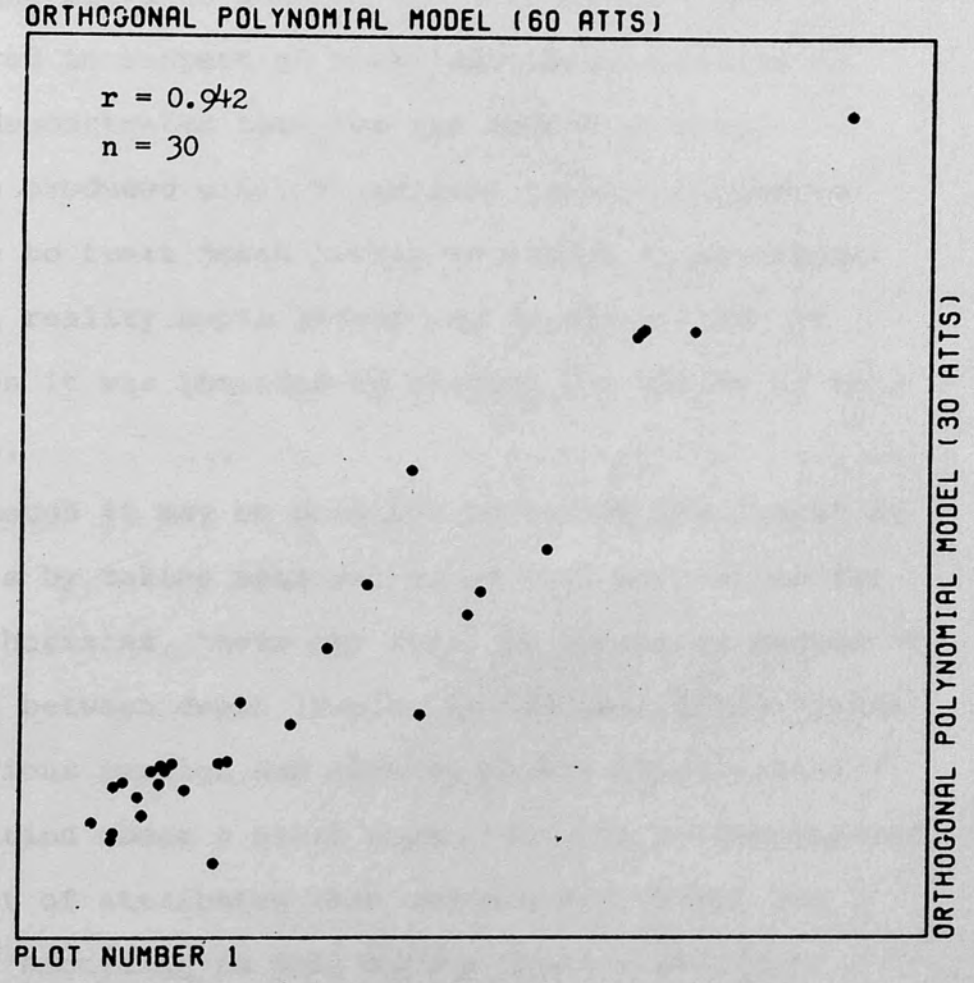
Although still there are differences between the two classifications, elimination of some attributes has improved the similarity between the two Euclidean matrices. This improvement is demonstrated by the scattergram (Fig. 4.3 c) and the higher correlation (r=0.791) between the two matrices. It is also interesting to note that the reduction of attributes has not changed the relative similarity between individuals in both models as shown by the high correlations. The correlation between the matrices based on all attributes and twenty two attributes of the 3-horizon model (r=0.942, Fig. 4.4. a) tends to confirm the suggestion of Sarkar et al. (1966) that a subset of attributes can be chosen in such a way that they contain the same amount of information about a population of soils when there are correlated attributes. The strong correlation between the matrices computed from all orthogonal polynomial coefficients and the first three coefficients (r=0.954, Fig. 4.3 e) suggests it is not necessary to use higher order polynomials to characterize soils for numerical classification. The improvement of similarity between the classifications when correlated attributes are eliminated from the 3-horizon model suggests that the differences may have caused by the unequal effect of inter-attribute correlations on the similarity matrix. Therefore, the use of mathematical functions to characterize soils for numerical classification is not necessary since similar results can

Fig. 4.4a



Relationship between two Euclidean distance matrices calculated for the 3-horizon model before and after masking attributes

Fig. 4.4b



Relationship between two Euclidean matrices obtained from the orthogonal polynomial model with all coefficients (c_0 - c_5) and 3 coefficients (c_0- c_2)

be obtained by a less complex model of the soil profile. The use of original observations raises the question of the number of such observations per property required to characterize a given soil.

4.3 Nature of Inter-attribute Correlation

In the previous section two soil profile models were compared in respect of numerical classification of soils and demonstrated that the two models of soil description produced similar results. It may, therefore, be possible to treat depth levels as arrays of attributes although in reality depth levels may be correlated. In this section it was intended to examine the nature of this correlation.

Although it may be possible to reduce the number of depth levels by taking mean values of soil properties for major soil horizons, there may still be a certain degree of correlation between depth levels. As has been demonstrated in the previous section and also by Sarkar et al.(1966) the information about a given population can be represented by a sub-set of attributes when certain attributes are correlated. According to Soil Survey Staff (1960) the attributes should be chosen in such a way that they should be the ones with the maximum number of accessory (correlated) attributes. Therefore, in a given data set, a certain number of redundant attributes may exist. These attributes can be eliminated in order to reduce the computational load. On the other hand, similarity measures defined in an Euclidean space require the attribute vectors

to be mutually orthogonal. Violation of this requirement would produce a distorted space in which relative distance between individuals is not accurate. Therefore, in this study attributes used in the classification of soils from USDA (1975) and the Soil Survey of England and Wales were classified on the basis of product-moment correlation.

Firstly, the thirty attributes used in the 3-horizon model were (table 3.1) classified by the average linkage method (Fig. 4.5). It can be seen from the dendrogram (Fig. 4.5) that in the majority of cases the clusters contain attributes related to different depth levels rather than totally different soil properties. This suggests that the correlation between the depth levels is greater than the correlation between soil properties. The same procedure was applied to the original observations for seven depth levels (Fig. 4.6). Again the groups produced by the classificatory strategy show the greater correlation between the depth levels except a few soil properties. As mentioned earlier, the Euclidean matrix requires the mutual orthogonality of attribute vectors but depth levels as arrays of attributes do not fulfil this requirement. On the other hand the use of correlated attributes is not necessary to characterize soils for numerical classification, since some attributes contain no additional information.

This analysis was extended to the data obtained from the Soil Survey of England and Wales. The classification obtained by the average linkage method (Fig. 4.7 a) does not show well defined clusters and therefore Ward's ESS method was also applied to the similarity matrix (Fig. 4.7 b).

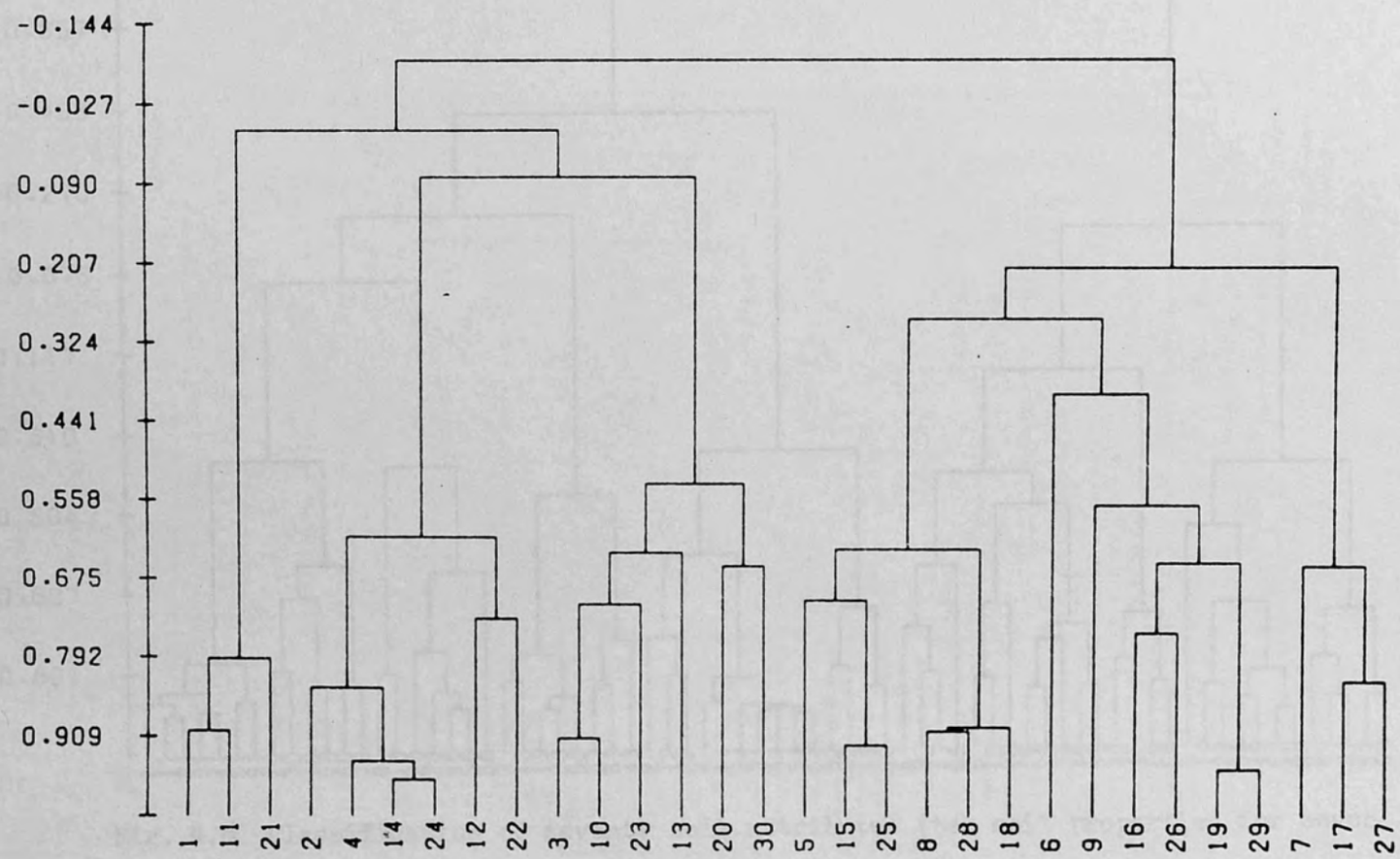


Fig. 4.5 Classification of 30 soil attributes (10 properties for 3 horizons) by average linkage method with product-moment correlation as the similarity measure

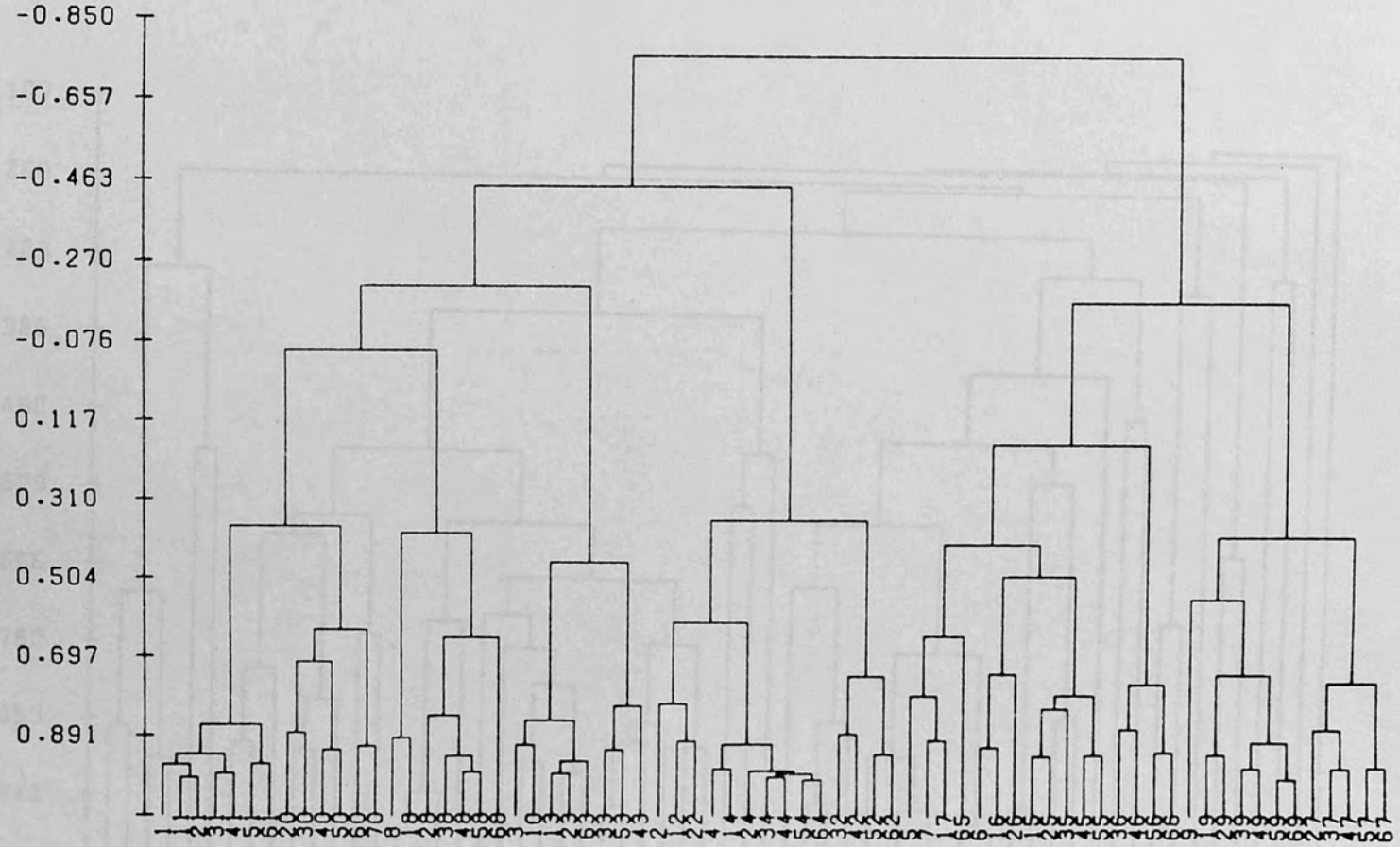


Fig. 4.6 Classification of seventy soil attributes (ten soil properties for seven horizons) by average linkage method with product-moment correlation as the similarity measure

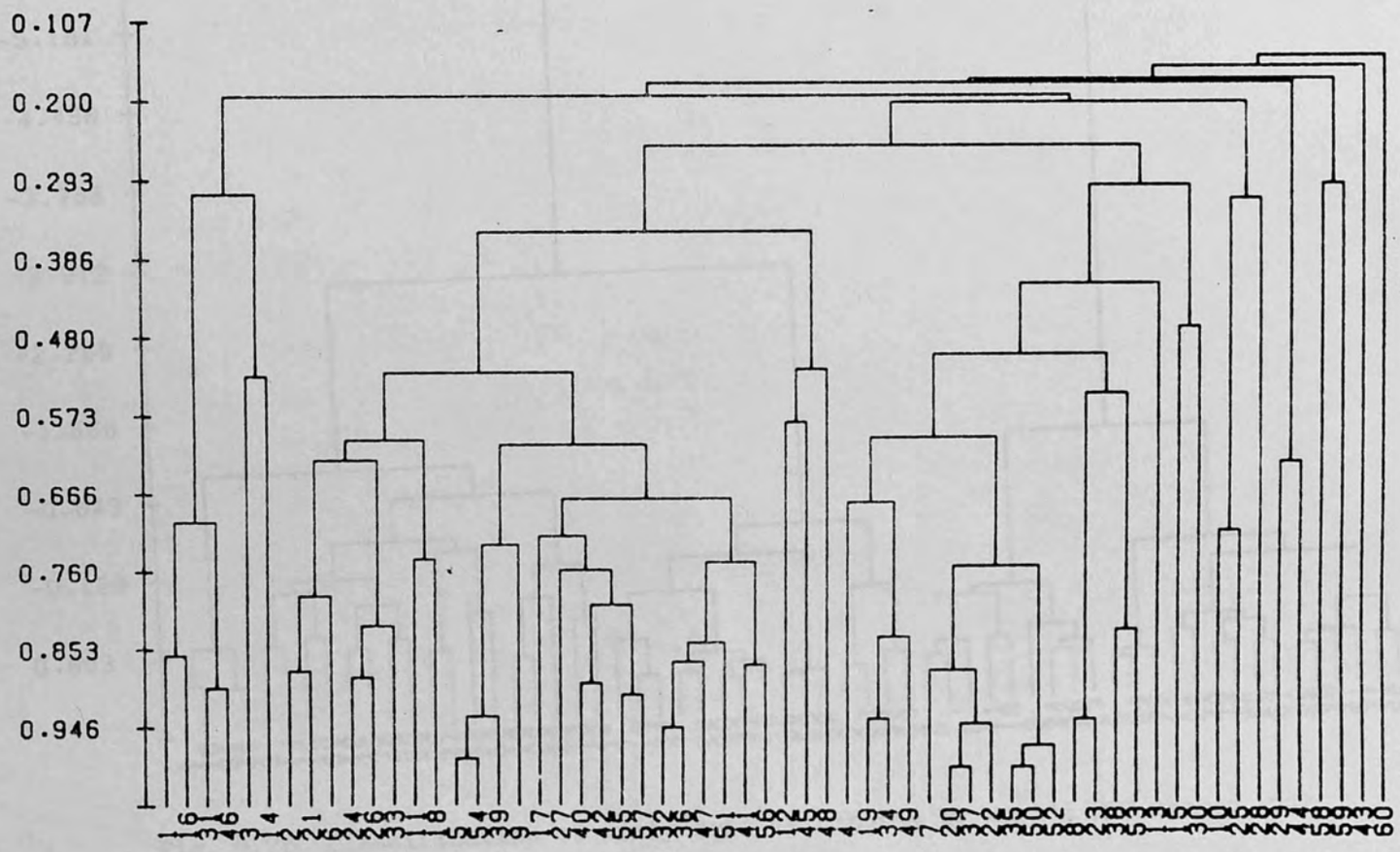


Fig. 4.7a Classification of sixty soil attributes by average linkage method with product-moment correlation as the similarity measure

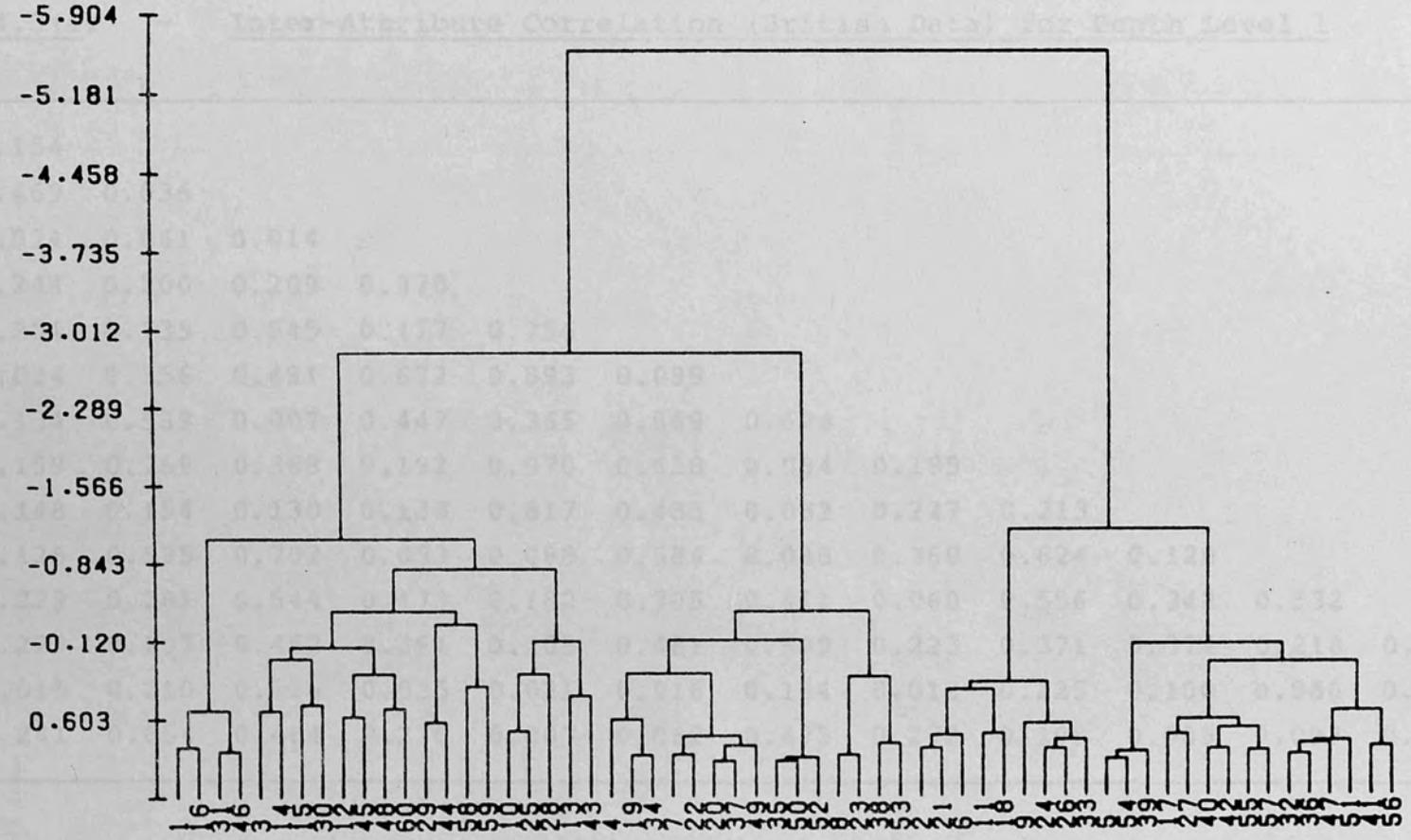


Fig. 4.7b Classification of sixty soil attributes by Ward's ESS method with product-moment correlation as the similarity measure

Table 4.3(a) - Inter-Attribute Correlation (British Data) for Depth Level 1

2	0.154													
3	0.469	0.036												
4	0.074	0.041	0.014											
5	0.244	0.200	0.209	0.370										
6	0.201	0.735	0.545	0.177	0.756									
7	0.024	0.156	0.491	0.672	0.893	0.099								
8	0.134	0.539	0.007	0.447	0.365	0.569	0.694							
9	0.159	0.769	0.368	0.192	0.970	0.628	0.084	0.185						
10	0.148	0.154	0.130	0.138	0.817	0.485	0.082	0.227	0.213					
11	0.126	0.595	0.702	0.053	0.088	0.584	0.045	0.360	0.624	0.120				
12	0.229	0.261	0.544	0.173	0.182	0.705	0.411	0.060	0.556	0.342	0.532			
13	0.223	0.107	0.462	0.361	0.209	0.481	0.609	0.223	0.371	0.374	0.218	0.116		
14	0.015	0.210	0.526	0.035	0.021	0.016	0.164	0.012	0.225	0.100	0.086	0.258	0.115	
15	0.241	0.054	0.464	0.270	0.067	0.062	0.475	0.272	0.108	0.005	0.097	0.233	0.406	0.205

Table 4.3(b) - Inter-Attribute Correlation (British Data) for Depth Level 2

2	0.070													
3	0.116	0.370												
4	0.170	0.112	0.279											
5	0.505	0.178	0.400	0.191										
6	0.183	0.079	0.761	0.729	0.040	0.538								
7	0.448	0.152	0.117	0.332	0.803	0.984	0.155							
8	0.308	0.024	0.437	0.223	0.554	0.632	0.656	0.740						
9	0.073	0.131	0.694	0.574	0.142	0.925	0.665	0.008	0.207					
10	0.009	0.178	0.351	0.356	0.035	0.692	0.489	0.046	0.177	0.515				
11	0.093	0.650	0.517	0.024	0.897	0.700	0.112	0.340	0.887	0.461				
12	0.231	0.758	0.610	0.010	0.180	0.506	0.223	0.070	0.763	0.556	0.364			
13	0.033	0.152	0.261	0.026	0.317	0.248	0.127	0.055	0.020	0.348	0.044	0.187		
14	0.127	0.166	0.181	0.204	0.231	0.347	0.360	0.361	0.153	0.143	0.268	0.059	0.195	
15	0.286	0.040	0.172	0.225	0.258	0.043	0.302	0.113	0.346	0.178	0.278	0.301	0.005	0.149

Table 4.3(c) - Inter-Attribute Correlation (British Data) for Depth Level 3

2	0.408													
3	0.397	0.813												
4	0.353	0.147	0.379											
5	0.178	0.013	0.803	0.259										
6	0.291	0.948	0.891	0.020	0.149									
7	0.175	0.137	0.007	0.826	0.986	0.220								
8	0.023	0.532	0.427	0.543	0.564	0.653	0.734							
9	0.387	0.752	0.528	0.227	0.741	0.735	0.014	0.192						
10	0.168	0.667	0.559	0.037	0.031	0.669	0.174	0.364	0.749					
11	0.311	0.752	0.649	0.053	0.712	0.767	0.334	0.500	0.762	0.670				
12	0.440	0.868	0.785	0.311	0.597	0.851	0.096	0.301	0.809	0.895	0.723			
13	0.011	0.256	0.408	0.082	0.293	0.368	0.267	0.260	0.278	0.117	0.432	0.214		
14	0.096	0.009	0.235	0.069	0.680	0.066	0.172	0.167	0.135	0.059	0.070	0.077	0.009	
15	0.159	0.020	0.026	0.174	0.180	0.112	0.226	0.223	0.231	0.314	0.193	0.307	0.055	0.203

TABLE 4.4 Correlation Between Depth Levels for British Data

Property	1/2	1/3	1/4	2/3	2/4	3/4
% Silt	0.860	0.650	0.610	0.783	0.612	0.899
% Clay	0.730	0.612	0.583	0.740	0.737	0.906
% Loss on Ignition	0.348	0.005	0.379	0.576	0.150	0.150
pH	0.646	0.735	0.656	0.940	0.796	0.887
% $CaCO_3$	0.932	0.551	0.983	0.644	0.989	0.866
CEC	0.842	0.537	0.461	0.638	0.323	0.877
% Base sat.	0.933	0.823	0.755	0.907	0.822	0.946
Exch. Ca	0.941	0.564	0.644	0.680	0.486	0.834
Mg	0.843	0.525	0.688	0.622	0.695	0.899
K	0.715	0.326	0.390	0.596	0.374	0.828
Na	0.671	0.422	0.190	0.667	0.522	0.874
Moisture	0.757	0.339	0.172	0.733	0.791	0.789
Thickness (mm)	0.052	0.389	0.094	0.228	0.058	0.102
Value	0.273	0.134	0.068	0.632	0.028	0.115
Chroma	0.470	0.157	0.080	0.519	0.295	0.183

Both dendrograms show a much more confused picture. The nature of relationship between depth levels and different soil properties can be demonstrated by considering the correlation matrices. Correlation between the fifteen soil properties for three horizons are listed in Table 4.3. Although there is a certain number of correlated properties, a great majority of them are not correlated. Table 4.4 shows that the correlation between depth levels for fifteen soil properties. All properties except the thickness of horizons, colour value and chroma are highly correlated depthwise. In some cases correlation between different soil properties is marginally higher than between depth levels, it is quite clear that soil properties are depthwise correlated and as a result depth levels cannot be considered as arrays of statistically independent attributes.

The findings of this analysis suggest that it may not be necessary to use a large number of depth levels or horizons to characterize soils for numerical classification. This helps reduce the number of attributes required when numerical methods are applied to classify soils. A sub set of soil attributes can be chosen as Sarkar et al (1966) did without a considerable loss of information.

CHAPTER 5

A COMPARATIVE STUDY OF SEVEN AGGLOMERATIVE CLUSTERING STRATEGIES

5.0 Introduction

The development of numerical taxonomy has led to the devising of a whole range of cluster seeking strategies. Among them, the hierarchical agglomerative strategies have been by far the most widely used strategies, especially in biological taxonomy (Sneath and Sokal, 1973, p.214). Certain properties of these methods were discussed in a previous chapter (chapter 2). Since those strategies do not always produce the same results, it was felt necessary to examine their properties empirically. Agglomerative strategies (table 5.1) are based on a similarity matrix and involve sorting of similar individuals into groups (clusters) by successive fusion. This process is generally continued until all individuals and groups are fused together to form a hierarchical tree which is presented graphically by a dendrogram. Different agglomerative methods differ from each other in the way in which similarity between individuals and groups or between groups is determined.

The agglomerative clustering strategies can be compared in two important ways:

(a) optimality of a classification in respect of some statistical criterion.

(b) goodness-of-fit.

The global optimality of a given classification may not be determined by existing methods but classifications obtained by different methods can be compared to choose a better classification. In this chapter the second property of agglomerative strategies was considered. The goodness-of-fit of a given strategy on the original space defined by a similarity matrix is important since certain strategies may lead to imposing artificial structures when the population is not well structured. Some agglomerative strategies tend to produce well separated clusters (P. 74-75 e.g. Ward's ESS method). At this point the properties of the similarity measure used are not important because all strategies were applied on the same similarity matrix.

The goodness-of-fit of a clustering strategy depends on its ability to preserve the original similarities between individuals. However, fusion of individuals or groups involve a certain degree of distortion of the original space and it tends to vary from strategy to strategy. The degree of distortion or the goodness-of-fit can be measured by cophenetic correlation defined by Sokal and Rohlf (1962). The cophenetic correlation r_c is defined as product-moment correlation between the original similarity matrix (S) and the similarity matrix (S*) of cophenetic values obtained from dendrograms. Sokal and Rohlf (1962) compared four agglomerative clustering

strategies and concluded that cophenetic correlation between them was greater than between the original similarity matrix and the cophenetic matrices. This study was extended to other widely available agglomerative strategies (table 5.1) to demonstrate the relationship between the goodness-of-fit of a strategy on the original space and the clarity of clusters, and also the relationship between different clustering strategies.

5.1 Data and method

The data of the 3-horizon model as applied to USDA data was used to obtain a series of classifications by seven agglomerative strategies (table 5.1). The similarity measure used was the squared Euclidean distance described in chapter 2. All seven classifications are presented by dendrograms. The similarity matrix was generated by the TAXON (CSIRO) program REMUL and classification was performed using the agglomerative methods available in the CLUSTAN 1C Release 2 (Wishart 1969a). Macquitty's similarity analysis strategy was not used because it is identical to the Lance-William's flexible sort with $\beta = 0.0$.

TABLE 5.1 Agglomerative Clustering Strategies used to Classify 32 Soil Profiles

1.	Single linkage method (nearest neighbour method)
2.	Complete linkage method (furthest neighbour method)
3.	Average linkage method (unweighted pair-group method)
4.	Centroid method
5.	Median sort method
6.	Ward's error sum of squares (ESS) method
7.	Lance-Williams flexible sort method ($\beta = 0.0$)

* The procedure used to obtain similarity coefficients involved measurements on the vertical axes of the dendrograms, which show the relative similarity between individuals. A new inter-individual similarity matrix can be obtained from the dendrograms and this matrix is called cophenetic matrix (S^*).

A sample of thirty pairs of individuals was chosen randomly from the original similarity matrix and the corresponding similarity coefficients were obtained from seven dendrograms (Fig. 5.1 a to g)* Product-moment correlation was computed between original similarity coefficients and the similarity coefficients obtained from seven dendrograms, and also between the cophenetic coefficients themselves. These relationships are presented by a series of scattergrams (Fig. 5.2).

5.2 Results

The classifications obtained by seven agglomerative strategies listed in table 5.1 are presented by dendrograms. It can be seen from the dendrograms that some of them are more similar to each other than to others with respect to cluster (group) memberships. For example, the dendrograms obtained by the single linkage method and the centroid method (Fig. 5.1 a and d). The relationship between strategies and cophenetic correlation for seven strategies are listed in table 5.2.

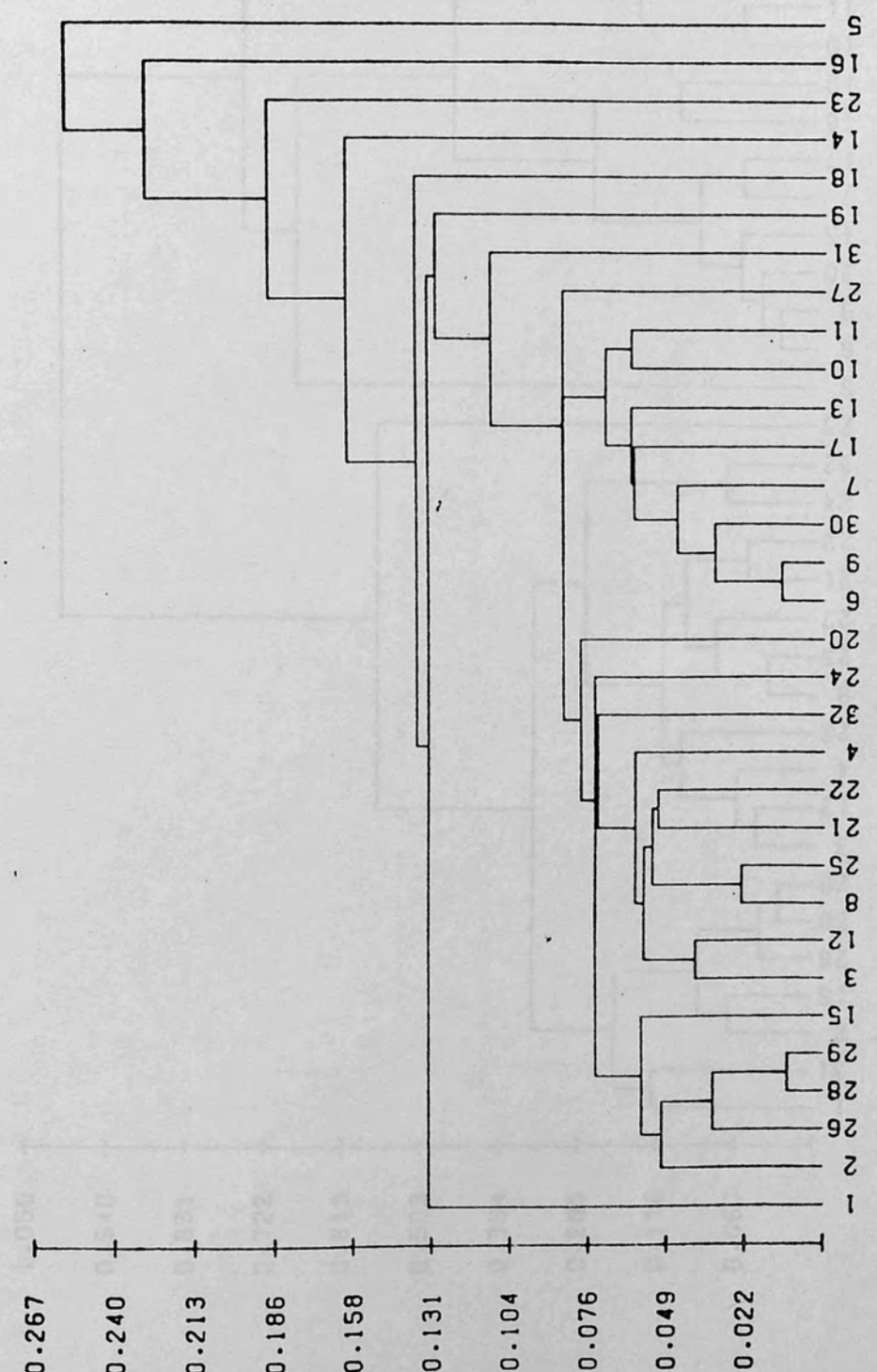


Fig. 5.1a Dendrogram produced by single linkage method

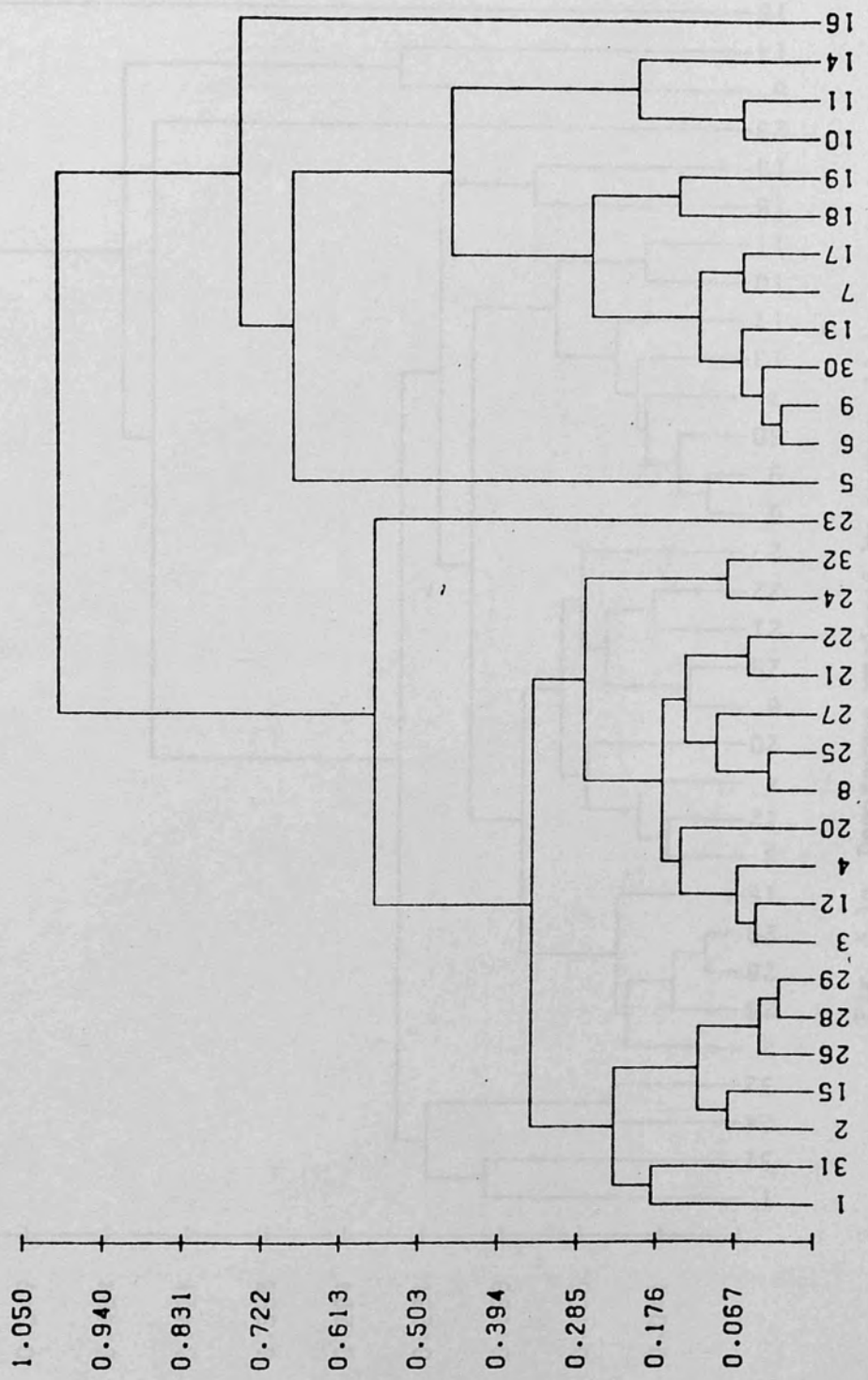


Fig. 5.1b Dendrogram produced by complete linkage method

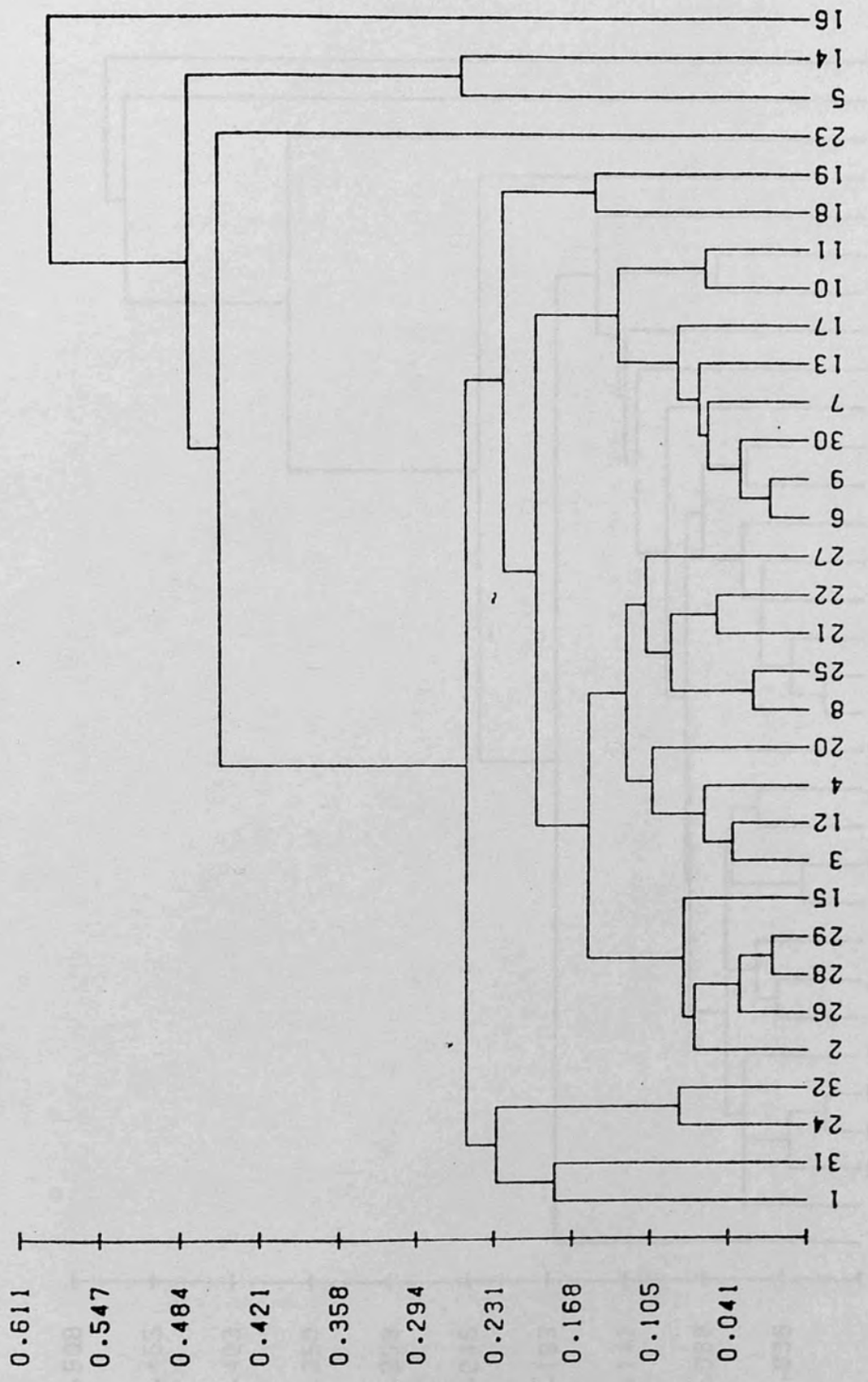


Fig. 5.1c Dendrogram produced by average linkage method

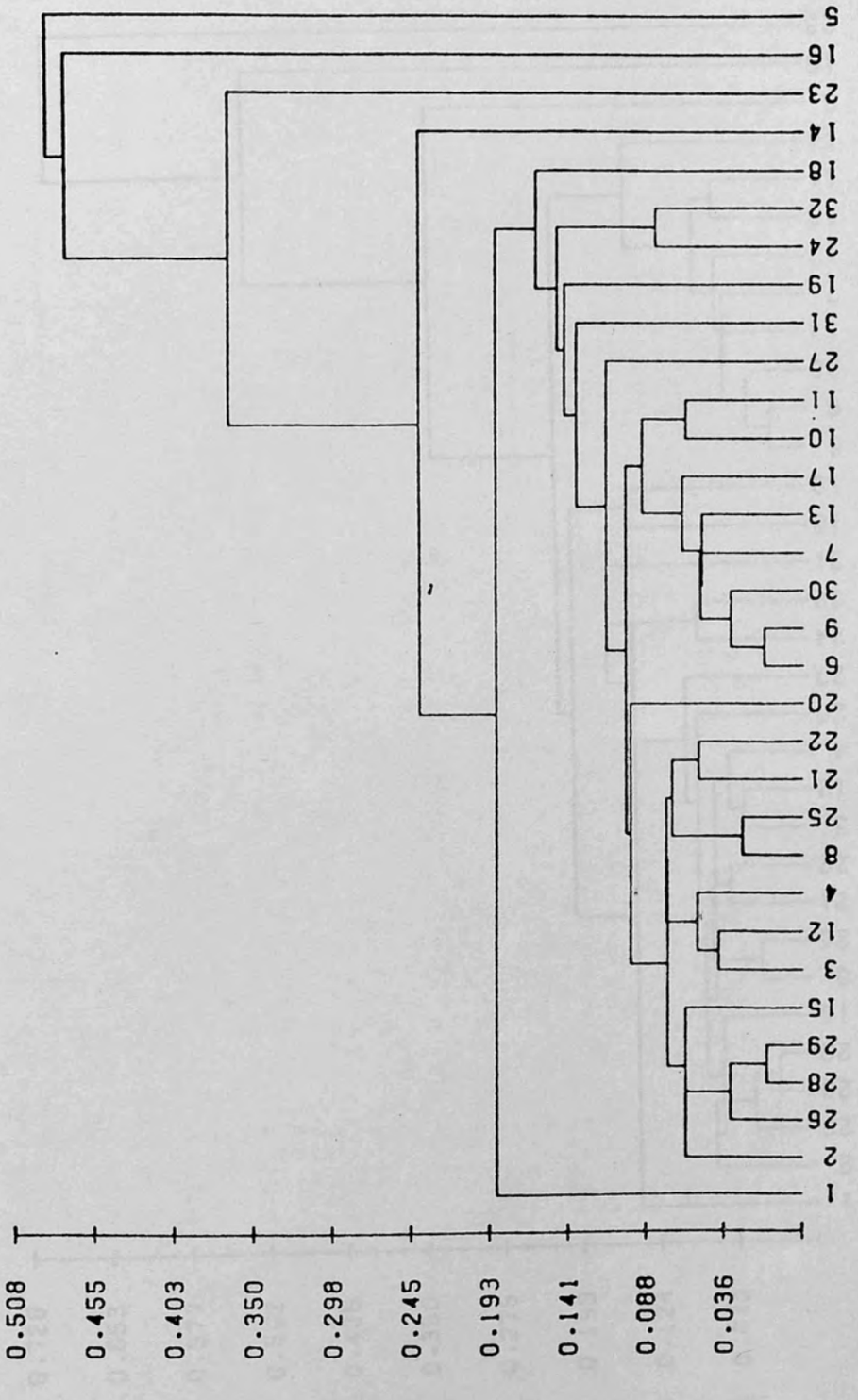


Fig. 5.1d Dendrogram produced by centroid method

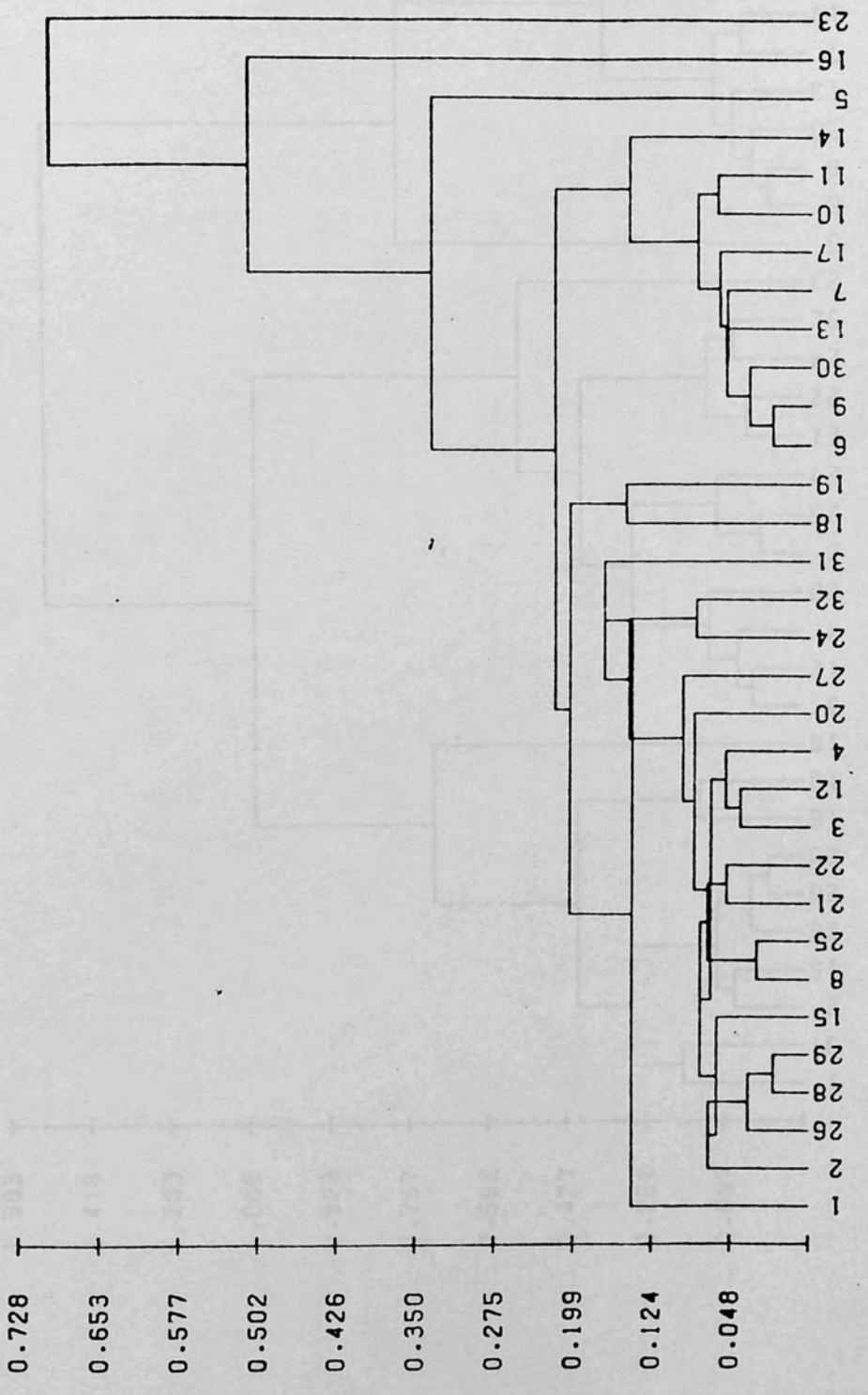


Fig. 5.1e Dendrogram produced by median sort

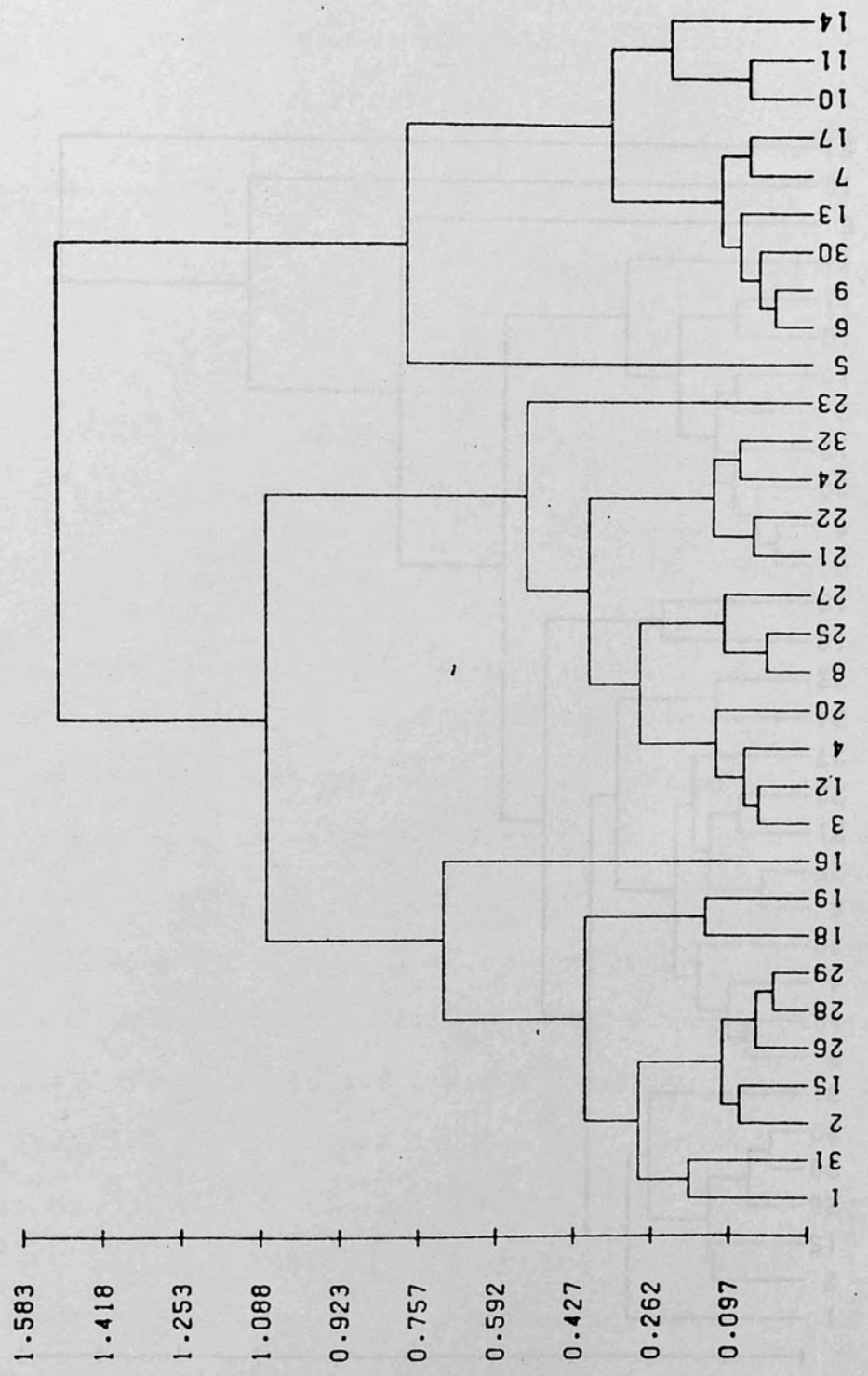


Fig. 5.1f Dendrogram produced by Ward's ESS method

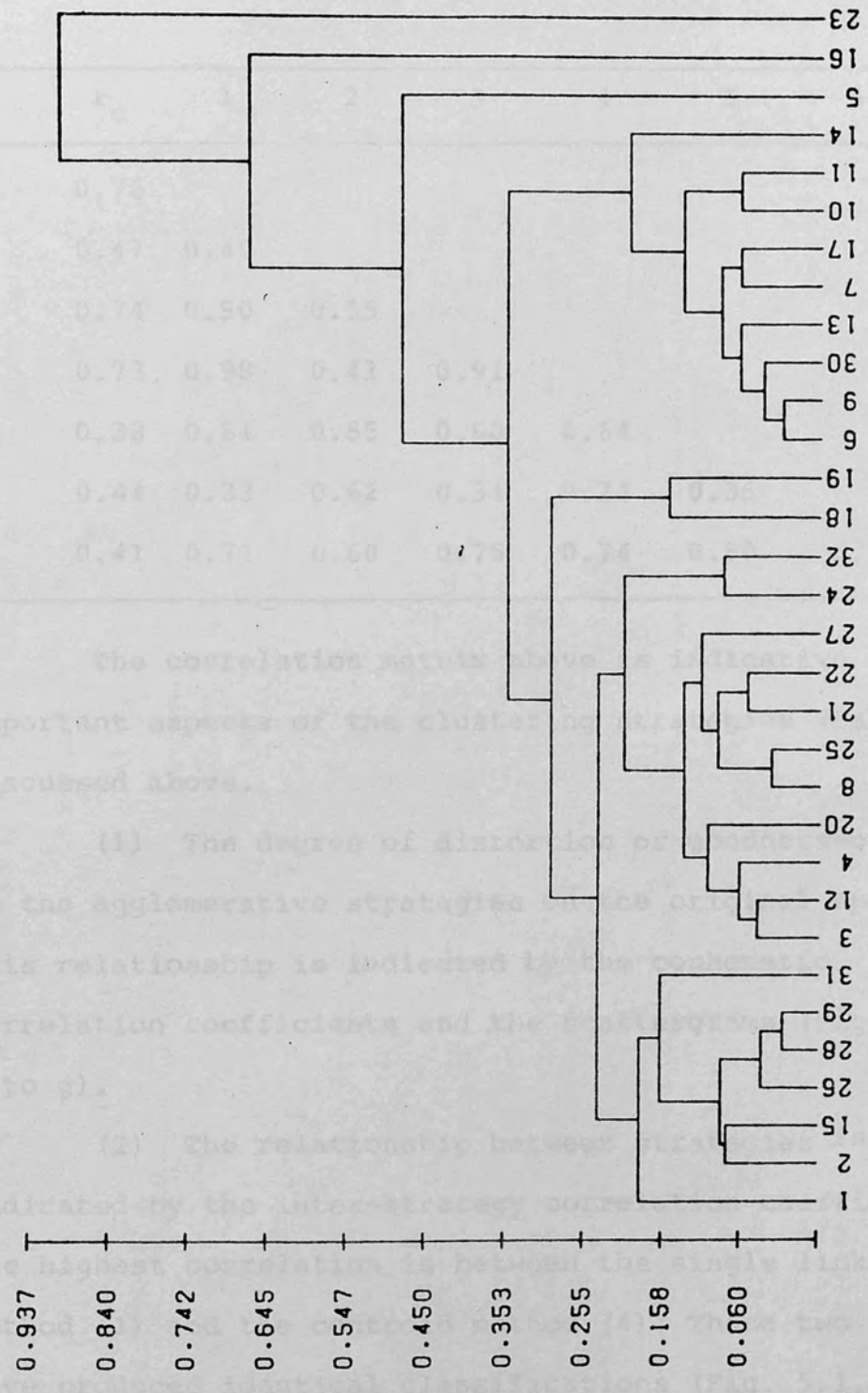


Fig. 5.1g Dendrogram produced by Lance-Williams flexible sort

TABLE 5.2 Matrix of Correlation Coefficients among Cophenetic Values and Original Similarity Matrix

	r_c	1	2	3	4	5	6
1	0.76						
2	0.47	0.40					
3	0.74	0.90	0.55				
4	0.73	0.98	0.42	0.91			
5	0.38	0.61	0.55	0.60	0.64		
6	0.44	0.33	0.62	0.31	0.24	0.36	
7	0.41	0.71	0.60	0.75	0.74	0.80	0.46

The correlation matrix above is indicative of two important aspects of the clustering strategies (table 5.1) discussed above.

(1) The degree of distortion or goodness-of-fit of the agglomerative strategies on the original space. This relationship is indicated by the cophenetic correlation coefficients and the scattergrams (Fig.5.2, a to g).

(2) The relationship between strategies is also indicated by the inter-strategy correlation coefficients. The highest correlation is between the single linkage method (1) and the centroid method (4). These two strategies have produced identical classifications (Fig. 5.1,a and d).

Fig. 5.2a Cophenetic correlation for the single linkage method

$r_c = 0.76$

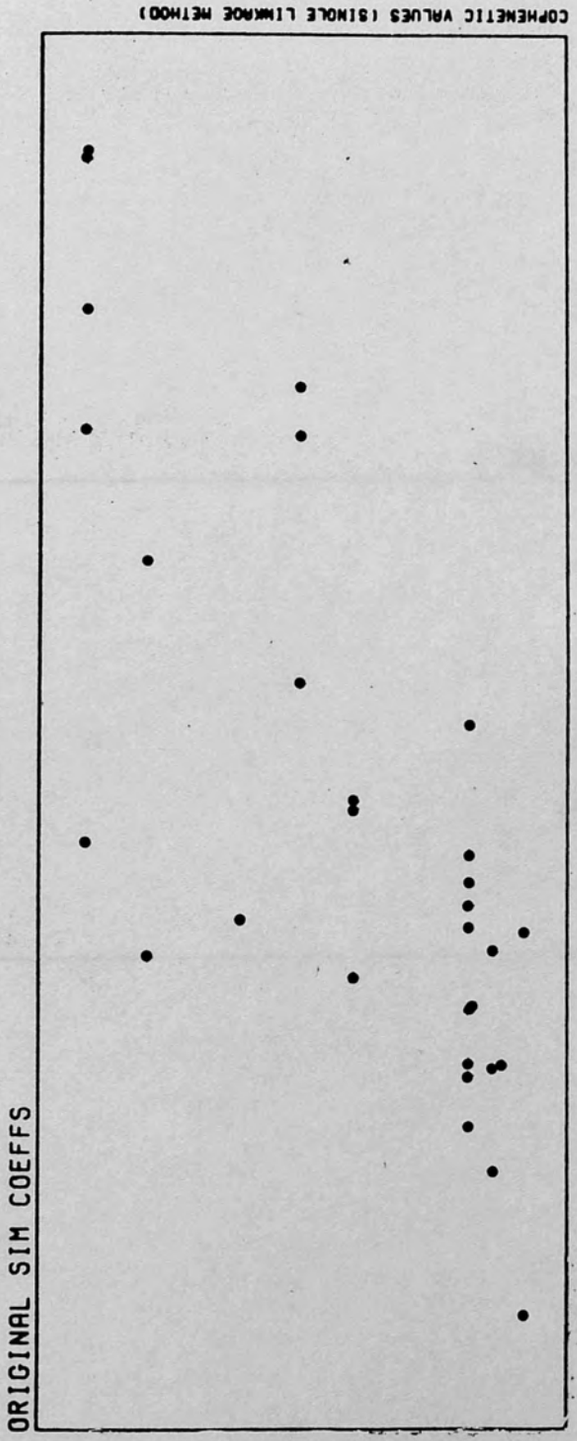
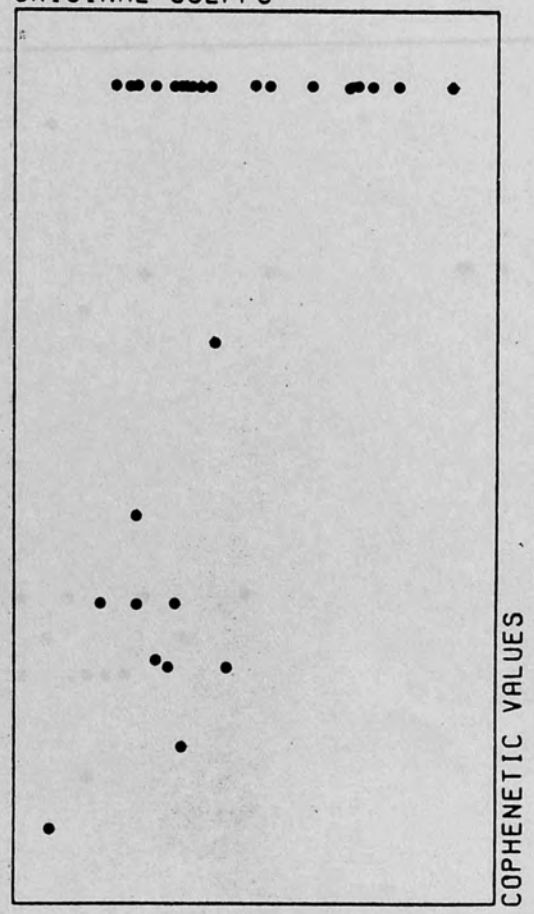


Fig. 5.2b Cophenetic correlation for the complete linkage method

$r_c = 0.47$



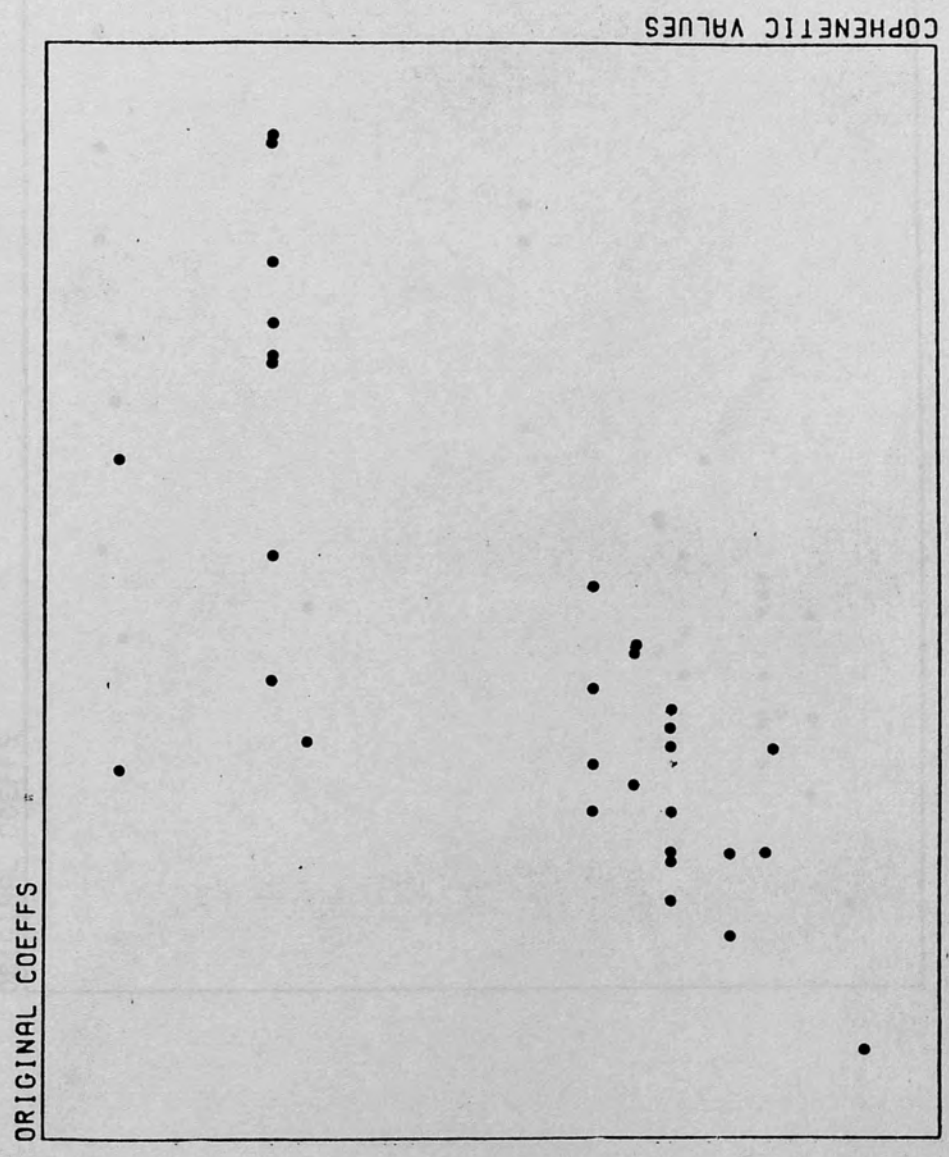
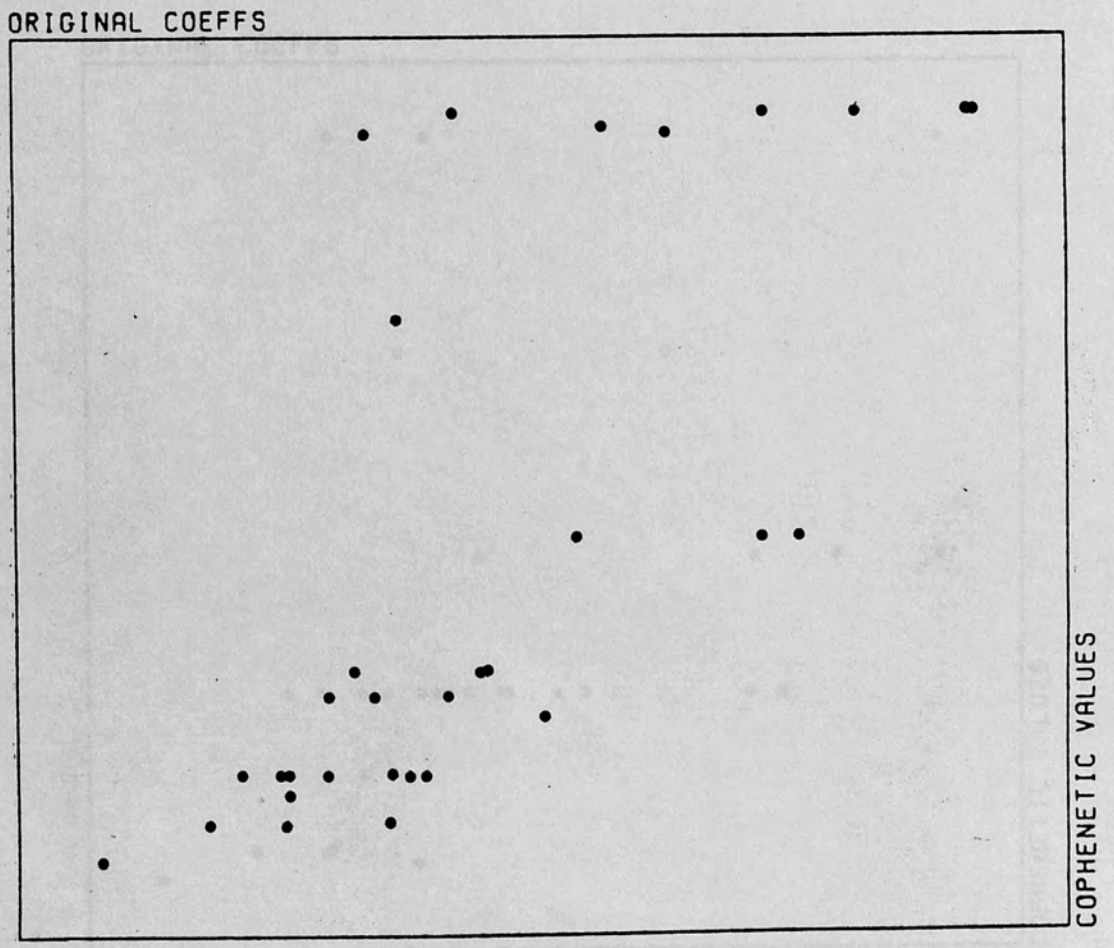


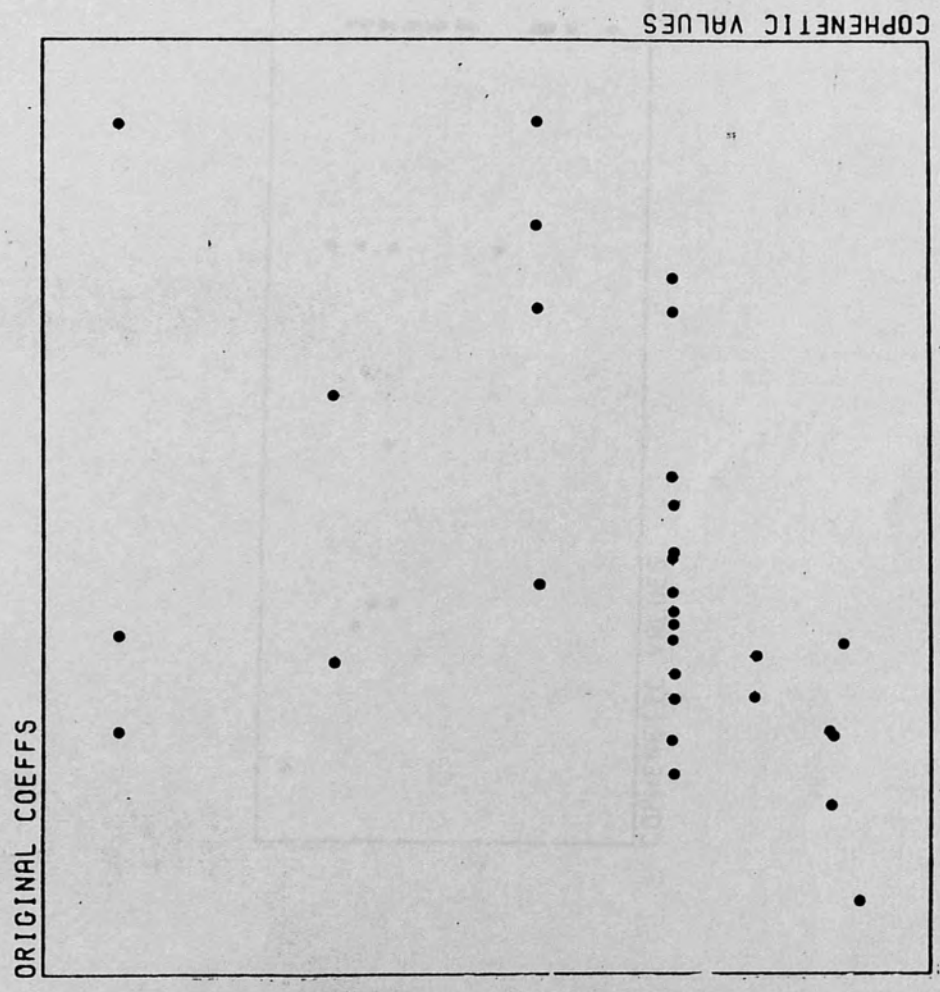
Fig. 5.2c Cophenetic correlation for the average linkage method

$r_c = 0.74$

Fig. 5.2d Cophenetic correlation for the centroid method

$r_c = 0.73$





$r_c = 0.38$

Fig. 5.2e Cophenetic correlation for the median sort

Fig. 5.2f Cophenetic correlation for the Ward's ESS method

$r_c = 0.44$

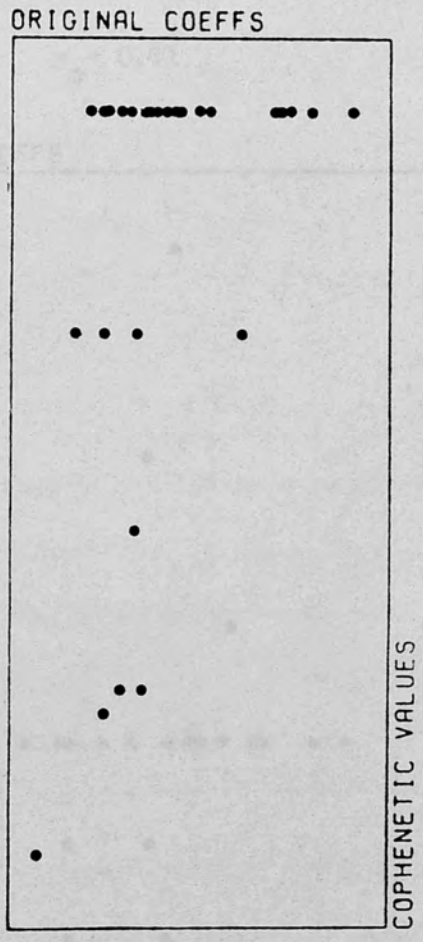
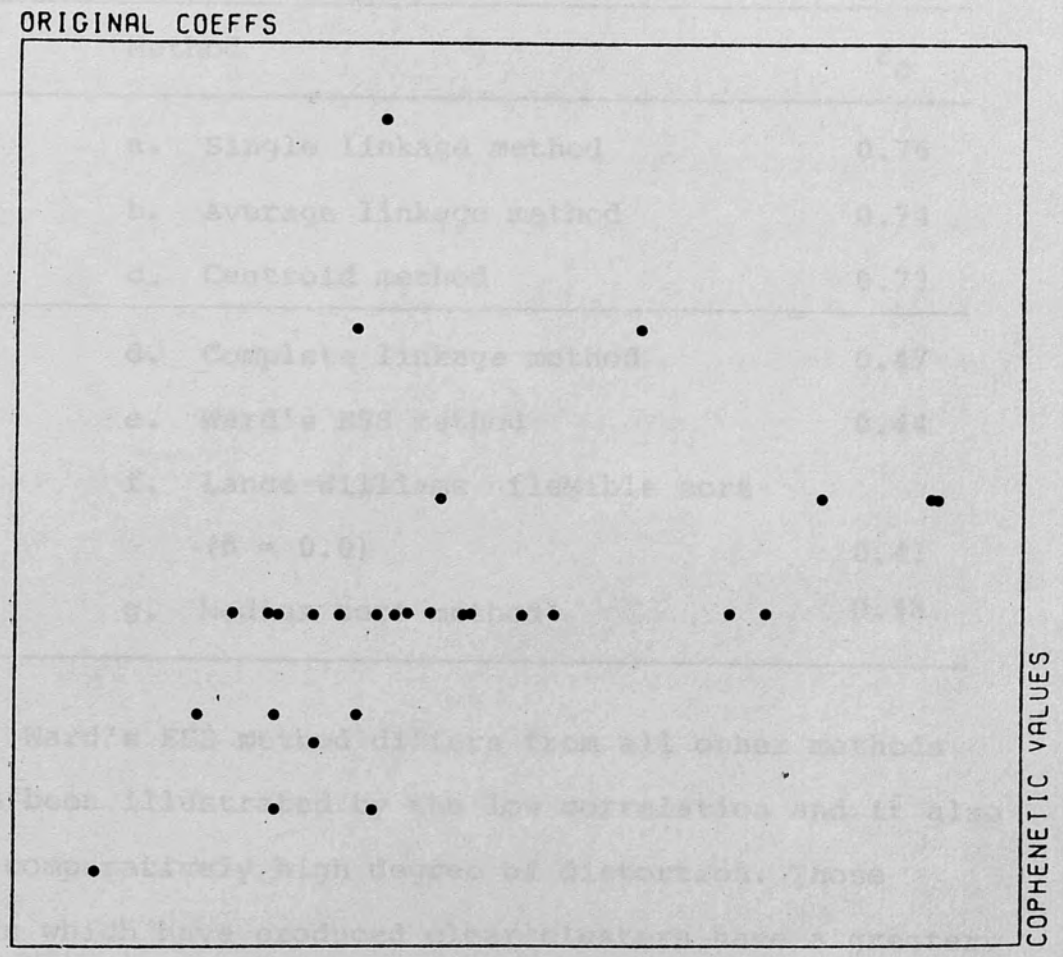


Fig. 5.2g Cophenetic correlation for the Lance-Williams method

$r_c = 0.41$



As far as cophenetic correlation is concerned the goodness-of-fit varies from strategy to strategy but a clear division of strategies into two classes can be seen (table 5.3).

TABLE 5.3 Ordering of Seven Agglomerative Strategies According to Cophenetic Correlation

	Method	r_c
I	a. Single linkage method	0.76
	b. Average linkage method	0.74
	c. Centroid method	0.73
II	d. Complete linkage method	0.47
	e. Ward's ESS method	0.44
	f. Lance-Williams flexible sort ($\beta = 0.0$)	0.41
	g. Median sort method	0.38

Ward's ESS method differs from all other methods as has been illustrated by the low correlation and it also has a comparatively high degree of distortion. Those methods which have produced clear clusters have a greater distortion.

5.3 Discussion and Conclusion

The results reported above show that goodness-of-fit can be obtained at the expense of the clarity of clusters. The single linkage method is the one with the least distortion but it has failed to show clusters. When

a greater emphasis is placed on the original relationships a certain degree of chaining of individuals cannot be avoided. The strategies which have the least distortion (first category in table 5.3) are known to suffer from chaining effect. The centroid method suffers also from reversing effect (Webster, 1977, pp.165-167). At lower levels of the hierarchy all dendrograms tend to be similar. Therefore, it may be more advantageous to use a strategy which is able to produce clearly defined clusters, especially, when the purpose of the strategy is to obtain an initial split of a population to be improved by subsequent steps of reallocation by a suitable method. Certain similarities in the original space can be ignored when they fall in the lower ranges of the similarity values. Therefore, the distortion due to a classificatory strategy should not always be a weakness.

Ward's ESS method and the complete linkage method have produced well defined clusters compared to the other methods. These two methods also show a certain degree of similarity as indicated by the highest correlation for these methods and also the two dendrograms (Fig. 5.1, b and f) are more similar to each other than to the rest at lower levels of the hierarchy. Ward's ESS method may have an added advantage of being the only agglomerative strategy which proceeds through the fusion of individuals by minimizing the within group variance.

It is possible to argue that the intensely clustering strategies may impose artificial structures on homogeneous populations. This probably can be investigated by using a strategy like the average linkage method which can be considered to suffer less from chaining than the single linkage method. The conclusion that can be drawn from the present study is that goodness-of-fit should not be considered as a fundamental criterion in the choice of agglomerative clustering strategies. The final objective of the taxonomist should be to partition a given population into homogeneous groups.

CHAPTER 6

CLASSIFICATION OF FORTY ONE SOIL PROFILES BY NUMERICAL TAXONOMIC METHODS

6.0 Data

This analysis is based on forty one soil profiles, of which ten properties (table 6.1) have been determined (Appendix 1). The soil profiles 1-32 are from the data published by USDA (1975) and the rest are from Sri Lanka (de Alwis 1971). The soil profile is a vertical cross-section through the soil and the soil properties have been determined at a series of depth levels (6 or 7), which are representative of genetic soil horizons. In this analysis, the soil horizon nomenclature was ignored, since the identification of soil horizons was done subjectively.

A series of soil profile models was compared in chapter 5 and the problems of soil description were discussed. Here, soils are described in two ways.

(1) Depth levels as arrays of independent attributes (soil profile model 1, chapter 4). Each soil profile is represented by an attribute vector.

(2) The distance (similarity) between pairs of soil profiles (individuals) is defined as the distance between the centroids of the soil profiles. The soil horizons were treated as members of the soil profiles and the soil profiles were considered as primary groups.

It has been possible to use the data from the two sources indicated as the laboratory determinations of all properties have been done using standard methods used by Soil Survey Staff (1951), but some properties have been determined by more than one method.

TABLE 6.1 Soil Attributes and their Code Numbers as used in this Analysis

Property	Attribute Code Numbers						
Percentage silt content	1	11	21	31	41	51	61
Percentage clay content	2	12	22	32	42	52	62
Percentage organic carbon	3	13	23	33	43	53	63
Diothionite extractable Fe(pct)	4	14	24	34	44	54	64
Exchangeable Ca me/100g soil	5	15	25	35	45	55	65
Exchangeable Mg me/100g soil	6	16	26	36	46	56	66
Exchangeable Na me/100g soil	7	17	27	37	47	57	67
Exchangeable K me/100g soil	8	18	28	38	48	58	68
pH(1:1 soil/water)	9	19	29	39	49	59	69
Cation exchange capacity (C.E.C.)	10	20	30	40	50	60	70

6.1 Methods

Three classificatory strategies were used to classify forty one soil profiles.

(1) Classification by agglomerative methods (average linkage method and Ward's method) using the Euclidean distance as the similarity measure.

(2) Classification by a divisive strategy, REMUL (TAXON computer package, CSIRO, Canberra, Australia). The procedure has been described by Lance and Williams (1975).

(3) Classification by agglomerative methods as before (1) but the similarity between soil profiles was determined by Mahalanobis D^2 (chapter 2).

It has been demonstrated that the agglomerative cluster analysis strategies produce comparable results (chapter 5), more important is the choice of the similarity measure. The methods 1 and 3 are different only in the choice of the similarity measure, and the method 2 is a clustering strategy, which begins the hierarchical division at the top of the hierarchy, but it must be noted here that the reallocation after each split contributes to the break down of the hierarchy; the properties of the similarity measure used in reallocation were discussed in chapter 2.

The classifications obtained by the three methods were evaluated by several multivariate statistical tests and canonical vector analysis described in chapter 2.

The main objective of the classification strategy was to produce a numerically optimum classification. The test used to determine the optimality of classifications was Wilk's Λ (section 2.7, chapter 2). Canonical vector analysis was performed on the population before and after the classification, the former of which was done by treating soil profiles as 'primary groups' with their depth levels as group members. In both cases scatter plots were drawn using the first two canonical vectors. The contribution of soil properties to the first two canonical vectors was illustrated by a vector diagram.

The classifications produced by hierarchical agglomerative classificatory strategies were illustrated by dendrograms.

6.2 Results

The forty one soil profiles were classified by the three strategies described. The average linkage method with the Euclidean distance as the similarity measure (Fig. 6.1a) did not produce clear clusters; by contrast clusters produced by the Ward's ESS method (Fig. 6.1b) were clearly separated. For the purpose of comparison, eight groups were obtained from Figure 6.1b, the members of the eight groups are listed in table 6.2.

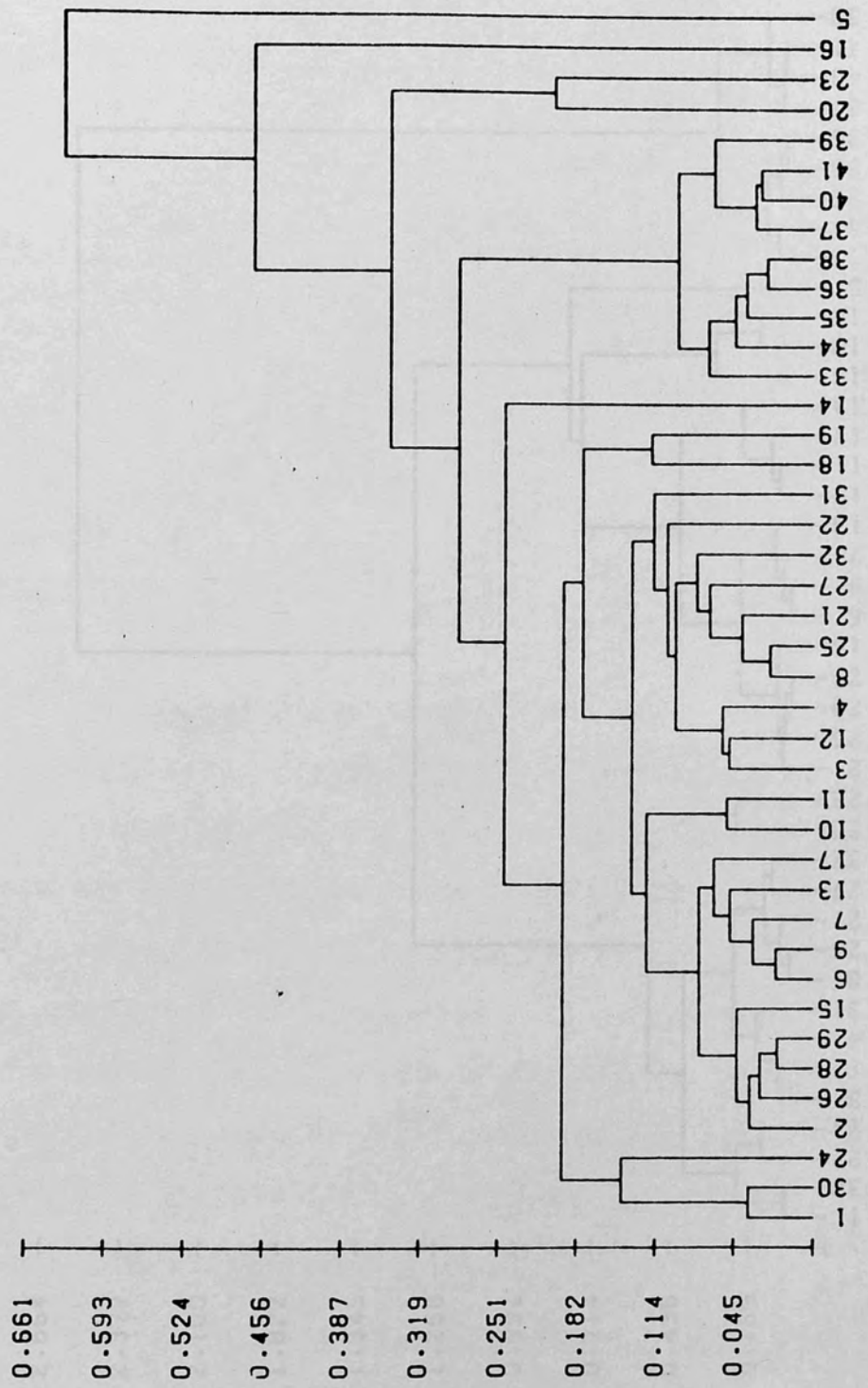


Fig. 6.1a Classification of forty one soil profiles by average linkage method with squared Euclidean distance as the similarity measure

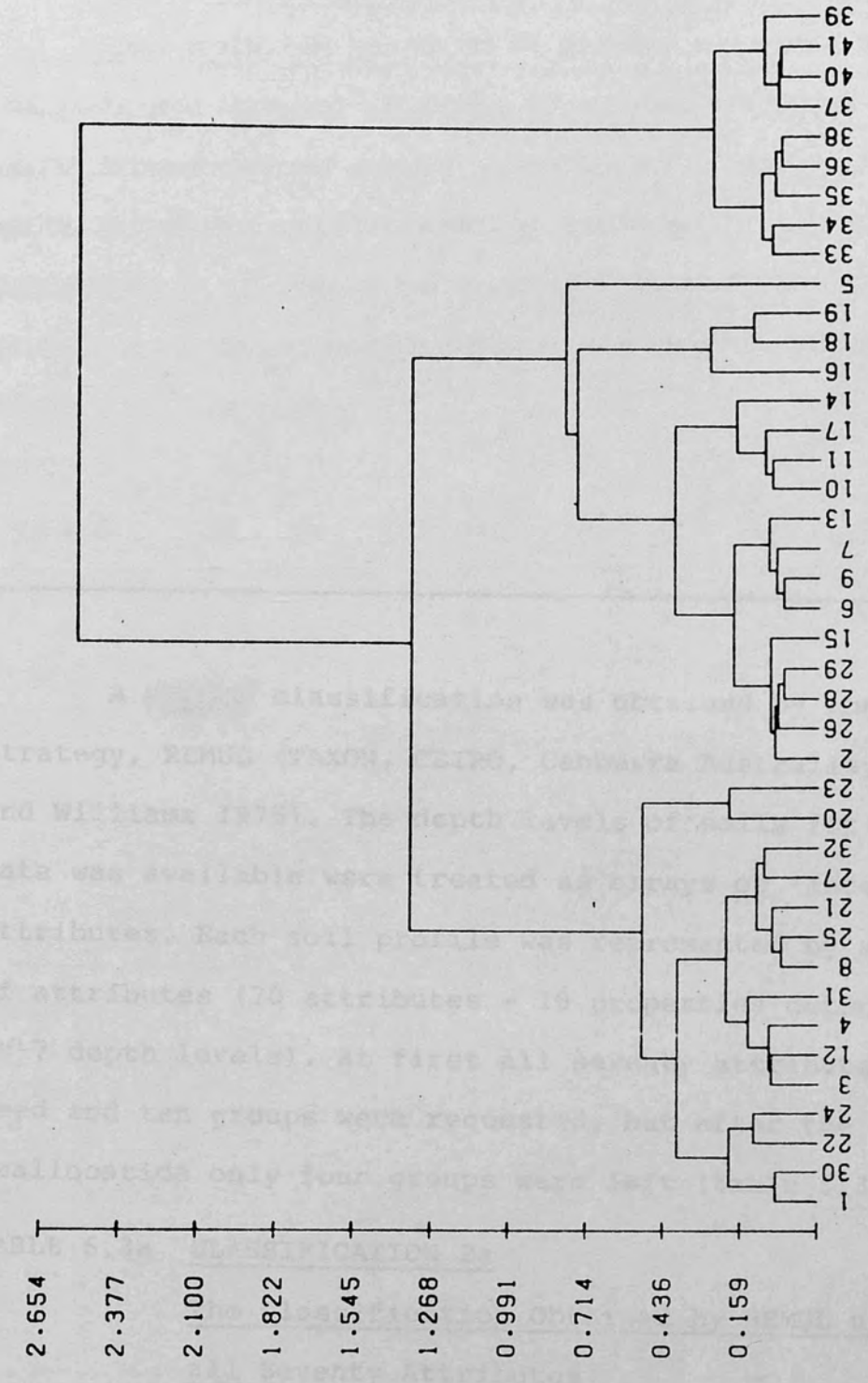


Fig. 6.1b Classification of forty one soil profiles by Ward's ESS method with squared Euclidean distance as the similarity measure

Table 6.2 may be usefully compared with Table 4.1. Both classifications were obtained from the same numerical strategy. However, the former was based on 30 attributes (3-horizon model) whereas the latter was based on 60 attributes (original measurements for six soil horizons). The difference between the two classifications may be in part due to the effect of inter-attribute correlation on the relative similarity between individuals and in part due to the sensitivity of the clustering strategy even to minor distortions of the similarity matrix.

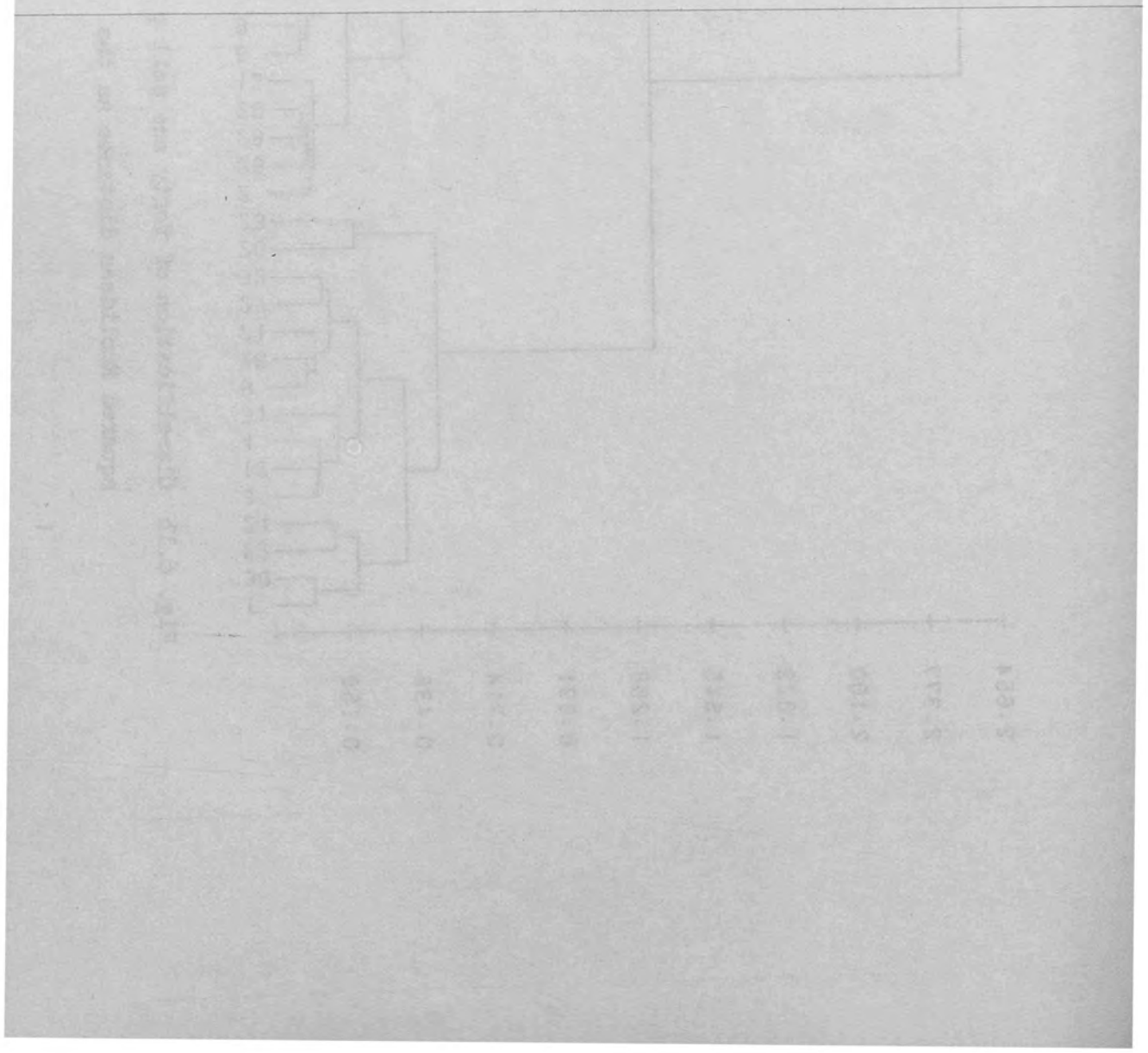


TABLE 6.2 CLASSIFICATION 1

Classification by Ward's ESS Method with the Euclidean Distance as the Similarity Measure

GROUP 1	1	22	24	30					
GROUP 2	3	4	8	12	21	25	27	31	32
GROUP 3	20	23							
GROUP 4	2	6	7	9	13	15	26	28	29
GROUP 5	10	11	14	17					
GROUP 6	16	18	19						
GROUP 7	5								
GROUP 8	33	34	35	36	37	38	39	40	41

A classification was obtained by the divisive strategy, REMUL (TAXON, CSIRO, Canberra Australia, Lance and Williams 1975). The depth levels of soils for which data was available were treated as arrays of 'independent' attributes. Each soil profile was represented by a vector of attributes (70 attributes - 10 properties determined at 7 depth levels). At first all seventy attributes were used and ten groups were requested, but after the final reallocation only four groups were left (table 6.3a).

TABLE 6.3a CLASSIFICATION 2a

The Classification Obtained by REMUL using all Seventy Attributes

GROUP 1	1	2	3	15	26	28	29	30	31	32
GROUP 2	33	34	35	36	37	38	39	40	41	
GROUP 3	6	7	9	17	18	19				
GROUP 4	4	8	12	13	20	21	22	23	24	25	26

It was demonstrated in chapter 4 that the seventy attributes used were depthwise correlated, and therefore, the effect of such correlations on the metric used in reallocation may be such that a large number of redundant attributes could distort the relative relationships between individuals. A sub set of attributes was selected to represent all attribute groups indicated by the dendrogram of attribute classification on the basis of inter-attribute correlation (Fig. 4.5 - Chapter 4). The case numbers of the attributes excluded from a second analysis were 1 11 14 21 24 27 29 31 33 34 38 39 41 45 46 47 52 54 55 58 61 62 64 67 68 71 73 74 76. As before ten groups were requested and after the final global reallocation the number of groups was reduced to eight, the composition of which is listed in table 6.3b.

TABLE 6.3b CLASSIFICATION 2b

Classification Obtained by REMUL After Masking Thirty Attributes

GROUP 1	1	3	26	29	30				
GROUP 2	2	6	7	9	15	19			
GROUP 3	5	10	11	14	16	17			
GROUP 4	8	13	25						
GROUP 5	22	23	24	32					
GROUP 6	4	12	27						
GROUP 7	18	20	21						
GROUP 8	33	34	35	36	37	38	39	40	41

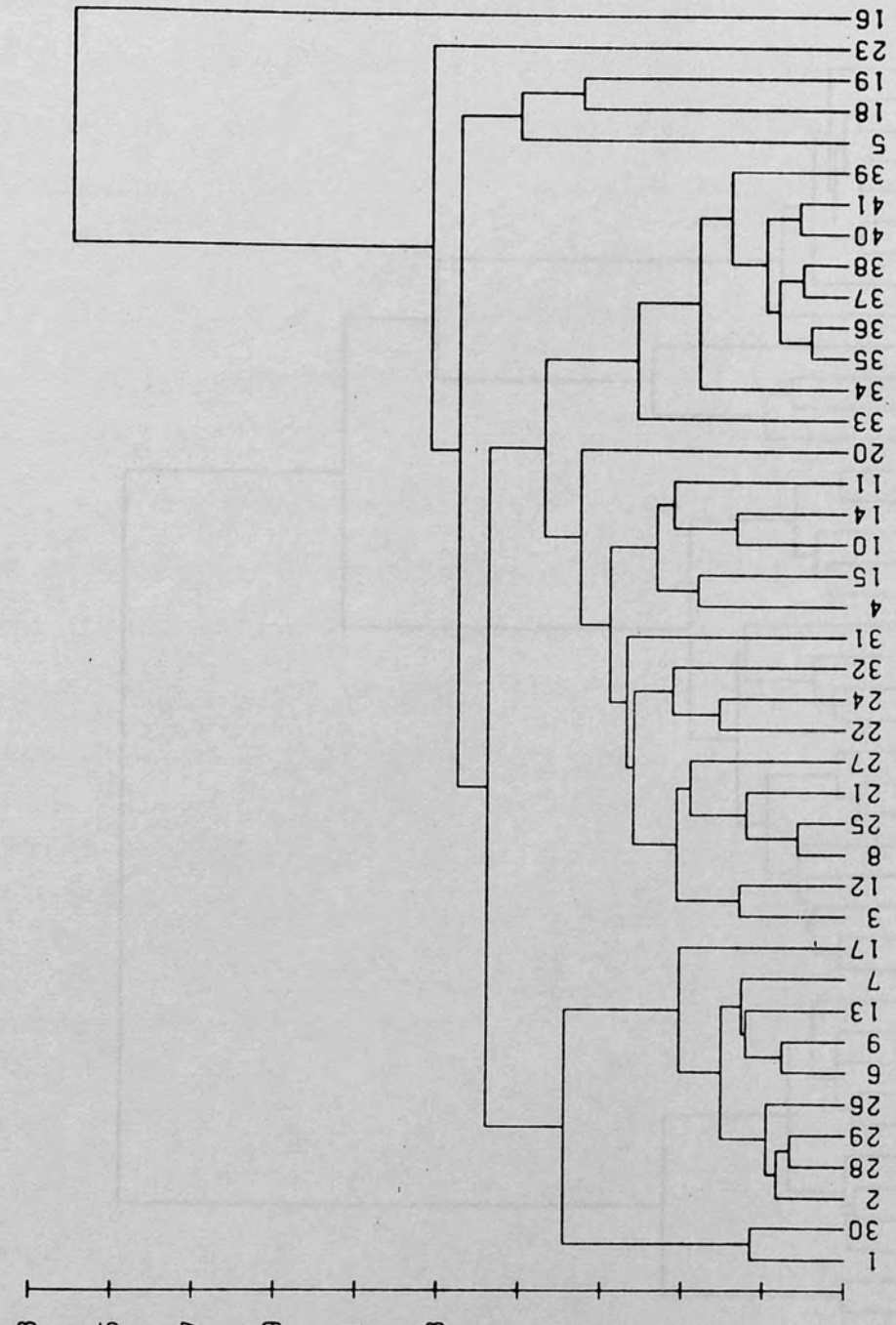


Fig. 6.2a Classification of forty one soil profiles by average linkage method with Mahalanobis distance as the similarity measure

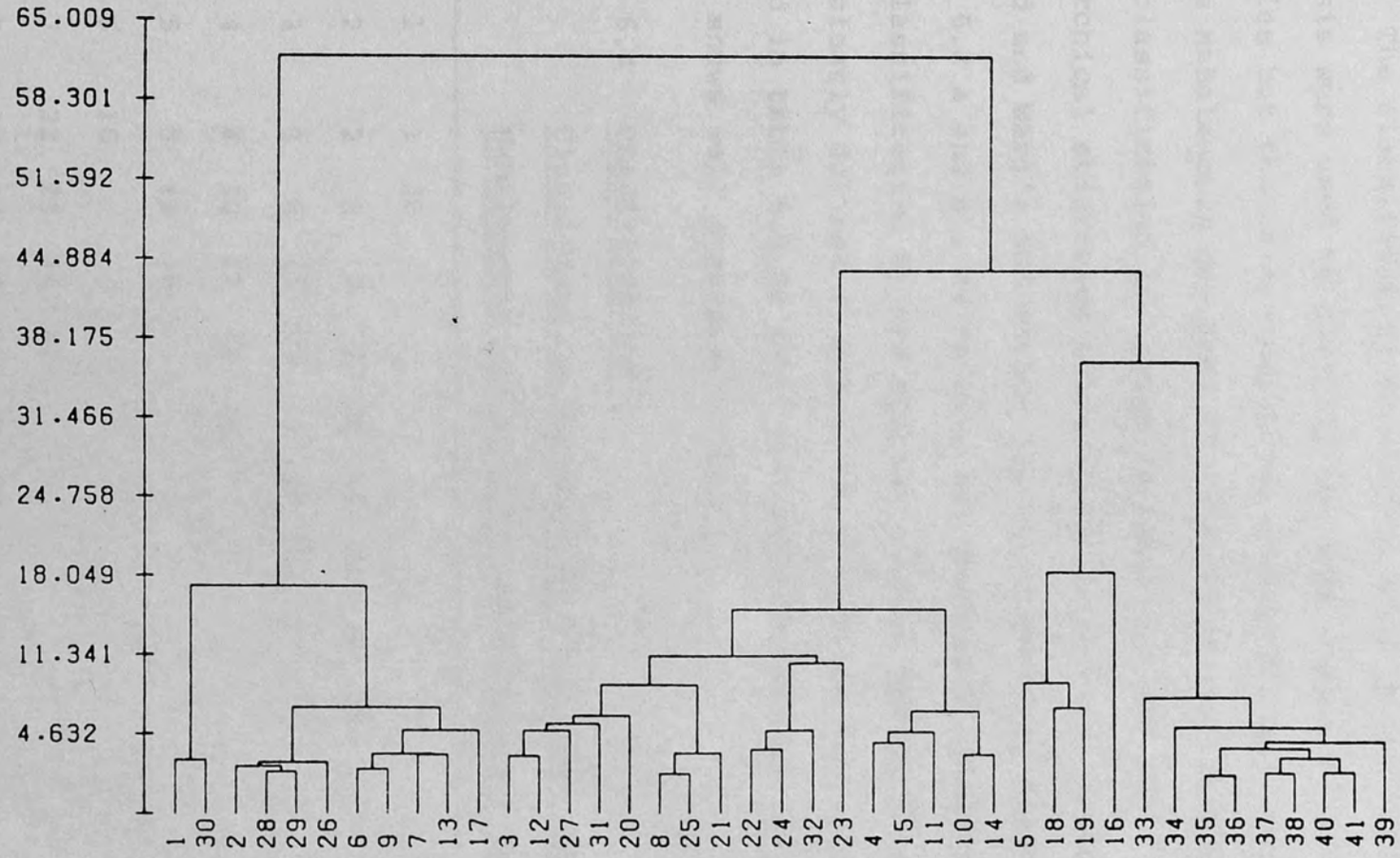


Fig. 6.2b Classification of forty one soil profiles by Ward's ESS method with Mahalanobis distance as the similarity measure

The effect of masking thirty attributes can be seen in an increase of the number of groups, and the composition of groups has also changed.

The classificatory strategies used in the first analysis were used to classify the same sample of soil profiles but the inter-individual similarity was measured by the Mahalanobis distance D^2, the group composition of this classification is listed in table 6.4 and the hierarchical structures produced by the average linkage method and Ward's ESS method are illustrated by dendrograms (Fig. 6.2 a and b). As before, the dendrogram produced for the classification by the average linkage method does not show clearly defined clusters, therefore the classification listed in table 6.4 is that of Ward's ESS method (Fig.6.2b) which shows well separated clusters.

TABLE 6.4 CLASSIFICATION 3

Classification Obtained by Ward's ESS Method

Mahalanobis Distance as the Similarity Measure

GROUP 1	1	30							
GROUP 2	2	6	7	9	13	17	26	28	29
GROUP 3	3	8	12	20	21	25	27	31	
GROUP 4	4	10	11	14	15				
GROUP 5	5	18	19						
GROUP 6	16								
GROUP 7	22	23	24	32					
GROUP 8	33	34	35	36	37	38	39	40	41

All classifications have identified the Red Latosols group (Group 8) and also the following combinations,

(a) 1 30

(b) 2 26 28 29

(c) 6 7 9 13 17

which remained together with the exception of profile 28 which separated from its most similar member, profile 29 in the second classification (table 6.3b).

Although the three classifications are not very similar to each other, they all have the Red Latosols group (Group 8) separated from the rest of the population. It can be seen from the classifications that the most similar soils tend to stay together, whereas the intermediate types can be variously classified depending on the classificatory strategy and the similarity measure used. The group nuclei can be easily identified from the dendrograms. The most similar soils in the population fused together at the initial stages of fusion and as the fusion proceeded groups grew but the probability of misclassifying individuals becomes very high.

The main objective of the taxonomic classification is to produce numerically optimum classifications and the statistical tests and other numerical criteria may be used to evaluate classifications. It can be seen from table 6.5 that the classification obtained by Ward's ESS method with the Mahalanobis D^2 as the similarity measure has attained the lowest value for Wilk's Criterion Λ, indicating greater between groups variance.

TABLE 6.5 Wilk's Criterion Λ Values, and χ^2 Values for Classifications Obtained by Numerical Strategies

Classification	Wilk's Λ	Chi-Square	DF
1a	0.00239	1647	70
1b	0.00084	1935	70
2	0.00842	1318	70
3	0.00039	2150	70

Classification 1a - Classification by average linkage method with Euclidean distance as the similarity measure.

1b - Classification by Ward's ESS method with Euclidean distance as the similarity measure.

2 - Classification by REMUL with thirty attributes masked.

3 - Classification by Ward's ESS method with Mahalanobis distance D^2 as the similarity measure.

Number of groups = 8 for all classifications

It can be seen from table 6.5 that Ward's ESS method produce more homogeneous groups than the average linkage method as judged by Wilk's Λ, despite the greater distortion.

Webster (1971, 1976) suggests that ΛG^2 can be used to determine the number of groups in the population. ΛG^2 can be plotted against the number of groups, the curve drops at or near the optimum number of groups as has been demonstrated by McBratney and Webster (1981). This method was used here as an exploratory tool. It can be seen from Figure 6.3 that the curve drops sharply when the number of groups increases from three to four and also a further sharp drop occurs as the number of groups increases upto eight and from there on flattening of the curve occurs. At this stage the number of groups in the population was treated as eight.

The three classifications can be compared graphically by plotting the group centroids using canonical vectors. The mean canonical points for the soil profiles were also plotted using the first two canonical vectors. A close relationship exists between the dendrogram drawn for the classification by Ward's ESS method (Fig. 6.1b) and the canonical plot (Fig. 6.4a). The first two canonical vectors account for 71 percent of the variance. Those soil profiles which are very similar to each other (as indicated by all classifications) are closer in the two dimensional space. The well defined clusters of the classification three can be identified from the canonical plot (Fig. 6.4), whereas outlying individuals are separated from the clusters (Fig. 6.4b). The canonical diagrams produced for the classifications (Fig. 6.5, a-c) also indicate the superiority of the classification three. The

group centroids of the classification three are better separated than the others (Fig. 6.5c).

Finally, canonical analysis was performed on the eight groups obtained by Ward's ESS method with the Mahalanobis distance as the similarity measure using the first three depth levels (thirty attributes) and the population was projected onto a two dimensional space using the first two canonical vectors. It can be seen from Figure 6.6 that the eight groups are not well separated but they occupy clearly defined areas and some groups are more separated than the others.

Reallocation of the classification three was attempted on the basis of the Mahalanobis distance between individuals and group centroids (D^2_{ik}) (table 6.7), which was calculated using thirty attributes (ten soil properties for the uppermost three horizons). D^2_{ik} is distributed approximately as χ^2 with p (number of attributes) degrees of freedom (Webster, 1977, p.207-208). When an individual is located considerably far away from all group centroids that individual can be treated as a separate group. However, in this study all individuals have low D^2_{ik} values for individuals and their parent groups and very high values for the other groups as seen in table 6.7. Canonical analysis on the classification (Fig. 6.6) shows that all eight groups can be identified easily though some of them are closer to each other than the others, but they occupy distinct regions in the canonical space. The groups 1, 2, 5, 6, 7 and 8 are well separated but there is a considerable overlap between groups 3 and 4. However, these two groups

The relationship between ΛG^2 and G (number of groups)

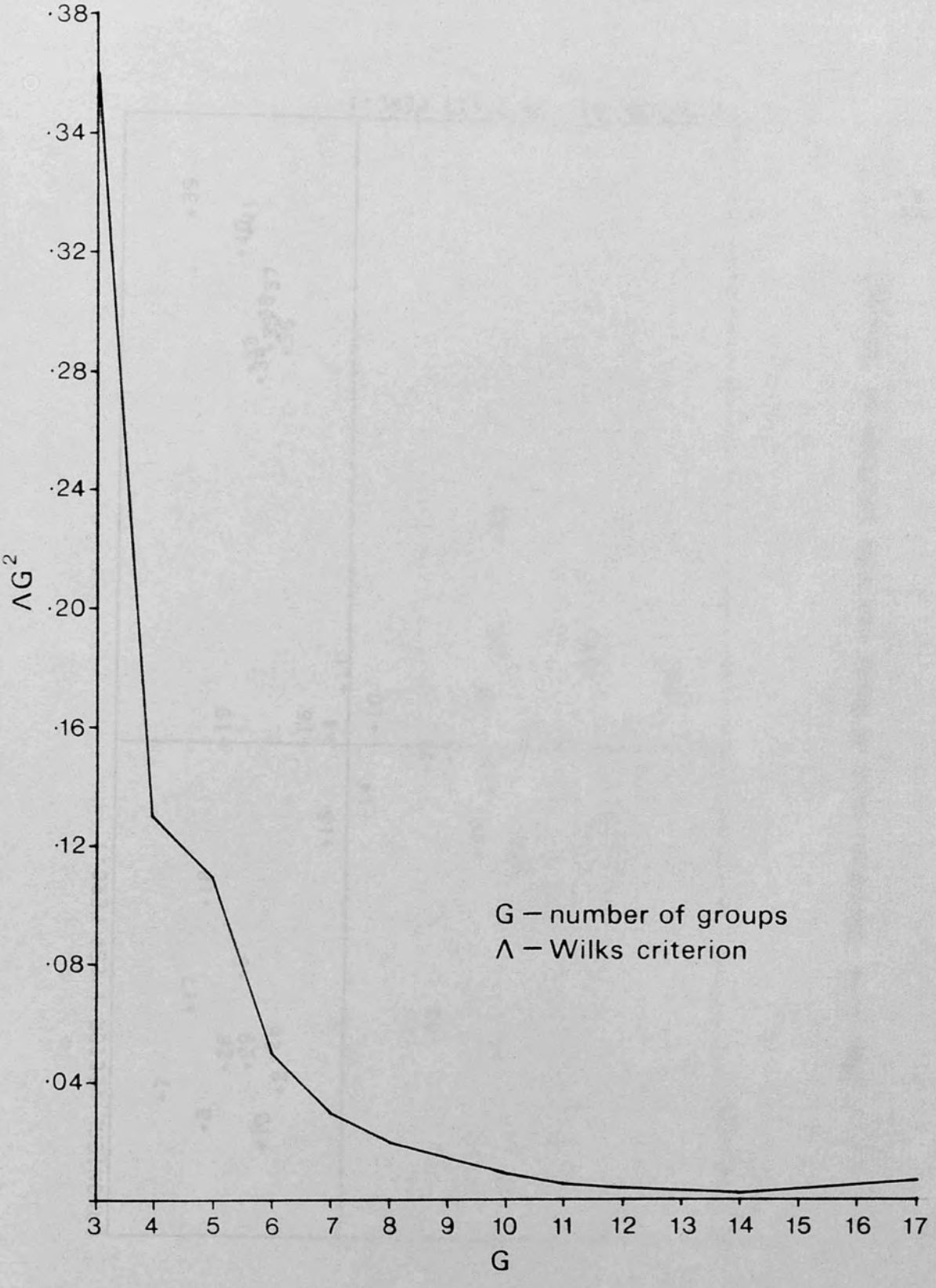


Figure 6·3

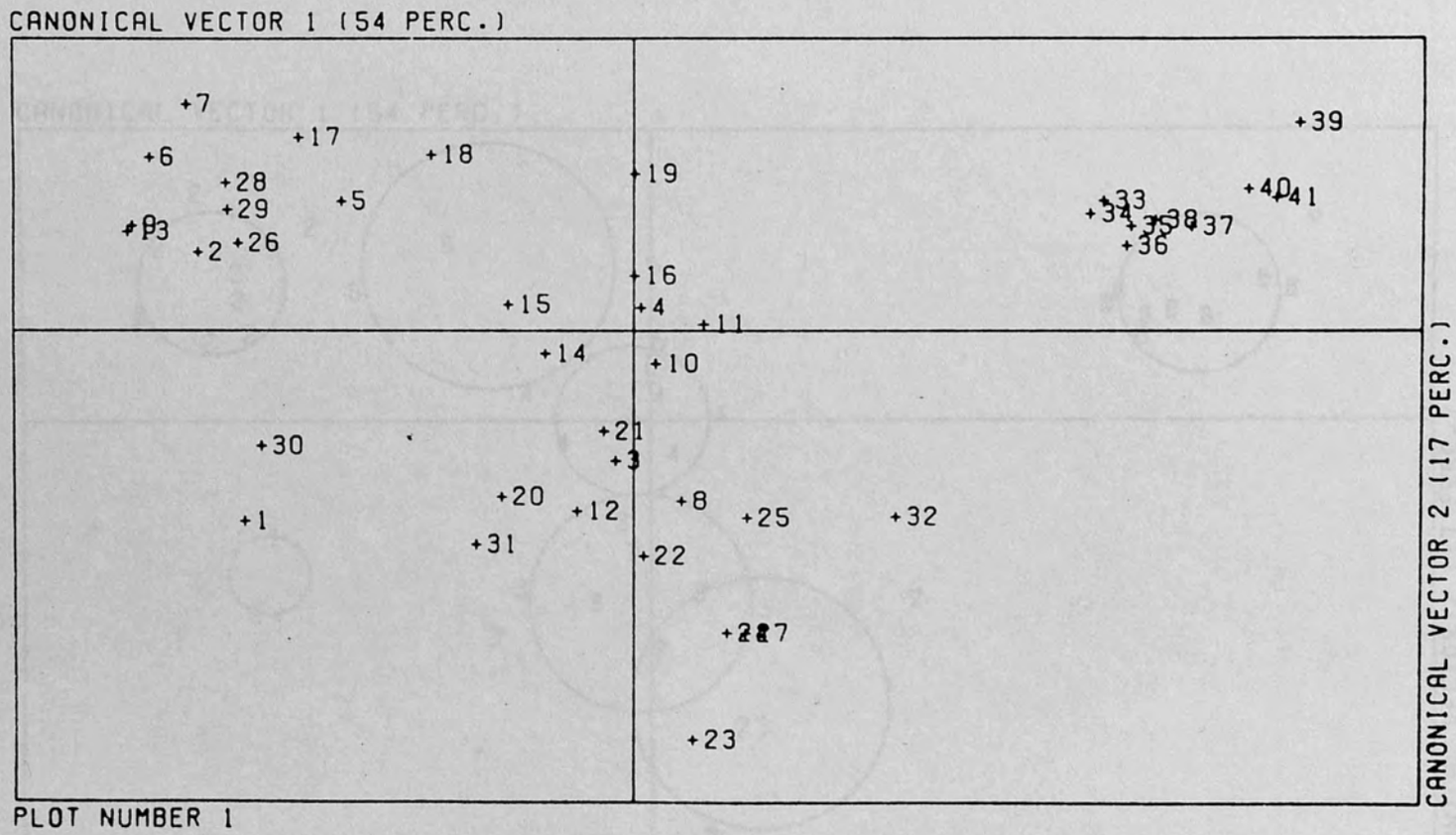


Fig. 6.4a Canonical plot of forty one soil profiles as 'groups'

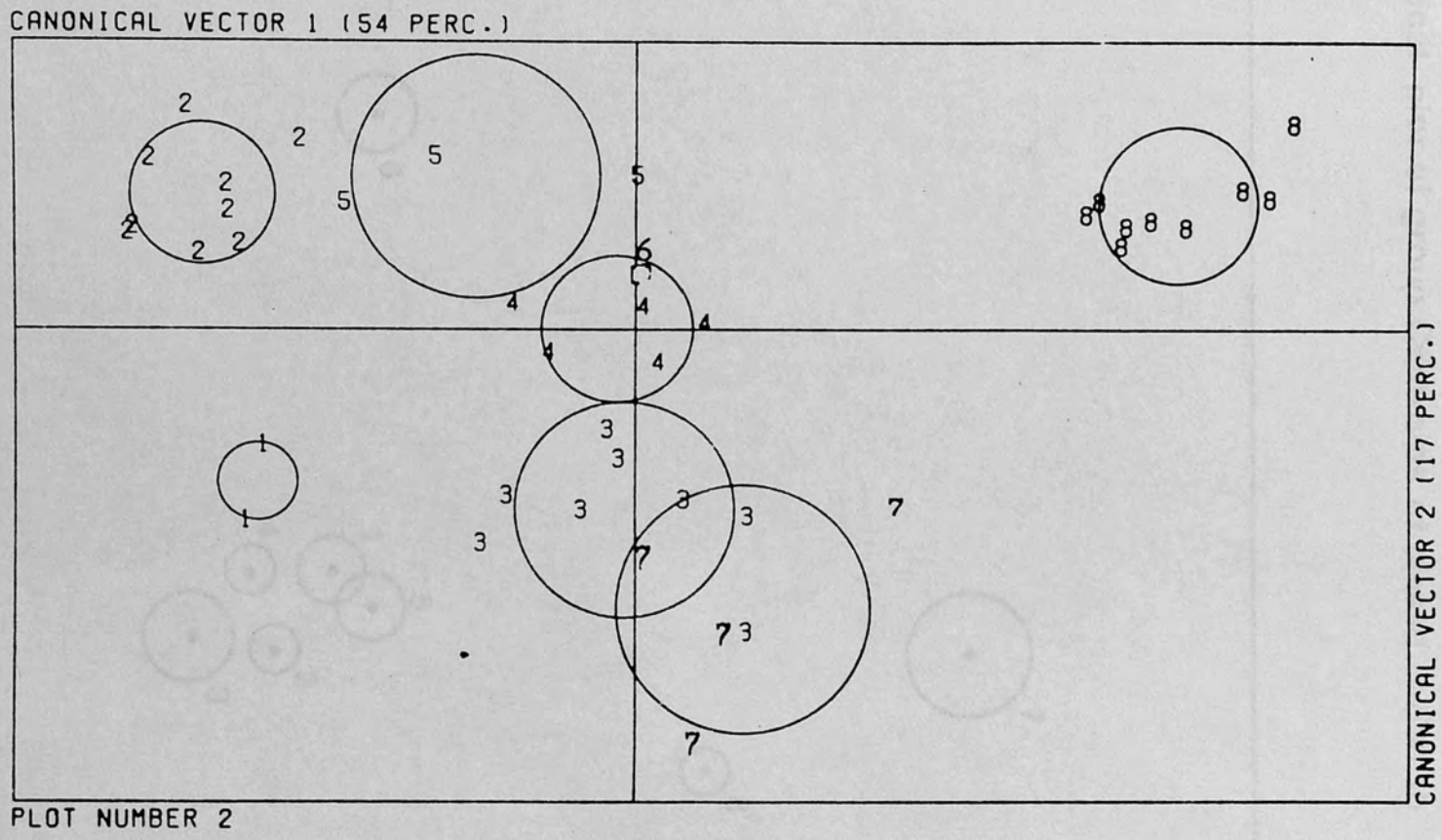


Fig. 6.4b Eight soil groups identified from the dendrogram produced by Ward's ESS method with Mahalanobis distance as the similarity measure

Canonical plot of group Centroids of classification 1b

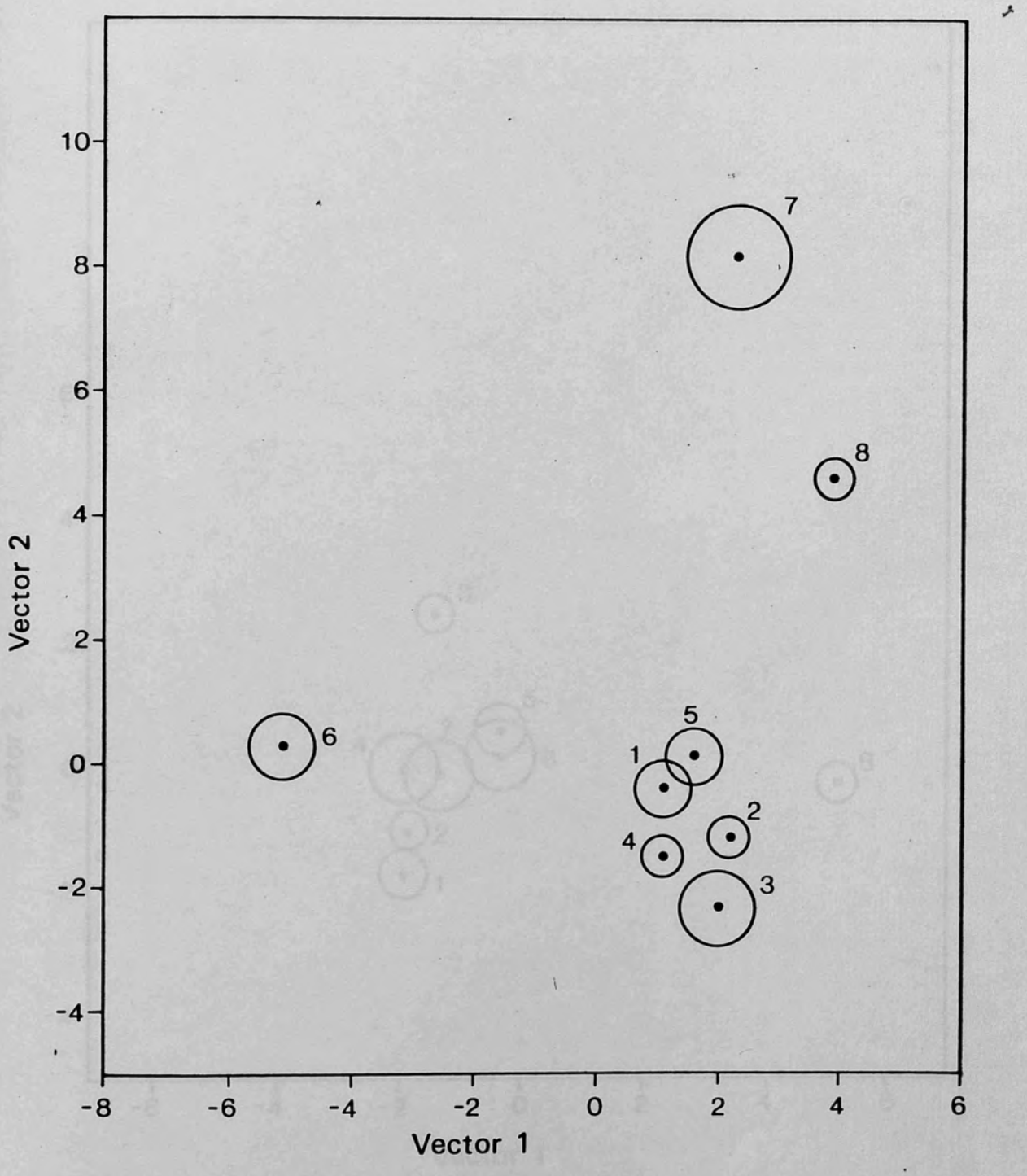


Figure 6·5a

Canonical plot of group Centroids of classification 2

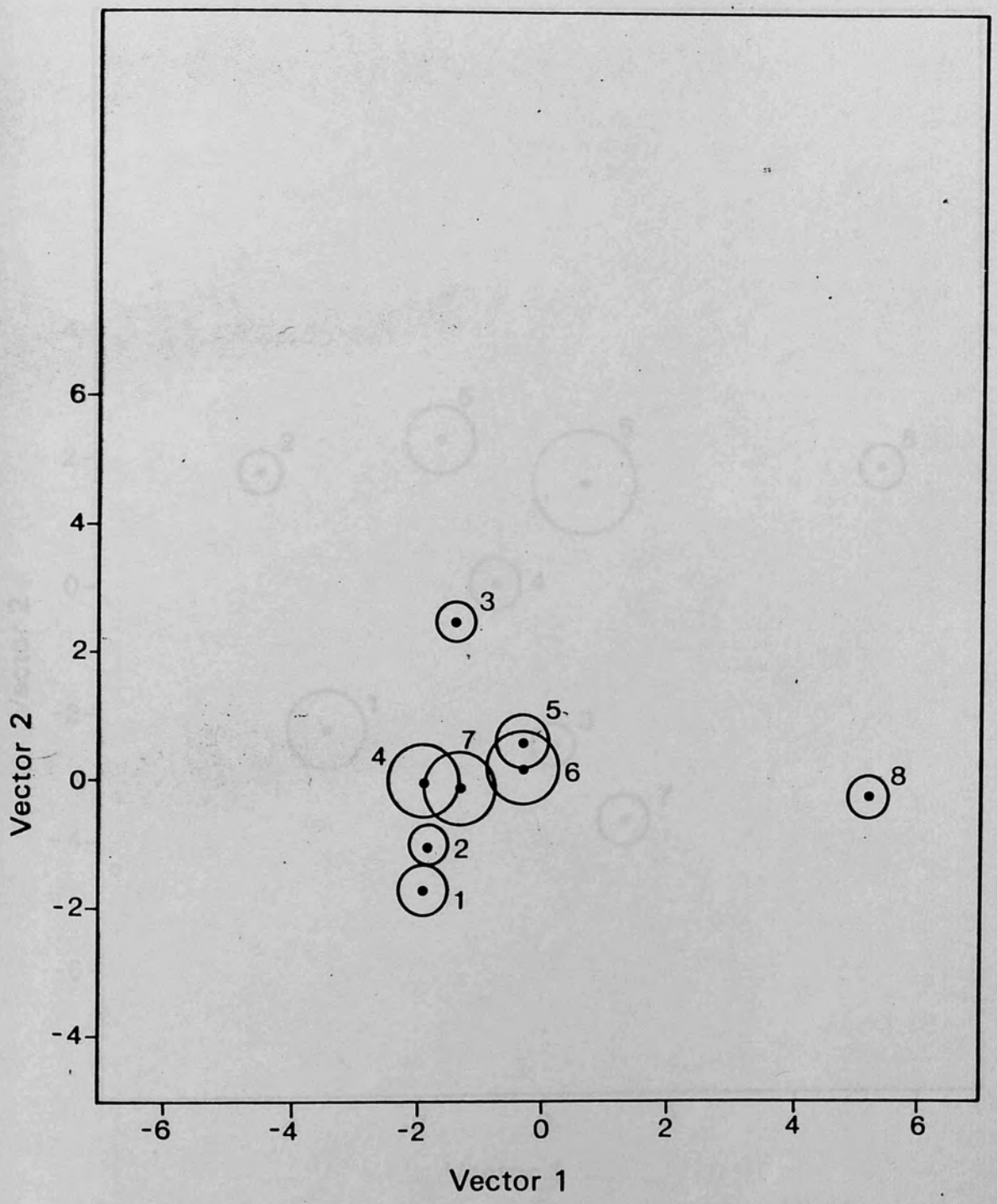


Figure 6·5b

Canonical plot of group Centroids of classification 3b

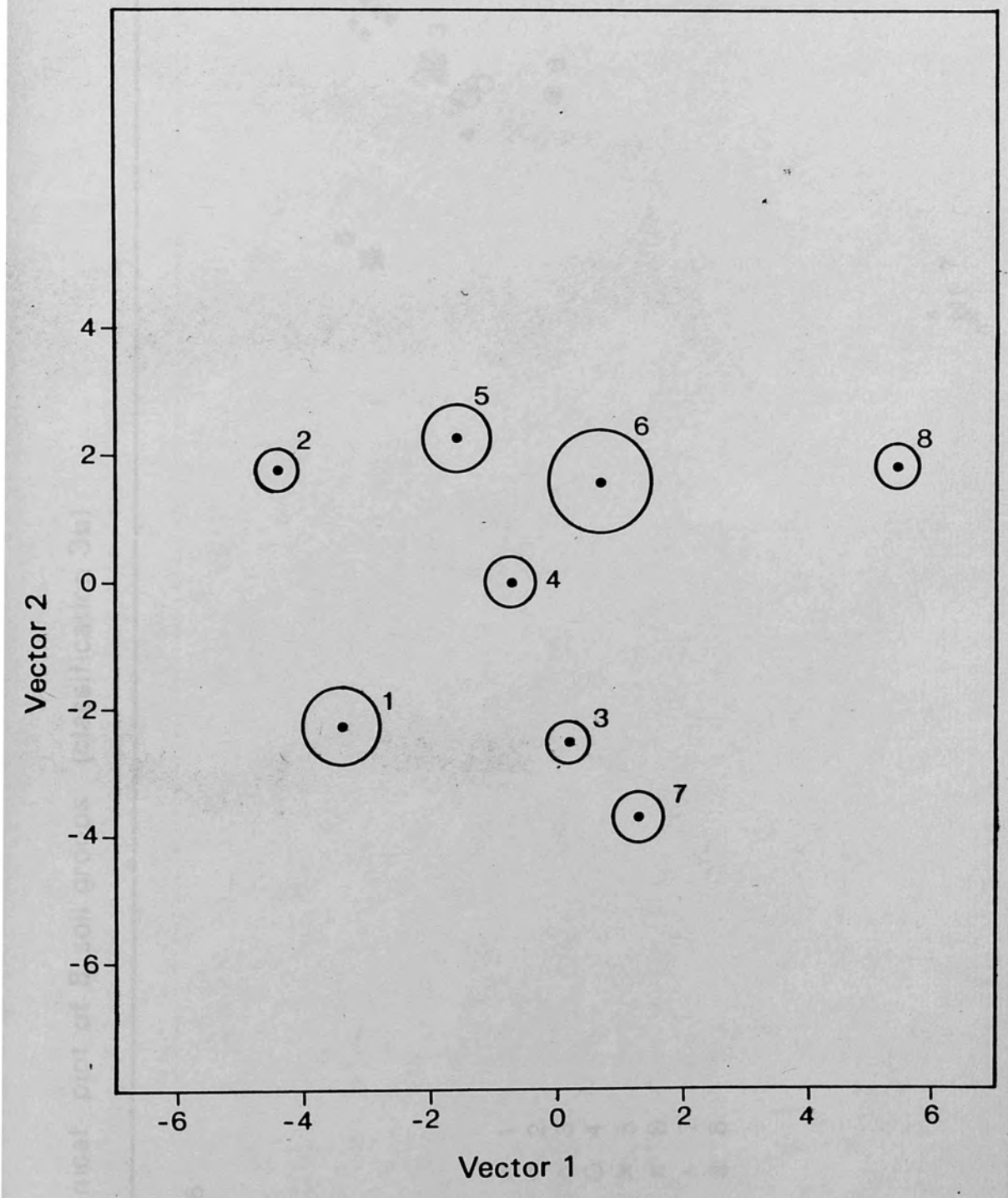


Figure 6·5c

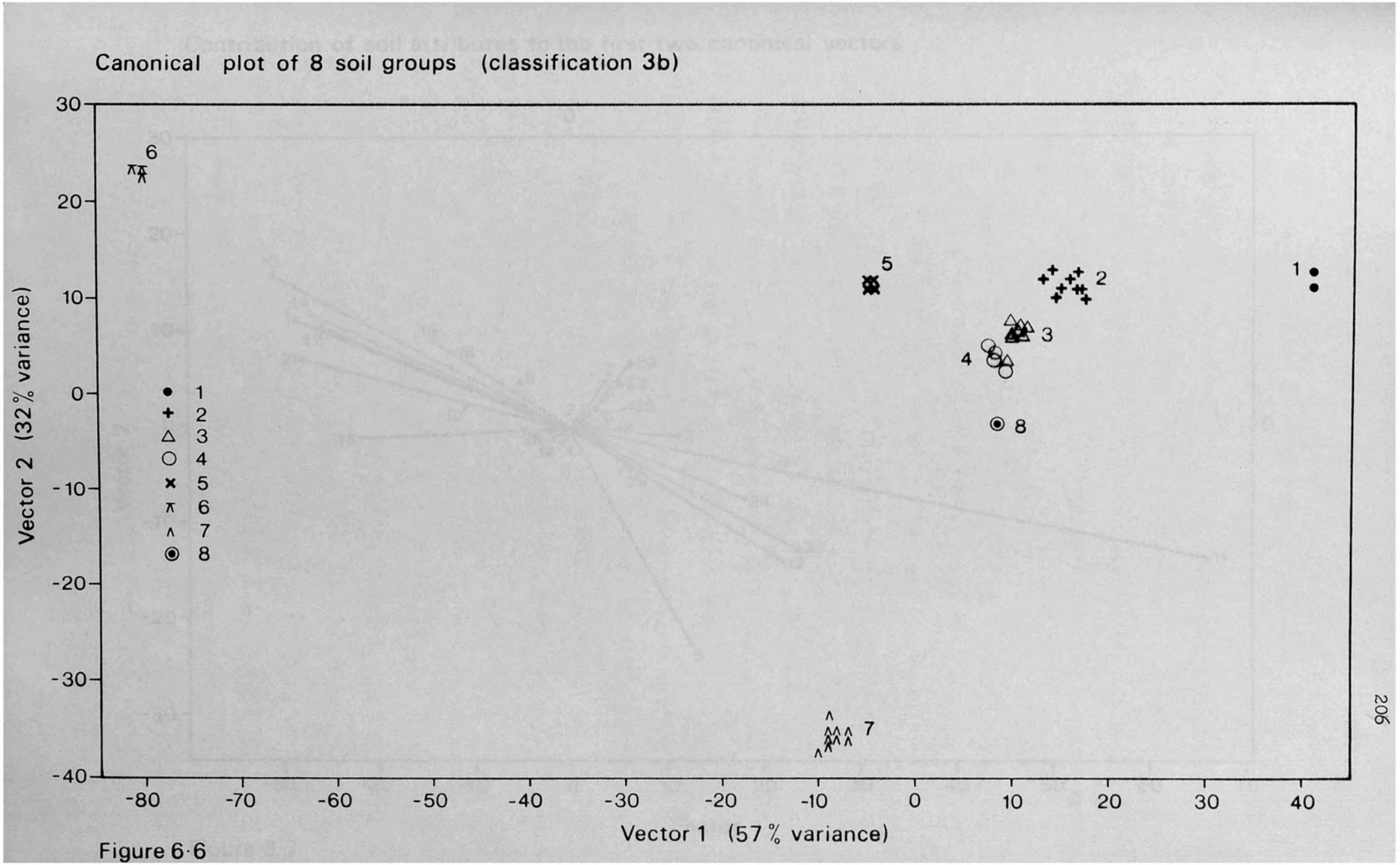

Canonical plot of 8 soil groups (classification 3b)
Vector 2 (32% variance)
Vector 1 (57% variance)
30
20
10
0
-10
-20
-30
-40
-80
-70
-60
-50
-40
-30
-20
-10
0
10
20
30
40
1
2
3
4
5
6
7
8

Figure 6·6

Contribution of soil attributes to the first two canonical vectors

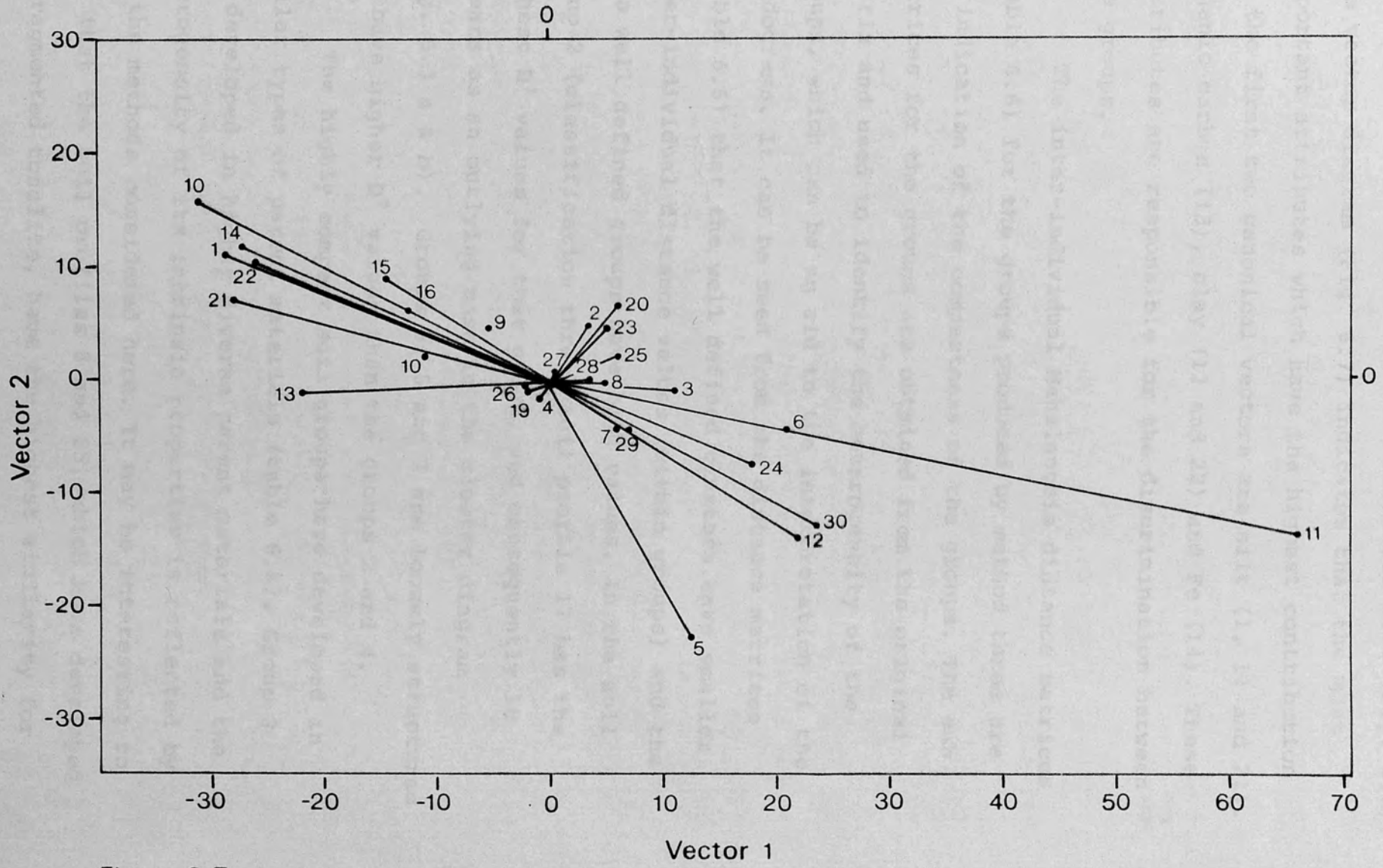


Figure 6·7

can be easily identified from the canonical plot (Fig. 6.6). The vector diagram (Fig. 6.7) indicates that the most important attributes which have the highest contribution to the first two canonical vectors are silt (1, 11 and 21). Organic carbon (13), clay (12 and 22) and Fe (14). These attributes are responsible for the discrimination between the groups.

The inter-individual Mahalanobis distance matrices (table 6.6) for the groups produced by method three are an indication of the compactness of the groups. The sub matrices for the groups are obtained from the original matrix and used to identify the heterogeneity of the groups, which can be an aid to the interpretation of the dendograms. It can be seen from the distance matrices (table 6.6) that the well defined clusters have smaller inter-individual distance values (within groups) and the less well defined groups have high values. In the soil group 2 (classification three) soil profile 17 has the highest D^2 values for that group, and consequently it appears as an outlying case in the cluster diagram (Fig. 6.3 a & b). Groups 3, 5 and 7 are loosely structured and have higher D^2 values than the groups 2 and 4.

The highly compact soil groups have developed in similar types of parent materials (table 6.8). Group 3 has developed in highly diverse parent materials and the heterogeneity of its intrinsic properties is reflected by all the methods considered here. It may be interesting to note that the soil profiles 8 and 25, which have developed in fragmented tonalite, have the highest similarity for group 3. Group 2 is mainly developed in Loess except soil

* This is probably because Aridisols are defined on the basis of present day climatic conditions rather than pedological considerations. Climatic data were not incorporated in the data set used here.

profile 17, and the soil profiles 26, 28 and 29 form a sub-group in group 2 and these soils have developed in Wisconsin Loess. Paradoxically, soil group 7 has developed in the same type of parent materials and also belongs to the same Order (Aridisols) and Sub Order (Argids) though it has very high D^2 values.*

It has been demonstrated in the previous section that the classification 3 is the numerically most optimum, and it is possible to examine the relation between the classification 3 and the U.S. Taxonomic system. Table 6.8 shows the relationship between the two classifications. The sample of soil profiles considered here belong to 7 Orders and 28 Sub-orders. Soil Orders have been split into different groups and different Orders have combined to form single groups. Therefore the two classifications are different to a great extent but some similarities can also be seen. The most similar soils in the sample (profiles 6, 9 and profiles 28 and 29) belong to the same Order as well as the same Sub-order. Therefore the identification of the intermediate individuals is the main problem to the taxonomist. The soil group 8 belongs to the Latosols Sub-order and Oxisols Order according to the USDA classification, and remains as a single group in the numerical classifications. The soil groups 1 and 7 also belong to single Orders.

TABLE 6.6

(a) Inter-individual D^2 matrix for group 1

	1
30	2.455

(b) Inter-individual D^2 matrix for group 2

	2	6	7	9	17	26	28	29
6	3.631							
7	3.813	2.254						
9	2.726	1.661	3.089					
13	2.662	2.747	3.025	2.439				
17	5.176	3.632	3.520	4.602	4.861			
26	2.167	4.151	4.254	3.795	3.993	4.715		
28	1.768	3.108	2.991	3.070	2.730	4.157	2.113	
29	1.835	2.829	3.020	2.583	3.064	4.327	1.867	1.430

(c) Inter-individual D^2 matrix for group 3

	3	8	12	20	21	25	27
8	4.234						
12	2.792	4.868					
20	5.202	6.205	3.975				
21	3.147	2.187	3.610	5.214			
25	4.945	1.334	5.643	7.246	3.122		
27	4.497	3.668	4.365	6.098	4.575	3.816	
31	4.730	6.209	5.033	6.896	5.732	7.098	5.756

(d) Inter-individual D^2 matrix for group 4

	4	10	11	14
10	5.600			
11	4.855	4.349		
14	4.443	2.932	4.717	
15	4.116	5.789	4.419	4.765

(e) Inter-individual D^2 matrix for group 5

	5	18
18	7.213	
19	9.990	6.955

(f) Inter-individual D^2 matrix for group 7

	22	23	24
23	8.226		
24	3.324	8.395	
32	4.867	9.046	4.964

(h) Inter-individual D^2 matrix for group 8

	33	34	35	36	37	38	39	40
34	5.768							
35	4.818	2.693						
36	5.018	3.162	0.970					
37	5.721	4.140	2.376	2.224				
38	5.499	3.703	1.581	1.373	1.257			
39	6.299	6.148	5.026	5.034	3.842	4.642		
40	5.876	5.098	3.706	3.470	2.345	3.048	1.486	

TABLE 6.7 - Mahalanobis Distance Between Individuals and Group Centroids

INDIVIDUAL	GROUP 1	GROUP 2	GROUP 3	GROUP 4	GROUP 5	GROUP 6	GROUP 7
Individual 1	4.o311	32.0708	37.9576	37.8754	49.3067	122.3668	69.0673
Individual 2	34.5695	5.1002	28.0855	33.0612	24.0711	96.4456	55.5591
Individual 3	40.2665	29.9516	5.2261	16.3987	23.7992	93.8810	46.6728
Individual 4	49.6629	25.0967	23.7495	26.8194	5.0444	77.7203	47.8381
Individual 5	123.6906	98.2888	94.5504	92.9857	77.6743	4.6872	93.5378
Individual 6	35.7742	5.2525	30.0747	33.0939	23.9768	95.4412	54.1838
Individual 7	32.1894	5.4111	30.0362	33.5486	27.0043	98.9184	56.0959
Individual 8	39.7733	30.4380	4.5792	17.3635	25.0313	95.4036	49.9023
Individual 9	31.0448	4.8239	26.7786	30.3370	25.5565	98.6128	53.9166
Individual 10	50.0505	25.6789	23.8985	27.3892	5.1170	77.7230	48.6017
Individual 11	50.3857	26.1431	23.6982	27.3058	5.1077	77.5563	48.1329
Individual 12	37.8124	28.1453	5.1836	16.3506	23.9605	94.9677	50.3382
Individual 13	31.0385	5.0788	29.6908	33.0664	27.5272	100.0861	54.5664
Individual 14	50.0785	25.7531	23.6686	27.3646	5.0739	77.2955	48.6156
Individual 15	50.1347	25.2388	23.0964	26.9289	5.0594	76.5983	48.5747
*Individual 16	151.0489	129.9077	116.0622	117.3657	108.3632	68.7984	107.1223
Individual 17	33.0007	5.3635	27.6234	31.1610	24.5912	97.1289	52.7543

* Individual 16 does not belong to any class p less than 0.01.

TABLE 6.7 (page 2) - Mahalanobis Distance Between Individuals and Group Centroids

Individual	GROUP 1	GROUP 2	GROUP 3	GROUP 4	GROUP 5	GROUP 6	GROUP 7
Individual 18	123.1935	97.7240	94.0240	92.4397	77.0521	4.6366	93.0960
Individual 19	123.3376	97.9643	94.1772	92.6278	77.3321	4.6516	93.0517
Individual 20	38.9720	28.5312	5.0666	16.1469	23.2733	93.8833	49.1067
Individual 21	38.9722	28.4993	5.1948	15.6612	23.2210	93.3621	49.9277
Individual 22	39.1931	32.3297	16.0811	4.8208	27.6648	93.6302	46.9597
Individual 23	39.3631	32.6565	16.2907	4.9201	27.6234	93.0421	47.5229
Individual 24	38.0695	30.0163	13.4755	4.9283	24.8907	91.9295	46.2627
Individual 25	38.6451	27.0173	4.9994	16.6163	22.3015	94.1081	48.7621
Individual 26	32.0708	4.8343	29.2568	32.0109	25.5659	97.4687	53.6942
Individual 27	39.5977	29.7832	4.9821	15.1714	24.2298	94.0091	49.1873
Individual 28	31.5276	5.0240	29.6101	32.9014	27.0911	99.6215	54.9069
Individual 29	31.3134	4.3209	27.2475	31.4823	24.5937	98.3072	55.0082
Individual 30	4.0311	32.7127	39.6836	40.0518	50.6195	124.3955	70.6326
Individual 31	37.4890	27.3156	5.2631	16.1833	23.0331	94.5433	49.7672
Individual 32	39.6140	33.7488	18.4241	4.4711	28.1385	92.1531	48.8184
Individual 33	69.3915	53.9060	43.6136	46.6508	47.6952	92.6261	5.4033
Individual 34	71.1073	55.6019	50.6230	48.9282	49.3417	93.3815	5.1875
Individual 35	68.3594	53.4019	47.5737	45.7783	47.5217	93.7161	5.3908
Individual 36	68.7813	53.1688	47.4985	45.6050	46.2118	91.7231	4.9741
Individual 37	69.2248	53.8317	49.5143	47.8605	48.7623	94.5729	4.7437
Individual 38	71.8767	56.2987	50.9462	48.9525	49.8645	93.0354	5.0467
Individual 39	70.5265	55.2567	49.6607	47.9650	48.8657	93.3840	5.3936
Individual 40	70.4131	55.0717	49.6487	47.9668	49.0110	93.8000	4.8116
Individual 41	69.5591	54.1491	48.8089	47.0548	47.8186	92.9984	4.8577

TABLE 6.8 Groups Produced by Numerical Classification their Parent Materials and Order and Sub Order of the USDA (1975) System

Group No.	Profile No.	Parent Materials	Soil Order	Sub Order
1	1	Loess	Mollisols	Ustolls
	30	Wisconsian Loess	Mollisols	Udolls
2	2	Loess	Mollisols	Udolls
	6	Deep Loess	Alfisols	Udalfs
	7	Loess on Loam	Alfisols	Aqualfs
	9	Loess	Alfisols	Udalfs
	13	Loess	Alfisols	Aqualfs
	17	Alluvium from Coastal Materials	Ultisols	Aquuls
	26	Wisconsian Loess	Mollisols	Aqualfs
	28	Wisconsian Loess	Mollisols	Udolls
	29	Wisconsian Loess	Mollisols	Udolls
3	3	Upland Glacial Till	Mollisols	Borolls
	8	Fragmented Tonalite	Alfisols	Xeralfs
	12	Calcareous Clay Loam	Aridisols	Argids
	20	Alluvium from Sed. rocks.	Alfisols	Xeralfs
	21	Alluvium from Sed. rocks	Alfisols	Xeralfs
	25	Fragmented Tonalite	Inceptisols	Ochrept
	27	Sed. from Alluv. Rocks	Mollisols	Borolls
	31	Loamy Alluvium	Mollisols	Ustolls

TABLE 6.8 (page 2)

Group No.	Profile No.	Parent Materials	Soil Order	Sub Order
4	4	Lacustrine materials	Mollisols	Borolls
	10	Sandy Deltaic Deposits	Alfisols	Udalfs
	11	- not given -	Ultisols	Udults
	14	Glacial Till primarily from Granite & Schist	Spodosols	Orthods
	15	Alluvium	Mollisols	Aquolls
5	5	Mixed Terrace Materials from Basic Rocks	Inceptisols	Andepts
	18	Coastal Alluvium	Alfisols	Aqualfs
	19	Loess	Alfisols	Udalfs
6	16	Chloritized Basaltic Andesite	Oxisols	Orthox
7	22	Alluvials	Aridisols	Argids
	23	Alluvials	Aridisols	Argids
	24	Alluvials	Aridisols	Argids
	32	Alluvials	Aridisols	Orthids
8	33	Beach Sand	Red Latosols	
	34	" "	Red Latosols	
	35	" "	Red Latosols	
	36	" "	Red Latosols	
	37	" "	Red Latosols	
	38	" "	Red Latosols	
	39	" "	Red Latosols	
	40	" "	Red Latosols	
	41	" "	Red Latosols	

6.3 Discussion

The US Taxonomic System of soil classification is hierarchical and the classification process begins at the highest level of the system. There are ten Orders, which are defined in terms of a few selected properties. The soils in a given Order can vary considerably in the properties which are not used in the definition of that Order. When an Order is defined sub-division is done internally,and the properties that are used for that purpose, do not necessarily have a meaning in other Orders. As the classification proceeds, lower categories are identified using an increasing number of properties, but at the same time the lower categories of different Orders can become closer. When numerical methods are used to classify soils, they are likely to fuse soils from different Orders as they are similar overall in terms of their measured properties.

The criteria used to separate Soil Orders have been described in Soil Taxonomy (USDA 1975,p.71) as "the presence or absence of diagnostic horizons or features, that are marks in the soil of differences in the degree and kind of the dominant sets of soil forming processes that have gone on". The marks in the soil of the soil forming processes can only be infered and therefore the processes themselves are not used in the differentiation of the Soil Orders (USDA, 1975, p.71). This bias towards the genetical homogeneity may have been a result of the influence from the previous systems of soil classification.

When a large number of intrinsic properties is used to classify soils, it is no longer possible to diagnose

the soils in the field without further empirical studies on the diagnostic features in relation to the soil groups which have been defined numerically. This will eliminate inconsistencies of the USDA system caused by the use of different properties to define different Orders.

Even the soil Orders are not defined using properties that have clearcut boundaries. A given Order may not be different from all the other Orders in terms of a given set of properties, but rather they are chosen only to separate Orders pairwise. For example, "to distinguish Alfisols from Aridisols,Alfisols must have either,

(a) an aquic, udic, ustic or xeric moisture regime

(b) an epipedon that is both massive and hard or massive and very hard when dry" (USDA,1975,p.96).

According to the features given, Alfisols cannot be distinguished from the other Orders; for that, another set of features have to be chosen.

Since a number of different properties are used at different levels of the classification system, at the lower categorical levels the total number of properties involved is considerably higher. But the sub-categories are defined using a few properties. The soil profiles 1 and 30 belong to two Sub orders of Mollisols (ustolls and udolls respectively). The distinction is made on the basis of two features:

i) type of moisture regime

ii) presence or absence of a calcic or gypsic horizon and concentration of powdery lime in spherical forms or coatings on peds disseminated in clay size particles.

But the numerical classification has shown that the soil profiles 1 and 30 are closer in Mahalanobis sense and fuse together to form one group.

The soil group 2 (classification 3) is composed of three soil Orders. It may be interesting to note that the members of the group 2 have comparable moisture regimes (udic and aquic). It is possible to subdivide the group along the line of Soil Orders, but it is not clear if they form separate groups at this level of information.

As suggested by Webster (1968) the main weakness of the US Taxonomic System is its hierarchical organization. Once the hierarchy is established, it is not possible to move misclassified individuals into appropriate groups. This may be true even for a hierarchical numerical classification, but the reallocation of misclassified individuals is possible to improve the classification when the hierarchy is not of interest. In the US Taxonomic system, there is the possibility of misclassification at Soil Orders level because of the limited number of features being used to define them. But it may be possible for an experienced soil surveyor to identify certain soils more accurately than others, however, it is not possible

to identify the intermediate types so easily using a few features. Therefore, the problem of soil classification involves identification of different soils and also assigning soil individuals to correct groups on the basis of a discriminant procedure.

The Mahalanobis distance was initially introduced (Mahalanobis, 1927) for two groups, but the method can be generalized to more groups, as demonstrated by Rao (1948). At the same time it is possible to use D^2 as a similarity measure to compare pairs of groups. The classification 3 obtained by this method appears to be numerically better than the other two methods. Since it is possible to identify the most similar soils in a given sample of soil profiles, an agglomerative strategy can be used to sort similar soils and subsequently discriminant analysis can be performed to improve the classification. The main problem of reallocation, encountered here is that when the soil horizon is treated as the basic unit of classification, there is a tendency for different horizons of a given soil profile to be allocated to several different groups. But this is not a serious problem since the distance from the group centroids can be estimated treating the soil profiles as the unit of classification.

The superiority of the Mahalanobis distance measure over other methods is shown by the results discussed so far. The effect of the inter-attribute correlation is such that the sample space is distorted

and the relative distance of the individuals is not accurate. The Euclidean metric requires the attribute vectors to be mutually orthogonal. But the soil attributes are correlated in varying degrees (chapter 4), and the resulting inter-individual distance matrix is not indicative of the structure of the sample under consideration. It is possible to transform the original space to an orthogonal space either by orthonormalizing the attribute vectors or by means of principal component analysis prior to calculation of the distance matrix. Gower (1966) has demonstrated an alternative strategy, in that the distance matrix can be transformed to an Euclidean space by principal coordinate analysis (Q-analysis). The principal component vectors (or principal coordinate vectors) define a new coordinate system along the dimensions of the maximum variance of the sample and the redundant vectors (attributes) can be eliminated.

The Mahalanobis distance is used here as an alternative distance measure because of its ability to cope with inter-attribute correlations and the statistical validity of the distance measure. The D^2 measure is a discriminant function, which is able to identify pairs of soils belonging to different groups. This method appears to perform better than the Euclidean metric.

The relationship between the cluster diagrams (Fig. 6.2, a and b) and the canonical plots (Fig. 6.4, a and b) can be considered as an important indication of the validity of the classification 3. The canonical vectors define the best discriminant axes, and best

separate the groups. This method has been used by Webster and McBratney (1981) to locate soil boundaries, and in the same way boundaries between soil groups can be identified by this method. As mentioned earlier, the soil groups are not mutually exclusive and one type of soil grades to another. By treating initially all soil profiles are separate groups, it may be possible to plot all individuals (profiles) and then compare the canonical plots with dendrograms for the purpose of interpretation.

In this exercise the Wilk's criterion is not used as a test of significance, as probabilistic decisions cannot be made on a classified population (Webster, 1971). Since the test is based on the total population, it is non-probabilistic.

The Mahalanobis distance between group centroids and individuals can be used to reallocate the misclassified individuals and also to remove those individuals, which are significantly different from all groups. When the soil horizon is used as the individual, there is a possibility of the soil profiles sometimes being split into several groups. But the results of this analysis indicate that the majority of the soil profiles do not split into several groups. The use of centroids of the soil profiles in this way is meaningful as it reflects the sub space occupied by the horizons of a given soil profile. The soil is defined here as the vertical cross-section down to the vertical boundary of that soil, in that all depth points have equal weight.

The relationship between the soil groups of the classification 3 and the parent material types is interesting. The soil series is defined within individual parent material types, but it has little relevance in defining Soil Orders. It is a well known fact that the soil parent materials have considerable influence upon the soil properties, but that relationship may be inversely related to the age of the soil, as pedogenetic processes alter the nature of the parent materials to such an extent, that no longer can the parent material be inferred from the soil properties. At this stage it is not possible to say whether the emerging relationship is significant or not. Further work may be needed in this direction. However, such relationships due to the classification processes are of great importance to the soil taxonomist as the validity of the classification is to a considerable extent dependent on the ability to generate new information about the population under consideration.

The majority of soils used in this study belong to the Orders of Alfisols and Mollisols when the Sri Lankan soils are excluded. These two Orders have split into several groups indicating pedological criteria used to define them are probably not sufficient under all circumstances. However, it is interesting to note that Alfisols and Mollisols in groups 2 and 3 have formed separate sub-clusters (Fig. 6.2b). The soil properties used in the numerical classification may be too limited to make a full comparison, but the strategy applied is capable of producing homogeneous groups. Use of more soil properties in numerical classification is required to fully study the difference between conventional and numerical classifications.

CHAPTER 7

CLASSIFICATION OF SOILS OF THE WEST SUSSEX COASTAL PLAIN BY NUMERICAL TAXONOMIC METHODS

7.0 Introduction

In Chapter 6 three classificatory strategies were used to classify a sample of soil profiles from geographically diverse areas. Emphasis was placed there on the idea of identifying groups of soils from a highly heterogeneous population (seven out of ten Orders defined by USDA). The soil profiles sampled from a small area may not show as great a difference one from another as when they come from a geographically and environmentally diverse area. It seems important to determine the effect of spatial scale on the classificatory strategies described in chapter 6.

The West Sussex Coastal Plain covers an area about 518 km^2 (1:25,000 O.S. Sheets SU 70, 80 and 90, SZ 89 and TQ 00 and 10) as defined by the Soil Survey of England and Wales (1967, pp.1-23). The land area has been divided into five units on the basis of the geomorphology of the area (Fig. 7.1).

(1) The Lower Coastal Plain. This is underlain by Pleistocene drift deposits and the maximum height of the area reaches 15m O.D. at its northern limit.

(2) The Upper Coastal Plain which lies at the foot of the South Downs and the surface level ranges from 22m to 46m O.D. The area is mainly covered by gravel.

(3) The South Downs. In the eastern section of the area, steep chalk escarpments occur. They rise in height up to 183m. The broad wooded interfluves are covered by drift deposits.

(4) The Eocene Outcrops. There are six separate outcrops of Eocene beds, which are predominantly clay and found below 46m O.D.

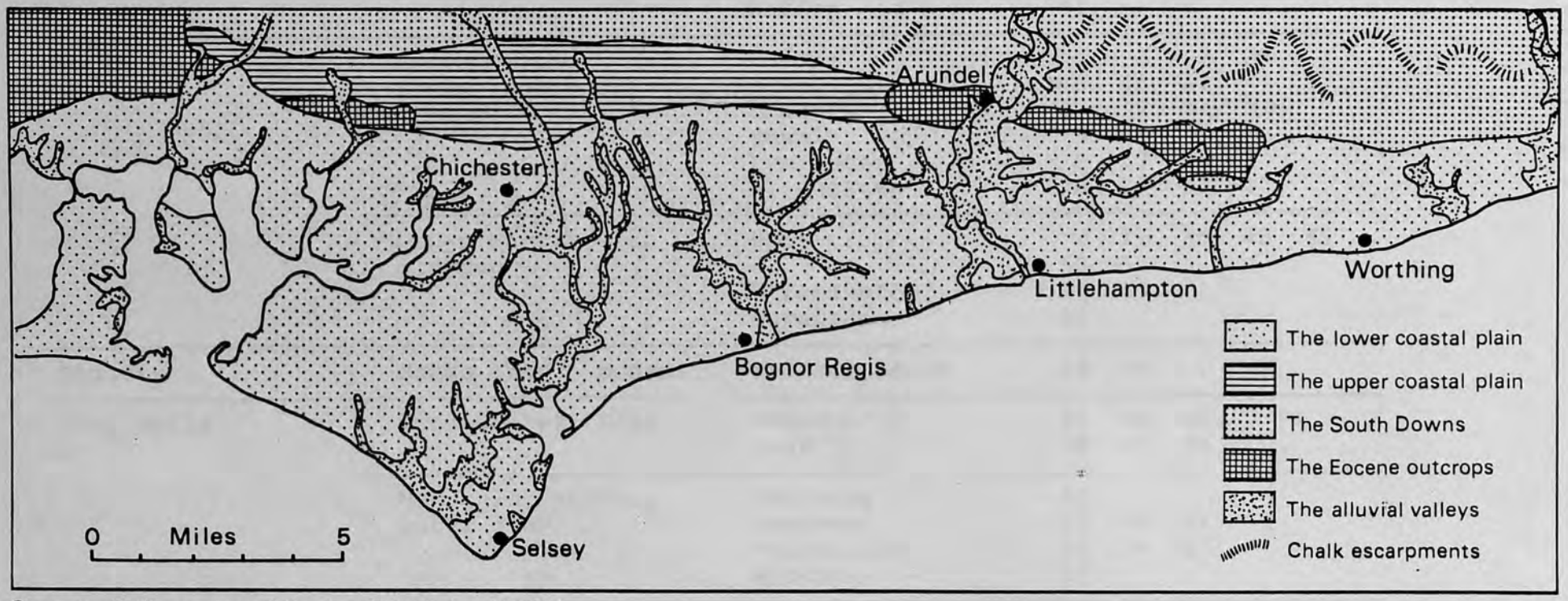
(5) The Alluvium Valleys. These are associated with the major streams of the area and are estuarine in character but partly protected by sea walls.

The identification of Soil Series has been based mainly on the type of parent materials and therefore, the classification of parent materials is a useful exercise prior to a soil classification for an area. Four major types of parent materials have been identified in the survey area (Fig. 7.2).

(1) Recent deposits
   a) dune sand
   b) estuarine and freshwater alluvium
   c) hill wash

(2) Pleistocene deposits
   a) brickearths
   b) head (Coombe) deposits
   c) marine beach deposits and sands
   d) riverine gravel
   e) clay-with-flints

Fig. 7·1 Geomorphological Regions of the West Sussex Coastal Plain



Source: Soil survey of England & Wales (1967)

TABLE 7.1 - Classification of Soils of the West Sussex Coastal Plain According to Soil Survey of England and Wales (1967)

MAJOR GROUP	GROUP	SERIES	PROFILE NUMBER
Calcareous Soils	Rendzinas	Icknield	1 10 33
	Brown Calcareous	Coombe	5 17 31
		Wallop	41
Brown Earths	Brown earths (sols lessives)	Hamble	4 6 7 25
			45 46 47
		Charity	28 32 38 54
		Rewel	9 43
		Lyminster	8 57 58
		Winchester	35 36 37 42
	Brown earths with gleying	Hook	2 11 18 48 49
			50 51
		Stretington	21 55
		Bursledon	62
Podzols	Humus iron podzol	Shirrel heath	29 64
Gley Soils	Groundwater Gley	Arundal	12 34 65
		Gade	22 23 24
	Non-calc. surface water gley	Patching	44
		Swanmore	20 60 61
		Titchfield	19 26 27
		Wickham	30

TABLE 7.1 (page 2)

MAJOR GROUP	GROUP	SERIES	PROFILE NUMBER
Gley Soils	Non-calc. gley	Parkgate	3 13 14 15 39 40 52 53
		Binsted	56
		Calcetto	16 59
		Curdridge	63

(3) Eocene deposits

a) Reading beds

b) London Clay

c) Bagshot Sands

(4) Cretaceous deposits

a) Upper Chalk

b) Middle Chalk

c) Lower Chalk

Broadly the whole area may be considered as climatically uniform. The mean monthly temperature ranges from 5° C in February to over 16° C in July and August. The mean annual rainfall is about 711mm with a slight rise landwards. Dry spells occur in the Spring and in the Summer. There is a soil moisture deficit from May to November (Soil Survey, England and Wales, 1967, pp.17-20). This area is noted for the highest average daily duration of sunshine for England and Wales.

According to the Soil Survey of England and Wales (1967, p.41), the soils of the area are divided into twenty four Series belonging to four Major Groups (Table 7.1). The Soil Series, which is identified as having similar profile characteristics and developing in lithologically similar parent materials, is used as the basic unit of classification and mapping by the Soil Survey of England and Wales. In the definition of the Soil Series, somewhat permanent features are used. The arrangement of the sub-horizons plays an important role in diagnosing the Soil

Soil parent materials

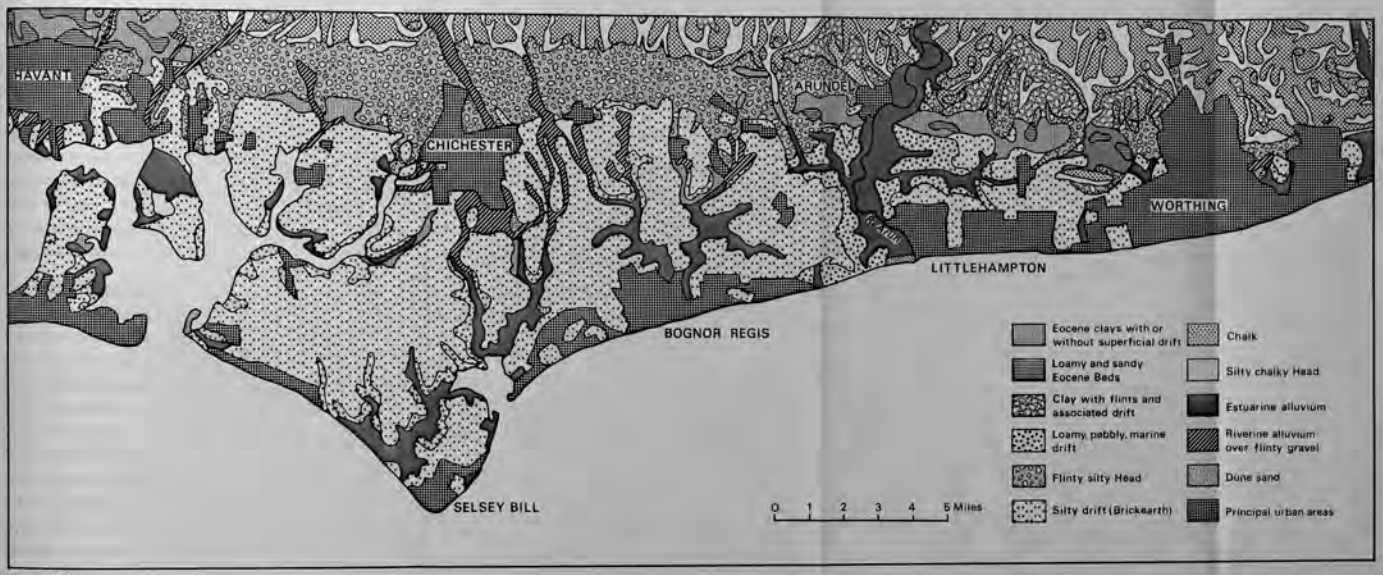


Figure 7·2

Series and consequently all observations on the soil are by the horizon. Each Soil Series is confined to one particular parent material class, which may be divided into a several Soil Series on the basis of soil profile characteristics, such as the arrangement and type of horizons, soil texture, etc.

It can be seen from the table 7.1 that the soils of any given Group or Major Group may have derived from different types of parent materials. The Major Groups are defined using a set of pre-determined diagnostic features known as 'keys'. Like USDA system, the soil classification of the Soil Survey of England and Wales is described as a hierarchical system (Avery, 1980).

7.1 Data and methods

The data used in this study was obtained from the Soil Survey of England and Wales (SSEW), a part of the data was listed in the Memoir for the West Sussex Coastal Plain Survey and the rest was obtained from the records of SSEW at Rothamsted. Because of the lack of comparable data only sixty five soil profiles were chosen, the geographical distribution of which is illustrated by Fig. 7.3. The main problem in the use of Soil Survey (SSEW) data is that a given property has often not been determined for all depth levels and all soil profiles. This could inevitably increase the number of missing cells in the data matrix. In order to minimize the number of missing values, a considerable number of attributes was excluded from this analysis. The attributes chosen can be described in three categories (Table 3.2).

Distribution of sampling sites

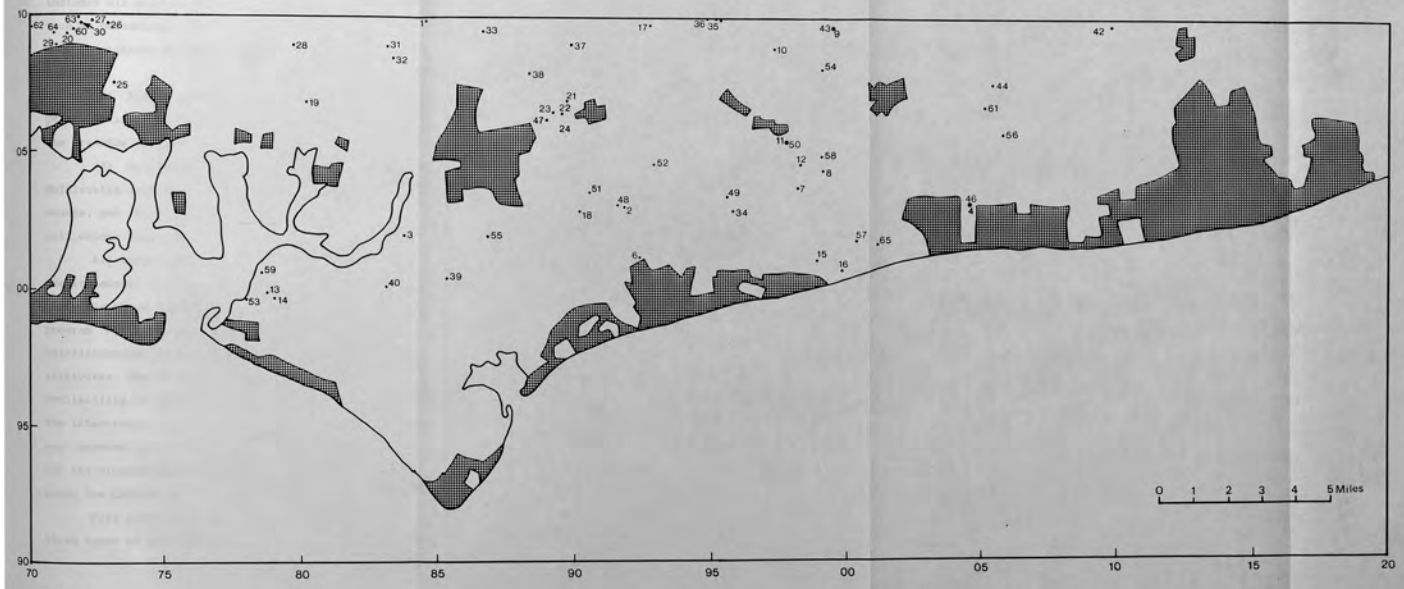


Figure 7·3

(1) Numerical (quantitative) attributes. This includes all measurements made on a continuous scale such as percentage silt, pH, percentage organic carbon etc. In statistical terms they can be described as variates.

(2) Binary attributes. They can take one of two possible states such as the presence or absence of mottling. The presence is coded as 1 and absence as 0.

(3) Multistate attributes. These may be ordered multistates such as the degree of mottling, abundance of stones, ped size etc., or disordered multistates such as soil colour (hue), rock type etc.

All three types of attributes can be used simultaneously to compute the similarity matrix. In this study the three attribute types were used only with the program REMUL (Lance and Williams, 1975) and the other classifications were obtained using only the numerical attributes. The choice of attributes was made on the availability of data with the minimum of missing values. The inter-individual squared Euclidean distance matrix was computed by the TAXON program, MULCLAS which allows for the missing values, and the cluster analysis was done using the CLUSTAN 1C (second release) computer package.

Five classifications were obtained by the following three types of strategy as before (chapter 4).

A. Method 1 Squared Euclidean Distance as the Similarity Measure

(a) classification by Average Linkage Method

(b) classification Ward's Error-Sum of Squares Method (ESS)

B. Method 2 Classification by the TAXON program REMUL

C. Method 3 Mahalanobis Distance D^2 as the Similarity Measure

(a) Classification by Average Linkage Method

(b) Classification Ward's Error Sum of Squares Method (ESS)

The Wilk's Criterion Λ was used to compare the classifications obtained by the numerical taxonomic methods and also the Soil Survey of England and Wales (1967,p.41) classification at the group level. Although this technique is generally used to test the significance of a null hypothesis, here it was used to compare the classifications for "numerical optimality" (chapter 2).

Canonical analysis (chapter 2) was performed on the sixty five soil profiles (individuals) treating them as primary 'groups' with the depth levels as the group members. The classifications obtained by the above taxonomic methods were compared using canonical plots representing the group centroids and the probablity circles, the radius of which is $\sqrt{x^2/n_k}$ (where n_k is the

size of the kth group). Finally, canonical analysis was performed on the best classification using the fifteen properties measured at the two uppermost depth levels (thirty attributes).

Once the best initial partition is established it is possible to reallocate the missclassified individuals in order to improve the classification. The Mahalanobis distance between the individuals and the group centroids was calculated using thirty attributes (fifteen numerical attributes measured at two depth levels. Table 3.2). The individuals which were closer to the group centroids other than those of their parent groups were transferred.

7.2 Results

A series of classifications were obtained by the three classificatory strategies. The effect of the similarity measure on the classification was examined.

Classification 1 (method 1)

The similarity between individuals was measured using squared Euclidean distance and the sorting of the similarity matrix was done by average linkage sort and Ward's error-sum of squares method. The two classifications, obtained by the two methods, are listed in Table 7.1, a and b respectively.

TABLE 7.2

(a) Classification by Average Linkage Sort with Squared Euclidean Distance as Similarity Measure

GROUP	INDIVIDUALS
GROUP 1	1 2 3 4 5 6 7 8 11 13 14 15
	16 17 18 19 21 22 23 24 25 28 31 32
	33 34 38 39 40 45 46 47 48 49 50 51
	52 53 55 57 58 59 62 63
GROUP 2	9 29 43 54 64
GROUP 3	10
GROUP 4	12 26 35 65
GROUP 5	20 27 30 56 60 61
GROUP 6	36 37 42 44
GROUP 7	41

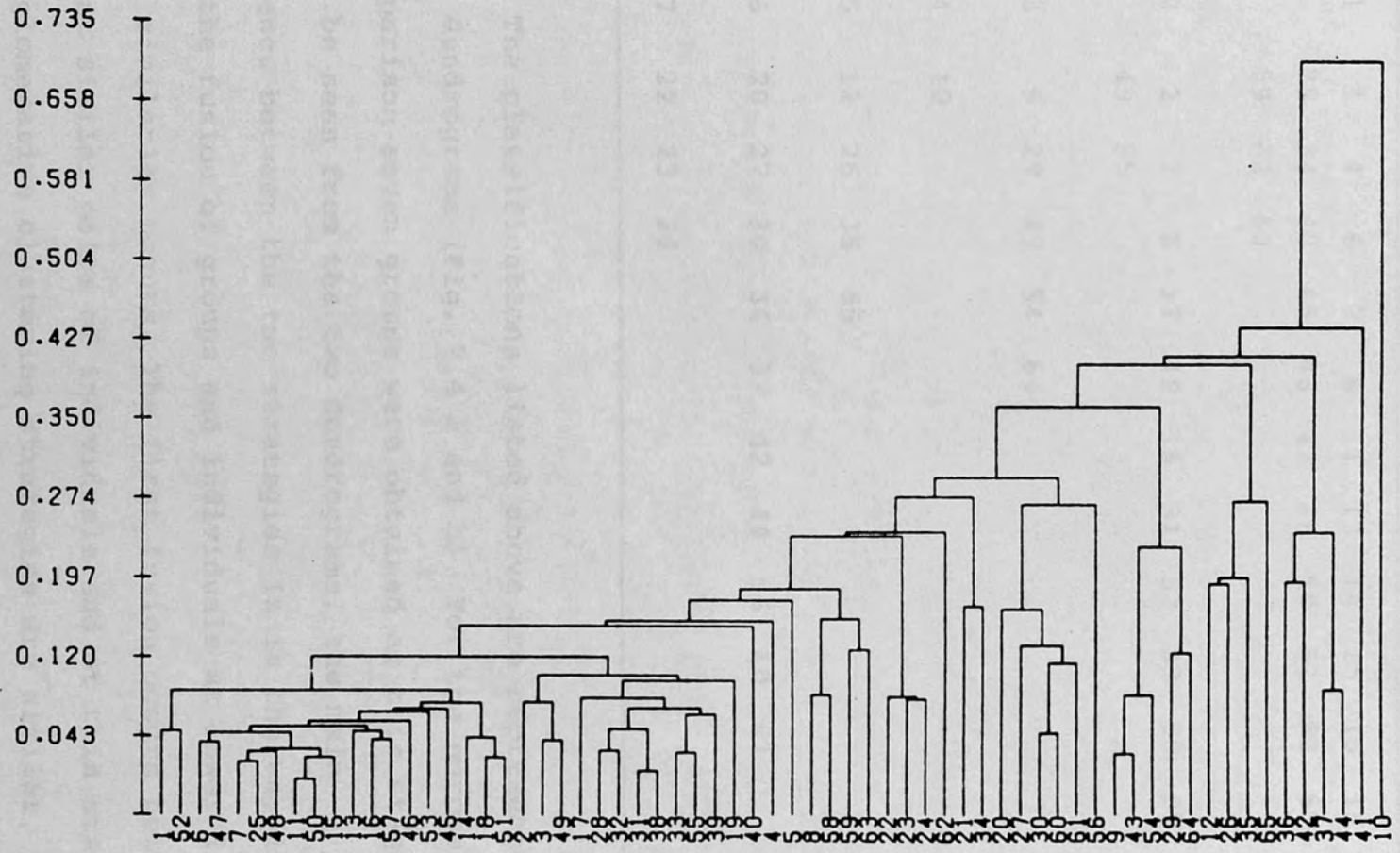


Fig. 7.4a Classification of soils of the West Sussex Coastal Plain by average linkage method with squared Euclidean distance as the similarity measure

(b) Classification by Ward's ESS Method with Squared Euclidean Distance as Similarity Measure

GROUP	INDIVIDUALS
GROUP 1	1 4 6 7 8 11 13 14 15 16 18 21 25 34 40 45 46 47 48 50 51 52 57 58 59 62 63
GROUP 2	2 3 5 17 19 28 31 32 33 38 39 41 49 55
GROUP 3	9 29 43 54 64
GROUP 4	10
GROUP 5	12 26 35 65
GROUP 6	20 27 30 36 37 42 44 56 60 61
GROUP 7	22 23 24

The classifications listed above are represented by two dendrograms (Fig. 7.4 a and b). For the purpose of comparison seven groups were obtained at this stage. As can be seen from the two dendrograms, the main difference between the two strategies is in the way in which the fusion of groups and individuals at lower levels of the similarity occurs. The first fusion occurs between the most similar pairs of individuals and at this stage all agglomerative clustering strategies are similar, but as fusion proceeds different methods take different courses.

The groups 2, 3 and 4 in the classification 1a are similar to the groups 3, 4 and 5 respectively in the classification 1b. The group 6 in 1b is divided into two groups in 1a (groups 5 and 6). The clusters produced by average linkage sort (Fig. 7.4(a)) are not clearly defined: in fact no clusters can be identified from the dendrogram. On the contrary, Ward's ESS method has shown the existence of well defined clusters (Fig.7.4 (b)).

Classification 2 (Method 2)

This classification (table 7.3) was obtained by the divisive strategy, REMUL, using all three types of attributes (60 numerical, 12 multistate and 4 binary). Ten groups were originally requested, but after the final reallocation only seven groups were left. The classification obtained by REMUL is not comparable with the others. Unlike the other classifications, the two podzol profiles (inds. 29 and 64) were separated to form two groups. All the other methods have separated the soil individual 10 from the rest of the sample to form a separate group, but in this classification it has joined group 1.

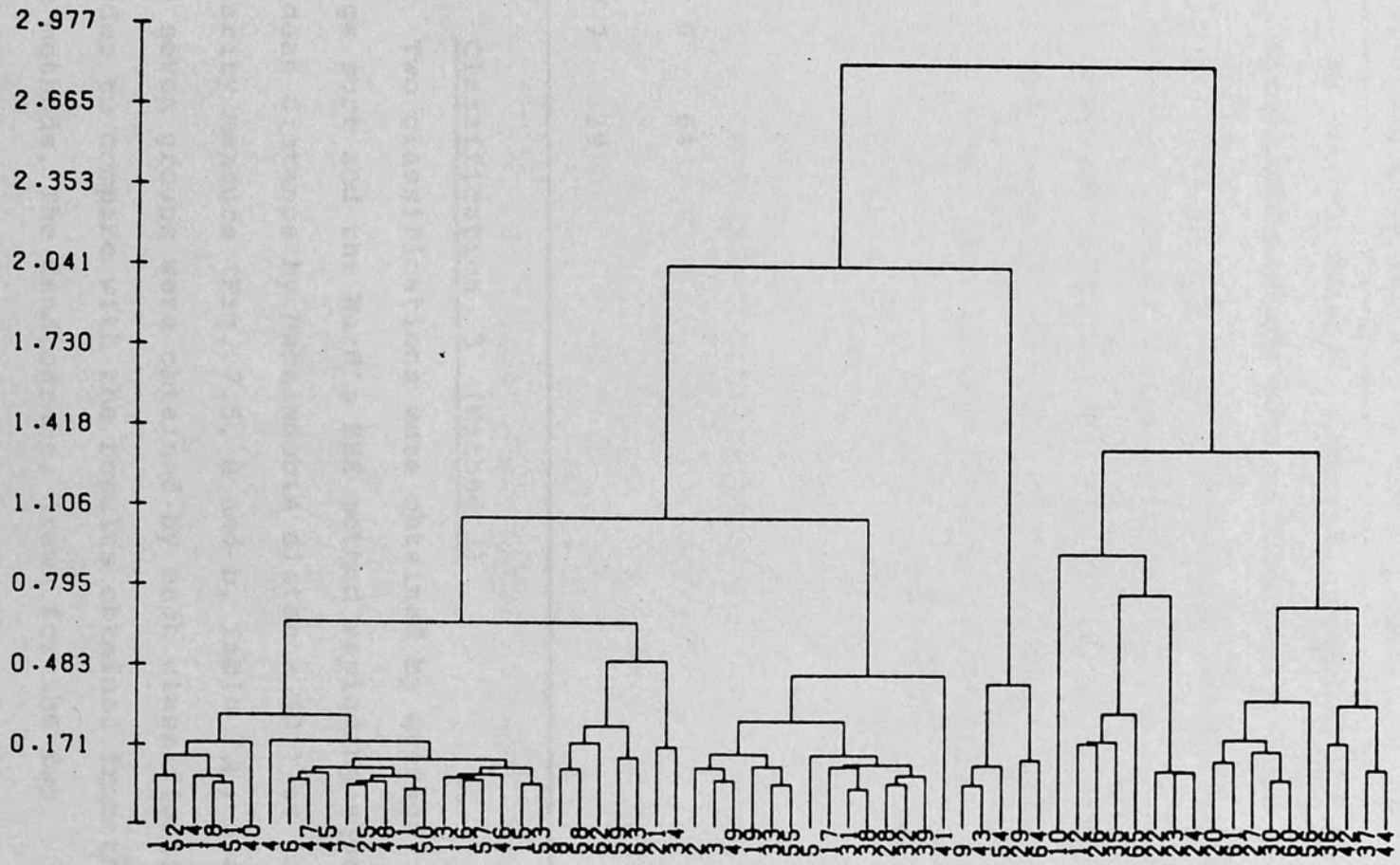


Fig. 7.4b Classification of soils of the West Sussex Coastal Plain by Ward's ESS method with squared Euclidean distance as the similarity measure

This classification (Table 7.3) has similarities to the Soil Survey of England and Wales (SSEW) classification into Major Groups. The majority of soils in group 1 are calcareous, derived from calcareous parent materials. Soil group 2 is made up of Brown Earths and Gleys, whereas group 3 is exclusively Gley soils, although they are not the only Gley soils in the population under consideration. Groups 4 and 5 are Brown Earths. This association between the SSEW classification and the classification produced by REMUL may be due to the inclusion in the data of certain morphological observations in addition to the fifteen numerical properties. Agreement with the conventional classification does not necessarily mean a good classification.

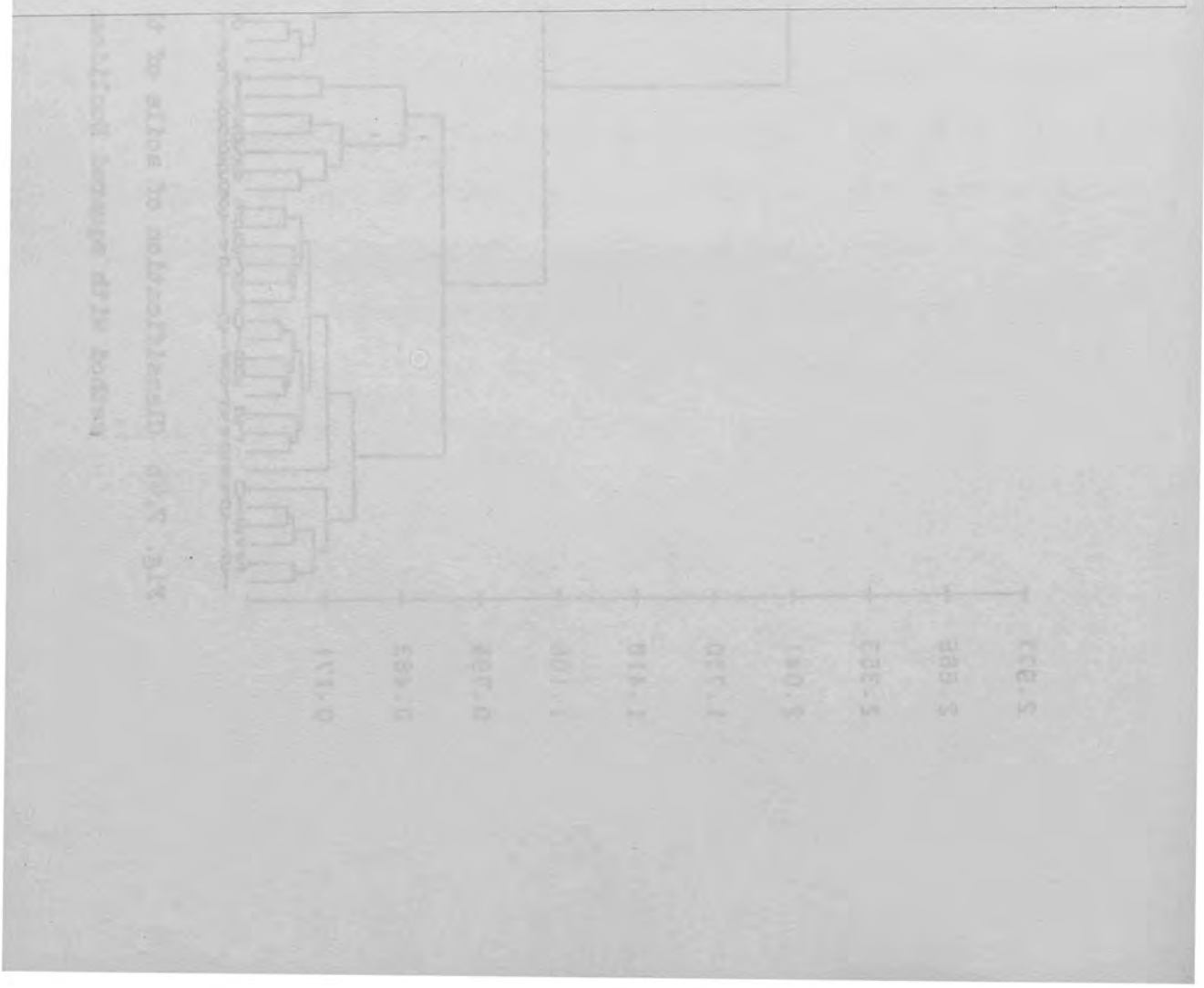


TABLE 7.3 Classification by REMUL using all Three Types of Attributes

GROUP	INDIVIDUALS
GROUP 1	1 5 10 17 19 22 23 24 28 31 32 33 34 38 39
GROUP 2	2 3 4 6 7 8 11 13 14 15 16 18 21 25 40 41 45 46 47 48 49 50 51 52 53 57 58 59 62
GROUP 3	12 20 26 27 30 35 44 56 60 61 63 65
GROUP 4	36 37 42
GROUP 5	9 43 54
GROUP 6	64
GROUP 7	29

Classification 3 (Method 3)

Two classifications were obtained by average linkage sort and the Ward's ESS method replacing squared Euclidean distance by Mahalanobis distance (D^2) as the similarity measure (Fig. 7.5, a and b, Table 7.4, a and b). Again seven groups were obtained by each classification in order to compare with the results obtained from the other methods. The dendrograms, drawn for the two classifications are considerably different in the clarity

of the clusters, but the smaller groups, produced at the initial stages of the fusion are similar (Fig. 7.5, a and b). The dendrogram by the average linkage sort does not indicate existence of clusters, a considerable chaining has occurred. The clusters produced by the Ward's ESS method are well defined.

TABLE 7.4

(a) Classification by Average Linkage Method with Mahalanobis Distance (D^2) as Similarity Measure

GROUP	INDIVIDUALS
GROUP 1	1 2 3 4 5 6 7 8 11 13 14
	15 16 17 18 19 22 23 24 25 28 31
	32 34 35 37 38 39 40 41 45 46 47
	48 49 50 51 52 53 55 56 57 59 63
	12 29 33
GROUP 2	26 27 30 36 42 60 61 65
GROUP 3	9 43 44 54
GROUP 4	21 58 62
GROUP 5	64
GROUP 6	20
GROUP 7	10

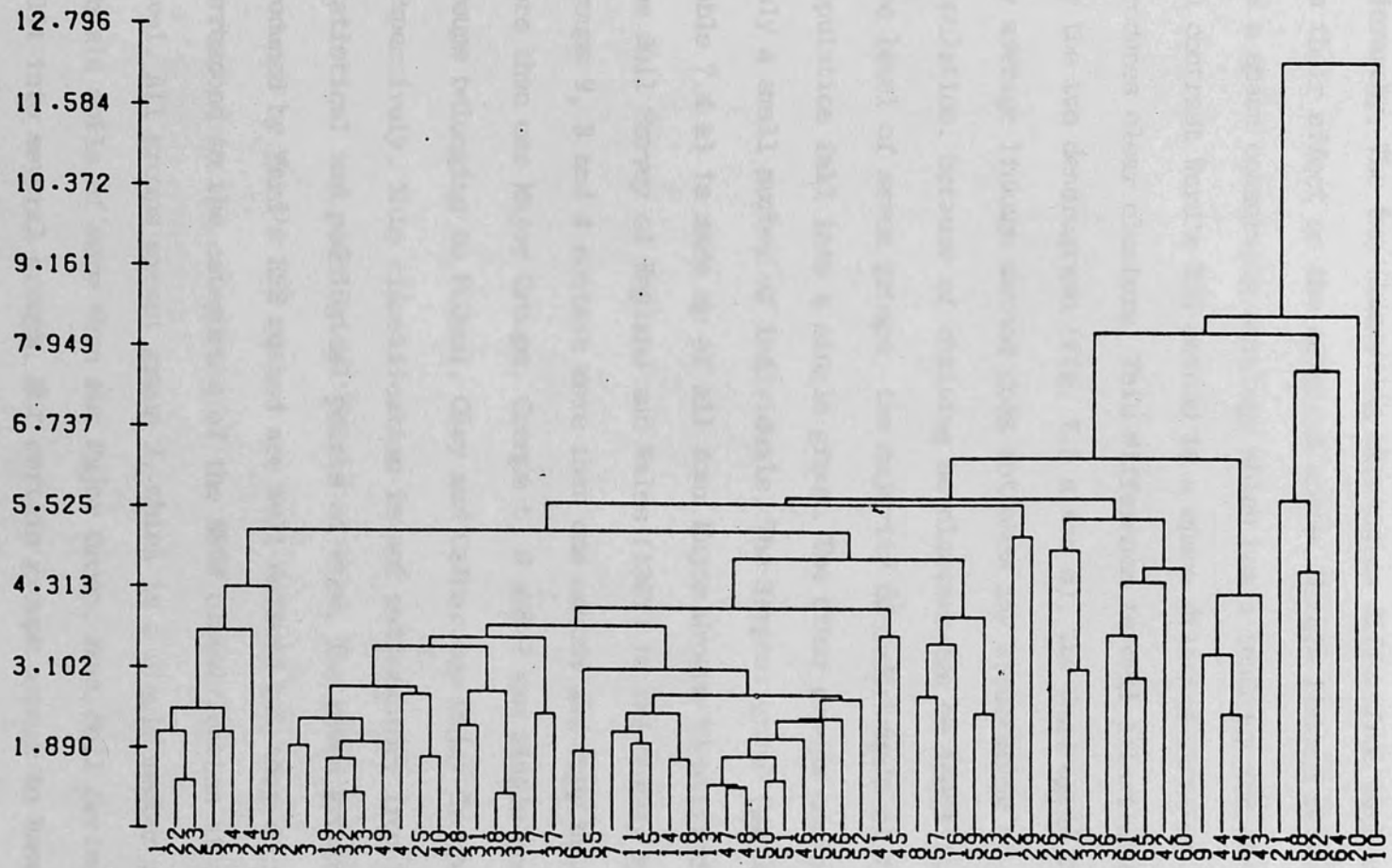


Fig. 7.5a Classification of soils of the West Sussex Coastal Plain by average linkage method with Mahalanobis distance as the similarity measure

The classifications 3 a and b are different from the other classifications and between themselves at high levels of the hierarchy. The difference between the classifications 3 a and b is due to the way in which groups fuse together at various levels of the hierarchy. The two clustering strategies differ from each other in their effect on the original space. Single linkage method (3a) is a space conserving strategy which has a tendency for chaining, in contrast Ward's ESS method is a space dilating strategy which produces clear clusters. This difference is well illustrated by the two dendrograms (Fig. 7.5 a and b). The dendrogram produced by average linkage method does not show any structuring of the soil population. Because of chaining no clusters can be identified. At the level of seven groups, the majority of individuals of the population fall into a single group. The other groups contain only a small number of individuals. The largest group (group 1, Table 7.4 a) is made up of all four Major Groups identified by the Soil Survey of England and Wales (1967) in the survey area. Groups 2, 3 and 4 contain more than one member and they belong to more than one Major Groups. Groups 5, 6 and 7 are single member groups belonging to Podzol, Gley and Calcareous Major Groups respectively. This classification is not satisfactory from both statistical and pedological points of view. The seven groups produced by Ward's ESS method are well defined but they do not correspond to the categories of the SSEW classification at any level. All groups except group 2, which is a single member group, contain soils of more than one Major Group. Most Soil Series have split into several groups. But certain groups appear to have common pedological features. For example, group 1 is made up of mainly calcareous soils. In this classification one of the Rendzina profiles separated from the rest of the population to form a single

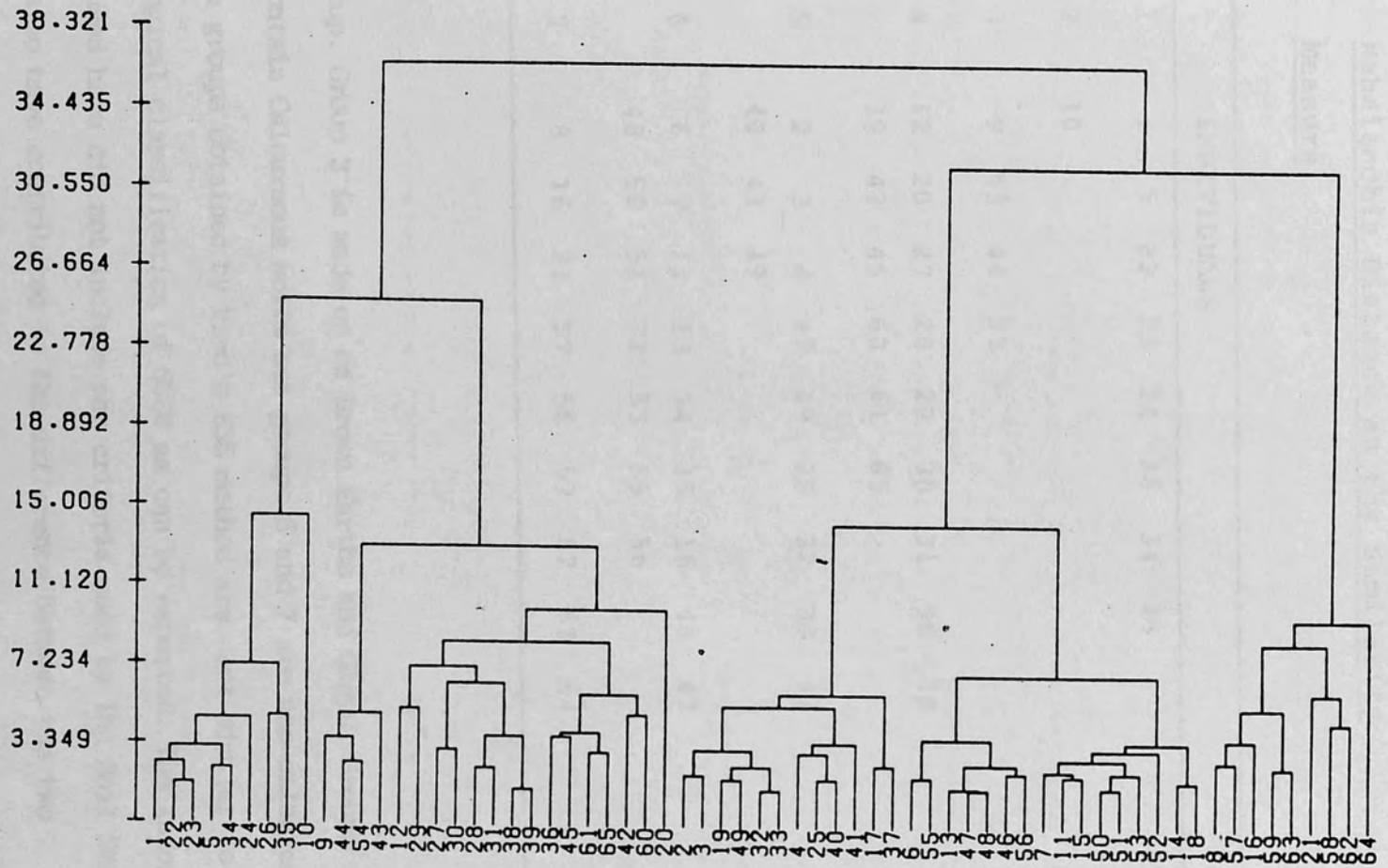


Fig. 7.5b Classification of soils of the West Sussex Coastal Plain by Ward's ESS method with Mahalanobis distance as the similarity measure

TABLE 7.4

(b) Classification by Ward's ESS Method with Mahalanobis Distance as the Similarity Measure

GROUP	INDIVIDUALS
GROUP 1	1 5 22 23 24 26 34 35
GROUP 2	10
GROUP 3	9 43 44 53
GROUP 4	12 20 27 28 29 30 31 36 38 39 42 45 60 61 65
GROUP 5	2 3 4 17 19 25 32 33 37 40 41 49
GROUP 6	6 7 11 13 14 15 18 46 47 48 50 51 52 53 55 56
GROUP 7	8 16 21 57 58 59 62 63 64

member group. Group 3 is made up of Brown Earths and Gleys. Groups 4 and 5 contain Calcareous soils but groups 6 and 7 are non-calcareous soils. The groups obtained by Ward's ESS method are not similar to the pedological classification of SSEW as can be expected. The information on soils used here did not include any criteria used by the Soil Survey. This may also have contributed to the difference between the two classifications.

TABLE 7.6 - Wilk's Criterion Λ, χ^2 and Degrees of Freedom for the Numerical Taxonomic and Soil Survey (1967) Classifications

CLASSIFICATION	G	Λ	χ^2	DF	P
1a	7	0.0835	561	90	< 0.001
1b	7	0.0433	709	90	< 0.001
2	7	0.0716	596	90	< 0.001
3a	7	0.0296	782	90	< 0.001
3b	7	0.0177	911	90	< 0.001
SOIL SURVEY CLASSIFICATION (1967)					
GROUPS	8	0.1672	404	105	< 0.001
SERIES	22	0.0169	901	315	< 0.001

G - number of groups

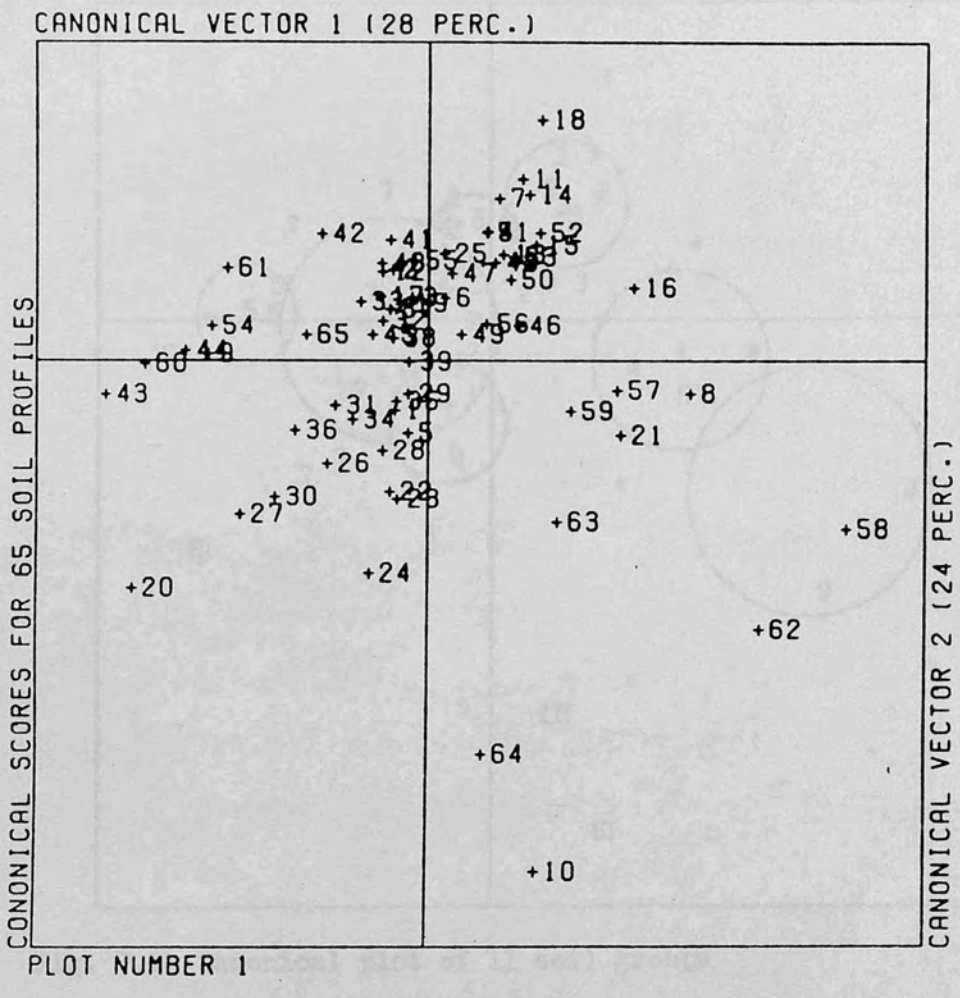


Fig. 7.6a Canonical plot of 65 soil profiles as 'groups'

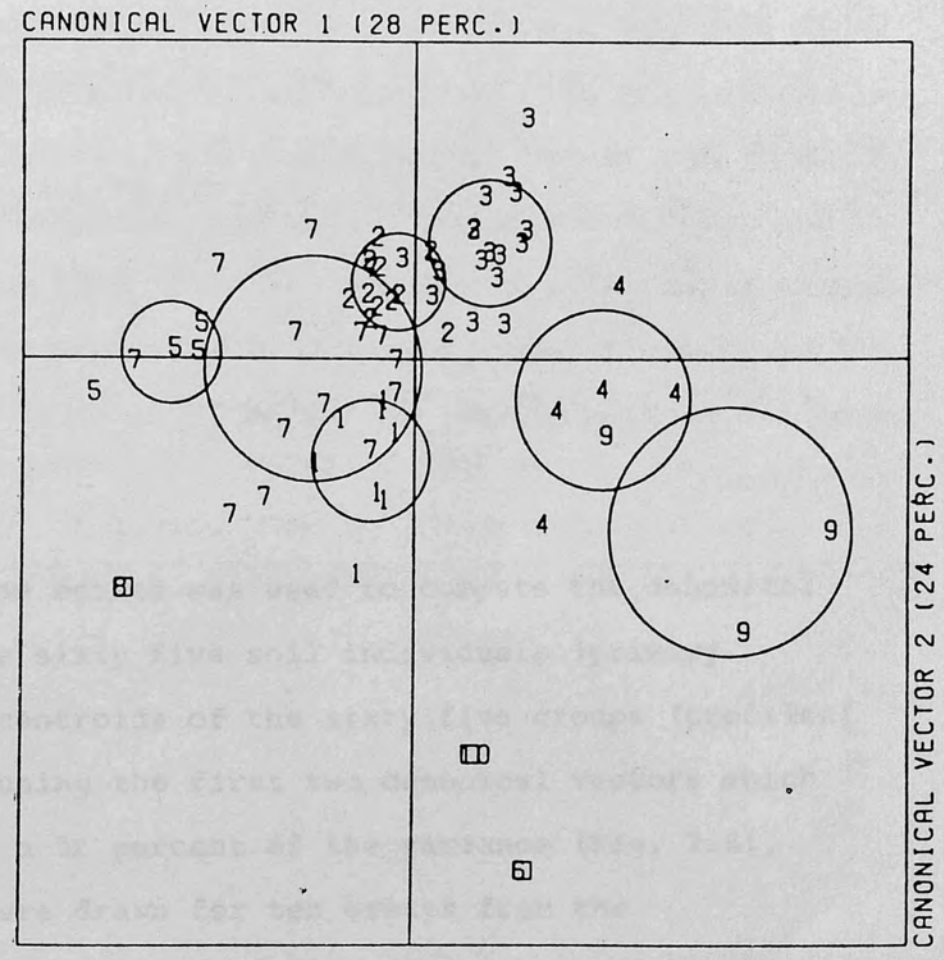


Fig. 7.6b Canonical plot of 11 soil groups

* If the classification used by the Soil Survey of England and Wales (1967) in "Soils of the West Sussex Coastal Plain" is considered to be a 'best' pedological classification then it may be argued that none of the numerical classifications appears satisfactory. However, conventional classifications weight criteria recognized as reflecting soil genesis, as well as not being explicit over the procedures used. If genetically important properties do not correlate with other soil properties their weighting in soil classification cannot be defended. On the other hand numerical classifications based on a small number of arbitrarily selected soil properties, as here, would not necessarily produce a true, general purpose classification. However, the objective of this study has been to classify soils into homogeneous groups with respect to the available information on soils. Until soils are sampled and described in a fashion compatible with the efficient use of numerical taxonomic methods, the true value of such methods may not be apparent.

The Soil Survey classification (Soil Survey of England and Wales, 1967) was considered at two categorical levels.

(1) Group level

(2) Soil Series level

It can be seen from Table 7.5 that the highest value of Λ was obtained for the eight Groups (Soil Survey classification) compared to the other classifications. Although the classification with 22 Soil Series has a low Λ value, it is closer to the classification 3b at the Seven group level.*

The Mahalanobis distance between soil individuals (profiles) was calculated assuming that the soil profile with several depth levels could be treated as a primary group. The same method was used to compute the canonical scores for the sixty five soil individuals (primary groups). The centroids of the sixty five groups (profiles) were plotted using the first two canonical vectors which accounted for a 52 percent of the variance (Fig. 7.6). The circles were drawn for ten groups from the classification 3b. The ten clusters represented by the circles (Fig. 7.6(b)) can be identified from the dendrogram (Fig. 7.5(b)) for the classification 3b. The cluster boundaries are not as clear as in the dendrogram. It may suggest that the soils of this area are not as diverse as the previous sample (chapter 6). However, the canonical plots may help determine the

relative position of the individuals in a two dimensional space.

The seven groups obtained by the numerical methods are represented by the canonical plots of the group centroids (Fig. 7.7). The group centroids of the classification 3b are better separated from each other as shown by the probability circles (Fig. 7.7). Although this does not mean the existence of seven groups, it is useful in determining the capability of the strategy to produce k (k<N) number of groups with a considerable degree of internal homogeneity.

Canonical plot of group Centroids of classification 1a

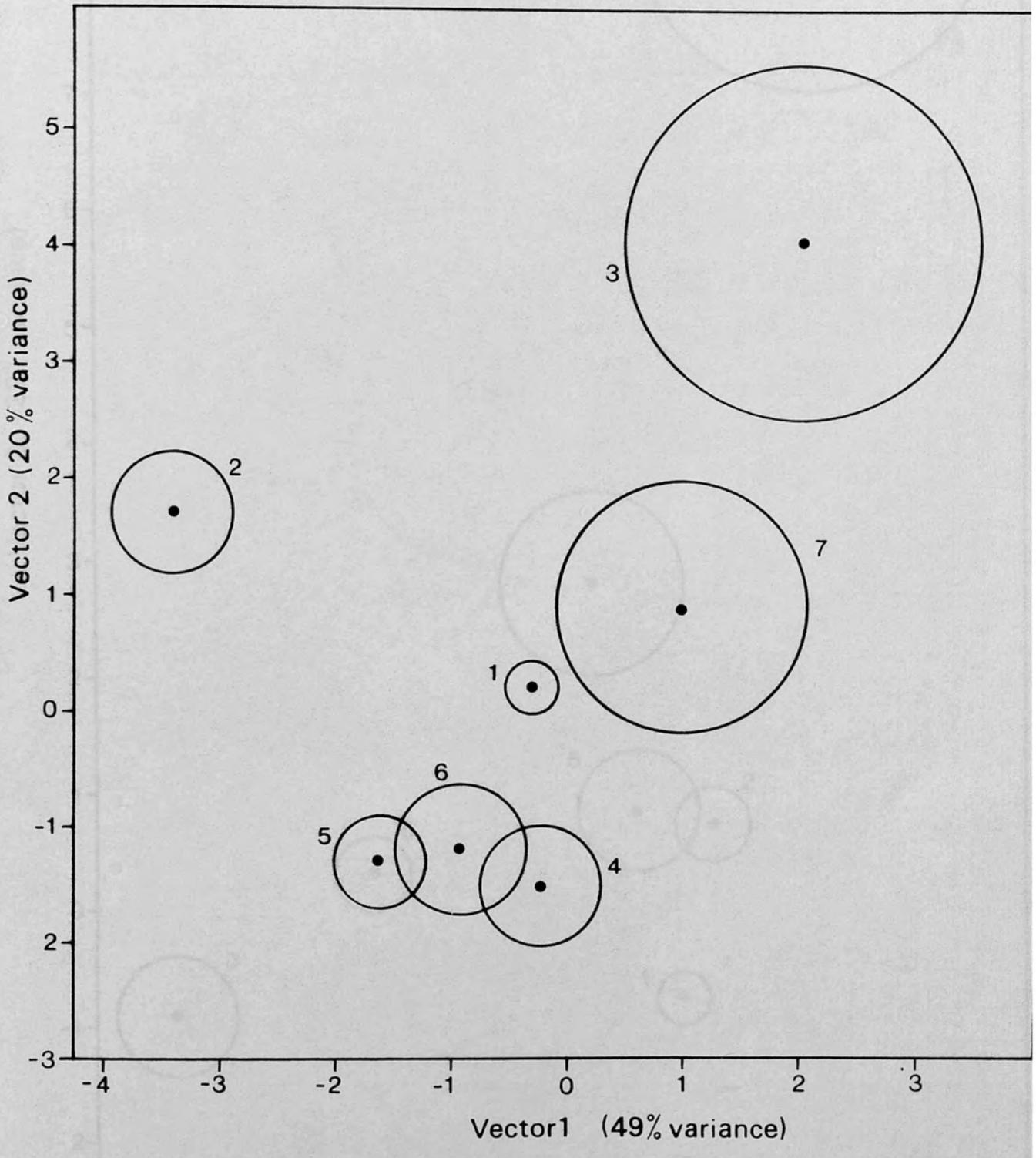


Figure 7·7a

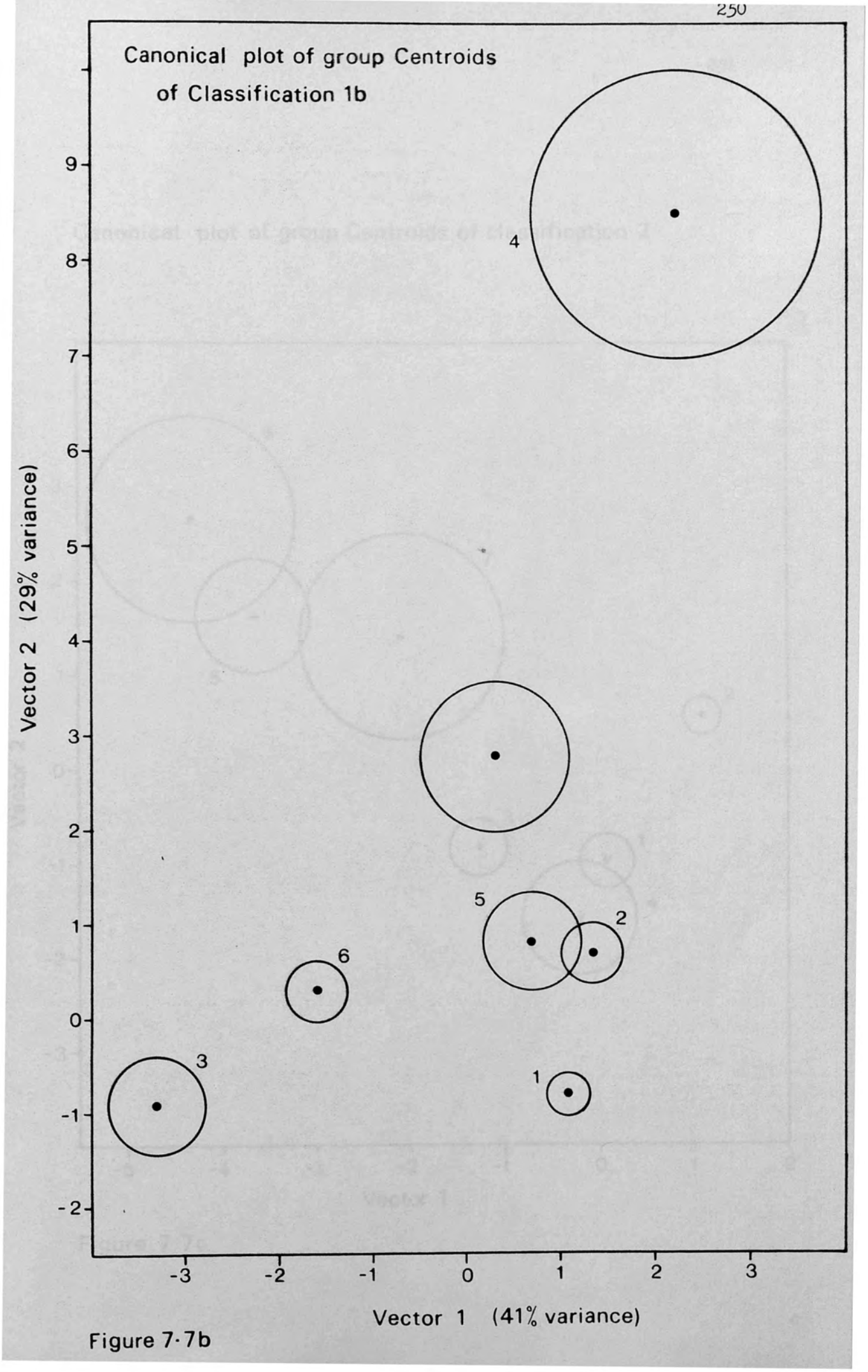

Canonical plot of group Centroids
of Classification 1b
Vector 2 (29% variance)
Vector 1 (41% variance)
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
4
5
6
1
2
3

Figure 7·7b

Canonical plot of group Centroids of classification 2

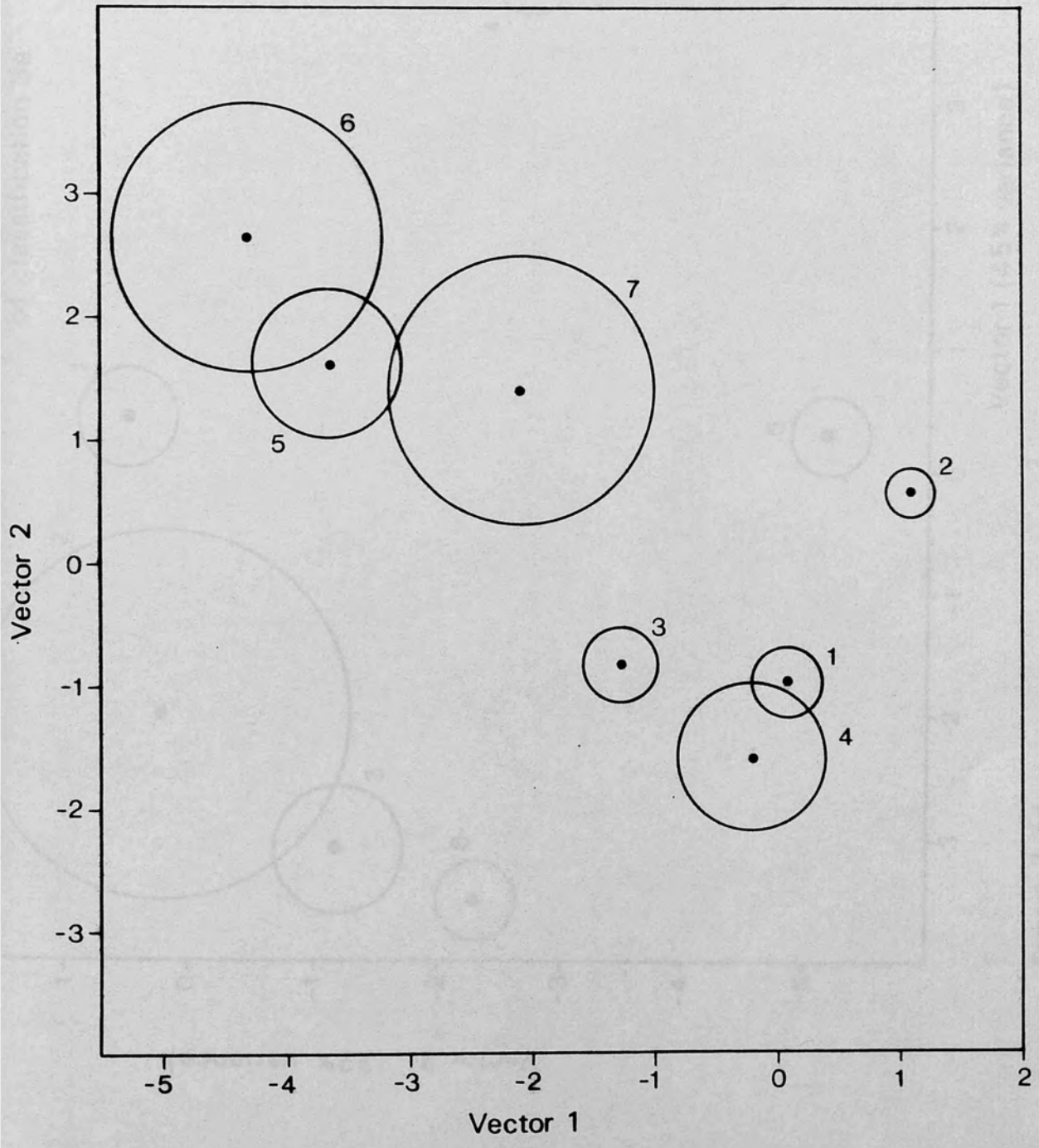


Figure 7·7c

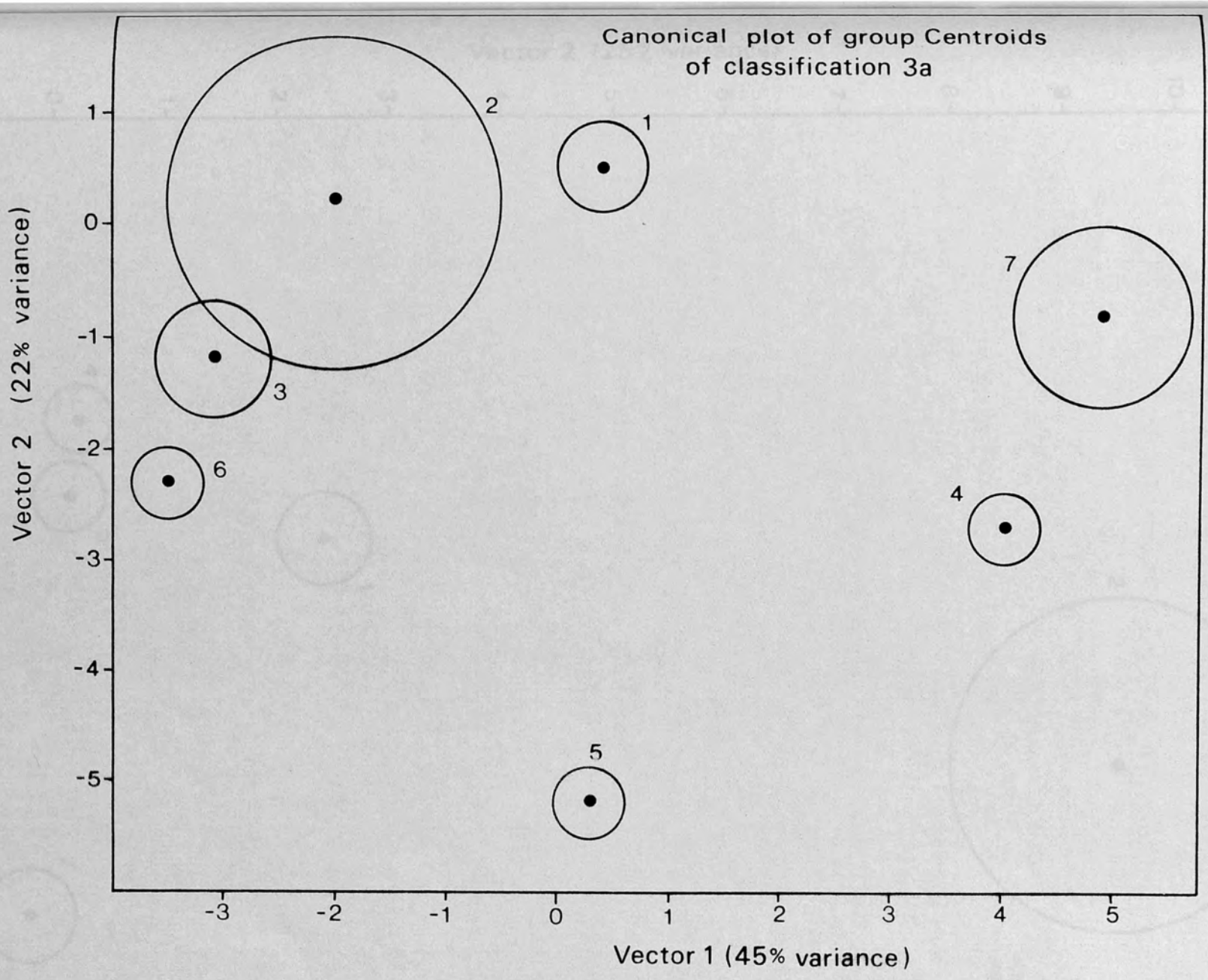

Canonical plot of group Centroids
of classification 3a
Vector 2 (22% variance)
Vector 1 (45% variance)
1
2
3
4
5
6
7
1
0
-1
-2
-3
-4
-5
-3
-2
-1
0
1
2
3
4
5

Figure 7·7d

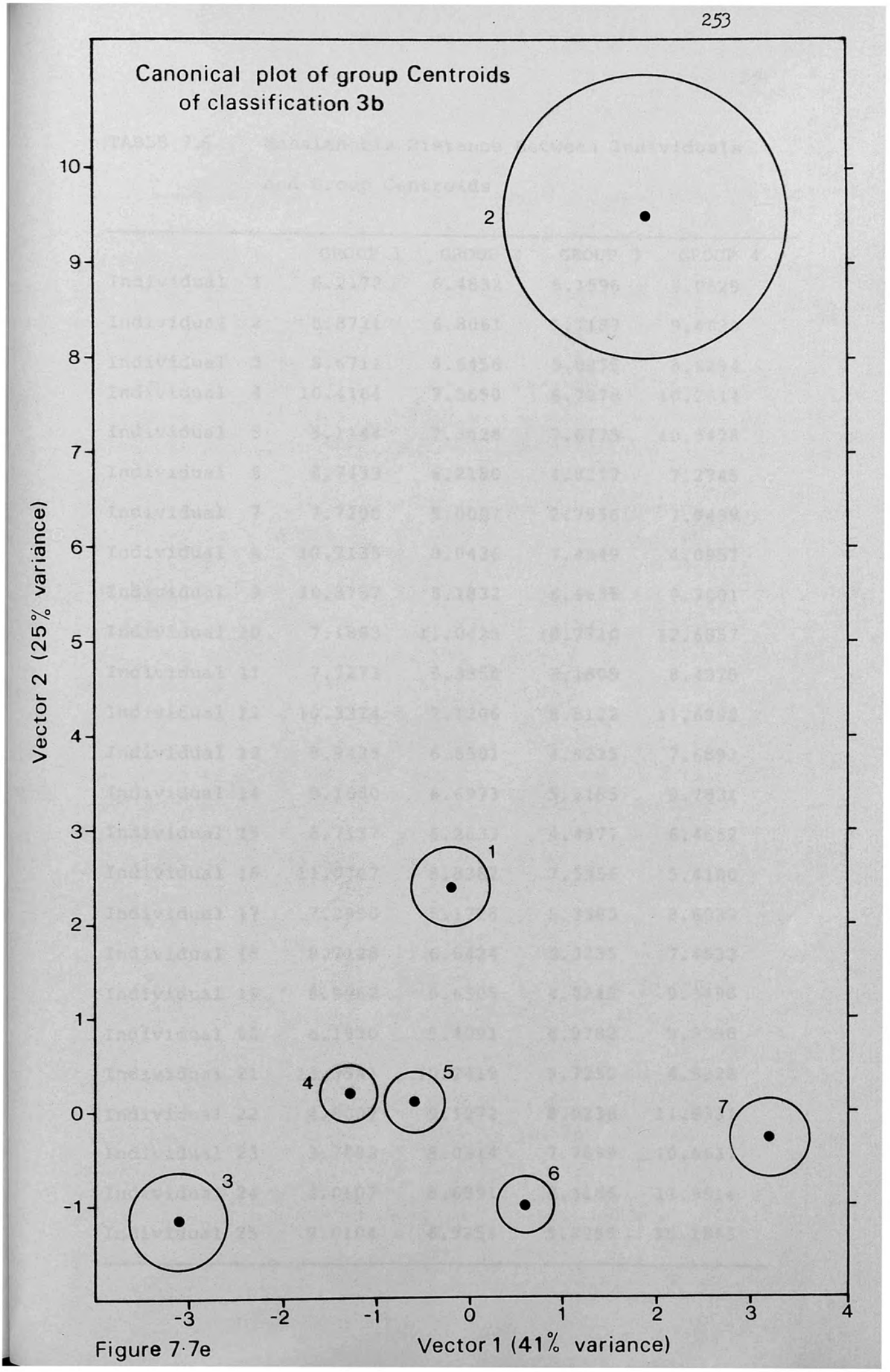

Canonical plot of group Centroids
of classification 3b
Vector 2 (25% variance)
Vector 1 (41% variance)
10
9
8
7
6
5
4
3
2
1
0
-1
-3
-2
-1
0
1
2
3
4
1
2
3
4
5
6
7

Figure 7·7e

TABLE 7.6 Mahalanobis Distance Between Individuals and Group Centroids

	GROUP 1	GROUP 2	GROUP 3	GROUP 4
Individual 1	6.2172	6.4832	6.1596	9.0529
Individual 2	8.8721	6.8061	5.7187	9.4630
Individual 3	8.6711	5.6456	5.0275	8.6294
Individual 4	10.4164	7.5650	6.7278	10.2514
Individual 5	5.1144	7.3628	7.6775	10.5478
Individual 6	8.7453	6.2180	4.8277	7.2745
Individual 7	7.7208	5.0007	2.7956	7.9439
Individual 8	10.7135	8.0436	7.4949	4.0957
Individual 9	10.8757	5.1832	6.6635	9.7001
Individual 10	7.1863	11.0435	10.7710	12.6857
Individual 11	7.7273	5.3358	3.1805	8.4375
Individual 12	10.3324	7.1206	8.5122	11.6998
Individual 13	8.9435	6.6501	4.9225	7.6893
Individual 14	9.1680	6.6973	5.1185	9.7831
Individual 15	8.7137	6.2633	4.4977	6.4652
Individual 16	11.0767	8.8382	7.5356	5.4100
Individual 17	7.2990	5.1728	5.3585	8.6033
Individual 18	8.7128	6.5424	5.3235	7.4533
Individual 19	8.9962	5.6505	4.8243	9.3490
Individual 20	8.1930	5.4091	6.9782	9.9398
Individual 21	12.3541	10.2419	9.7250	6.5328
Individual 22	4.8007	9.1272	8.9238	11.9331
Individual 23	3.7893	8.0314	7.7699	10.6637
Individual 24	4.0607	8.6891	8.5186	11.9914
Individual 25	9.6104	6.9254	5.8255	10.1865

TABLE 7.6 (page 2)

Individual 26	6.6040	9.7182	9.8618	11.9332
Individual 27	9.9394	6.1031	6.8020	9.3126
Individual 28	8.0339	5.0227	5.6208	10.6895
Individual 29	10.6567	6.9449	7.8273	10.1215
Individual 30	9.7027	6.8204	8.2071	10.9025
Individual 31	7.9801	5.8974	6.0277	8.0371
Individual 32	6.7988	3.4101	3.4582	9.0856
Individual 33	8.2698	3.8554	3.9466	8.7119
Individual 34	6.7639	11.8745	11.5293	14.0713
Individual 35	5.2681	9.2554	8.6429	11.2436
Individual 36	10.0141	6.4924	7.4461	11.2588
Individual 37	8.9475	4.4063	4.6939	9.2790
Individual 38	7.7913	3.7052	3.9076	8.7449
Individual 39	7.9573	4.1761	4.8657	7.9271
Individual 40	6.8928	5.2871	4.9225	10.0120
Individual 41	10.6142	8.1681	7.1013	10.5324
Individual 42	10.2144	6.6601	7.7625	10.7679
Individual 43	8.8092	4.7401	5.5706	11.0358
Individual 44	7.6291	3.8186	4.6888	9.7216
Individual 45	8.4197	5.4982	5.7011	9.0587
Individual 46	7.4054	5.3724	4.1628	8.3621
Individual 47	7.9989	5.1477	3.4613	9.2100
Individual 48	8.0756	4.9612	3.5525	8.2846
Individual 49	8.0278	5.3522	4.3749	9.1204
Individual 50	7.3074	5.3004	3.0002	8.0327
Individual 51	8.3624	5.7997	4.1340	9.3729
Individual 52	7.1648	6.1701	4.3922	9.6003
Individual 53	8.9858	5.6323	4.5046	7.8436
Individual 54	8.9945	5.2397	6.5979	10.5891
Individual 55	9.7018	8.4210	6.9133	10.6316
Individual 56	10.0078	7.8989	6.9554	10.4914
Individual 57	10.9742	8.7063	7.9843	3.8927
Individual 58	12.4367	11.4053	10.0566	5.4038
Individual 59	10.2140	8.6118	7.8713	3.8235
Individual 60	9.3561	4.5248	6.6376	9.5281

TABLE 7.6 (cont.)

Individual 61	8.1288	4.7200	5.9854	9.0091
Individual 62	12.2200	11.1717	10.9078	6.7778
Individual 63	11.5157	10.0393	9.7395	5.6303
Individual 64	12.1057	11.3125	11.2704	6.2675
Individual 65	10.3801	6.8835	7.8770	11.0541

7.2.1 Optimal number of groups

The determination of the optimum number of groups in a given sample of soil individuals (profiles) classified by an agglomerative strategy may be done in two ways.

(1) On the basis of the structure of the dendrograms produced by the classificatory strategy.

(2) Using a statistical criterion.

The structure of the dendrogram produced by the Ward's ESS method (Fig. 7.5(b)) suggests the possible existence of seven groups, which can be easily identified. But the dendrogram produced by the average linkage method (Fig. 7.5(a)) does not show well defined clusters and cannot be used effectively to determine the number of groups. The main danger in the use of Ward's ESS method is that it could impose artificial structures when the whole sample is from the same population. Therefore the second method is of considerable importance as an independent criterion. When the ΔG^2 is plotted against G(X axis) two breaking points at four and seven group levels can be seen (Fig. 7.8). The first minima of ΔG^2 was considered here as to indicate the number of groups.

7.2.2 Reallocation

The Mahalanobis distance D^2 between the individuals and the group centroids (Table 7.6) was calculated using thirty numerical attributes (15 properties measured at the two uppermost depth levels). The individuals which are closer to the centroids of groups other than their parent

* The individuals transferred are not considerably different from their parent groups, as can be seen from Table 7.6. They can be considered as lying between the two groups (parent group and the new group) near the half-way point in descriminant space. The soil group that most of the individuals transferred from is mainly Brown Earths and soils with gleying features.

The four groups produced by numerical methods are not distinct with respect to pedological criteria defined by the Soil Survey of England and Wales (1967). All groups identified by the Soil Survey have split into several groups. and also different soils have combined. Only the Soil Series Gade, Charity, Swanmore, Lyminster and Calcetto remained undivided. The main conclusion that can be drawn from these results is that the pedologically distinct soil groups as defined by the Soil Survey of England and Wales (1967) are not homogeneous with respect to the soil properties used in this study.

groups were transferred. Only five individuals were transferred as the rest remained unchanged. Here the soil profile was considered as an individual (operational taxonomic unit - Sneath and Sokal, 1973, pp.68-71) rather than a group of a series of depth levels which could make it difficult to perform reallocation since the depth levels of a given soil profile could be allocated to more than one group. The following individuals were transferred.

INDIVIDUAL 1 TRANSFERRED FROM GROUP 1 TO GROUP 3
INDIVIDUAL 17 TRANSFERRED FROM GROUP 3 TO GROUP 2
INDIVIDUAL 32 TRANSFERRED FROM GROUP 3 TO GROUP 2
INDIVIDUAL 33 TRANSFERRED FROM GROUP 3 TO GROUP 2
INDIVIDUAL 37 TRANSFERRED FROM GROUP 3 TO GROUP 2

*Canonical analysis was performed on the new classification and the members of the four groups were plotted using the first two canonical vectors which account for 87 percent of the variance (Fig. 7.9). It can be seen from the canonical diagram that the groups 1 and 4 are well separated along the first canonical axis while the groups 2 and 3 show a considerable overlap. The vector diagram (Fig. 7.10) indicates that the attributes 1 (perc. silt), 8 (exch. Ca), 12 (perc. moisture), 17 (perc. clay), 22 (perc. base saturation) and 23 (exch. Ca) have the highest contribution to the first two canonical vectors. Therefore, these attributes are the best discriminators for this population. Although the groups 2 and 3 tend to overlap, they occupy distinct regions of the canonical space.

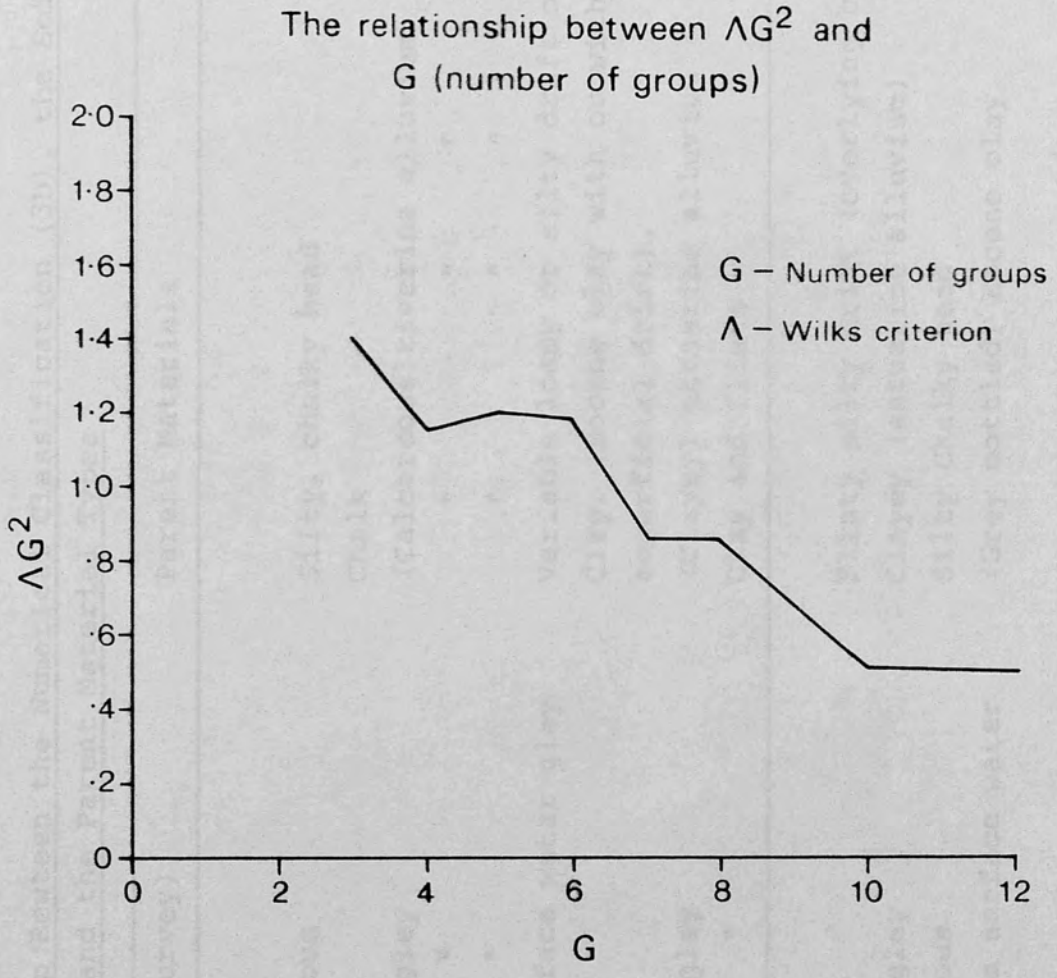

The relationship between ΛG2 and
G (number of groups)
G – Number of groups
Λ – Wilks criterion
ΛG2
2·0
1·8
1·6
1·4
1·2
1·0
·8
·6
·4
·2
0
0
2
4
6
8
10
12
G

Figure 7·8

TABLE 7.7 - The Relationship Bewteen the Numerical Classification (3b), the Soil Survey Classification and the Parent Material Types

Individual No.	Group (Soil Survey)	Parent Materials
	GROUP 1	
5	Brown Calcareous	Silty, chalky head
10	Rendzinas	Chalk
22	Ground-water gley	(Calcareous) riverine alluvium
23	" " "	" " "
24	" " "	" " "
26	Non-calc. surface water gley	Variable loamy or silty drift over Eocene Clay. (Eocene clay with or without superficial drift).
34	Ground-water gley	(Clayey) estuarine alluvium
35	" " "	Clay and flints
	GROUP 2	
9	Brown earths	Flinty silty drift (overlying clay + flint)
12	Ground-water gley	Clayey (estuarine alluvium)
17	Brown Calcareous	Silty Chalky head
20	Non-calcareous surface water gley	(Grey mottled) Eocene clay
27	Non-calc. gley	(Loamy) Eocene beds

TABLE 7.7 (page 2)

Individual No.	Group (Soil Survey)	Parent Materials
	GROUP 2 cont.	
28	Brown earths	Flinty silty head
29	Podzols	Sandy Eocene beds
30	Non-calc. gley	Variable loamy or silty materials (Eocene Clay)
31	Brown calcareous	Flinty, silty head
32	Brown Earths	Chalk
33	Rendzinas	"
36	Brown Earths	Clay with flint
37	" "	Flinty, silty head
38	" "	" " "
39	Non-calc. gley	Brick earths
42	Brown earths	Clay with flints overlying chalk
43	" "	Flinty silty heads overlying clay with flints
44	Non-calc. surface water gley	Disturbed flinty remnants (flinty silty head)
45	Brown earths	Clay + flints
54	" "	Flinty silty head
60	Non-calc. surface water gley	Grey with mottled Eocene clay
61	" " " " "	Eocene clays
65	Ground-water gley	Clayey estuarine alluvium

TABLE 7.7 (page 3)

Individual No.	Group (Soil Survey)	Parent Materials
	GROUP 3	
1	Rendzinas	Chalk
2	Brown Earths	Brickearths
3	Non-calc. gley	"
4	Brown Earths	"
6	" "	"
7	" "	"
11	Non-calc. with gleying	"
13	Non-calc. gley	"
14	" "	"
15	" "	"
18	Brown earths + Gleying	"
19	Non-calc. surface water gley	Variable flinty silty drift over Eocene clay
25	Brown earths	Brickearths
40	Non-chalk gley	"
41	Brown calcareous	Clay with flint over chalk
46	Brown earths	Brickearths
47	" "	"
48	Brown earths with gleying	"

TABLE 7.7 (page 4)

Individual No.	Group (Soil Survey)	Parent Materials
	GROUP 3 cont.	
49	Brown earths	Brickearths
50	" "	"
51	" "	"
52	Non-calc. with gleying	"
53	" " "	Loamy, pebbly marine drift
55	" " "	Brickearths
56	" " "	"
	GROUP 4	
8	Brown earths	Loamy pebbly marine drift
16	Non-calcareous gley	" " " "
21	Brown earths	Flinty, silty drift
57	Non-calcareous gley	Loamy, pebbly, marine drift
58	Brown earths	" " " "
59	Non-calcareous gley	" " " "
62	Brown earths with gleying	Loamy Eocene beds
63	Non-calcareous gley	" " "
64	Podzols	Sandy Eocene beds

Four groups obtained from Ward's ESS method with Mahalanobis distance as the similarity measure after reallocation

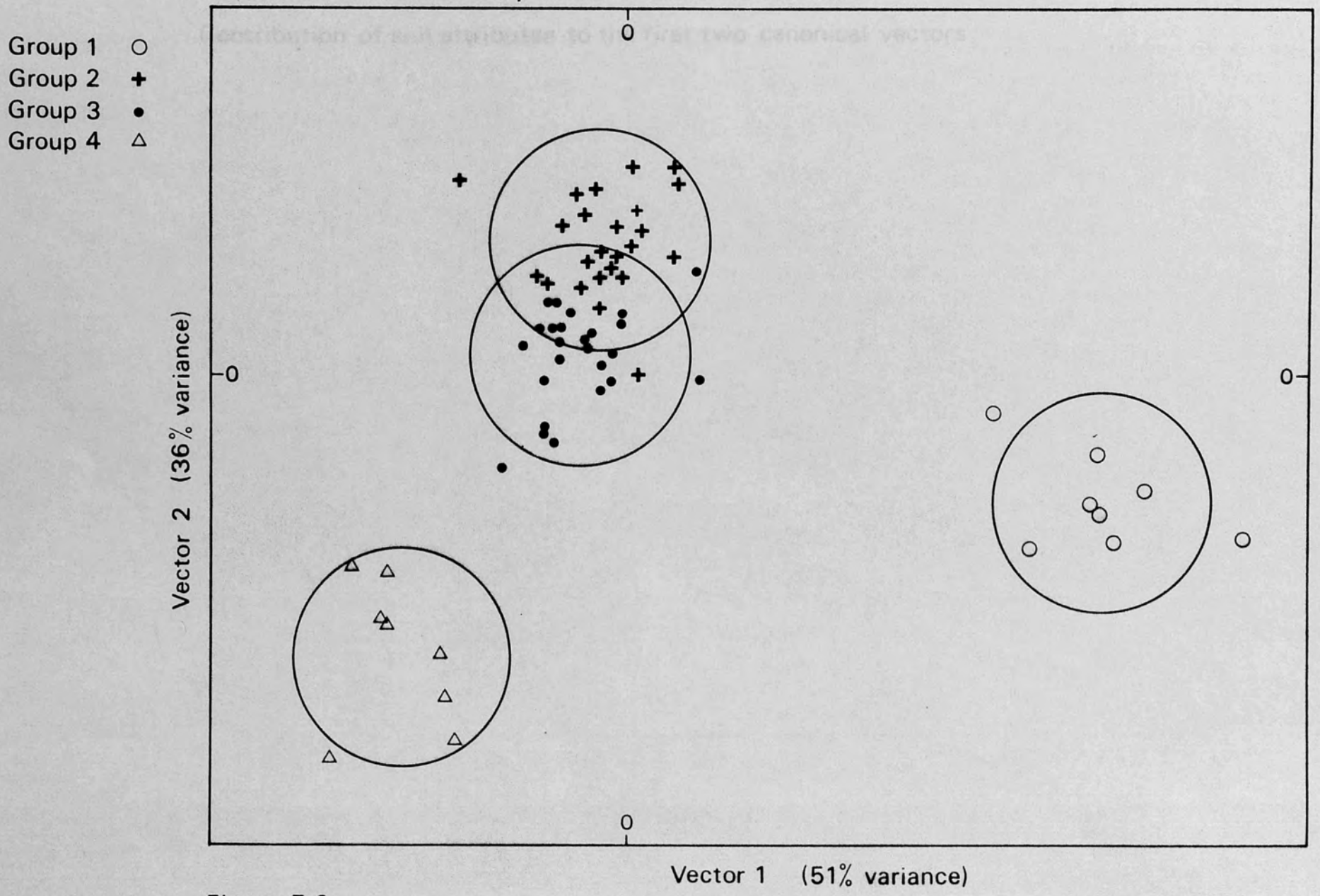


Figure 7·9

Contribution of soil attributes to the first two canonical vectors

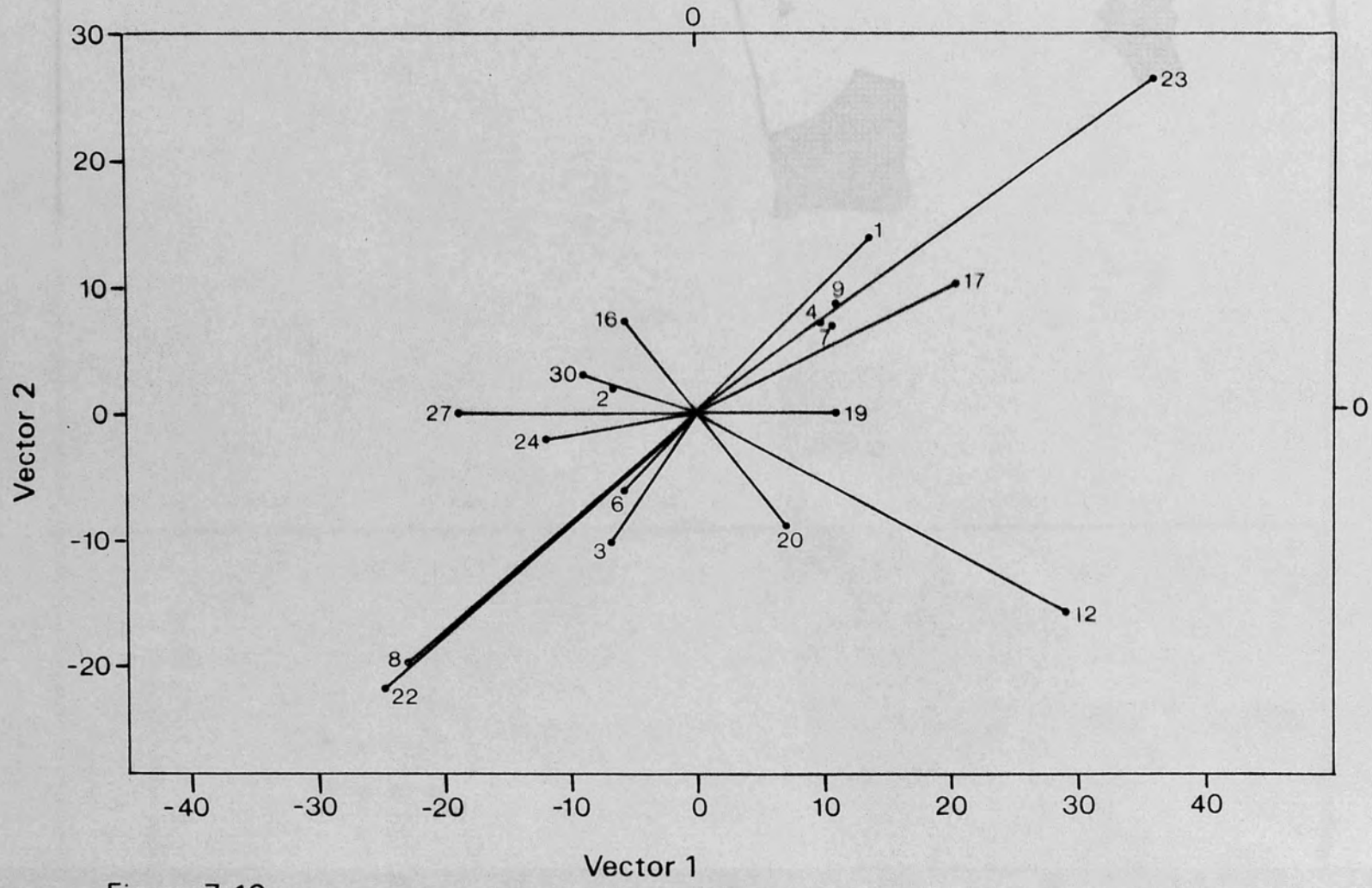


Figure 7·10

Distribution of soil groups produced by numerical classification

Figure 7·11

7.3 Discussion and conclusion

The Soil Series is the basic unit of classification and mapping as used by the Soil Survey of England and Wales (Avery, 1973). It is assumed that the soil series is made up of homogeneous soils derived from homogeneous parent material occupying a geographical area. But as shown by the Wilk's Criterion Λ, the twenty two Soil Series of the West Sussex Coastal Plain are not composed of homogeneous soils as far as the fifteen numerical attributes are concerned. The classifications obtained by the numerical taxonomic methods form more homogeneous groups at a level of producing fewer groups. This may suggest that the use of Soil Series as a mapping unit in the West Sussex Coastal plain was not satisfactory if the homogeneity of the mapping unit was a major priority.

The main objective of a numerical taxonomic classification is to identify a manageable number of groups which are homogeneous in their intrinsic properties. Therefore the most satisfactory classification is the one which has the lowest within-group variance. In this numerical analysis the Soil Series concept was ignored and the soil profile was used as the basic unit of classification. The soil profile has an area of one square metre and therefore the identification of the soil boundaries cannot be done without further field investigations. Since the soils are three dimensional bodies the same weight is given to all depth levels.

The Wilk's Criterion Λ values computed for the classifications (Table 7.5) reveals four things.

(1) The clustering strategy, Ward's ESS method produced more homogeneous groups than the average linkage method irrespective of the similarity measure, despite the high degree of distortion to the similarity matrix.

(2) The Mahalanobis distance D^2 is more successful in measuring the relative similarity between soil individuals (profiles) than the squared Euclidean distance and the Canberra metric.

(3) The classification obtained by Ward's ESS method with the Mahalanobis distance D^2 as the similarity measure is capable of identifying more homogeneous groups than the other methods discussed.

(4) All the classifications obtained by the numerical taxonomic methods have a lower Λ than the Soil Survey (1967, p.41) classification into eight Groups.

The similarity measure is of fundamental importance to any agglomerative strategy. The Euclidean distance can be used to measure the relative distance between individuals if the attribute vectors define an Euclidean space (attribute vectors should be mutually orthogonal), but with real data this can rarely be achieved. Therefore the Mahalanobis distance is not affected by the inter-attribute correlation is far more effective in determining the relative similarity between soil individuals.

It has been demonstrated (in chapter 5) that the two agglomerative classificatory strategies (average linkage method and Ward's ESS method) are different from each

other with respect to the degree of distortion introduced to the original similarity matrix. The soil universe is a continuum with groups or sub-sets grading from one type to another through an intermediate phase which would give rise to an overlap between groups. As has been suggested by Webster (1968) it is not possible to define mutually exclusive soil groups. Therefore the classificatory strategy must be capable of isolating the clusters as clearly as possible. The average linkage method appears to have been affected by the intermediate types of soil individuals giving rise to a confused structure to the dendrograms. The Ward's ESS method proceeds with the fusion of individuals and groups by minimizing the loss of information due to fusion of groups. The information loss is quantified by an objective function (Ward, 1963). This method tends to minimize the within group variance as indicated by the Wilk's Criterion Λ. The soils of the West Sussex Coastal Plain are less diverse than the previous sample (chapter 6) and therefore it can be expected that there would be a considerable overlap between groups. In such a situation identification of groups can only be done by an intensely clustering strategy. It can be seen from Figure 7.5(b) that there are seven clearly defined clusters. But it is not possible to say that this represents the optimal number of groups in the sample. Therefore, it is necessary to examine the classification and try to improve it using a reallocation procedure. The two-dimensional projection of the sample (canonical plots representing sixty five soil individuals Fig. 7.6) does not show well separated clusters.

But the relative position of the individuals of the canonical plots (Fig. 7.6(a)) is very similar to that of the dendrogram (Fig. 7.5(b)). Since the first two canonical vectors account for a greater part of the variance of the sample this relationship is useful to examine the validity of the classification. The lack of separation of the clusters on the canonical plot (Fig. 7.6(b)) may well suggest that the majority of soil individuals (profiles) are similar.

It was demonstrated that perhaps there were four or seven groups in the sample. This is comparable with the Soil Survey classification at the Major Group and Group level despite the dissimilarity in the composition of the groups. The seven groups obtained by the numerical method 3b have very much lower Λ compared to the eight Groups of the Soil Survey classification indicating the superiority of the numerical classification with respect to the homogeneity of the groups as measured by $W_k (k=1,2,...)$.

The number of attributes should be used in a natural classification is theoretically unlimited, but because of the correlation between attributes it is possible to choose a small number of attributes (Sarkar et al, 1966) to produce a natural classification. In general, soil properties are measured at a series of depth levels and it could increase the number of attributes considerably. As shown earlier (chapter 4) the soil attributes considered here are correlated depthwise, and therefore some attributes are redundant.

Therefore the use of a sub-set of attributes in classifying soils would not seriously distort the classification. In this classification fifteen soil properties (attributes) were chosen as described earlier on the basis of the availability of data.

The success of a soil classification cannot be judged purely in statistical terms. If the new classification leads to discovering more information about the soil, that classification has a greater use for the pedologist as a frame of reference leading to further understanding of soil. In this context it is possible to examine the relationship between the soil classification 3b and the soil parent materials. The soil properties chosen for this study were expected to reflect the nature of soils. In the previous sample (chapter 6) the soil profiles consisted of upto seven depth levels extending to a depth of 1 - 2m but only four depth levels within the uppermost one metre layer were used here. Therefore the West Sussex data are less likely to have included depth levels reflecting the parent material. The relationship between the soil groups and parent materials is of great importance. The alteration of parent materials by pedogenetic processes takes place over a long period of time and older the soil is the greater the difference from the parent material. Although the soil parent material plays an important role in determining the properties of the soil,the influence of the other environmental factors, notably the climate, can obscure

such a relationship. The identification and the definition of the Soil Series is done taking the parent material into consideration, but the higher categories are defined independently of the parent material and thus the Major Groups and the Groups can contain soils derived from various types of parent materials. This study suggests that it is possible to consider parent materials even when the higher categories are defined although not all soils derived from a given type of parent materials necessarily belong to the same group. Therefore a primary classification of soils is necessary to assess the importance of the parent material to identification of the soils.

The relationship between the final classification and the parent material types is illustrated by Table 7.7. There is a tendency to group soils derived from the similar parent materials. For example, the soil group 3 is mainly derived from Brickearths, and the majority of soils in the group 4 have developed in loamy pebbly marine drift. Group 1 is dominated by chalky parent materials.

CHAPTER 8

CONCLUSIONS

Classification of soils by subjective methods has been a major problem in communicating soil information among pedologists in different parts of the world. The recent development of numerical taxonomy and the ready availability of powerful electronic computers have provided an alternative approach to soil taxonomy. The use of numerical taxonomic methods in soil classification involves two important decisions:

(1) a method of soil characterization compatible with numerical taxonomic strategies,

(2) a classificatory strategy.

In this study, these two important aspects were investigated. The soil profile is generally used as the basic unit of pedological studies and therefore, can be used as the basic unit of soil classification. It was shown that the fitting of mathematical curves produced little additional information about soils. Multilevel observations of soil properties can be used as attributes to characterize soils (soil profiles) but the use of all such attributes may not be necessary because of inter-attribute correlation. Therefore, a sub set of attributes can be selected for classification and also for further statistical analysis. Correlation between the same property measured at different depth levels is higher than between

totally different soil properties. As a result, it is possible to use mean attribute vectors of soil profiles to characterize soil individuals (profiles of soil) with little information loss. However, correlation between different soil properties is high enough to distort the vector space and therefore the similarity measure selected should be insensitive to the inter-attribute correlation.

Classifications produced by hierarchical agglomerative strategies may not be suitable for soil classification because soil individuals are not hierarchically related and also there is no way of correcting misclassified individuals while retaining the hierarchical nature of the classification. However, a hierarchical agglomerative strategy can be applied on the population to obtain an initial partition which can be more satisfactory than an arbitrary partition. The most important aspects of these strategies are (a) goodness-of-fit, and (b) the cla rity of clusters. Goodness-of-fit of an agglomerative strategy can be measured by cophenetic correlation r_c. The seven agglomerative strategies compared in this study (chapter 5) fall into two classes on the basis of goodness-of-fit:

(a) the strategies with comparatively high degree of distortion, and (b) the strategies with minimum of distortion. An inverse relationship exists between goodness-of-fit and the clarity of clusters. Therefore, goodness-of-fit should not be the most important criterion on the choice of the clustering strategy. Ward's ESS method and average linkage method were chosen respectively from the two classes of agglomerative strategies mentioned.

Ward's ESS method produced the best initial partition of both populations of soils irrespective of the similarity measure. However, the similarity measure is proven to be of great importance since the relative similarity between soil individuals could be distorted by the presence of correlated attributes. Mahalanobis D^2 performed better than the Euclidean distance which is sensitive to inter-attribute correlation. The classifications obtained from the Ward's ESS method with the Mahalanobis D^2 as the similarity measure had the lowest Wilk's Λ, indicating greater within group homogeneity.

The final classifications obtained for both populations showed a good association between soil groups as defined in this study and parent material classes. USDA data shows that soils developed in Loess parent material tend to group together when numerical taxonomic methods were applied. It is also interesting to note that soils developed in uniform parent materials are more homogeneous in their characteristics; they are tightly clustered in dendrograms and have comparatively high inter-individual similarity. This relationship is clearer in the classification obtained from numerical methods for the West Sussex Coastal Plain. The majority of soil profiles sampled from Brickearth areas have grouped together, and also they have somewhat tightly clustered in the dendrograms and canonical plots. Although these results cannot be extrapolated to all situations they nevertheless are of great importance to the soil taxonomist. There is a

* When the relationship between the four soil groups and parent material types is translated into geographical terms, a spatial relationship can be established. The soil groups (constructed by numerical taxonomic methods) appear to occupy geographical areas determined by the distribution of parent materials (Fig. 7.11). The full distribution of the four soil groups in the West Sussex Coastal Plain cannot be established on the basis of the available information. It would be necessary to have a spatially representative sample of soil profiles to test the validity of this suggested relationship.

possibility that uniform parent materials in a comparatively small area could produce uniform soils. Therefore, the study of parent materials prior to the soil survey may be very important.

Previous workers, Webster (1976) for example, have shown that soil groups are spatially dependent, in that geographically closer soils tend to have high chance to fall into the same group . This relationship may have resulted from the dominance of local factors in the formation of soils. The classification obtained from the numerical methods for the West Sussex Coastal Plain shows such a relationship.

The numerical classification obtained for the West Sussex Coastal Plain is superior to that of Soil Survey of England and Wales (1967) classification at Group and Major Group level as far as the within group homogeneity is concerned. It shows that it is possible to reduce the number of groups while retaining a high degree of internal homogeneity. This was supported by the fact that Wilk's Λ for seven groups produced by the Ward's ESS method was lower than that of the classification of soils into twenty two Soil Series.*

Spatial dependency of soil groups of the West Sussex Coastal Plain illustrates an important area for further research. In this study, results of the numerical classification of soils of the region mentioned points at a possible association between soils, parent materials and the local geomorphology. This may not be applicable to all areas and all soils but it is of great interest as it emerged from the

classification of soils. The nature of soil parent materials may or may not be related to the present state of land forms of an area. The dominance of such factors on the genesis of soils, recent soils in particular, can be expected.

The main problem of applying numerical taxonomic methods to soil classification has been the lack of data compatible with numerical strategies. Therefore, sampling and determination of soil properties require particular attention. The conventional procedure of sampling from genetic horizons makes it difficult to compare different soils. Sampling from standard depth levels is more useful to study the vertical variation of soil properties.and mathematical curves may be fitted to describe such variations. Although curve fitting has value in describing variation, the present study indicates that such a complex method is seldom justified by any improvement in soil classification. This method, however, is a useful tool for characterization of soil groups produced by numerical methods. The use of numerical methods highlights the need for data on soils to be both complete and compatible in terms of the methods of determination which are used; this is more important than achieving a great accuracy of determination. Simple, quick and standardized methods should be used for soil description in the field and the laboratory.

Usually, at present, soil profiles are sampled by most soil surveys to represent the mapping units such as Soil Series subjectively identified in the field through a 'free-survey' procedure. This makes it difficult if not impossible to establish the real distribution of soil properties, information which is required for unbiased soil classification and soil mapping. Data collected through traditional free-survey and used for intuitive classifications

cannot form a satisfactory basis for modern, numerical taxonomic procedures. Many of the differences apparent between numerical and non-numerical classification (discussed in Chapters 6 and 7) result from the particular form that the soil data take, and from the assumptions built into the conventional classification.

In the recent past, soil survey authorities have taken steps to computerize soil information storage increasing the flexibility of information retreival and processing. But the way in which such information is gathered has changed little. Both site sampling and the sampling from the soil profile should be stadardized. Unbiased site sampling should be representative of the survey area under consideration and the soil description should include as many properties as possible. Soil properties need to be determined on samples collected from standard depth levels in contrast to the present method of sampling by horizon. Field observations and to some extent laboratory data are used to define soil mapping units (e.g. Soil Series). Such mapping units are useful only if they are homogeneous with respect to the intrinsic properties of soils. Numerical methods can be used to study the spatial variation of soils if soils are sampled to cover the survey area adequately and objectively. But free-survey methods, now generally in use, do not provide sufficient coverage of the area to be mapped. A knowledge of the variability of soils with respect to their intrinsic properties is invaluable for the numerical taxonomist to determine the variability of soil properties that can be allowed in defining the taxa of the classification.

APPENDIX I

Data for 7 soil horizons (USDA 1975 and De Alwis 1971)

I I I	Silt %	Clay %	Org. C %	Fe %	Exchangeable Cations me/100g soil			pH 1:1	C.E.C. me/100g soil	I I I
	56.9	33.2	1.17	.7	14.7	1.6	.1	1.9	6.1	22.58	1A
	48.3	46.5	.83	.8	23.4	7.3	.3	1.8	6.9	30.69	1B
	52.8	42.7	.54	.7	22.7	6.2	.6	1.7	7.8	29.46	1C
	62.2	34.1	.35	.7	20.8	5.9	1.2	1.6	8.1	25.58	1D
	65.1	29.2	.27	.7	21.4	5.3	2.2	1.7	7.9	24.24	1E
	65.2	27.9	.24	.6	22.6	4.2	2.7	1.7	7.9	22.32	1F
	62.9	28.2	.28	.8	20.5	3.5	2.2	1.7	7.9	22.56	1G
	72.3	25.2	2.17	.9	17.1	3.1	.1	.4	6.7	20.16	2A
	68.7	28.5	1.21	1.0	12.6	4.7	.1	.3	5.6	20.81	2B
	61.2	36.4	.92	1.1	14.4	7.9	.1	.4	5.7	24.39	2C
	55.9	42.5	.56	1.4	17.3	11.3	.1	.6	5.8	30.17	2D
	60.3	37.1	.39	1.6	15.9	10.9	.1	.5	6.4	27.45	2E
	72.8	24.9	.21	1.4	11.3	7.5	.2	.3	7.5	18.68	2F
	80.9	14.0	.08	1.2	6.4	3.8	.1	.2	7.8	10.22	2G
	29.2	22.1	2.06	*	10.1	4.7	.0	1.1	6.6	16.40	3A
	27.4	25.3	1.29	*	7.8	5.7	.0	.9	6.4	16.80	3B
	24.7	32.4	.97	*	11.6	8.3	.1	.9	6.5	20.80	3C
	21.1	37.8	.37	*	14.0	11.4	.2	.8	6.8	25.10	3D
	27.3	36.8	.74	*	14.9	11.7	.2	.6	7.5	24.20	3E
	34.6	33.5	.67	*	13.3	10.4	.3	.4	8.0	20.10	3F
	36.2	31.1	.50	*	10.5	11.2	.5	.4	8.2	17.90	3G
	30.0	14.7	2.71	*	13.7	4.8	.6	1.0	*	19.30	4A
	33.6	10.8	2.11	*	9.6	4.6	1.8	.05	*	15.40	4B
	34.1	20.0	1.25	*	10.0	8.1	4.9	.7	*	20.50	4C
	27.6	26.1	1.09	*	11.0	10.8	10.0	.7	*	23.20	4D
	25.8	24.8	.61	*	11.6	12.5	8.1	.7	*	19.20	4E
	37.4	36.4	.14	*	*	12.9	6.0	.5	*	14.70	4F
	58.1	25.9	.30	*	*	10.7	4.8	.6	*	14.90	4G
	53.6	31.4	15.43	4.8	6.7	1.2	.4	.4	4.2	78.10	5A
	55.2	32.7	14.77	4.0	2.0	1.1	.3	.5	4.0	*	5B
	58.7	27.0	10.96	4.2	.5	.6	.2	.4	4.1	70.20	5D
	65.6	16.8	6.34	4.1	.5	.4	.2	.2	4.2	*	5D
	52.7	18.1	1.56	4.0	.6	.3	.2	.1	4.4	*	5E
	47.6	15.4	1.76	4.5	.5	.3	.2	.1	4.5	37.90	5F
	43.9	13.3	2.23	5.1	.4	.3	.3	.1	4.5	*	5G
	85.9	12.6	.96	.7	4.0	.6	.0	.4	5.9	7.00	6A
	85.5	13.2	.64	.8	3.9	.5	.0	.2	5.8	6.50	6B
	78.4	20.6	.28	1.2	4.5	1.0	.0	.2	5.3	8.70	6C
	70.0	29.1	.22	1.7	6.0	2.0	.0	.3	4.8	13.20	6D
	69.0	30.0	.18	1.8	7.0	3.3	.1	.4	4.9	15.50	6E

71.4	27.4	.14	1.8	6.3	3.6	.1	.4	4.7	14.90	6F
75.8	23.0	.10	1.8	5.2	3.2	.1	.3	4.8	13.40	6G
78.9	14.3	5.54	1.1	8.0	2.4	.0	.4	4.9	20.00	7A
84.8	8.5	.76	1.0	2.0	.8	.0	.1	4.9	8.10	7B
82.7	12.2	.13	.9	2.6	2.2	.0	.1	5.0	8.90	7C
71.1	24.2	.16	1.0	5.7	5.7	.1	.2	4.8	17.60	7D
67.7	27.6	.13	.9	7.6	7.6	.2	.2	4.8	21.10	7E
67.2	24.6	.12	1.0	7.9	7.5	.2	.2	4.8	20.60	7F
66.6	18.3	.06	1.0	7.2	7.1	.2	.2	4.9	17.70	7G
26.4	8.8	2.00	.6	8.8	4.2	.5	.6	7.2	16.20	8A
26.2	12.2	.83	.8	6.6	3.1	.7	.1	7.3	12.40	8B
25.2	16.3	.30	1.3	10.2	4.0	1.3	.1	7.2	18.10	8C
22.6	24.6	.23	1.2	11.0	4.6	1.6	.1	7.0	20.90	8D
23.7	19.4	.16	1.1	11.3	5.1	1.8	.1	7.0	18.70	8E
14.7	11.4	.03	1.1	16.1	6.1	1.6	.1	7.2	25.10	8F
16.4	7.0	.01	1.0	17.0	4.9	2.0	.1	7.3	24.80	8G
84.2	12.8	.86	1.0	6.7	1.6	*	.4	6.7	8.20	9A
84.7	12.4	.85	.8	5.6	1.8	*	.3	6.3	8.60	9B
76.1	21.7	.33	1.3	7.0	2.6	.1	.3	6.5	10.40	9C
69.3	29.5	.17	1.8	8.2	4.6	.1	.4	6.5	15.60	9D
71.3	27.2	.12	1.7	5.4	4.3	.1	.4	5.6	14.40	9E
76.2	22.5	.10	1.8	3.6	3.4	.1	.3	4.8	12.00	9F
79.2	19.9	.08	1.7	3.5	3.7	.1	.3	4.8	11.50	9G
46.2	4.6	1.42	.6	6.9	1.4	.1	.3	5.4	*	10A
25.6	2.4	.31	.4	1.2	.7	.1	.0	4.7	*	10B
25.7	1.5	.09	.4	.7	.2	.0	.0	4.5	*	10C
27.7	1.9	.10	.6	.9	.3	.0	.0	4.4	*	10D
30.4	.5	.04	.4	.7	.1	.0	.0	4.4	*	10E
21.4	3.8	.05	.6	1.6	.2	.0	.0	4.4	*	10F
19.9	1.0	.04	.5	1.2	.2	.0	.0	4.4	*	10G
27.2	8.7	.89	.4	1.4	.2	.0	.2	4.9	*	11A
26.8	7.4	.38	.3	.7	.1	.0	.1	4.9	*	11B
19.0	33.9	.24	2.6	1.3	.4	.0	.2	4.7	*	11C
19.8	34.4	.19	2.8	1.3	.6	.0	.1	4.6	*	11D
14.8	30.2	.10	2.5	.6	.6	.0	.1	4.7	*	11E
15.7	31.5	.08	2.8	.3	.3	.0	.1	5.1	*	11F
18.7	27.6	*	3.3	.0	.3	.0	.0	4.9	*	11G
42.4	13.8	.82	.6	3.6	3.1	.3	.9	5.6	10.80	12A
35.1	36.8	.97	.7	8.0	11.7	2.5	.9	7.3	22.00	12B
31.3	35.7	.92	.7	9.5	14.0	3.0	1.0	8.0	24.50	12C
25.4	26.0	.54	.5	10.1	10.6	2.7	.7	8.3	17.80	12D
26.6	21.8	.31	.6	7.7	8.0	2.6	.5	8.6	13.90	12E
35.6	28.6	.31	.6	7.4	10.8	4.5	.5	8.3	17.10	12F
35.3	29.3	.31	.7	8.1	12.0	5.5	.6	8.1	19.30	12G
84.0	15.2	.51	.7	7.1	1.2	.2	.1	6.2	11.80	13A
82.2	16.1	.57	.3	2.6	1.0	.3	.1	4.9	10.00	13B
70.3	28.7	.38	.6	4.4	3.2	1.3	.2	4.5	20.20	13C
70.5	28.4	.16	.5	5.2	5.4	3.0	.2	5.8	17.90	13D
69.1	27.8	.05	.7	6.4	7.4	5.8	.2	7.0	23.20	13E
71.1	24.9	.10	.8	7.0	7.8	5.8	.3	7.4	24.00	13F
75.7	21.3	*	.6	9.4	7.1	4.9	.2	7.6	21.00	13G
41.1	3.2	3.54	.9	2.9	.2	.0	.0	4.9	*	14A
41.8	3.1	3.34	1.0	3.6	.2	.1	.0	5.1	*	14B

39.4	3.3	.76	.2	1.9	.2	.0	.0	5.1	*	14C
44.7	2.7	3.86	1.9	2.4	.3	.0	.0	5.0	*	14D
48.1	2.3	2.90	1.6	1.7	.1	.0	.0	4.9	*	14E
42.8	1.7	1.23	.3	.5	.1	.0	.0	4.9	*	14F
34.0	1.3	.27	.3	.5	.1	.0	.0	4.9	*	14G
49.7	30.5	1.51	1.5	15.1	4.4	.1	.4	6.4	21.50	15A
46.2	32.8	1.42	1.5	14.9	4.3	.2	.4	6.4	21.40	15B
43.8	32.0	.94	1.4	11.8	4.1	.2	.3	5.3	20.20	14C
45.9	33.3	.70	1.7	12.2	4.5	.2	.4	5.4	20.30	15D
45.4	33.7	.59	2.0	12.6	5.0	.2	.4	5.6	21.40	15E
19.2	24.6	.29	1.7	7.7	3.3	.2	.2	5.7	13.50	15F
19.5	23.2	.24	1.6	7.2	3.0	.2	.3	5.9	12.10	15G
17.6	77.1	4.30	9.4	8.3	1.5	.1	.9	5.6	20.50	16A
17.2	76.8	1.75	9.8	1.8	.2	.0	.5	4.7	10.60	16B
13.0	83.9	.70	10.1	1.4	.2	.0	.1	4.8	7.50	16C
14.9	83.2	.50	11.5	.5	.6	.0	.0	5.0	7.70	16D
23.9	70.4	.25	14.8	.1	.1	.0	.0	5.1	7.60	16E
35.7	52.7	.16	14.3	.0	.2	.0	.0	5.1	8.40	16F
36.6	48.4	.12	14.0	.1	.1	.1	.0	4.9	8.90	16G
71.4	14.3	3.58	*	.2	.3	.0	.2	3.9	*	17A
70.2	19.7	.58	*	.1	.1	.0	.1	4.3	*	17B
65.2	26.1	.51	*	.1	.2	.0	.1	4.4	*	17C
57.2	35.5	.35	*	.1	.6	.0	.1	4.4	*	17D
49.3	43.4	.35	*	.1	1.2	.0	.2	4.3	*	17E
52.0	34.2	.45	*	.1	1.5	.1	.1	4.4	*	17F
47.3	23.6	.20	*	.1	1.2	.0	.1	4.4	*	17G
43.8	40.5	1.48	5.9	9.4	3.0	.3	.1	4.6	20.70	18A
43.2	42.0	1.19	6.5	9.0	3.6	.4	.1	5.1	19.90	18B
40.6	48.4	.54	4.9	5.1	4.4	.6	.2	4.8	18.50	18C
39.7	53.2	.32	5.2	4.1	7.8	1.1	.3	4.8	21.70	18D
42.9	38.0	.10	5.7	4.5	11.9	1.7	.3	5.1	21.81	18E
44.9	49.2	.09	4.8	6.4	15.4	2.2	.4	5.6	25.00	18F
52.4	41.0	.06	5.2	6.5	14.6	2.1	.3	5.8	23.80	18G
55.5	24.5	1.79	2.0	7.8	.9	.0	.3	5.0	*	19A
22.6	65.0	.61	4.3	23.0	1.6	.0	.3	5.5	*	19B
22.5	66.3	.24	4.5	20.0	1.7	.0	.3	5.0	*	19C
19.3	68.0	.17	4.8	16.3	1.2	.1	.3	4.8	*	19D
20.9	64.4	.10	4.8	17.3	1.2	.0	.3	4.7	*	19E
22.7	63.3	.02	4.3	20.2	1.2	.0	.3	4.8	*	19F
27.3	61.7	.12	4.3	24.0	1.2	.1	.3	5.2	*	19G
45.9	12.5	1.13	.8	3.6	4.6	.8	.2	6.0	12.60	20A
41.8	21.3	.41	.9	1.6	9.4	.5	.2	8.0	17.50	20B
40.6	24.0	.15	.9	1.6	10.4	11.8	.2	9.1	17.30	20C
41.5	34.7	.11	1.2	4.3	17.4	17.7	.3	9.4	24.70	20D
44.5	42.2	.10	1.4	4.5	18.5	17.6	.3	9.5	28.10	20E
37.7	29.5	.06	1.0	8.5	15.2	10.6	.3	9.8	23.60	20F
33.0	30.5	.05	1.0	15.3	15.2	6.4	.3	9.7	25.60	20G
33.8	10.4	.69	.6	4.0	2.5	.2	.5	6.4	9.80	21A
32.4	11.5	.36	.6	4.7	1.8	.2	.2	5.7	9.50	21B
24.2	37.2	.32	1.2	13.9	9.5	1.0	.4	6.6	25.30	21C
22.3	30.6	.14	1.3	12.8	8.8	1.4	.3	7.0	23.80	21D
21.5	23.4	.07	1.1	10.6	7.9	1.8	.2	7.2	26.70	21E
43.2	28.1	.07	1.3	16.5	11.8	2.9	.2	7.4	32.40	21F

19.6	18.6	.03	.8	10.2	6.9	1.6	.2	7.7	19.50	21G
18.7	5.6	.35	.5	2.2	1.1	.2	.8	6.3	4.20	22A
23.6	10.3	.29	.6	2.8	1.6	.2	1.0	6.0	5.70	22B
28.7	13.0	.26	.7	4.0	2.2	.2	.8	6.6	6.90	22C
32.5	17.2	.26	.8	5.0	2.3	.3	.9	7.1	8.50	22D
22.5	50.4	.20	1.2	13.4	9.0	1.2	1.4	7.4	24.00	22E
25.0	47.7	.28	1.1	16.7	10.9	1.7	1.5	7.9	24.90	22F
37.9	25.9	.36	.8	24.2	12.3	3.0	1.9	7.9	26.00	22G
23.0	8.3	.40	.3	16.8	1.6	2.0	1.0	8.7	8.20	23A
34.4	9.6	.29	.4	13.0	1.8	4.1	.8	9.2	9.50	23B
29.3	29.3	.38	.4	16.3	2.0	17.2	.9	9.0	20.80	23C
36.5	24.3	.13	.2	11.2	.5	27.9	.9	9.9	20.30	23D
30.1	28.3	.07	.1	10.6	.6	21.7	.6	10.0	13.00	23E
14.0	16.2	.03	.2	8.4	.8	19.8	.5	9.9	14.70	23F
14.1	10.2	.01	.3	3.6	.4	12.4	.2	9.7	11.30	23G
28.7	5.4	.25	*	2.5	1.2	1.0	.8	5.7	5.20	24A
28.5	9.2	.33	*	3.9	1.5	1.0	.7	6.3	6.70	24B
27.2	16.7	.43	*	7.1	2.6	1.0	.9	7.0	10.20	24C
15.0	58.7	.34	*	26.0	9.5	2.9	2.4	7.5	32.60	24D
10.4	48.1	.20	*	28.3	8.3	4.2	2.6	7.9	29.60	24E
15.2	23.7	.04	*	15.3	4.5	4.4	2.0	7.7	19.10	24F
6.1	10.9	.02	*	7.8	2.1	4.0	1.8	7.6	14.00	24G
17.1	10.2	.96	1.0	10.5	3.4	1.4	.2	6.9	16.90	25A
18.5	10.6	.56	.9	10.5	2.8	.4	.1	6.6	18.60	25B
18.2	11.5	.34	1.0	13.0	4.0	.4	.1	6.8	19.80	25C
19.4	11.9	.28	.9	11.1	4.0	.4	.1	6.7	19.10	25D
18.5	10.9	.10	1.1	13.5	5.1	.4	.1	6.9	22.80	25E
12.5	5.2	.03	.7	9.1	3.7	.5	.0	7.1	14.90	25F
12.5	4.6	.03	.8	9.0	3.7	.7	.2	7.4	16.20	25G
69.7	27.9	2.78	*	21.4	4.1	.0	.4	6.9	25.30	26A
65.7	31.6	1.88	*	15.8	5.1	.1	.5	6.1	24.50	26B
62.1	35.2	1.32	*	15.8	6.3	.1	.6	6.0	25.30	26C
55.6	40.4	.92	*	18.0	8.1	.2	.7	6.0	30.40	26D
55.1	42.6	.48	*	19.5	8.5	.3	.8	6.2	33.40	26E
60.0	37.6	.25	*	18.4	8.4	.3	.7	5.8	30.40	26F
60.8	37.6	.18	*	18.2	8.8	.3	.7	6.2	29.80	26G
29.9	9.1	1.82	.7	7.2	2.8	.1	.8	6.4	11.60	27A
27.6	10.5	.93	.9	6.5	3.5	.4	.6	6.9	10.70	27B
24.1	9.5	.68	.8	5.6	3.1	1.2	.4	7.7	9.70	27C
21.4	6.1	.47	.9	3.5	2.4	2.0	.3	8.4	7.20	27D
17.1	14.0	.41	1.3	3.7	4.5	5.9	.5	9.0	11.80	27E
14.3	19.5	.48	1.2	8.2	6.9	10.4	.6	8.6	15.60	27F
8.3	9.3	.21	.9	5.2	3.2	5.9	.3	8.5	8.40	27G
69.2	28.6	2.35	.6	13.9	3.4	.1	.5	5.7	20.60	28A
66.0	32.2	1.95	1.1	13.8	4.2	.1	.4	5.8	21.40	28B
64.0	34.2	1.42	1.1	14.3	5.8	.1	.4	5.7	23.80	28C
62.8	35.6	.97	1.1	14.6	6.4	.1	.4	5.8	24.00	28D
63.0	35.4	.68	1.2	15.0	6.8	.1	.4	5.7	23.60	28E
65.2	33.2	.45	1.2	14.7	6.6	.1	.3	5.7	22.80	28F
67.9	30.5	.34	1.2	14.3	6.6	.1	.3	5.7	21.70	28G
66.6	30.4	2.20	1.3	13.9	4.3	.1	.8	5.6	22.00	29A
63.9	33.5	1.87	1.3	14.7	5.9	.1	.6	5.7	22.90	29B
64.4	32.8	1.11	1.3	14.8	6.3	.1	.6	5.8	21.60	29C

66.4	30.4	.58	1.3	14.8	6.6	.1	.6	5.8	20.00	29D
67.7	28.2	.33	1.4	14.7	6.6	.1	.5	5.9	20.70	29E
69.0	26.9	.21	1.3	14.6	6.8	.2	.6	5.9	20.40	29F
68.2	28.0	.17	1.3	15.2	6.9	.2	.6	6.0	20.70	29G
67.4	23.1	2.54	.5	12.3	3.7	.0	1.1	5.9	18.20	30A
59.3	33.2	1.80	.6	15.4	5.7	.1	1.3	5.9	22.40	30B
51.8	42.6	1.24	.7	18.7	7.6	.5	1.8	6.0	27.20	30C
50.2	44.7	.75	.6	17.5	8.3	.5	2.0	6.6	28.30	30D
56.8	36.8	.38	.5	16.2	8.1	.6	1.9	7.1	25.70	30E
69.5	23.2	.23	.5	14.6	8.1	1.2	1.9	8.1	23.00	30F
67.2	25.0	.18	.5	13.3	7.8	2.1	2.0	8.1	22.90	30G
37.4	26.4	1.43	*	15.4	2.0	.1	1.9	7.8	16.20	31A
35.7	31.6	1.02	*	16.3	1.8	.1	1.4	7.7	15.90	31B
36.1	36.3	.69	*	15.8	1.5	.1	.6	7.8	15.40	31C
37.3	38.0	.48	*	15.3	1.7	.3	.5	7.8	14.80	31D
40.1	37.5	.29	*	13.6	2.2	.3	.5	7.6	12.10	31E
44.0	37.8	.18	*	10.8	2.1	.3	.4	7.6	9.20	31F
45.2	38.0	.13	*	8.2	2.6	.3	.4	7.7	8.00	31G
11.9	2.3	.11	.7	11.7	1.2	.2	.5	*	5.60	32A
14.5	4.9	.08	.7	15.5	1.6	.2	.7	*	7.30	32B
14.0	10.8	.15	.6	18.9	2.9	.2	1.1	*	11.60	32C
6.5	6.5	.20	.1	15.6	3.6	.5	.9	*	12.10	32D
12.8	6.5	.03	.3	20.9	3.2	1.1	1.7	*	14.30	32E
9.4	3.7	.01	.3	16.9	3.1	1.1	1.3	*	12.20	32F
11.7	4.1	.01	.3	19.0	4.5	1.0	.8	*	16.30	32G
6.0	15.0	.62	2.0	37.7	20.8	5.7	0.0	6.6	2.90	T5A1
5.0	26.0	.38	3.0	18.0	18.0	6.0	0.0	5.7	2.60	TPA1
5.0	41.0	.27	4.5	10.8	9.2	6.2	0.0	5.1	3.20	TPB1
4.0	45.0	.18	4.8	9.7	8.1	4.8	0.0	5.3	3.00	TPB2
4.0	51.0	.18	5.0	10.0	12.9	4.3	0.0	4.3	3.90	TPB3
3.0	47.0	.11	4.8	18.2	7.6	4.5	0.0	5.5	3.60	TPB4
7.0	49.0	.12	5.0	32.9	7.1	2.9	1.4	5.7	4.20	TPB5
6.0	21.0	.91	.2	35.7	12.9	4.3	0.0	6.2	3.80	TP2A1
5.6	25.0	.70	.2	17.7	11.3	3.2	0.0	5.5	3.20	TP2A2
4.0	38.0	.23	.3	6.6	6.6	3.3	0.0	5.0	3.00	TP2A3
3.0	44.0	*	*	3.4	15.3	3.4	0.0	5.2	3.40	TP2B1
3.0	50.0	*	*	14.3	11.4	2.9	0.0	5.5	4.20	TP2B2
2.0	50.0	*	*	18.6	10.0	2.9	0.0	5.2	3.90	TP2B3
4.0	49.0	.12	*	26.8	11.3	2.8	0.0	5.4	4.30	TP2B4
3.0	15.0	.68	*	31.4	7.8	5.9	0.0	5.9	2.60	TP3A1
3.0	17.0	.36	*	15.8	5.3	2.6	0.0	5.2	1.70	TP3A2
3.0	31.0	*	*	10.0	6.0	4.0	0.0	5.1	2.70	TP3B1
3.0	41.0	*	*	15.3	15.3	3.4	0.0	5.2	3.50	TP3B2
4.0	41.0	*	*	32.3	9.7	3.2	0.0	5.6	4.0	TP3B3
5.0	40.0	*	*	37.5	8.9	1.8	0.0	5.8	3.60	TP3B4
6.0	43.0	*	*	42.4	9.1	1.5	1.5	5.7	4.30	TP3B5
4.0	11.0	.84	.8	34.0	8.5	4.3	0.0	6.1	2.40	MP1A1
4.0	16.0	.68	1.0	20.8	6.3	4.2	0.0	5.8	2.30	MP1A2
4.0	26.0	.48	1.8	11.4	4.5	2.3	0.0	5.3	2.20	MP1B1
2.0	33.0	.27	2.2	14.9	8.5	2.1	0.0	5.4	2.60	MP1B2
4.0	35.0	.24	2.5	29.2	8.3	0.0	0.0	5.5	2.8	MP1B3
4.0	36.0	.15	2.4	31.9	8.5	0.0	0.0	5.6	2.7	MP1B4
5.0	35.0	.07	2.7	41.9	11.6	0.0	0.0	5.7	2.80	MP1B5

4.0	11.0	1.28	1.0	48.5	12.1	4.6	0.0	6.4	3.80	MP2A1
4.0	16.0	.95	1.3	38.9	11.9	5.1	0.0	6.4	3.00	MP2A2
4.0	25.0	*	2.4	28.3	8.7	4.3	0.0	6.4	2.50	MP2B1
4.0	34.0	*	*	30.6	10.2	4.1	0.0	6.3	2.90	MP2B2
5.0	35.0	*	*	34.0	10.6	4.3	0.0	6.4	2.70	MP2B3
4.0	36.0	.17	*	36.2	10.6	4.3	0.0	6.5	2.80	MP2B4
4.0	39.0	*	*	37.8	13.3	2.2	0.0	6.4	2.80	MP2B5
3.0	12.0	.91	*	46.4	10.4	3.6	0.0	6.5	3.20	MP3A1
4.0	21.0	.78	*	30.0	8.3	3.3	0.0	6.1	3.1	MP3A2
5.0	30.0	*	*	23.5	9.8	3.9	0.0	6.1	2.7	MP3B1
5.0	35.0	*	*	24.0	10.0	2.0	0.0	5.6	2.80	MP3B2
1.0	40.0	*	*	28.9	8.9	2.2	0.0	5.6	2.80	MP3B3
4.0	40.0	*	*	37.2	5.9	0.0	0.0	5.6	3.10	MP3B4
6.0	37.0	*	*	39.1	8.7	0.0	0.0	5.7	2.80	MP3B5
2.7	6.0	.55	*	62.2	21.6	5.4	0.0	6.8	2.40	PP1A1
3.0	7.0	.56	*	53.0	22.0	5.0	0.0	6.8	2.60	PP1A2
3.0	14.0	*	*	35.7	21.4	4.8	0.0	6.4	2.50	PP1B1
3.0	19.0	*	*	27.0	13.5	2.7	0.0	5.5	2.20	PP1B2
6.0	26.0	*	*	40.5	7.1	0.0	0.0	5.6	2.70	PP1B3
4.0	27.0	*	*	35.9	15.4	0.0	0.0	5.8	2.00	PP1B4
*	*	*	*	*	*	*	*	*	*	PP1-5
2.0	8.0	.62	*	57.1	11.9	2.4	0.0	6.3	2.60	PP2A1
2.0	12.0	.49	*	42.5	12.5	2.5	0.0	6.3	2.30	PP2A2
2.0	20.0	.16	*	35.3	11.8	5.9	0.0	6.2	2.20	PP2B1
3.0	21.0	.14	*	31.3	12.5	3.1	0.0	5.9	1.70	PP2B2
2.0	27.0	.11	*	33.3	16.7	2.8	0.0	6.0	2.10	PP2B3
2.0	30.0	.08	*	38.9	8.3	2.8	0.0	5.9	2.20	PP2B4
*	*	*	*	*	*	*	*	*	*	PP2-5
2.0	9.0	.86	*	60.6	12.1	3.0	0.0	6.6	4.30	PP3A1
2.0	12.0	.78	*	51.9	11.5	3.8	0.0	6.7	3.00	PP3A2
2.0	14.0	.92	*	38.5	11.5	11.5	0.0	6.7	1.30	PP3B1
1.0	18.0	*	*	29.0	9.7	6.5	0.0	5.7	1.40	PP3B2
3.0	23.0	*	*	29.0	9.7	6.5	0.0	5.7	1.40	PP3B3
1.0	26.0	.07	*	32.4	11.8	2.9	0.0	5.6	1.70	PP3B4
*	*	*	*	*	*	*	*	*	*	PP3-5

* missing values

APPENDIX II

3-Horizon Model (USDA data)

Texture					Exchangeable cations me/100g soil					
%silt	%clay	%Org.C	%Fe	pH (1:1)	Ca	Mg	Na	K	C.E.C	
56.9	33.2	1.17	.70	6.1	14.70	1.60	0.10	1.90	22.60	1 A
54.1	41.1	0.57	0.73	7.6	22.30	6.50	0.70	1.70	28.70	1 B
64.4	28.4	0.26	0.80	7.9	21.50	4.30	2.40	1.70	23.00	1 C
70.5	26.9	1.69	0.95	6.2	14.90	3.90	0.10	0.40	20.60	2 A
62.6	35.2	0.52	1.40	6.4	14.70	9.40	0.10	0.50	25.10	2 B
83.4	11.5	0.06	1.10	7.9	5.40	3.30	0.10	0.20	8.10	2 C
29.2	22.1	2.06	*	6.6	10.10	4.70	0.00	1.10	16.40	3 A
27.0	33.2	0.89	*	7.0	12.30	9.50	0.20	0.70	21.40	3 C
39.0	29.4	0.39	*	8.4	9.00	11.90	1.00	0.40	17.70	3 C
31.8	12.8	2.41	*	*	11.70	4.70	1.20	0.80	17.40	4 A
29.2	23.6	0.98	*	*	10.90	10.50	7.70	0.70	21.10	4 B
58.9	27.4	0.26	*	*	*	12.20	5.60	0.50	17.10	4 C
58.3	27.0	11.88	4.28	5.2	2.40	0.80	0.30	0.40	74.10	5 A
46.4	14.6	1.95	4.60	5.5	0.40	0.30	0.20	0.10	37.90	5 B
31.6	7.2	1.26	3.40	5.7	0.80	0.40	0.20	.10	38.00	5 C
85.7	12.9	0.80	0.75	5.9	4.00	00.60	0.00	0.30	6.80	6 A
73.9	25.0	0.17	1.65	4.9	5.60	2.70	0.10	0.30	12.90	6 B
81.3	17.7	0.06	1.60	4.7	4.60	3.00	0.10	0.20	10.90	6 C
82.1	11.7	2.14	1.00	4.9	4.20	1.80	0.00	0.20	12.30	7 A
68.2	23.7	0.12	0.98	4.8	7.10	7.00	0.20	0.20	19.30	7 B
34.3	8.4	0.04	1.15	6.0	3.70	2.80	0.10	0.10	7.00	7 C
26.3	11.5	1.47	0.70	7.3	7.70	3.70	0.60	0.40	14.30	8 A
23.8	20.1	0.23	1.20	7.1	10.80	4.60	1.60	0.10	17.10	8 B
15.6	9.2	0.02	1.05	7.3	16.70	5.50	1.80	0.10	24.10	8 C
84.5	12.6	0.86	0.90	6.5	6.20	1.70	*	0.40	8.40	9 A
73.2	25.2	0.18	1.70	5.9	6.10	3.70	0.10	0.40	13.10	9 B
79.2	19.9	0.08	1.70	4.8	3.50	3.70	0.10	0.30	11.50	9 C
46.2	4.6	1.42	0.60	6.0	6.90	1.40	0.10	0.30	15.18	10A
20.4	3.0	0.07	0.55	5.5	1.40	0.30	0.03	0.03	7.43	10B
*	*	*	*	*	*	*	*	*	*	10C
27.0	8.1	0.64	0.38	4.9	1.10	0.15	0.01	0.17	*	11A
14.6	26.6	0.15	2.76	4.9	0.40	0.32	0.03	0.06	*	11B
7.4	17.9	*	0.59	4.7	0.30	0.07	0.02	0.03	*	11C
42.4	13.8	2.82	0.60	6.2	3.60	3.10	0.30	0.90	10.80	12A
30.6	32.8	0.81	0.63	8.4	9.20	12.10	2.73	0.87	21.43	12B
34.6	29.6	0.33	0.77	8.6	7.73	10.85	5.40	0.50	18.53	12C
83.1	15.7	0.54	0.50	5.6	4.85	1.10	0.25	0.10	10.90	13A
70.3	27.2	0.17	0.65	6.2	5.75	5.95	3.98	0.23	21.33	13B
75.7	21.3	*	0.60	7.6	9.40	7.10	4.90	0.20	21.00	13C
40.8	3.2	2.55	0.71	5.0	2.80	0.20	0.03	0	*	14A
45.2	2.2	2.66	1.27	4.9	1.38	0.17	0	0	*	14B
28.0	1.2	0.17	0.25	4.9	0.30	0.10	0	0	*	14C
48.0	31.7	1.47	1.50	6.4	15.00	4.35	0.15	0.40	21.45	15A

38.6	30.9	0.63	1.70	5.5	11.08	4.23	0.20	0.33	18.85	15B
19.5	23.2	0.24	1.60	5.9	7.20	3.00	0.20	0.30	12.10	15[
17.6	77.1	4.30	9.40	5.6	8.30	1.50	0.10	0.90	20.50	16A
23.6	69.2	0.58	12.42	4.9	0.65	0.23	0.02	0.10	8.45	16B
42.2	44.2	0.08	13.43	4.8	0.10	0.13	0.10	0	8.80	16C
70.8	17.0	2.06	1.05	4.1	0.15	0.20	0	0.15	*	17A
57.2	35.0	0.40	2.23	4.4	0.10	0.67	0	0.13	*	17B
40.8	25.3	0.20	1.00	4.4	0.10	0.90	0	0.10	*	17C
43.5	41.3	1.34	6.20	4.9	9.20	3.30	0.35	0.10	19.90	18A
41.0	46.5	0.32	5.27	4.9	4.57	8.03	1.13	0.27	20.67	18B
48.7	45.1	0.75	5.00	5.7	6.45	15.00	2.15	0.35	24.40	18C
55.5	24.5	1.79	2.00	5.0	7.80	0.90	0	0.30	*	19A
22.6	64.8	0.21	4.50	5.0	20.13	1.35	0.03	0.30	*	19B
*	*	*	*	*	*	*	*	*	*	19C
45.9	12.5	1.13	0.80	5.5	3.60	4.60	0.80	0.20	12.63	20A
41.2	30.3	0.17	1.08	8.2	4.10	14.20	11.60	0.30	22.54	20B
34.3	18.3	0.03	1.03	8.5	10.00	14.30	3.80	0.30	16.10	20C
33.1	11.0	0.53	0.60	6.1	4.40	2.20	0.20	0.40	9.74	21A
29.3	28.3	0.11	1.15	7.3	13.50	9.50	1.80	0.30	27.30	21!
*	*	*	*	*	*	*	*	*	*	21C
23.7	9.6	0.30	0.60	6.3	3.00	1.80	0.20	0.90	5.60	22A
29.5	35.3	0.28	0.98	7.6	14.80	8.60	1.60	1.40	20.90	22B
34.2	13.6	0.07	0.54	8.3	18.40	5.90	2.30	1.20	15.90	22C
28.7	9.0	0.35	0.32	9.0	14.90	1.45	3.05	0.90	8.91	23A
29.3	29.3	0.38	0.40	9.0	16.30	2.00	17.20	0.90	20.80	23B
23.7	19.8	0.06	0.23	9.9	8.50	0.58	20.45	0.55	16.43	23C
28.1	10.4	0.34	1.50	6.3	4.50	1.80	1.00	0.80	7.97	24A
13.5	43.5	0.19	1.29	7.7	23.20	7.40	3.80	2.30	28.85	24B
6.1	10.9	0.02	1.38	7.6	7.80	2.10	4.00	1.80	13.95	24C
17.9	10.8	0.62	0.97	6.8	11.33	3.40	0.73	0.13	17.28	25A
19.0	11.4	0.19	1.00	6.8	12.30	4.55	0.40	0.10	19.21	25B
12.5	4.9	0.03	0.75	7.3	9.05	3.70	0.60	0.10	17.71	25C
63.3	33.8	1.73	*	6.3	8.88	5.90	0.10	0.55	26.38	26A
59.9	38.2	0.26	*	6.1	18.48	8.43	0.30	0.70	30.45	26B
67.0	31.9	0.10	*	7.0	17.10	7.55	0.25	0.60	25.43	26C
25.7	8.8	0.98	0.83	7.4	5.70	2.95	0.93	0.53	9.80	27A
12.2	12.9	0.30	1.28	8.9	5.30	4.30	6.93	0.43	10.72	27B
17.2	17.3	0.19	1.13	8.5	5.00	5.67	11.67	0.50	16.40	27C
66.5	32.0	1.91	0.93	5.7	14.00	4.47	0.10	0.43	21.93	28A
66.4	31.9	0.47	1.17	5.8	14.70	6.32	0.10	0.33	22.22	28B
69.0	27.6	0.12	1.10	6.5	16.10	5.60	0.20	0.40	20.10	28C
65.0	32.2	1.73	1.30	5.7	14.47	5.37	0.10	0.67	22.17	29A
67.8	28.4	0.32	1.33	5.9	14.82	6.73	0.15	0.58	20.45	29B
69.8	26.0	0.10	1.37	6.2	14.03	6.87	0.20	0.60	19.60	29C
66.6	15.9	0.85	1.00	6.5	5.70	0.65	*	0.40	6.75	30A
47.3	36.3	0.43	2.22	5.5	5.28	3.30	*	0.32	8.90	30B
*	*	*	*	*	*	*	*	*	*	30C
36.6	29.0	1.23	*	7.8	15.85	1.90	0.10	1.65	16.05	31A
39.4	37.4	0.41	*	7.7	13.88	1.88	0.25	0.50	12.88	31B
45.2	38.0	0.13	*	7.7	8.20	2.60	0.30	0.40	8.00	31C
13.5	6.0	0.11	0.67	*	15.37	1.90	0.20	0.77	8.17	32A
*	*	*	*	*	*	*	*	*	*	32B
9.6	4.6	0.06	0.26	*	16.84	3.44	0.98	1.58	12.56	32C

* Missing data

APPENDIX III

West Sussex Coastal Plain

(a) Quantitative data

I I I %Silt	%Clay	%Org. C	C/N Ratio	%Loss on Ignition	N %	pH 1:2.5	%CaCo 3	C.E.C me/100g soil	%Base Sat.	I I I
58.0	21.0	*	*	6.9	*	7.3	25.8	*	*	Q1.1A
60.0	22.0	*	*	9.0	*	7.3	22.3	*	*	Q1.1B
*	*	*	*	*	*	*	*	*	*	Q1.1C
*	*	*	*	*	*	*	*	*	*	Q1.1D
65.0	19.0	*	*	5.8	*	8.0	.7	*	*	Q1.2A
65.0	22.0	*	*	3.3	*	8.1	.2	*	*	Q1.2B
58.0	27.0	*	*	3.7	*	8.0	.1	*	*	Q1.2C
59.0	32.0	*	*	3.9	*	8.1	.1	*	*	Q1.2D
54.0	26.0	*	*	6.0	*	7.0	.0	*	*	Q1.3A
51.0	29.0	*	*	5.2	*	7.3	.1	*	*	Q1.3B
49.0	31.0	*	*	4.7	*	7.7	.1	*	*	Q1.3C
50.0	35.0	*	*	4.2	*	8.0	.7	*	*	Q1.3D
59.0	17.0	*	*	8.1	*	5.7	*	*	86.0	Q1.4A
60.0	18.0	*	*	4.5	*	6.2	*	*	57.0	Q1.4B
60.0	19.0	*	*	4.1	*	7.3	.2	*	*	Q1.4C
61.0	20.0	*	*	3.4	*	7.2	.1	*	*	Q1.4D
59.0	22.0	*	*	6.7	*	8.0	15.3	*	*	Q1.5A
46.0	28.0	*	*	3.1	*	8.3	44.0	*	*	Q1.5B
46.0	29.0	*	*	.5	*	8.3	47.2	*	*	Q1.5C
*	*	*	*	*	*	*	*	*	*	Q1.5D
59.0	18.0	*	*	8.7	*	5.2	*	20.0	54.0	Q1.6A
65.0	21.0	*	*	3.7	*	5.8	*	12.8	66.0	Q1.62
59.0	34.0	*	*	4.3	*	6.4	*	17.5	83.0	Q1.6C
55.0	33.0	*	*	4.1	*	6.5	*	15.7	83.0	Q1.6D
64.0	20.0	*	*	5.3	*	6.4	*	14.5	78.0	Q1.7A
63.0	23.0	*	*	3.8	*	6.8	*	11.7	81.0	Q1.7B
59.0	31.0	*	*	4.3	*	6.8	*	14.8	85.0	Q1.7C
54.0	35.0	*	*	4.6	*	6.8	*	16.6	86.0	Q1.7D
34.0	27.0	*	*	6.5	*	6.2	*	18.3	78.0	Q1.8A
26.0	28.0	*	*	3.7	*	6.3	*	13.1	79.0	Q1.8B
8.0	33.0	*	*	3.9	*	6.6	*	16.2	85.0	Q1.8C
7.0	30.0	*	*	3.6	*	7.0	*	14.1	87.0	Q1.8D
54.0	16.0	5.32	17.8	10.4	.30	4.5	*	17.4	14.0	Q1.9A
60.0	11.0	2.41	16.6	5.5	.14	4.2	*	10.2	8.0	Q1.9B
52.0	17.0	.95	14.6	3.3	.07	4.2	*	7.7	5.0	Q1.9C
50.0	28.0	.44	7.6	3.5	.06	4.1	*	9.6	6.0	Q1.9D
37.0	12.0	8.15	10.5	18.9	.77	7.2	34.6	*	*	Q1.10A
27.0	10.0	3.29	9.0	4.6	.37	7.5	62.6	*	*	Q1.10B
*	*	*	*	*	*	*	*	*	*	Q1.10C

*	*	*	*	*	*	*	*	*	*	Q1.10D
63.0	21.0	1.15	8.6	4.2	.13	6.7	*	15.5	88.0	Q1.11A
62.0	23.0	.85	7.7	3.9	.11	6.5	*	14.0	87.0	Q1.11b
58.0	33.0	*	*	4.1	*	6.7	*	16.1	86.0	Q1.11C
54.0	33.0	*	*	3.8	*	6.8	*	16.9	86.0	Q1.11D
31.0	40.0	6.28	9.8	17.0	.64	5.8	*	46.3	79.0	Q1.12A
33.0	41.0	3.02	8.3	10.9	.37	6.5	*	40.0	90.0	Q1.12B
37.0	56.0	*	*	*	*	7.3	1.0	*	*	Q1.12C
43.0	51.0	*	*	*	*	7.4	.1	*	*	Q1.12D
60.0	19.0	1.37	9.3	4.5	.15	6.2	*	13.3	78.0	Q1.13A
58.0	29.0	.50	6.8	3.5	.07	6.4	*	13.6	84.0	Q1.13B
55.0	37.0	*	*	4.2	*	6.3	*	17.3	84.0	Q1.13C
56.0	33.0	*	*	3.8	*	6.6	*	17.2	90.0	Q1.13D
62.0	19.0	1.12	8.5	4.2	.13	6.6	*	15.5	85.0	Q1.14A
65.0	20.0	.57	7.1	3.1	.08	6.3	*	11.2	79.0	Q1.14B
60.0	24.0	*	*	3.2	*	6.2	*	12.4	81.0	Q1.14C
57.0	33.0	*	*	4.0	*	6.1	*	16.6	84.0	Q1.14D
53.0	18.0	2.52	9.5	6.9	.26	5.9	*	21.9	84.0	Q1.15A
54.0	25.0	1.08	8.1	4.5	.13	6.2	*	16.8	87.0	Q1.15B
49.0	32.0	*	*	4.2	*	6.5	*	19.3	88.0	Q1.15C
57.0	32.0	*	*	3.6	*	6.7	*	19.3	93.	Q1.15D
44.0	17.0	1.02	*	3.5	.12	7.9	.9	14.3	92.0	Q1.16A
48.0	20.0	.50	*	3.0	.08	6.2	*	10.2	78.0	Q1.16B
39.0	23.0	*	*	3.1	*	6.8	*	13.2	86.0	Q1.16C
44.0	31.0	*	*	3.6	*	6.3	*	16.5	86.0	Q1.16D
51.0	34.0	2.28	9.0	6.6	.25	7.6	22.9	*	*	Q1.17A
56.0	28.0	1.74	8.2	6.0	.21	7.6	26.5	*	*	Q1.17B
57.0	30.0	*	*	5.1	*	7.5	37.6	*	*	Q1.17C
*	*	*	*	*	*	*	*	*	*	Q1.17D
68.0	20.0	1.49	8.5	5.0	.17	6.8	.4	19.0	88.0	Q1.18A
66.0	24.0	.57	7.1	3.5	.08	6.8	0	13.0	85.0	Q1.18B
60.0	30.0	*	*	3.9	*	6.8	0	15.4	87.0	Q1.18C
57.0	34.0	*	*	4.2	*	7.0	0	17.9	91.0	Q1.18D
62.0	20.0	3.90	*	11.4	*	6.2	*	*	*	Q1.19A
67.0	21.0	*	*	5.6	*	7.8	2.4	*	*	Q1.19B
66.0	18.0	*	*	4.7	*	7.6	1.4	*	*	Q1.19C
62.0	19.0	*	*	4.1	*	7.6	1.0	*	*	Q1.19D
*	*	5.00	*	11.6	*	4.5	*	24.9	18.0	Q1.20A
44.0	43.0	*	*	8.3	*	4.4	*	21.1	20.0	Q1.20B
30.0	51.0	*	*	6.8	*	4.8	*	28.3	53.0	QU.20C
27.0	68.0	*	*	6.2	*	5.2	*	30.8	81.0	Q1.20D
52.0	21.0	2.60	*	7.0	*	5.1	*	24.7	62.0	Q1.21A
56.0	19.0	*	*	3.9	*	5.4	*	13.3	61.0	Q1.21B
36.0	25.0	*	*	3.0	*	6.2	*	14.0	84.0	Q1.21C
1.0	16.0	*	*	2.2	*	7.4	.1	*	*	Q1.21D
*	*	6.40	*	15.9	*	6.8	.1	*	*	Q1.22A
56.0	31.0	*	*	4.8	*	7.9	11.0		*	Q1.22B
39.0	33.0	*	*	.8	*	8.0	78.5	*	*	Q1.22C
*	*	*	*	*	*	*	*	*	*	Q1.22D
*	*	*	6.0	15.0	*	7.6	13.6	*	*	Q1.23A
52.0	26.0	*	*	3.2	*	7.9	37.3	*	*	Q1.23B
*	*	*	*	*	*	*	*	*	*	Q1.23C
*	*	*	*	*	*	*	*	*	*	Q1.23D

*	*	*	7.8	18.0	*	7.4	27.7	*	*	Q1.24A
48.0	31.0	*	*	6.0	*	7.7	42.8	*	*	Q1.24B
*	*	*	*	1.2	*	8.0	77.8	*	*	Q1.24C
*	*	*	*	*	*	*	*	*	*	Q1.24D
74.0	13.0	1.70	*	4.1	*	7.3	.4	11.8	*	Q1.25A
66.0	23.0	*	*	3.6	*	6.5	*	12.6	72.0	Q1.25B
60.0	23.0	*	*	3.3	*	6.3	*	16.3	73.0	Q1.25C
60.0	31.0	*	*	4.0	*	6.3	*	15.7	79.0	Q1.25D
54.0	31.0	2.40	*	8.1	*	7.9	1.0	*	*	Q1.26A
28.0	69.0	3.30	*	7.0	*	6.2	*	*	*	Q1.26B
30.0	70.0	*	*	7.2	*	5.6	*	*	*	Q1.26C
26.0	71.0	*	*	7.5	*	7.7	*	*	*	Q1.26D
49.0	14.0	*	*	7.6	*	4.5	*	14.2	11.0	Q1.27A
44.0	31.0	*	*	6.0	*	5.0	*	14.7	20.0	Q1.272B
21.0	47.0	*	*	5.8	*	5.0	*	20.6	31.0	Q1.27C
14.0	40.0	*	*	4.2	*	4.9	*	18.8	40.0	Q1.27D
57.0	26.0	2.40	*	7.7	*	8.1	2.7	*	*	Q1.28A
61.0	41.0	*	*	6.0	*	8.2	2.5	*	*	Q1.28B
*	*	*	*	*	*	*	*	*	*	Q1.28C
*	*	*	*	*	*	*	*	*	*	Q1.28D
*	*	*	*	28.3	*	4.0	*	*	*	Q1.29A
23.0	4.0	*	*	2.5	*	4.1	*	5.7	14.0	Q1.29B
12.0	20.0	*	*	2.0	*	4.1	*	6.5	14.0	Q1.29C
9.0	17.0	*	*	7.5	*	4.1	*	24.8	6.0	Q1.29D
*	*	*	*	19.5	*	4.8	*	*	*	Q1.30A
*	*	*	*	11.9	*	4.7	*	22.7	16.0	Q1.30B
43.0	40.0	*	*	6.8	*	4.7	*	17.9	17.0	Q1.30B
29.0	59.0	*	*	5.8	*	5.1	*	24.0	56.0	Q1.30D
46.0	35.0	2.70	*	8.0	*	8.0	3.1	*	*	Q1.31A
63.0	25.0	*	*	4.4	*	8.2	17.7	*	*	Q1.31B
*	*	*	*	*	*	*	*	*	*	Q1.31C
*	*	*	*	*	*	*	*	*	*	Q1.31D
64.0	20.0	2.30	*	7.2	*	7.7	.7	*	*	Q1.32A
60.0	26.0	*	*	4.9	*	8.0	5.8	*	*	Q1.32B
55.0	33.0	*	*	*	*	8.3	36.6	*	*	Q1.32C
*	*	*	*	*	*	*	*	*	*	Q1.32D
64.0	21.0	1.90	*	6.1	*	7.7	.3	*	*	Q1.33A
67.0	24.0	*	*	4.3	*	7.2	0	*	*	Q1.33B
55.0	36.0	*	*	4.6	*	7.3	0	*	*	Q1.33C
48.0	43.0	*	*	5.1	*	7.5	.1	*	*	Q1.33D
52.0	30.0	2.80	*	4.3	*	8.0	61.8	*	*	Q1.34A
38.0	40.0	*	*	.9	*	8.3	82.0	*	*	Q1.34B
*	*	*	*	*	*	*	*	*	*	Q1.34C
*	*	*	*	*	*	*	*	*	*	Q1.34D
58.0	29.0	4.00	*	1.8	*	6.1	*	31.2	78.0	Q1.35A
56.0	37.0	*	*	5.4	*	7.0	*	24.8	90.0	Q1.35B
49.0	48.0	*	*	5.2	*	7.2	0	28.7	93.0	Q1.35C
40.0	57.0	*	*	5.3	*	7.7	0	30.9	94.0	Q1.35D
*	*	5.80	*	14.9	*	5.0	*	29.9	36.0	Q1.36A
44.0	36.0	*	*	9.5	*	5.2	*	24.2	46.0	Q1.36B
14.0	63.0	*	*	10.0	*	6.0	*	39.3	83.0	Q1.36C
35.0	53.0	*	*	2.0	*	8.1	77.6	*	*	Q1.36D
*	*	3.80	*	10.9	*	5.5	*	*	*	Q1.37A

51.0	30.0	*	*	4.7	*	4.6	*	*	*	Q1.37B
11.0	74.0	*	*	10.3	*	5.1	*	*	*	Q1.37C
*	*	*	*	*	*	*	*	*	*	Q1.37D
*	*	*	*	7.6	*	7.4	3.9	*	*	Q1.38A
59.0	27.0	2.70	*	4.6	*	8.0	6.1	*	*	Q1.38B
34.0	51.0	*	*	6.6	*	7.6	0	*	*	Q1.38C
23.0	56.0	*	*	7.1	*	7.7	0	*	*	Q1.38D
60.0	23.0	1.60	*	5.5	*	7.3	.6	*	*	Q1.39A
43.0	36.0	*	*	5.7	*	7.1	*	*	*	Q1.39B
52.0	34.0	*	*	4.4	*	7.3	*	*	*	Q1.39C
*	*	*	*	*	*	*	*	*	*	Q1.39D
69.0	24.0	4.80	*	*	*	6.9	1.0	*	*	Q1.40A
69.0	24.0	3.00	*	*	*	5.9	0	*	*	Q1.40B
68.0	25.0	1.40	*	*	*	5.9	0	*	*	Q1.40C
66.0	26.0	.50	*	*	*	7.1	0	*	*	Q1.40D
72.0	24.0	1.30	*	*	*	7.1	.2	*	*	Q1.41A
65.0	33.0	1.20	*	*	*	7.6	0	*	*	Q1.41B
64.0	32.0	.20	*	*	*	7.8	.2	*	*	Q1.41C
71.0	26.0	.30	*	*	*	8.2	1.5	*	*	Q1.41D
43.0	48.0	8.30	12.0	*	.71	5.3	*	29.0	*	Q1.42A
32.0	45.0	*	*	*	*	5.3	*	31.0	47.0	Q1.42B
15.0	78.0	*	*	*	*	4.9	*	45.0	62.0	Q1.42C
3.0	93.0	*	*	*	*	6.3	*	*	*	Q1.42D
72.0	17.0	5.30	18.0	*	.30	4.5	*	17.4	14.0	Q1.43A
72.0	17.0	2.40	17.0	*	.14	4.2	*	10.2	8.0	Q1.43B
63.0	13.0	.90	15.0	*	.06	4.2	*	7.7	5.0	Q1.43C
57.0	26.0	.40	6.7	*	.06	4.1	*	9.6	6.0	Q1.43D
70.0	17.0	3.70	13.0	*	.28	5.1	*	19.1	37.0	Q1.44A
58.0	21.0	*	*	*	*	4.8	*	10.9	25.0	Q1.44B
12.0	77.0	*	*	*	*	5.0	*	48.0	62.0	Q1.44C
17.0	71.0	*	*	*	*	5.3	*	61.0	80.0	Q1.44D
69.0	17.0	2.10	*	8.9	.24	6.0	*	17.4	71.0	Q1.45A
70.0	17.0	.86	*	7.5	.11	6.0	*	13.0	66.0	Q1.45B
71.0	19.0	*	*	*	*	6.1	*	11.9	66.0	Q1.45C
70.0	25.0	*	*	*	*	6.1	*	14.8	81.0	Q1.45D
68.0	15.0	1.40	*	*	*	5.4	*	9.4	71.0	Q1.46A
53.0	21.0	*	*	*	*	5.9	*	9.1	56.0	Q1.46B
27.0	26.0	*	*	*	*	6.1	*	9.9	86.0	Q1.46C
44.0	36.0	*	*	*	*	6.1	*	14.8	93.0	Q1.46D
70.0	23.0	2.10	*	*	*	6.0	*	17.8	67.0	Q1.47A
66.0	22.0	*	*	*	*	6.0	*	15.1	72.0	Q1.47B
56.0	31.0	*	*	*	*	6.2	*	19.6	84.0	Q1.47C
*	*	*	*	*	*	*	*	*	*	Q1.47D
62.0	25.0	1.80	9.0	*	.20	6.5	*	18.3	81.0	Q1.48A
67.0	21.0	.58	6.4	*	.09	6.4	*	15.5	86.0	Q1.48B
57.0	26.0	*	*	*	*	6.6	*	16.6	85.0	Q1.48C
63.0	26.0	*	*	*	*	6.8	*	15.6	87.0	Q1.48D
61.0	21.0	3.40	*	*	*	6.1	0	*	*	Q1.49A
59.0	17.0	*	*	*	*	7.8	.4	*	*	Q1.49B
38.0	33.0	*	*	*	*	7.7	0	*	*	Q1.49C
26.0	40.0	*	*	*	*	8.0	.3	*	*	Q1.49D
63.0	21.0	1.20	9.2	*	.13	6.7	*	15.5	88.0	Q1.50A
62.0	23.0	.85	7.7	*	.11	6.5	*	14.0	87.0	Q1.50B

58.0	33.0	*	*	*	*	6.5	*	16.1	86.0	Q1.50C
54.0	33.0	*	*	*	*	6.8	*	16.9	86.0	Q1.50D
61.0	22.0	1.20	8.0	*	.15	6.6	0	10.6	75.0	Q1.51A
66.0	27.0	.59	7.3	*	.08	6.6	0	12.0	82.0	Q1.51B
60.0	35.0	*	*	*	*	6.8	0	16.5	86.0	Q1.51C
58.0	32.0	*	*	*	*	7.0	0	17.4	91.0	Q1.51D
71.0	17.0	1.70	*	*	*	6.6	*	18.4	85.0	Q1.52A
66.0	19.0	*	*	*	*	7.0	*	15.6	92.0	Q1.52B
64.0	22.0	*	*	*	*	6.9	*	12.5	86.0	Q1.52C
56.0	28.0	*	*	*	*	7.0	*	14.4	87.0	Q1.52D
56.0	20.0	1.40	10.0	*	.14	5.3	*	15.2	68.0	Q1.53A
54.0	26.0	.57	7.1	*	.08	5.9	*	13.9	76.0	Q1.53B
53.0	33.0	*	*	*	*	5.9	*	16.2	81.0	Q1.53C
52.0	34.0	*	*	*	*	6.3	*	17.5	83.0	Q1.53D
75.0	16.0	5.90	16.0	*	.36	4.3	*	24.0	20.0	Q1.54A
66.0	19.0	1.30	16.0	*	.08	4.0	*	10.8	6.0	Q1.54B
62.0	25.0	.72	12.0	*	.06	4.1	*	9.2	8.0	Q1.54C
51.0	38.0	.30	6.0	*	.05	4.2	*	15.5	16.0	Q1.54D
72.0	19.0	1.70	*	*	*	6.6	.3	23.0	92.0	Q1.54A
69.0	22.0	*	*	*	*	7.4	*	16.4	91.0	Q1.55B
57.0	29.0	*	*	*	*	6.9	*	21.0	91.0	Q1.55C
52.0	38.0	*	*	*	*	6.8	*	23.0	90.0	Q1.55D
66.0	19.0	10.50	17.0	*	.61	3.9	*	10.4	68.0	Q1.56A
56.0	25.0	*	*	*	*	6.0	*	16.5	84.0	Q1.56B
24.0	38.0	*	*	*	*	6.7	*	*	*	Q1.56C
*	*	*	*	*	*	*	*	*	*	Q1.56D
41.0	16.0	1.50	10.0	*	.15	5.7	0	13.1	67.0	Q1.57A
37.0	17.0	.58	7.2	*	.08	6.2	0	10.3	80.0	Q1.57B
42.0	30.0	*	*	*	*	6.7	0	16.1	85.0	Q1.57C
42.0	26.0	*	*	*	*	6.9	0	14.2	87.0	Q1.57D
21.0	14.0	1.20	9.2	*	.13	6.4	0	11.9	81.0	Q1.58A
20.0	13.0	.43	6.1	*	.07	7.2	.2	10.7	91.0	Q1.58B
18.0	12.0	*	*	*	*	6.8	0	6.9	83.0	Q1.58C
16.0	7.0	*	*	*	*	7.1	0	5.3	81.0	Q1.58D
42.0	20.0	1.80	9.5	*	.19	6.3	*	15.1	77.0	Q1.59A
34.0	21.0	.66	8.0	*	.08	6.4	*	12.4	72.0	Q1.59B
38.0	37.0	*	*	*	*	5.9	*	17.1	82.0	Q1.59C
*	*	*	*	*	*	*	*	*	*	Q1.59D
52.0	36.0	6.30	13.0	*	.48	4.9	*	30.0	32.0	Q1.60A
53.0	36.0	*	*	*	*	4.6	*	21.0	23.0	Q1.60B
40.0	48.0	*	*	*	*	4.7	*	19.5	27.0	Q1.60C
26.0	70.0	*	*	*	*	4.6	*	32.0	49.0	Q1.60D
71.0	23.0	5.80	14.0	*	.40	4.5	*	26.0	39.0	Q1.61A
57.0	33.0	*	*	*	*	4.7	*	14.1	45.0	Q1.61B
38.0	59.0	*	*	*	*	4.8	*	29.0	62.0	Q1.61C
33.0	66.0	*	*	*	*	4.8	*	34.0	73.0	Q1.61D
25.0	10.0	3.10	*	*	*	6.4	*	19.6	82.0	Q1.62A
25.0	10.0	*	*	*	*	6.4	*	11.5	73.0	Q1.62B
23.0	10.0	*	*	*	*	6.2	*	7.9	65.0	Q1.62C
15.0	7.0	*	*	*	*	6.1	*	6.4	59.0	Q1.62D
29.0	17.0	3.40	16.0	*	.21	4.9	*	15.1	30.0	Q1.63A
35.0	18.0	*	*	*	*	5.4	*	9.0	44.0	Q1.63B
28.0	22.0	*	*	*	*	5.9	*	12.3	76.0	Q1.63C

23.0	28.0	*	*	*	*	6.4	*	16.8	87.0	Q1.63D
18.0	6.0	5.20	17.0	*	.31	4.3	*	16.6	18.0	Q1.64A
17.0	5.0	*	*	*	*	4.2	*	4.2	12.0	Q1.64B
13.0	5.0	*	*	*	*	4.2	*	5.2	10.0	Q1.64C
7.0	10.0	*	*	*	*	4.1	*	11.6	9.0	Q1.64D
31.0	59.0	5.40	9.3	*	.58	5.3	*	37.0	71.0	Q1.65A
45.0	46.0	4.30	8.3	*	.52	5.1	*	35.0	67.0	Q1.65B
41.0	50.0	*	*	*	*	6.3	*	25.0	89.0	Q1.65C
53.0	39.0	*	*	*	*	7.5	13.0	*	*	Q1.65D

* Missing values

(b) Binary data

Presence/absence of mottling

depth levels				
1	2	3	4	Prof. No.
0	0	0	*	B1
0	0	1	*	B2
0	0	0	0	B3
1	0	0	1	B4
0	0	*	*	B5
1	0	0	1	B6
0	0	0	0	B7
0	0	0	0	B8
0	1	1	1	B9
0	0	*	*	B10
0	0	1	1	B11
1	1	1	1	B12
1	1	1	1	B13
0	1	1	1	B14
0	1	1	1	B15
1	1	1	1	B16
0	0	0	*	B17
0	0	1	1	B18
0	0	0	0	B19
1	1	1	1	B20
0	0	0	1	B21
0	0	0	*	B22

0	1	*	*	B23
1	0	0	*	B24
0	0	1	1	B25
0	1	1	0	B26
0	1	1	1	B27
0	0	*	*	B28
*	0	0	0	B29
0	1	1	1	B30
0	0	0	*	B31
0	0	0	*	B32
0	0	0	0	B33
0	0	0	*	B34
0	1	1	1	B35
*	0	0	0	B36
0	0	0	*	B37
0	0	0	0	B38
0	0	0	*	B39
1	1	1	1	B40
1	1	1	1	B41
0	0	0	0	B42
0	1	1	1	B43
1	1	0	0	B44
1	1	0	0	B45
0	0	0	1	B46
0	0	0	0	B47
0	0	0	*	B48

*	*	*	*	B49
0	0	1	1	B50
0	0	1	1	B51
1	1	1	1	B52
*	*	*	*	B53
0	1	1	1	B54
0	0	1	1	B55
0	1	1	*	B56
0	0	0	0	B57
0	0	0	1	B58
1	1	1	*	B59
1	1	1	1	B60
1	1	1	1	B61
0	0	1	1	B62
1	1	1	1	B63
0	0	0	0	B64
1	1	1	1	B65

---

* Missing values

(c) Ordered multistate data

Ped size Ranks

Depth Levels				
1	2	3	4	Prof. No.
1	*	*	*	01
1	1	1	2	02
3	3	2	2	03
1	3	2	1	04
2	1	*	*	05
1	2	3	3	06
1	1	1	1	07
2	1	1	1	08
*	*	*	1	09
*	*	*	*	010
1	*	1	*	011
1	1	2	2	012
1	2	2	*	013
1	*	1	1	014
1	*	2	1	015
1	1	1	1	016
1	1	*	*	017
1	*	2	2	018
1	1	2	2	019
1	1	1	1	020
1	*	*	*	021

1	1	2	*	022
*	*	*	*	023
1	1	1	*	024
1	*	1	*	025
1	1	2	3	026
1	1	2	2	027
1	2	*	*	028
*	*	*	*	029
*	1	1	3	030
1	1	*	*	031
1	1	*	*	032
2	2	2	2	033
1	1	1	*	034
1	2	3	3	035
*	1	*	4	036
*	2	*	2	037
1	1	2	2	038
1	*	2	*	039
2	2	3	3	040
2	2	2	*	041
1	1	2	2	042
1	*	*	2	043
2	2	2	3	044
1	1	*	1	045
*	1	2	*	046
1	2	2	2	047
1	1	2	2	048

*	*	*	*	049
1	*	1	1	050
1	1	2	*	051
2	2	1	2	052
1	*	*	2	053
1	1	1	1	054
1	2	2	2	055
2	2	*	*	056
1	1	1	1	057
1	*	*	*	058
1	*	2	*	059
1	2	2	2	060
1	3	3	3	061
1	1	*	*	062
1	2	3	2	063
*	3	*	*	064
1	1	2	3	065

Missing values

(d) Disordered multistates

ped shape				colour (hue)				
1	2	3	4	1	2	3	4	prof. no.
*	*	*	*	4	*	*	*	M1
4	4	4	4	4	4	4	2	M2
4	4	4	4	4	4	4	2	M3
4	4	3	3	4	4	4	2	M4
4	3	4	*	4	4	*	*	M5
4	4	3	3	4	4	4	2	M6
4	4	4	3	4	4	4	2	M7
4	4	3	3	4	4	2	2	M8
4	4	4	3	*	6	6	4	M9
4	4	4	*	*	5	4	*	M10
4	*	4	4	4	6	4	*	M11
4	4	6	4	4	4	2	2	M12
4	4	4	4	4	4	2	*	M13
4	4	4	4	4	6	4	4	M14
4	4	4	4	4	6	2	4	M15
4	4	4	4	4	4	4	4	M16
4	4	4	*	4	4	6	*	M17
4	4	4	4	4	6	2	2	M18
4	4	4	*	4	4	4	4	M19
4	4	4	6	4	4	3	4	M20
4	4	4	3	4	*	*	*	M21
4	6	*	*	4	4	*	*	M22
4	4	*	*	*	6	*	*	M23

4	4	4	*	4	4	4	*	M24
4	4	4	*	4	6	2	*	M25
4	4	6	6	4	4	4	2	M26
4	4	4	4	4	4	4	4	M27
4	3	*	*	4	4	*	*	M28
*	2	2	3	*	6	6	6	M29
4	4	4	4	*	4	4	2	M30
4	4	4	*	4	4	*	*	M31
4	3	4	4	4	4	6	*	M32
4	4	3	3	4	4	4	4	M33
4	4	4	*	4	4	4	*	M34
4	6	6	4	4	4	2	2	M35
4	4	3	3	*	*	6	4	M36
*	4	4	2	*	*	6	*	M37
4	3	3	3	4	4	4	4	M38
4	3	3	*	4	*	4	*	M39
3	4	4	5	4	4	4	2	M40
4	5	4	3	2	2	2	*	M41
4	4	2	2	4	4	4	4	M42
4	4	4	4	*	6	6	4	M43
4	4	4	4	4	4	*	3	M44
4	4	4	4	4	4	6	2	M45
4	4	3	3	6	4	2	6	M46
4	4	3	3	4	4	2	4	M47
4	4	4	4	4	4	2	2	M48
*	*	*	*	*	*	*	*	M49

4	3	4	*	4	*	4	4	M50
4	4	4	*	4	4	2	*	M51
4	4	4	4	4	4	2	2	M52
4	4	4	4	4	6	6	2	M53
4	4	4	3	*	5	5	5	M54
4	3	4	3	4	4	4	4	M55
4	4	4	*	4	4	6	*	M56
4	4	4	4	4	1	2	2	M57
4	4	2	4	4	*	*	*	M58
4	4	4	*	4	*	2	*	M59
4	4	2	4	4	4	4	4	M60
4	*	*	*	4	*	2	2	M61
4	4	4	4	4	4	*	*	M62
4	4	1	4	4	4	3	2	M63
4	4	4	2	*	1	6	6	M64
1	4	2	2	4	4	2	2	M65

* Mssing values

Colour codes

1	2.5YR
2	5YR
3	7.5YR
4	10YR
5	2.2Y
6	5Y

REFERENCES

Anderberg, M. R. (1973) Cluster Analysis for Applications Academic Press, 359 pp.

Arkley, R. J. (1968) "Statistical methods in soil classification". In Ninth International Congress of Soil Science (ed) Transactions, Vol. IV, p.187-92, Adelaide, Australia.

Avery, B. W. (1956) "A classification of British soils". In Sixth International Congress of Soil Science (ed) Transactions, E, p.279-85, Paris, France.

_____ (1968) "General soil classification: Hierarchical and co-ordinate systems". In Ninth International Congress of Soil Science (ed) Transactions, Vol. IV, p.169-175, Adelaide, Australia.

_____ (1973) "Soil classification in the Soil Survey of England and Wales". J. Soil Sci., 24, p.324-38.

_____ (1980) System of Soil Classification for England and Wales. (Higher Categories), Technical Monograph No.14, Soil Survey of Great Britain, Harpenden, 67pp.

Basinski, J. J. (1959) "The Russian approach to soil classification and its recent development". J. Soil Sci. 10,pp.14-26.

Bidwell, O. W. and Hole, F. D. (1964a) "Numerical taxonomy and soil classification". Soil Sci., 97, pp.58-62.

_____ (1964b) "An experiment in the numerical classification of some Kansas soils". Soil Sci. Soc. Am. Proc., 28,pp.263-68.

Blackith, R. E. and Reyment, R. A. (1971) Multivariate Morphometrics, Acad. Press, London and New York, 412 pp.

Bonner, R. E. (1964) "On some clustering techniques". IBM J. Res. Develop., 8, pp.22-32.

Bottomley, J. (1971) "Some statistical problems arising from the use of information statistics in numerical classification". J. Ecol., 59, pp.339-42.

Boulton, D. M. and Wallace, L. S. (1970). "A program for numerical classification". Computer J. 13, pp.63-9.

_____ (1973) "An information measure for hierarchic classification". Computer J., 16, pp.254-261.

Boyce, A. J. (1969) "Mapping diversity: a comparative study of some numerical methods". In A. J. Cole (ed) Numerical Taxonomy. (Procedings of the colloquium in Numerical Taxonomy Held in the University of St. Andrews, September, pp.1-31. Academic Press, London,pp.324.

Buol, S. W. et al (1973) Soil Genesis and Classification
Hole, F. D. and McCracken, R. J.
The Iowa State Univ. Press, Ames. 360pp.

Burr, E. J. (1968) "Cluster Sorting with Mixed Character Types". Aus. Comp. J. 1(2)97-9.

Burrough, P. A. and Webster, R. (1976) "Improving a reconnaissance soil classification by multivariate methods". J. Soil Sci., 27, pp.554-571.

Cain, A. J. (1958) "Logic and memory in Linnaeus's system of taxonomy". In Proc. Linn. Soc. Lond., 169th Session,pp.144-163.

Campbell, N. A., Mulcahy, M. J. and McArthur, W. M. (1970) "Numerical classification of soil profiles on the basis of field morphological properties". Aust. J. Soil Res. 8, pp.43-58.

Cassetti, E. (1964) Classificatory and Regional Analysis by Discriminant Iterations TR12, Contract No. 1228(26), AD 608093. Northwestern University, Illinois.

Cattel, R. B. (1952) Factor Analysis, Harper, New York, 462 pp.

Cheetham, A. H. and Hazel, J. E. (1969). "Binary (presence-absence) similarity coefficients"., J. Paleontol., 43, pp.1130-1136.

Clark, P. J. (1952) "An extention of the coefficient of divergence for use with multiple characters". Copea, 2, pp.61-64.

Clifford, H. T. and Williams, W. T. (1976) "Similarity measures". In Williams, W. T. (ed) Pattern Analysis in Agricultural Science, pp.37-46, CSIRO, Elsevier Scientific Pub. Comp., 331pp.

Colwell, J. D. (1970) "A statistical-chemical characterization of four soil groups in southern New South Wales, based on orthogonal polynomials"., Aus. J. Soil Res., 20, pp.221-38.

Crowson, R. A. (1970) Classification and Biology. Atherton, New York, 350pp.

Crowther, E. M. (1953) "The Sceptical Soil Chemist". J. Soil Sci., 4, pp.107-122

Cruickshank, J. C. (1972) Soil Geography, David and Charles: Newton Abbott, 256 pp.

Cuanalo de la G., H.E. and Webster, R. (1970) "A comparative study of numerical classification and ordination of soil profiles in a locality near Oxford, part I, Analysis of 85 sites". J. Soil Sci., 21, 340-352.

De Alwis, K. A. (1971) Pedology of Red Latosols of Ceylon, Dept. of Soil Sci., University of Atlanta (Unpublished Ph.D. thesis).

Demirmen, F. (1969) "Multivariate procedures and FORTRAN IV program for evaluation and improvement of classifications". COMPUTER CONTRIBUTIONS No.31 STATE GEOL. SURVEY,UNIV. of KANSAS,LAWRENCE.

Dubes, R. C. (1970) Information compression, structure analysis and decision making with a correlation matrix. AD 720811 Michigan State Univ., East Lansing, Michigan.

Duchaufour, P. (1963) "Soil classification: A comparison of the American and the French system". J. Soil Sci., 14,pp.149-55.

Eades, D. C. (1965) "The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance". Syst. Zool. 14, pp.98-100.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). "A method for cluster analysis". Biometrics, 21, pp.362-375.

Everitt, B. (1974) Cluster Analysis published on behalf of the Social Science Research Council by Heinemann Educational Books, 121pp.

Fitzpatrick, E. A. (1967) "Soil nomenclature and classification". Geoderma, 1, pp.91-105.

_____ (1980) Soils, their formation, classification and distribution. Longman, London and New York, 353pp.

Forgy, E. W. (1965) "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications". Biometric Soc. Meetings, Riverside, California (Abstract in Biometrics,21,No.3,pp.768.

Fortier, J. J. and Solomon, H. (1966) Clustering Procedures. In Proc. Symp. Multiv. Analysis, Dayton, Ohio, (P. R. Krishnaiah,ed.) pp.493-506.

Friedman, H. P. and Rubin, J. (1967) "On some invariant criteria for grouping data". J. Am. Statist. Ass. 62, pp.1159-78.

Gerasimov, I. P. (1964) "Discussion of the new American Soil classification system". Soviet Soil Science pp.572-575.

Gibbons, F. R. (1968) "Limitations to soil classification" in Transactions of International Congress of Soil Science (ed). Transactions, pp.159-167, Adelaide, Australia, 776pp.

Gilmour, J. S. L. (1936) "Wither Taxonomy". A paper given by G. S. L. Gilmour to section K (Botany) of the British Association for the Advancement of Science at its meeting in Blackpool. Classification Soc. Bull. vol.3, No.4, 1976.

_____ (1937) "A taxonomic problem", Nature, London, 139, pp.1040-42.

_____ (1940) "The development of taxonomic thought since 1851". Nature, 168, pp.400-402.

_____ (1961) "Taxonomy". In A. M. Macleod and L. S. Cobley (eds.) Contemporary Botanical

Thought pp.27-45. Oliver and Boyd, Edinburgh and Quadrangle Books, Chicago, 197pp.

Glinka, K. D. (1928) The Great Soil Groups of the World and their Development (Translated by Marbut, C. F.) Edwards Bros., Ann. Arbor, Mich., 255pp.

Gower, J. C. (1966) "Some distance properties of latent roots and vector methods used in multivariate analysis". Biometrika, 53, 3 and 4, pp.325. 338

_____ (1967) "Multivariate analysis and multidimensional geometry" Statistician 17, pp.13-28.

_____ (1967) "Comparison of some methods of cluster analysis". Biometrics,23, pp.857-71.

_____ (1971) "A general coefficient of similarity and some of its properties". Biometrics, 27, pp.857-71.

Harman, H. H. (1976) Modern Factor Analysis,3rd Ed., Univ. of Chicago Press, 487pp.

Harris, R. J. (1975) A Primer of Multivariate Statistics, Acad. Press, London, 332pp.

Hess, P. R. (1971) A Text Book of Soil Chemical Analysis, John Murray (Pub.) Ltd., London,520pp.

Hodgson, J. M. (1978) Soil Sampling and Soil Description, (Monograph of Soil Survey), Clarendon Press, Oxford, England, 241pp.

Hole, F. D. and Hironaka, M. (1960) "An experiment in ordination of some soil profiles", Proc. Soil Sci. Soc. Am.,24, pp.309-312.

Ito, K. (1969) "On the effect of heteroscedasticity and non-normality upon some multivariate test procedures". In Krishnaiah, P. R. (ed) Multivariate Analysis, II,pp.87-120, Acad. Press, New York.

Kawaguchi, K. and Kyuma, K. (1977) Paddy Soils in Tropical Asia. Their material nature and fertility. (Monograph of Centre for South East Asian Studies, Kyoto Univ. (English language series No.10) Honolulu, 258pp.

Kendal, M. G. and Stuart, A. (1961). The Advanced Theory of Statistics, vol. 2, pp.356-363,Charles Grifiths and Co. Ltd.,London, 676pp.

Kesseba, A. et al (1972). "Trends of soil classification in Tanzania - The experimental use of 7th approximation". J. Soil Sci.,23(2) pp.235-247.

Knox, E. G. (1965) "Soil individuals and soil classification" Soil Sci. Soc. Am. Proc.,29,pp.79-84.

Kubiena, W. L. (1953). The soils of Europe, Murby, London,317pp.

_____ (1958) "The classification of soils".J. Soil Sci.,9, pp.9-19.

Lambert, J. M. (1973) "AXOR and MONIT: two new
Meacock, S.E., Barrs, J. and Smartt, P. F. polythetic-divisive strategies for hierarchical classification". Taxon, 22, p. 173-176.

Lance, G. N. and Williams, W. T. (1966). "Computer programs for hierarchic polythetic classification (similarity analysis)" Comp. J., 9, pp.60-64.

_____ (1967a) "A note on the classification of multilevel data". Comp. J.,pp.381-383.

_____ (1967b) "A general theory of classificatory sorting strategies: I Hierarchical systems". Comp. J.,pp.373-380.

_____ (1968) "Mixed data classificatory programs" Divisive systems". Aust. Comp. J., Vol. 1, pp.82-85.

_____ (1975). "REMUL: A new divisive polythetic classificatory programs". Aust. Comp. J. Vol. 7, No.3, pp.109-112.

_____ (1977). "Hierarchical classificatory methods". In Englein, K., Rohlston, A. and Wilf, H. (eds) Statistical Methods for Digital Computers, pp.269-339.

Leeper, G. W. (1956) "The soil classification". J. Soil. Sci., 7, pp.59-64.

McBratney, A. B. and Webster, R. (1981). "Spatial dependence and classification of the soil along a transect in northeast Scotland". Geoderma, 26, pp.63-82.

McDonald, R. C. (1981) "Proposed field description sheet associated with the Australian Soil and Land Survey Handbook". In Moore, A. W., Cook, B. G. and Lynch, L. G. (eds). Information Systems for Soil and Related Data, (Proceedings of the Second Australian Meeting of the ISSS Working Group on Soil Information Systems Canberra, Australia, 1 - 21 Feb. 1980), Centre for Agricultural Publishing and Documentation, Wageningen, the Netherlands.

Mcrae, D. J. (1971) "MICKA: A FORTRAN IV iterative K means cluster analysis programs". Behavioural Sci., 16, No. 4, pp.423-24.

McQueen, J. B. (1967) "Some methods for classification and analysis of multivariate observations" Proc. Symp. Math. Statist. and Probability, 5th, Berkley, 1, pp.281-297.

Mahalanobis, P. C. (1927). "Analysis of race mixture in Bengal". J. Asiatic Soc. Bengal.,23, pp.301-33.

_____ (1936) "On the generalized distance in Statistics". Proc. Nat. Inst. Sci. India, 2, pp.49-55.

Mather, P. M. (1975) "Use of orthogonal polynomials in least-squares problems" Computer Applications. Natural and Social Science New Series, 2(1) , pp.310-321.

_____ (1976) Computational methods of Multivariate Analysis in Physical Geography. Willey & Sons, London, 532pp.

McNaughton-Smith, P. Williams, W. T., Dale, M. B. and Mockett, L. G. (1964). "Dissimilarity analysis: A new technique of hierarchical sub-division" Nature, 202, pp.1034-5.

Mincoff, E. C. (1965) "The effects on classification of slight alterations in numerical technique". Sys. Zool., 14, pp.196-213.

Moore, A. W. and Russell, J. S. (1967). "Comparison of coefficient and grouping procedures in the numerical analysis of soil trace element data". Geoderma, 1, pp.139-58.

Moore, A. W., Russel, J. S. and Ward, W. T. (1972). "Numerical analysis of soils. A comparison of three soil profile models with field classification". J. Soil Sci. , 23, pp.193-209.

Muir, J. W. (1962). "The general principles of classification with reference to soils". J. Soil Sci., 13, pp.22-30.

Muir, J. W., Hardie, H. G. Inkson, R. H. E. and Anderson, A. J. B. (1970) "The classification of soil profiles by traditional and numerical methods". Geoderma, 4, pp.81-90.

Norris, J. M. (1971) "The application of multivariate methods to soil studies" J. Soil Sci., 22(1), pp.69-80.

Pearson, K. (1926) "On the coefficient of racial likeness". Biometrika, 18, pp.105-117.

Ragg, J. M. and Clayden, B. (1973). The classification of some British Soils according to the comprehensive system of the United States. Soil Survey Technical Monograph, No.3, Harpenden, 227pp.

Rao, C. R. (1948) "The utilization of multiple measurements in problems of biological classification". J. Roy. Stat. Soc., 10, pp.159-193.

_____ (1952) Advanced Statistical Methods in Biometric Research, Wiley, New York,390pp.

Rayner, J. H. (1966). "Classification of soils by numerical methods". J. Soil Sci. vol. 17, pp.79-92.

_____ (1969) "The numerical approach to soil systematics" in Sheal, J. G. (ed.) The Soil Ecosystems: A Symposium,pp.31-39. (A Systematic Association Pub. No.8).

Robinson, G. W. (1932) Soils, their Origin, Constitution and classification, Murby, London.

Robson, D. S. (1959) "A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced". Biometrics, pp.187-191.

Rohlf, R. J. (1968). "Stereograms in taxonomy" Sys. Zool. 17, pp.246-255.

Rubin, J. (1967). "optimal Classification into groups. An approach to solving the taxonomy problem". J. Theo. Biol. 15,pp.103-44.

Russell, J. S. and Moore, A. W. (1967). "Use of numerical method in determining afinities between deep sandy soils". Geoderma 1,pp.47-48.

Sarkar, P. K., Bidwell, O. W. and Marcus, L. F. (1966). "Selection of characteristics for numerical classification of soils" Soil Sci. Soc. Am. proc. 30, pp.269-272.

Simonson, R. W. (1952) "Lessons from the first half-century of soil survey. I classification of soils". Soil Sci., 74,pp.249-257.

Sneath, P. H. A. (1957) "Application of computers to taxonomy" J. Gen. Microbiol. 17, pp.201-226.

_____ (1964) Mathematics and classification from Adanson to the present. In G. H. M. Lawrence (ed) Adanson. The Bicentennial of Michael Adanson's "Familles des plantes", part 2, pp.471-498. Hunt Botanical Library. Carnegie Inst. of Technol., Pittsburgh, Penna. 635pp.

Sneath, P. H. A. and Sokal, R. R. (1973). Numerical Taxonomy, Freeman, San Francisco, 573pp.

Soil Survey of England and Wales (1967). Soils of the West Sussex Coastal Plain. Soil Survey of Great Britain, England and Wales Bulletin, No.3, Harpenden.

(Hodgson, J. M.)

_____ (1974) Soil Survey Laboratory Methods. Technical Monograph No.6, Harpenden 83pp.

_____ (1976) Soil Survey Field Handbook Describing and Sampling Soil Profiles. Technical Monograph No.5, Harpenden 99pp.

Soil Survey Staff (1951). Soil Survey Manual. U.S. Dept. Agri. Handbook 18, U.S. Govnt. Printing Office, Washington, D. C.

_____ (1960) Soil Classification, A Comprehensive System, 7th Approximation. Soil conserv. serv., U.S. Dept. Agri., U.S. Govt. Printing Office, Washington D.C. 503pp.

_____ (1972) "Soil Survey Laboratory Methods and Procedures for Collecting Soil Samples" (Soil Survey Investigation Report No.1).

Sokal, R. R. (1961) "Distance as a measure of taxonomic similarity" Sys. Zool. 10, pp.70-79.

Sokal, R. R. and Mitchener, C. D. (1958). "A systematical method for evaluating systematic relationships". Univ. Kansas, Sci. Bull. 38, pp.1409-1438.

Sokal, R. R. and Rohlf, F. J. (1962). "The Comparison of Dendrograms by objective methods". Taxon, vol. 11, pp.33-40.

Solomon, H. (1971) "Numerical Taxonomy" in Hodson, F. R. et al., Mathematics in the Archaeological and Historical Sciences. Edinburgh, Univ. Press.

Talkington, L. (1967) "A method of scaling for a mixed set of discrete and continuous variables". Sys. Zool, 16, pp.149-152.

Thorndyke, R. L. (1953) "Who belongs in a family?" Psychometrika 18, pp.267-276.

USDA (1975) Soil taxonomy: A basic system of soil classification for making and interpreting soil survey. Government Printing Office (1976). Agriculture Handbook 436pp.

Ward, J. H. (1963) "Hierarchical grouping to minimize an objective function". J. Am. Statist. Ass., 58, pp.236-244.

Webb, L. J., Tracy, J. G. Williams, W. T. and Lance, G. N. (1967). "Studies in numerical analysis of complex rainforest communities I. A comparison of methods applied to site/species data". J. Ecol. 55, pp.171-191.

Webster, R. (1968). "Fundamental objections to 7th approximations". J. Soil Sci.,19, pp.354-66.

_____ (1971) "Wilk's Criterion: A Measure for comparing the value of general purpose soil classification". J. Soil Sci., 22, pp.254-260.

_____ (1975) "Intuition and rational choice in the application of mathematics to soil systematics" Soil Sci. 119, pp.394-404.

_____ (1976) "The nature of soil variations" Classification Soc. Bulletin, vol.3, No.4, pp.43-55.

_____ (1977) Numerical and Quantitative Methods in Soil Classification, Oxford Univ. Press, 279pp.

_____ (1979) "Exploratory and descriptive uses of multivariate analysis in soil survey" In Wrigley, N. (ed.) Statistical Applications in the Spatial Sciences pp.286-306. Pion Ltd., London 310pp.

Webster, R. and Burrough, P.A. (1972). "Computer based soil mapping of small areas from sample data I. Multivariate classification and Ordination". J. Soil Sci. 23, pp.222-234.

Webster, R. and Butler, B. E. (1976). "Soil Classification and survey studies at Gininderra" Aust. J. Soil Res. 14(1), pp.1-24.

Webster, R. and McBratney, A. B. (1981). "Soil segment overlap in character space and its implications for soil classification" J. Soil. Sci. 32, pp.133-147.

Williams, W. T. (1976a) "Hierarchical agglomerative strategies". In Williams, W. T. (ed) Pattern Analysis in Agricultural Science, pp.84-90, CSIRO, Elsevier Sci. Pub. Comp. Amsterdam, 331pp.

_____ (1976b) "Hierarchical divisive strategies". In Williams, W. T. (ed) Pattern Analysis in Agricultural Science, pp.91-94, CSIRO, Elsevier sci. pub. comp., Amsterdam, 331pp.

_____ (1976c) "Other ordination procedures" . In Williams, W. T. (ed). Pattern Analysis in Agricultural Science, pp.59-69, CSIRO, Elsevier Sci. Pub. Comp.,331pp.

Williams, W. T. and Dale, M. B. (1965). "Fundamental problems in munerical taxonomy". Adv. Botan. Res.,2, pp.35-68.

Williams, W. T. and Lambert, J. M. (1959). "Multivariate methods in plant ecology:Association analysis in plant communities". J. Ecol., 47, pp.83-101.

_____ (1960) "Multivariate methods in plant ecology: the use of an electronic digital computer for association analysis. J. Ecol. 48, pp.689-710.

Wishart, D. (1969a) FORTRAN 11 Programs for 8 methods of cluster analysis (CLUSTAN 1).Computer Contributions 38, Kansas Geological Survey, Lawrence.

_____ (1969b) Mode analysis: a generalization of nearest neighbour which reduces chaining effect. In Cole, A. J. (ed) A Numerical Taxonomy pp.282-311, Academic press, London, 324pp.