Numerical classification of soils and its application in survey

Dit proefschrift met stellingen van Jacobus Jacob de Gruijter, landbouwkundig ingenieur, geboren te Geldrop op 14 mei 1943, is goedgekeurd door de promotoren, dr. ir. J. Bennema, hoogleraar in de tropische bodemkunde, en dr. ir. L.C.A. Corsten, hoogleraar in de wiskundige statistiek. Dit proefschrift is bewerkt onder leiding van dr. ir. J. Schelling, adjunct-directeur van de Stichting voor Bodemkartering.

De rector magnificus van de Landbouwhogeschool,
J.P.H. van der Want

Wageningen, 22 november 1976

J.J. de Gruijter

Numerical classification of soils and its application in survey

Proefschrift
ter verkrijging van de graad van
doctor in de landbouwwetenschappen,
op gezag van de rector magnificus,
dr. ir. J.P.H. van der Want, hoogleraar in de virologie,
in het openbaar te verdedigen
op vrijdag 11 maart 1977 des namiddags te vier uur
in de aula van de Landbouwhogeschool te Wageningen

Centre for Agricultural Publishing and Documentation
Wageningen – 1977

ISN = 104115-03

Abstract

Gruijter, J.J. de (1977). Numerical classification of soils and its application in survey. Agric. Res. Rep. (Versl. landbouwk. Onderz.) 855, ISBN 90 220 0608 5, (ix) + 117 p., 18 tables, 23 figs, 176 refs, Eng. and Dutch summaries.
Also: Doctoral thesis. Wageningen; Soil Survey Papers 12.

Numerical classification of soils was studied with emphasis on methodology and feasibility in survey. A procedure was designed for construction of classes sufficiently homogeneous in terms of relevant properties and handlable by the surveyor. In the procedure 'central' depth-profiles are calculated separately for each property (e.g. clay content), from a sample of depth-profiles, with a relocation method minimizing within-class variances. Any soil profile can thus be identified in the field by allocating its constituent depth-profiles to the central depth-profile that is most similar for the respective properties. Resulting strings of class labels serve for interim data recording. If too many combinations of central depth-profiles arise to map all individually, they are fused into larger classes and within-class variances are again minimized. This procedure was applied to survey data from a marine clay area in the Netherlands: field estimates for 6 properties in 2212 profiles divided into 20 depth intervals. A new method was used to map classes automatically. Tests showed that: samples of several hundred profiles were needed; order of profiles and initial solution for relocation had little effect on results; only extreme weighting significantly affected homogeneity for different variables. Choice of weights and number of classes should be related and supported by sensitivity analysis.

Keywords: numerical classification, numerical taxonomy, cluster analysis, depth profile, soil classification, soil survey, marine clay, the Netherlands, line-printer map, automated cartography.

This thesis will be published as Agricultural Research Reports 855 and as Soil Survey Papers No 12 of the Soil Survey Institute, Wageningen.

Stellingen

1. Het kan de beschrijving en de classificatie van bodemprofielen ten goede komen als, in plaats van een beperkt aantal horizonten, voor elke relevante eigenschap afzonderlijk het verloop met de diepte wordt beschouwd.

dit proefschrift

2. Bij vele bodemclassificaties ligt de nadruk op eigenschappen van de bovengrond. Men dient zich te realiseren dat men door het toepassen van dergelijke classificaties informatie over de ondergrond welke relevant is, bijvoorbeeld voor het ontwerpen van drainage-systemen, kan verliezen.

dit proefschrift

3. De groei van onze kennis en de ontwikkelingen in de informatiebehoefte van kaartgebruikers maken het ongewenst dat een systematische aardwetenschappelijke kartering met vaste legenda meer dan 15 à 20 jaar in beslag neemt.

4. Het in kaart brengen van specifieke fysische eigenschappen van de bodem kan, mits voldoende gedetailleerd, de huidige en toekomstige bedrijfsvoering in landbouwbedrijven belangrijk ondersteunen. Het verdient aanbeveling dergelijke karteringen uit te voeren.

5. Bij het Nederlandse universitaire onderwijs in regionale bodemkunde en fysische geografie dient meer aandacht te worden geschonken aan toepassing van statistische methoden en automatische gegevensverwerking in die vakgebieden.

6. In de bodemkunde komen vaak andere classificatie-problemen voor dan die waarbij het bestaan van twee of meer verschillende populaties wordt vooropgesteld. Het is wenselijk dat statistici hieraan aandacht besteden.

R.M. Cormack, 1971. A review of classification. J. Roy. Stat. Soc. A, 134: 321-367.

7. Indien bij numerieke classificatie een nominale variabele wordt gebruikt, dan is een kwantificering van de verschillen tussen de onderscheiden categorieën van deze variabele noodzakelijk. Het zonder meer toekennen van dezelfde waarde aan deze verschillen is echter misleidend en veelal niet juist.

J.A. Hartigan, 1975. Clustering algorithms. Wiley, New York p. 143.

8. De overheid zou het begrip voor de door haar gefinancierde ontwikkelingshulp kunnen bevorderen door de gemeenschap beter in te lichten over uitvoering en resultaten van die hulp.

Proefschrift van J. J. de Gruijter
Numerical classification of soils and its application in survey

Voorwoord

Graag dank ik hier dr. ir. J. Schelling, die het initiatief tot dit onderzoek heeft genomen en mij bij de uitvoering ervan steeds met raad en daad terzijde heeft gestaan. Ik heb zijn voortdurende steun hoog gewaardeerd; zonder deze zou dit proefschrift niet tot stand zijn gekomen.

Speciale dank komt toe aan prof. dr. ir. J. Bennema. Ik heb het zeer op prijs gesteld dat hij bereid was zich in korte tijd in te werken in een voor hem 'vreemd' onderwerp. Zijn kritiek en adviezen zijn dit proefschrift zeer ten goede gekomen. Ook prof. dr. ir. L.C.A. Corsten ben ik zeer erkentelijk. Met de vele waardevolle suggesties die ik van hem mocht ontvangen, is hij mij zeer bijzonder van dienst geweest.

Ik dank de directie van de Stichting voor Bodemkartering voor de vele faciliteiten die zij mij heeft geboden en waardoor de uitvoering en voltooiing van dit onderzoek mogelijk werd. De heer J.C.A. Zaat (IWIS-TNO) betuig ik mijn dank voor zijn wiskundige adviezen. De heer J.A.J. Schasfoort (IWIS-TNO) komt lof toe voor de voortvarendheid en accuratesse waarmee hij de benodigde computer-programma's heeft vervaardigd.

De stimulerende discussies met dr. S.W. Bie en de grote moeite die hij zich getroostte bij het verbeteren van mijn 'Duglish', heb ik zeer gewaardeerd

Met dank memoreer ik dat de Afdeling Geobotanie van de Universiteit van Nijmegen programmatuur en rekentijd ter beschikking stelde.

Bij het gereedmaken van het manuscript heb ik veel medewerking ontvangen van de afdelingen Redactie en Kartografie en de typekamer van de Stichting voor Bodemkartering. In dit verband noem ik in het bijzonder de heer J.W. Zwolschen. Ook de medewerkers van het Pudoc betuig ik hiervoor mijn dank.

Curriculum vitae

De auteur werd op 14 mei 1943 te Geldrop geboren. Nadat hij in 1960 de HBS-B had doorlopen, studeerde hij aan de Landbouwhogeschool te Wageningen (hoofdvak bodemfysica en -chemie, keuzevakken regionale bodemkunde (tropische specialisatie), cultuurtechniek en wiskunde). In 1969 studeerde hij af, waarna hij in dienst trad bij de Stichting voor Bodemkartering te Wageningen. Daar begon hij een onderzoek naar de toepassing van numerieke classificatie in de bodemkartering. In 1973 werd hij aangesteld bij het Instituut TNO voor Wiskunde, Informatieverwerking en Statistiek, en gedetacheerd bij de Stichting voor Bodemkartering. Zijn taak is inmiddels uitgebreid met wiskundige advisering van de diverse afdelingen van genoemde stichting.

Contents

1 Introduction

Right from the early days of soil science in the last century, considerable effort has been directed towards classification. Two types of activity may be distinguished: arranging soil individuals (e.g. profiles) in classes ('classification'), and assigning an individual to an existing class ('identification'). Both classification and identification may be performed by numerical methods.

The reason for the present study was classification problems arising from surveys done by the Netherlands Soil Survey Institute. The classification of Dutch soils, developed at this institute (de Bakker & Schelling, 1966), has formed a basis for surveys since the early 1960s. The principles underlying this classification are partly similar to the new classification used in the United States (Soil Survey Staff, 1975), but adapted to Dutch circumstances. It has a pedogenetic background, and the classes are morphometrically defined where possible. So far, four levels exist in the Dutch system: order, suborder, group and subgroup. The system has been extensively used in soil surveys since its introduction. (It is the framework for the legend of the Dutch soil map of scale 1:50 000.) Apart from this system, some special classifications have been devised to cope with particular aspects, such as the contents of clay and carbonate in relation to depth (see Bodemkaart van Nederland, 1:50 000, 1964). However, problems remained and new ones have arisen. There is a need to discriminate at levels lower than subgroup, and some of the existing divisions proved to be unsatisfactory for some purposes. Also, a pedogenetic approach to disturbed soil profiles is not always fruitful.

Numerical methods commonly involve large and time-consuming calculations. When computers became readily accessible, research workers in biology and the social sciences began in the 1950s to approach their classification problems by numerical methods. Application of these methods to soil data has been reported in the literature since 1960.

The numerical approach has several attractions. More intensive and consistent use can be made of the original soil data. Also, when a computer is used to support classification, alternative solutions can be easily generated and tested. The whole process of classification may then require less time and effort.

Published studies on numerical soil classification do not tell us everything about which data should be used, and which of the numerous methods is likely to be most appropriate in a given situation. Above all, little allowance is found in the literature that the usual purpose of a soil classification is as a basis for soil survey, and that this may create additional requirements and constraints. Thus the present study concentrates on the choice of a numerical method, giving special attention to applicability in practical soil survey.

This study considers firstly the main problems in soil classification from the viewpoint of a numerical approach. Thus Chapter 2 deals with the purposes of classification, data collection and preliminary processing, types of classifications, identification, and assessment of classifications.

The problem of choice of a numerical method of classification from the vast array of possibilities is separately treated in Chapter 3. Three basic approaches are distinguished and discussed: the heuristic approach, the approach by imposing mathematical requirements and by objective functions. In Chapter 4 the rationale is given for a numerical procedure, which can be integrated in the normal survey procedures, and which aims at homogeneous classes that can be handled in the field. The method is described in detail and applied to profile descriptions from a routine soil survey in the Netherlands. In addition, experiments are reported on some particular aspects, including sample size, the number of classes and weighting of the variables. General conclusions from this study and suggestions for further investigations are presented in Chapter 5.

2 General problems of soil classification

> 'This is the most elementary fact about classification – that we classify for a purpose' (Leeper, 1963)

This chapter deals with general questions related to soil classification. They concern the purpose of classification, collection and pre-processing of the data, the choice of an appropriate type of classification, identification, and assessment of classifications. This applies whether conventional or numerical approaches are used. In the former case, the decisions are often not explicitly stated, in the latter they must be.

In the following sections, we shall discuss these problems only where they are relevant to a numerical approach.

2.1 Purposes of soil classification

Hallsworth (1965) saw soil classification as primarily directed towards 'the mental satisfaction that follows the logical organisation of knowledge in a coherent and mutually consistent scheme'. How ever gratifying, in general it is not the reason for classifying. Reviewing the literature de Bakker (1970) concluded that those who made soil classifications had little to say about their purposes. However a dichotomy according to 'theoretical purposes' and 'purposes of practical importance' seemed obvious. 'Theoretical' conveys the transmission of comprehension about soils, especially their genesis and mutual relations. 'Practical' here relates to communication about soils, prediction of their behaviour or their survey. This distinction may be useful, though mixtures frequently occur. Soil classifications exist that result from genetic considerations only. There are also purely pragmatic single-purpose classifications. Then there are intermediate forms. Many classifications reflect genetic theory but are intended as frameworks for predicting suitability for practical soil uses.

Intended use may vary, but the handling of soil information is a common central element. An essential function of a classification is that it facilitates the description of the soil in a given area. This is achieved by substituting a unified description for a class, covering many slightly different profile descriptions. The simplification reflects technical and psychological desires. A division into classes is indispensable for the simultaneous graphic display of the spatial variations of several soil properties on one map. Also soil information is better memorized and, consequently, its transfer to, for instance, planners of land-use or to students is easier, if it is restricted to a few classes.

In this study, the construction of a legend for soil survey is considered as the main purpose of soil classification. The area to be surveyed, the method of soil survey and the aim of the map are all further specifications of that purpose in a particular case. For

instance, when some form of 'free survey' (*sensu* Steur, 1961) is to be used, an important condition will be that the classification can be satisfactorily employed in the field.

The above can be more formally expressed. In a classification, one can store information about individual soil profiles by allocating the individuals to their proper classes. Information will be retrieved in the form of knowledge about the class to which its name refers. As indicated in Fig. 1, the knowledge of a class in general entails two types of information. Firstly, the definition of a class represents the *differentiating characteristics* (*sensu* Cline, 1949) of the class members. Hereafter this is called *primary information*. Secondly, one usually knows more about a class then its mere definition. This additional knowledge may be either empirical (e.g. observed soil properties: *accessory characteristics*, Cline 1949, reactions to various treatments, geographical distribution) or it may be theoretical (e.g. about genesis or relations between classes and the environment). This is called *secondary information*.

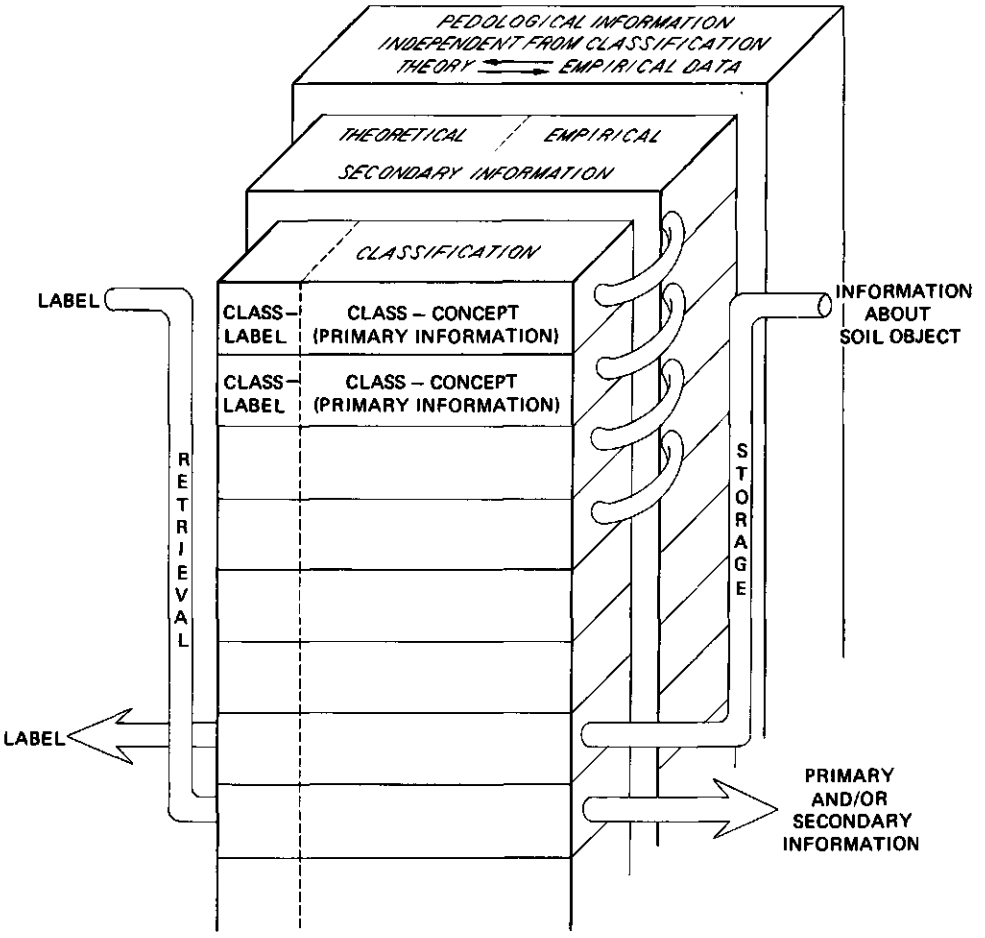


Fig. 1. Storage and retrieval by a classification.

Naturally, empirical information about a class will increase by further observation on already known members or by observation of new members. Advance of pedological theory may update specific theories related to classes. Neither of these two processes as such will alter or extend the classification, but both could make this desirable.

For simplicity a non-hierarchical classification is indicated in Fig. 1. In the case of a hierarchical system, the scheme should be adapted and would have a tree structure, but the principle would remain the same.

If a classification has been derived from theory, it may conveniently represent the essentials of that theory. If mainly empirically derived, the classification need not correspond the existing theory, but it may help to generate hypotheses and thus direct the development of theory. In short: a classification can be seen as a medium through which theory may affect collecting and manipulating empirical data, and *vice versa*. This gives rise to the question which direction of influence should prevail. Biologists have extensively discussed the similar question of choosing between the genetic and the phenetic approach (Johnson, 1970).

Typically, traditional soil classification rests, at least partly, on genetic theory. However, one does not need to be dogmatic here. Norris (1972) recommended to avoid 'the definition of soil types' being 'influenced by hypotheses about the causes of soil differences', because otherwise they 'cannot be used subsequently to justify the hypotheses'. This kind of hypotheses need not be the main concern of applied pedological research, and the requirement seems excessive. A hypothesis should not be statistically tested on the basis of data from which it arose, but directing data collection by preconceptions is an accepted practice. Science often proceeds this way. But there are risks. The view on the object may gradually become biased. It is therefore said that genetic classification ultimately boils down to a circular argument. We consider these risks not sufficient to abandon the principle, but rather stress the need for intensive confrontation of data with theory, i.e. frequent and effective feed-back.

On the other hand, one should admit that a theoretic basis might not be appropriate, or even available. Firstly, theory may be insufficiently established to generate, reliably, as detailed a classification as required. Secondly, a considerable body of established theory may exist, which however cannot be translated into terms relevant to the given purpose of classification. It is therefore recommended to decide pragmatically on the choice between theory and empirical information as the basis for classification.

2.2 Data collection

This section deals with the collection of data to be used, possibly after pre-processing, for the construction of a soil classification. Emphasis lies on fundamental aspects, rather than on the practicalities of data collecting. Although other types of pedological data exist, the discussion here is confined to data contained in profile descriptions.

In the following, any number, code or term used to describe a profile with respect to a given property is considered as a basic element of the data. In the discussion, this is referred to as a *value*.

It is inherent in numerical classification that, at least conceptually, the values are arranged in an $n \times m$ data matrix X, where n and m are the numbers of rows and columns respectively. Each row refers to what is called an entity, individual or object, for instance

a soil profile. We often will call it an *object*; in applications no distinction will be made between the object itself and the corresponding row of values. Each column consist of values regarding the same characteristic or *variable*. The value recorded for the ith object and the jth variable will thus be denoted by x_{ij}.

When the objects are soil profiles, divided into genetic horizons or fixed intervals, examples of variables are:

- percentage of clay between 50 and 60 cm depth, estimated by finger test (*value*, for instance, 18),
- number of mottles in A2 horizon (*value*, for instance, 'few'),
- colour of A1 horizon in Munsell code, when moist (*value*, for instance, 10YR 4/3),
- kind of structure in B horizon (*value*, for instance, 'prismatic').

2.2.1 Choice of variables

The variables on which a classification is based determine by definition the nature of the primary information that can be stored and retrieved by the classification. Indirectly these variables also partly determine the secondary information related to the classes. The choice of variables is thus of paramount importance; the usefulness of the classification heavily depends on it.

The number of morphological, physical, chemical and biological variables by which soil classes can be defined is immense. Application of numerical computer techniques enables one to include many variables in an analysis. This has revived interest in the taxonomic principles of Adanson (e.g. Sokal & Sneath, 1963), which were hitherto hardly practicable. According to these principles, a classification must be based on as many variables as possible, chosen without preconceived opinions about their significance. Whatever the merits of these principles may be for biological classification, their initial identification with the numerical approach of soil classification (e.g. Bidwell & Hole, 1964b) seems a futile effort towards 'objectivity'. Even if the purpose of the classification is only vaguely defined, one could still think of variables being irrelevant. If these are still used in classification, they may detrimentally influence the storage and retrieval of relevant information. Numerical classification with many variables is technically possible. But if classes are defined on many variables, new profiles may be difficult to identify. Thus also for practical reasons, a limited number of well-chosen variables is desirable.

This implies that the variables ought to be chosen in relation to the purpose of the classification. Suppose that the purpose of a soil map of a region is to display suitability for a given type of agriculture. The way we chose the variables may be described as follows. Using existing theory as well as experience, one tries to establish a number of conceptual properties that together determine the suitability of the soil for the type of agriculture in question. These may be referred to as assessment factors, for instance 'availability of water', 'availability of oxygen', 'availability of nutrients', and 'penetrability for roots'. Since these factors are not easily measured, we seek others that may be assumed to be good predictors of the assessment factors, for instance 'texture of top-soil', 'structure of subsoil', 'groundwater regime'. The latter often relate to several assessment factors, and also to one another. The search for variables that are technically and economically acceptable results in a restricted set of relevant variables.

Two aspects that are more specific deserve to be mentioned. Laboratory facilities

allow for accurate measurements that are more closely related to at least part of the relevant conceptual properties than field data. On the other hand, the costs and effort involved are usually much higher. But reduction in the number of samples lowers the reliability of estimates. A rational strategy is possible only in so far as the predictive power with respect to the conceptual characteristics is known for both laboratory and field variables. The effect of field and laboratory variables on classification can be conveniently studied by numerical classification methods. This was done by Sarkar (1965), Grigal & Arneman (1969) and Norris (1971).

The second aspect occurs in literature on numerical soil classification as the problem of *vertical anisotropy*. It arises when the same property is measured at various depths in the profile. The recorded values may then concern fixed depth-intervals or varying intervals such as genetic horizons. In both cases, there is the question which interval of the one profile is to be compared with a given interval of the other. Imagine for instance that clay contents are estimated at various depths in a number of profiles. Even if these depths are the same for all profiles, comparing contents at the same depth is not obvious if one thinks of the possibility that some of the profiles have been buried or eroded.

The problem might be seen as a special case of establishing comparability of data, rather than specifically one of numerical classification. Another form is encountered when, for instance, chemical data are to be analysed that arise from slightly different methods of analysis. A related problem in biology is to establish homologies.

Just as with the other aspects of defining the variables, the solution of the present problem depends on the purpose of the classification. When a genetic system is desired, homologies between soil horizons or layers have to be established. Rayner (1966) attempted to accomplish this by a numerical procedure, later modified by Grigal & Arneman (1969). The idea is to consider, order constraints apart, the most similar pairs of horizons as homologous. If, however, a classification if primarily meant for planning soil use, the approach of Russel and Moore (1968) might be better. They divided profiles into fixed depth-intervals, and then compared intervals at the same depth. The same line has been followed in the experiments of Chapter 4. See also Lamp (1972) for a discussion of this matter.

2.2.2 Choice of profiles

The choice of variables, discussed in the previous section, embodies the decision on *how* to describe the profiles. This section deals with the question of *which* profiles are to be described where to allocate the observation points in the area.

This is largely a matter of sampling design.[1] With respect to sampling in soil survey, it is useful to distinguish between data collection for the construction of classes, and for the geographical delineation of existing classes. Although both aims are in practice often realized more or less simultaneously, they are different and may in principle require different sampling designs.

1. There is no sampling problem if soil bodies are previously delineated and each one is to be treated as an object in subsequent classification. An advantage of this approach is that undue fragmentation of the map can be avoided from the beginning. On the other hand, control of heterogeniety within the classes is lost as far as this is due to variation within these delineated soil bodies. Therefore this approach is not discussed further.

As far as delineation of classes is concerned, whether carried out manually or automatically, strictly random sampling is not usual nor essential. In fact, as in free survey, the surveyor may sample sequentially, and deliberately site each new observation point there where he expects most information. It is commonly assumed that such a directed search may be more efficient in class delineation than a random search. This efficiency depends of course on the true pattern of the classes, the relations between soil properties and landscape features, sample density and the experience of the surveyor. A comparative study of soil survey methods is being conducted by the Oxford School (e.g. Burrough et al. (1971), Bie (1972) and Bie & Beckett (1973). The outcome of such studies are unclear at the moment. So in divising a classification procedure, some form of directed search for delineation will be assumed in this study and it will be required that surveyors can use the classification in the field.

As distinct from delineation, sampling for classificatory analysis should produce data that represent the variations in the area sufficiently well. A random sample sufficient in size to represent adequately the multivariate distribution would be best. There are three main categories of random sampling: strictly random, stratified random and systematic. Each type has its own merits; for sampling theory see, for instance, Raj (1968) and Yamane (1967). Classification of modal profile descriptions, originally selected to represent already established classes, and extracted haphazardly from the literature, is a dubious exercise (but see Hole & Hironaka 1960, and Cipra et al. 1970).

The first numerical soil classifications were with only some tens of objects. This has gradually grown to some hundreds, which is certainly more realistic in view of the intricate variations usually involved. Sample size is, like number of variables, of computational concern. High numbers of objects may rule out certain methods as requiring too much computer time or storage.

2.3 Data pre-processing

It may be desirable to pre-process the data in some way before they are used for classification. Apart from choosing the data and the method of classification, pre-processing constitutes another main category of decisions that have to be taken, and that generally affect the final classification.

Pre-processing may be undertaken for different reasons. For instance, a data transformation may be necessary to a form required for classification. Pre-processing could also be used to obtain a better classification or a more manageable set of data. When its effect is to reduce the amount of data, it is henceforth called *data reduction*. Where the data themselves change but not the number of data, it is referred to as *data transformation*.

2.3.1 Data transformation

By transformation, the data matrix X will be changed into a matrix Y, according to a more or less intricate procedure. Column-wise defined transformations are, for instance

- all values 'not', 'half' and 'fully' in a given column are replaced by 0, 5 and 10, respectively,
- all values in a given column are multiplied by a constant, or replaced by their logarithm,

– reduction of the columns (subtracting the column mean),
– standardization of the columns: dividing all values in a column by the square root of the sum of squares.
Examples of row-wise transformation are:
– all values in a row are replaced by their proportion of the corresponding row total,
– reduction of the rows.
Various other possibilities exist, for example
– all values x_{i1} in Column 1 and x_{i2} in Column 2 are replaced by their sum and difference, respectively,
– the matrix X is replaced by a lower rank approximation calculated by principal components,
– reduction of the columns followed by reduction of the rows (double centring).

Because of the implications for the choice of a classification method and for preliminary transformations, first some distinct types of variables are discussed.

If the set of possible values of a variable is finite, or at least countable, that variable is called *discrete*. An example is type of epipedon as defined in the US soil classification system (Soil Survey Staff, 1975). In particular, counting gives rise to discrete variables, like number of worm-holes. In the special case where only two values are possible, one speaks of *binary* or *dichotomous* variables, like presence or absence of hydromorphic characteristics. Some classification methods can only be used with dichotomous variables.

If the concept of a variable is such that all possible values within a certain range constitute an (uncountably) infinite set, that variable is conceptually *continuous*. Examples are C/N ratio or clay content. Due to coarseness of measurement and rounding, each variable is discrete in practice. The concept of continuity, in cases where precision of measurement may be increased ever further, may facilitate mathematical considerations, e.g. for application of linear vector spaces or calculus, but in fact is an approximating model of reality. Handling strictly discrete variables requires discrete mathematics, which is much more difficult. Classificatory concepts based on strictly discrete variables have been developed by van Emden (1971).

Besides the number of possible values, the kind of relations between the values is also important. In this respect, the following subdivision seems useful (Siegel, 1956).
Nominal variables: the values have no natural order. The only relation between the values is that of equivalence: they are equal or unequal to each other. An example is type of epipedon, with values 'mollic', 'anthropic', 'umbric', etc.
Ordinal variables: the values have a natural order, but only equivalence and order relations between them exist. An example is degree of mottling, with values 'no', 'few', 'moderate', 'many' and 'abundant'.
Metrical variables:[2] assignment of numerical values is at least definite up to a linear transformation. Examples are mass fraction of clay and Celsius temperature.

2. Includes interval variables, ratio variables and counts. For counts, the only reasonable choice is the identity transformation.

2.3.1.1 Transformation of nominal and ordinal variables

The purpose of soil classification normally implies that classes be defined such that members of the same class in some sense resemble each other more than members of different classes. This, in turn, implies the concept of difference or distance between two soils, or between a soil and the typical representative of a class. Whether such differences are established quantitatively or qualitatively, assumptions must be made about the magnitude or significance of the difference between any pair of values, relative to those of other pairs. For nominal and ordinal variables this information is by definition absent and the use of such variables for classification thus seems paradoxical.

Suppose a 3-valued nominal variable with values a, b and c has been recorded for a set of objects which is required to be partitioned into two classes. Are the a's to be lumped with the b's or the c's? Or should the b's go together with the c's? A rational choice does not seem possible unless we know something about the differences between the values. The same difficulty exists for an ordinal variable, where it is known that, for instance, $a > b > c$. One of the alternatives (a combined with c) may then be discarded as being inferior, but the rating of the other two remains uncertain.

The paradox does not exist in practice. With non-mathematical classification, the values of a nominal variable are generally not used as meaningless arbitrary labels. Rather, differences between values are, at least implicitly, weighted against each other according to what is known about them. The same holds for ordinal variables.

In numerical classification, the values are often handled as if they were equidistant. Burr (1968) suggested, as an alternative, to decompose an m-valued nominal variable into m binary variables, each denoting the presence or absence of a particular value, and to assign numerical values to these variables by 'reciprocal proportions'. This means that the non-zero values of the binary variables are made proportional to the square root of the reciprocals of the corresponding relative frequencies. With classification under the least-squares criterion (to be discussed in 3.4.1.1), this standardization has the effect that a variable with many values has greater influence than one with few values. Another tendency, at least if the variables are statistically independent, is that fusions of objects with rare values receive high priority. It is unlikely that these effects would always lead to a useful soil classification. More generally, it seems difficult to devise one rigid scheme for value assignment which is useful for all ends. Therefore, as a more pragmatic strategy, it is advised that the user deliberately chooses the non-zero values of the binary variables, thus controlling their influence on the classification according to what he knows about them. In practice, a nominal variable usually refers to a complex of soil properties and could be conceived as a previously established classification or typology. If data are available on the content of the already established classes, these could be used to evaluate the mutual differences (examples in Ch. 4). If not, the differences have to be estimated subjectively. Even then, however, the transformation may be in better agreement with the purpose of the classification than if the values are assumed equidistant.

A similar argument applies to ordinal variables, except that this type need not be decomposed into binary variables. Suppose, for instance, that the perceived soil reaction to 10% HCl has been recorded with the values 'no', 'weak' and 'strong'. These values could be replaced by numerical ones, proportional to the estimated contents of carbonate with which the reactions on average correspond. The resulting variable is then treated as

being metrical. In addition, statistical and mathematical methods for converting ordinal into metrical variables exist, called 'scaling methods' (Kruskal, 1964a,b).

2.3.1.2 Transformation of metrical variables; weighting

If the original values x are transformed to:
$y = a + b.x$ (a and b constant for a given variable), the transformation is said to be *linear*. Under such transformations the ratios of differences between values are preserved. Any other mode of transformation, like logarithmic, is termed *non-linear*.

2.3.1.2.1 Non-linear transformations This type of transformation is sometimes applied to obtain normal frequency-distributions, the latter being considered indispensable for a valid application of numerical classification. We see no reason for this requirement. It is true that some methods presuppose that the sample which is to be partitioned arises from different, normally distributed populations. However, classification methods based on the assumption that the union of such populations is also normally distributed have not been encountered and would also seem unlikely.

On the other hand, just as with nominal and ordinal variables, non-linear transformations could be desirable for pragmatic reasons, to produce a more useful classification. If, for instance, a certain difference in clay content is judged to be more important in the lower end of the scale than in the higher end, that could be accounted for in the classification process by using, for instance, the square root or the logarithm of the clay content.

2.3.1.2.2 Linear transformations; weighting of variables If a set of objects is conceived of as points in a space of which the co-ordinate axes correspond with the variables, it is easy to see that multiplying the values by a factor and adding a constant have quite different effects. Addition of a constant shifts the points relative to the origin, without affecting the distances between the points. Classification methods, however, are nearly always insensitive to such translations. If, on the other hand, the values of one variable are multiplied by a constant, the group of points will stretch or shrink in the corresponding direction, the distances between the points will change and the resulting classification usually too. The general tendency is that the larger the factor, the more 'weight' attached to the variable, so the more the classification will be determined by that variable.

As Williams (1971) pointed out, the concept of weight is rather vague and ambiguous. Both the multiplication factor and the influence of a variable on a classification are sometimes referred to as weight. Hereafter, the multiplication factor will be termed *scale factor*. The latter concept has been given a more precise meaning by Burr (1968), who referred to the average contribution of a variable to all $\binom{n}{2}$ inter-object distances as the *effective weight* of that variable. When, for instance, squared Euclidean distances (3.2.1.) are used, the effective weight of a variable equals $2/(n-1)$ times the overall sum of squares.

Burr's effective weight seems to be a useful measure. It is defined for the unpartitioned set of objects, though, and therefore confined to the situation before classification. It is generally related, but not identical with the degree to which a classification is actually determined by a variable. The latter, however, may be of direct interest for the

usefulness of a soil classification. For this reason a second concept of weight could be defined analogously, as the average contribution of a variable to the distances between the objects when replaced by the representative (e.g. centroid) of their respective classes. For squared Euclidean distance, this contribution equals $2/(n-1)$ times the between-class sum of squares.

Now the basic question arises whether the initial weights should be accepted as they are in the raw data and, if not, how they are to be changed. From the beginning, these questions were among the main issues in numerical classification.

The choice of measurement units is often partly a matter of convenience. Direct processing of raw data could thus lead to arbitrary weights, to classifications arbitrarily governed by a minority of variables.

An obvious remedy, often advocated, is standardization. The variables are then transformed to equal range or variance. (Note that transformation to equal overall variance results in equality of effective weights if squared Euclidean distance is used!). One of the Adansonian principles (see also 2.2.1) indeed prescribe equal weighting. In my opinion, this is not acceptable as a general principle for soil classification. Here too, decisions should rather consider the purpose of the classification, the method by which this will be established, and the raw data. We may not expect that the quality of a classification will go beyond one's ability to specify adequately the required accuracies of the different kinds of information to retrieve. The study of Russell & Moore (1968) on effects of different depth weightings on numerical soil classification, may be seen in that light. For a clear expression of the same viewpoint in an econometric context, see Morrison (1967).

If a soil map is intended for predicting the suitability for a particular type of land-use, the classification on which the survey is to be based must be constructed such that it is correlated as strongly as possible with suitability. The more the suitability depends on a given variable, the more important it is that information on this variable is preserved by the classification: the more homogeneous the classes should be with respect to that variable. Ideally, if adequate data on suitability were available, optimum scale factors could be objectively established by multiple regression analysis. If that be impossible, the scale factors have to be estimated subjectively.

Only a general approach to the problem of weighting is outlined in this section. The actual procedure depends on the chosen method of classification, and further discussion is therefore postponed to Section 3.2.1 and 4.2.3.3. Effects of different weightings on within-class variances were investigated (4.3.4).

Special problems of weighting may arise for 'hierarchical' variables. Hierarchical variables are, for instance, the presence or absence of a certain type of horizon (primary variable) and the content of clay in this (secondary variable; only applicable if that horizon is present). Without special provision, the differences in secondary variables could preponderate over the differences in primary variables. Kendrick (1965), Williams (1969) and Gower (1971) examined this problem.

Standardization is sometimes applied row-wise instead of column-wise. The values for each object are then transformed, for instance, to zero mean and unit variance or total value 1 for the values or their squares. Row-wise standardization might be appropriate for special purposes, for instance if the average of the values of an object is immaterial for comparison with other objects. It is sometimes applied for that reason by biologists and psychologists.

Soil data are frequently transformed to percentages of an object total, for instance of mineral constituents or adsorbed cations. The use of percentages in numerical classification is dealt with in 3.2.1.8. Standardization of object values has been discussed by, for instance, Cronbach & Gleser (1953) and Orlóci (1967a, b).

2.3.2 Data reduction

2.3.2.1 Reduction of the number of variables

The simplest reduction is deleting one or more variables of minor importance. The choice could be made by inspection of the correlation coefficients, as in the procedure of Sarkar et al. (1966). However, this is still subjective. Principal component analysis, sometimes preceded by factor analysis in order to find a suitable scale transformation, is a better established technique for selection from covariance or correlation matrices. This results in a reduced number of new variables, each of which is a linear combination of the original variables. These methods indeed are frequently applied before classification. They are treated in textbooks on multivariate analysis. The SELFIC/CLAFIC procedure of Watanabe (1969a) is designed for classificatory problems. See also Arkley (1971) and Lamp (1972) for examples of factor analysis and principal component analysis preceding numerical soil classification.

In many instances, these methods of reducing the number of variables will not save computer time. Usually calculation of eigenvectors and eigenvalues from large matrices is involved, which is apt to outweight the lower number of variables, especially if the time required for a classification procedure is only linearly dependent on that number. As a theoretical end, however, factor analysis may provide information alongside that obtained by classification methods. This is clearly so when only few dimensions are retained, so that visual inspection of scatter diagrams is feasible. Marked clustering of objects could already be detected in that stage, if it exists.

Especially if only one factor is used for subsequent analysis or description, as in contour mapping, the loss of information may be serious and caution is needed (e.g. Lamp, 1972; Norris, 1972; Webster & Burrough, 1972a).

If new objects are to be identified it is necessary to express the observations in terms of factors on which the classification is based. This transformation renders manual identification difficult.

When soil profiles have been described by depth interval, for instance by horizon, an obvious way to reduce the number of variables is to reduce the number of intervals. The values of the new variables are averages over two or more previous intervals. If necessary, differences in bulk density and non-linearity of scales (as with pH) must be taken considered in calculating an average. The original units of measurement are preserved by this procedure. One matter to be considered is the extent to which the inter-profile similarities are distorted by this simplification. In tests of my own, a high correlation coefficient (0.99) was found between Euclidean distances based on 5 layers of 40 cm and those based on 40 layers of 5 cm.

Another method of reducing the number of variables is to represent the value of a property (y) as a polynomial function of depth below surface (x):

$$y = \sum_{i=0}^{n} a_i x^i$$

The coefficients a_i are calculated by least squares approximation of the values y recorded at different depths. Each coefficient a_i is then taken as a new variable. As n increases, the approximation improves, but reduction will be less. As n decreases, the danger increases that the polynominal assumption is untrue.

Approximation by polynomials is treated in textbooks on numerical analysis and statistics. Applications in numerical soil classification are found in Campbell et al. (1970) and Moore et al. (1972). Although superficially attractive, the method raises problems. Firstly, if the degree of the polynomials is chosen too small, a considerable distortion may result for irregular profiles. Secondly, if the total depth of the profiles varies, the polynomials are difficult to compare. If, for instance, a shallow profile is similar to the upper part of a deeper one, the calculated coefficients may differ considerably. Thirdly, it is difficult to choose appropriate weights for the new variables. How important is cubic trend of, for instance, phosphate concentration for plant growth, compared with quartic trend? The unsatisfactory results obtained by Campbell et al. (1970) and Lamp (1972) are probably due to these difficulties.

Finally, a strategy frequently followed in conventional soil classification is to replace the values of a subset of the original variables by a reduced number of classes, which form a special classification or typology. This classification serves as a new variable for the final classification. One example is the definition of diagnostic horizons as a preliminary to the US soil taxonomy. This principle is a main element of the numerical classification procedure, designed and tested in this study (Ch. 4).

2.3.2.2 Reduction of the number of objects

Reduction of the number of objects is of special interest when the classification method is such that the computational effort increases proportional to the square of the number of objects, or faster. That is so for agglomerative methods (3.2.2.1.1), for instance.

The simplest and usual method of reduction is to use a random sample from the original set as classification input. Little attention has yet been given to the question of the sample size. As described in 4.3.1, I attempted to acquire some evidence on this.

Watanabe (1969a) suggested a procedure (REPREX) for extraction of a subset of objects representing the whole set as well as possible. This method is theoretically advanced, but the computational effort required is apt to outweigh the advantage in subsequent classification.

2.4 Major types of classification

This section is concerned with some general problems of choice involved in classifying itself, i.e. starting from a given purpose and a set of possibly pre-processed data. Five issues are discussed below. The first two are primarily related to the purpose; the next three concern the structure of the resulting classification.

2.4.1 *Intrinsic versus extrinsic*

These terms are used in the sense of Lance (1973); they are synonymous with 'descriptive' and 'predictive', respectively, as used by Macnaughton-Smith (1965).

In general, the specification of an object as member of a certain class carries primary and secondary information (2.1, Fig. 1). The primary information tells something about the object in terms of the same variables as used for its identification, and the secondary information may predict other variables. They are further called *primary* and *secondary*[3] variables respectively. An *intrinsic* classification is only based on information about primary variables. If, for a subset of the objects, information exists on the secondary variables and this has been used for the construction of the classification, the latter is called *extrinsic*.

Of course also with intrinsic classification one should aim at high predictive value through the choice and transformation of data (2.1 and 2.3). The idea of explicit usage of selected data for this purpose seems of great potential interest. However, on extrinsic classification only the work of Macnaughton-Smith (1963) is known to me; this is restricted to presence-absence variables and only one secondary variable. In the following we shall therefore confine the discussion to the intrinsic approach.

2.4.2 *Distribution fitting versus homogeneity optimizing*

Many arguments among numerical taxonomists about the suitability of their methods seem to be caused by fundamental disagreement as to whether a classification should reflect the distribution of objects in multivariate space as well as possible, or should consist of classes that are as homogeneous as possible. Beside the vagueness of these concepts, it is confusing that they are not mutually exclusive. On the contrary, distribution fitting seems often to imply optimization of homogeneity to a certain extent, and *vice versa*. On the other hand, when the objects form elongated groups of points in multivariate space, classes that correspond to these groups may be too heterogeneous.

The concept of distribution fitting has always had a strong appeal to taxonomists. Several classification methods have this explicit aim (see 3.4.1.2). It is related to the idea of a 'natural' classification, of which the classes are different populations. Undoubtedly there are many situations, for instance in pedogenetic research, in which it is important to know whether a given set of objects should be regarded as a mixture of samples out of different populations; and if so, to indicate which objects belong to each population, and to estimate the population parameters.

If the area to be surveyed is genetically heterogeneous, then it might be worth-while trying first to separate some broad classes with soils having similar histories, by means of distribution-fitting classification. If such classes are still too heterogeneous with respect to the primary variables, they could be further split by homogeneity optimizing classification. The classes resulting from such a strategy are perhaps better mappable and more homogeneous for secondary variables than by homogeneity optimizing alone. As this study is primarily directed to the mapping of genetically fairly homogeneous areas, the

3. Not to be confused with 'primary' and 'secondary' in relation to hierarchical variables (2.3.1.2).

survey of numerical methods (Ch. 3) as well as the experiments (Ch. 4) have however been concentrated upon optimizing homogeneity.

Other discussions of this topic are by Forgy (1965), Cattell and Coulter (1966), Wishart (1969c) and Spence and Taylor (1970).

2.4.3 Fuzzy, overlapping or disjoint classes

A major choice in classification is whether disjoint or overlapping classes have to be constructed. If the latter, an object may be a member of more than one class. If a set is divided into disjoint classes one speaks of a *partition*. Fuzzy classes (*sensu* Zadeh, 1965) are a third alternative. There one can no longer speak of an object being member of a class, but only of its degree of membership. In practice, fuzzy classes arise when a series of central concepts is defined and no unambiguous rules for identification are given.

The information that an object is near the boundary between two classes is lost if one is working with disjoint classes. Through overlapping or fuzzy classes, it can be preserved, by specifying the object's multiple membership or its low degree of membership. Thus with overlapping as well as with fuzzy classes, more detailed data about the objects can theoretically be passed on to a user than with disjoint classes.

Even if fuzzy classes are used for soil survey, then each point of the map has still to be definitely allocated to a class when drawing the (non-fuzzy) geographical boundaries. It is true that, in this case, the definition of the classes can be adapted to the situation in the field. However, a disadvantage of this strategy is that the concept of a class is likely to shift when going from one part of the area to another. The final classes might then be too heterogeneous.

To avoid excessive fragmentation of the map it is sometimes desirable to have overlap between the classes. On the other hand, overlap must be avoided as much as possible when homogeneity is to be optimized. Therefore, a soil survey can better start from disjoint rather than overlapping classes, overlap being introduced only where, and to the degree, it is necessary.

In summary, variations within classes can be better controlled if disjoint classes are taken as a starting point for soil survey, and possible adaptations of the classes are well recorded. For this reason the following will be confined to construction of disjoint classes. Methods leading to overlapping classes have been discussed by Jones & Jackson (1967), Cole & Wishart (1970) and Jardine & Sibson (1971). Bezdek (1974) gives an example of fuzzy classes being used in a mathematical model.

2.4.4 Hierarchic versus non-hierarchic

When it is decided that the classes should be disjoint, one has the choice between a single partition and a series of hierarchically related partitions. Usually these alternatives are called respectively non-hierarchical and hierarchical classification. Intuitively, it will be clear what is meant by hierarchical classification. A precise definition is as follows.

Definition 1. Partition A is at least as fine as partition B (denoted by: $A > B$) if and only if each class of A is a subset of a class of B.

If $A > B$ and $B > A$, then $A = B$. If $A > B$ and $B \not> A$, then A is finer than B. If so, one also says that A is at hierarchical lower level than B. Note that if $A > B$ and $B > C$, then $A > C$.

Definition 2. A hierarchical classification is a set of partitions that can be ordered in the sense of Definition 1.

Sets of partitions which cannot be ordered in the sense of Definition 1 are called reticulate classifications; they are of theoretical interest only.

The advantage of hierarchical classifications over non-hierarchical ones is that both storage and retrieval of information are easier. Any new object can be identified stepwise, allocating it to classes of decreasing levels. In this way many redundant comparisons between the object and definitions of classes may be avoided, and the identification may proceed more efficiently. Furthermore, the geographical boundaries in an area between the classes of a given partition form a subset of those between the classes of any finer partition in the same area. So if soil maps at different scales are requested, the classes can be more efficiently delineated if a hierarchical classification is used instead of a reticulate one. Also, due to the structure of the classes, a hierarchical system is more comprehensible. Without the constraint of a hierarchical structure, the homogeneity within classes could in general be further optimized. However, the importance of easy storage and retrieval will often override this drawback. Especially if the total variability is large, many classes will be needed to achieve sufficient homogeneity and then the advantage of a hierarchical structure will be greatest. Examples are the Linnaean system, the Universal Decimal Classification system for documents, and various national and international soil classifications. If, during a soil survey, the profiles must be easily identifiable, a hierarchical system seems indispensable.

Special numerical methods exist for constructing hierarchical classifications; these are briefly discussed in 3.2.2.1. Other methods lead in principle to a single partition but when applied again to the subsets a hierarchical classification will result. Alternatively, one could create beforehand two or more partitions independently from each other, based on different sets of variables. These partitions could then be combined into one, such that every resulting class consists only of objects in the same classes of the respective original partitions. This so-called *product partition* is at a hierarchical lower level than each of the original partitions. The latter strategy is often practised conventionally. It has also been followed in the numerical experiments described in Chapter 4.

The choice between hierarchical and non-hierarchical classifications has been discussed by, for instance, Williams & Dale (1965) and Pielou (1969).

2.4.5 Monothetic versus polythetic

These terms were introduced by Sneath (1962). They refer to the kind of distinction made between classes.

Definition 3. If a partition is such that for any pair of classes the values of at least one variable are mutually exclusive, then the partition is monothetic.

In geometrical terms, each class boundary can be represented by a plane perpendicular to one of the coordinate axes. Otherwise the partition is *polythetic.*

Although this is not inherent in the concept, the construction of a monothetic hierarchical classification is in practice always a divisive procedure, i.e. successively dividing of the complete set into finer partitions. Each new partition requires one variable.

The advantage of monothetic classification is its simplicity: the construction procedure is straightforward, both conventionally and by computer; definitions of the resulting classes are simple and clear, often to the extent that they can be used directly as class labels. This, of course, enables quick storage and retrieval, especially with a hierarchical system, which could directly be used as a key for identification.

However, just as with hierarchies, the advantage can in general only be achieved at the price of optimality of the partition. Without the constraint of perpendicular boundaries, more homogeneous classes might generally be possible, while the idea of fitting distributions is hardly compatible with monothetic division. This suboptimality is probably the reason for bad experience with monothetic classification. Polythetic methods will therefore be of major concern in this study.

The choice between monothetic and polythetic classification is discussed, for instance, by Williams (1971).

2.5 Identification

The concepts of classification and identification as described in Chapter 1, are not always clearly distinguished from each other. Identification is basically the allocation of an object to one or more already established classes. Classification must precede identification. Watanabe (1969b) discussed this issue in detail.

Much of the confusion is probably because classification methods may be used in some stage of the construction of identification devices (e.g. Firschein & Fischler, 1963), and conversely, identification techniques may be involved in a classification procedure. Various other terms are used in this connection, for instance pattern cognition and pattern recognition (Watanabe, 1969b).

The problem of identification arises when the objects on which a classification is based are only part of the total universe considered. In soil science, this is mostly so. We argued in 2.2.2 for adapted sample allocation (free survey) for the estimation of the geographical distribution of the classes. Though not necessarily in definitive form, such a strategy assumes those classes to be established beforehand on the basis of only a limited sample. Also the condition arises that identifications should be carried out in the field. This in turn implies that identification should not involve more than simple diagram or a short calculation, if any. For this reason we will not go into the field of multiple discrimination analysis, although this might be of interest for other purposes in soil science, such as automated analysis of air and thin-section photographs. See Sebestyen (1962) and Watanabe (1969c).

The use of a key could be an interesting alternative. Despite the recent progress in automated key generation (e.g. Pankhurst, 1975), the present methods would not serve our needs adequately, and this line will not be pursued here.

A suitable structure of the classification itself could in principle solve the identification problem most directly. A hierarchical system would therefore be appropriate. As indicated already in 2.4.4, this line has actually been followed in the experiments of Chapter 4.

Special attention to identification of soil profiles was paid by Norris & Loveday (1971).

2.6 Assessment of classifications

It is evident already from the preceding sections that the construction of soil classifications is not at all straightforward. Several problems of choice exist for non-numerical methods of classification; they are clearly stated by Schelling (1970). For a numerical approach one must in addition choose the actual classification method; Chapter 3 is entirely devoted to that subject.

The assessment of classifications has only recently received more than superficial attention. For numerical classification, the literature shows that method and practice of assessment are still in their infancy. The possibilities for such assessment are summarized below.

2.6.1 Direct subjective assessment

As a first approximation the quality of a classification may be subjectively assessed by informally forecasting how far it could fulfil its purpose. Various aspects may then be relevant: suitability as a basis for soil survey, homogeneity of the classes and interpretability in terms of pedogenetic theory. The flaw of this procedure is clear: only evidently bad solutions can be spotted with certainty, the remainder can be rated only roughly and with unknown reliability.

Williams et al. (1966) indicated how a small step could be made towards formalization of the above procedure. Starting from the same considerations, a grouping could be erected subjectively beforehand as a standard for comparison with numerical solutions. If a conventional classification existed already, this could play the same role. In fact, these are special cases of a more general one, as discussed below.

2.6.2 Stability of the result

Many miscellaneous statements in the literature suggest that as evidence for the goodness of a classification, one might take its stability against changes in either data or procedure. For instance Campbell et al. (1970) took explicitly the latter line: if one starts from different points and arrives at similar solutions, then they consider such a classification more reliable. At least two questions arise.

Firstly, is the conclusion justified? If similar classifications result from different classification procedures, then probably a clear-cut clustering of the objects exists in the multivariate space. However, it depends on the purpose whether such classifications are the best ones. Conversely, also if the resulting classifications are different, it is still possible that one of them is suitable.

Secondly, stability will be judged in general on the basis of classifications that differ only moderately. Such differences, however, are often assessed in a subjective way. Demands for objectivity give rise to the quest for an appropriate method of comparing classifications; this is a difficult problem in itself. (See Rand (1971) for a quantitative approach.) These remarks need not lead to the conclusion that empirical research on

classification is necessarily futile. If this yields further insight into classificatory processes, it may *indirectly* contribute to a better strategy.

Comparison of a numerical classification may be with either other numerical ones or with conventional ones. Many workers have compared with conventional but the inherent difficulties seem sometimes to be overlooked. If the reason for searching for numerical solutions is suspicion about the optimality of a conventional classification, it is hardly right to adopt the suspect as a standard.

2.6.3 Assessment by mathematical criteria

Many attempts have been made to assess classifications objectively. For that purpose mathematical criteria have been defined by which the goodness of a classification, once established, can be measured and possible alternatives rated. Such criteria are surveyed briefly in the following.

2.6.3.1 Criteria for hierarchical classifications

Numerical methods for hierarchical classification will be treated in 3.2.2.1. The process of lumping or splitting subsets of objects, is usually displayed with a treelike diagram called a dendrogram, dendrograph or phenogram. An example is given in Fig. 2. The vertices represent the single objects. The level of each horizontal line may be interpreted

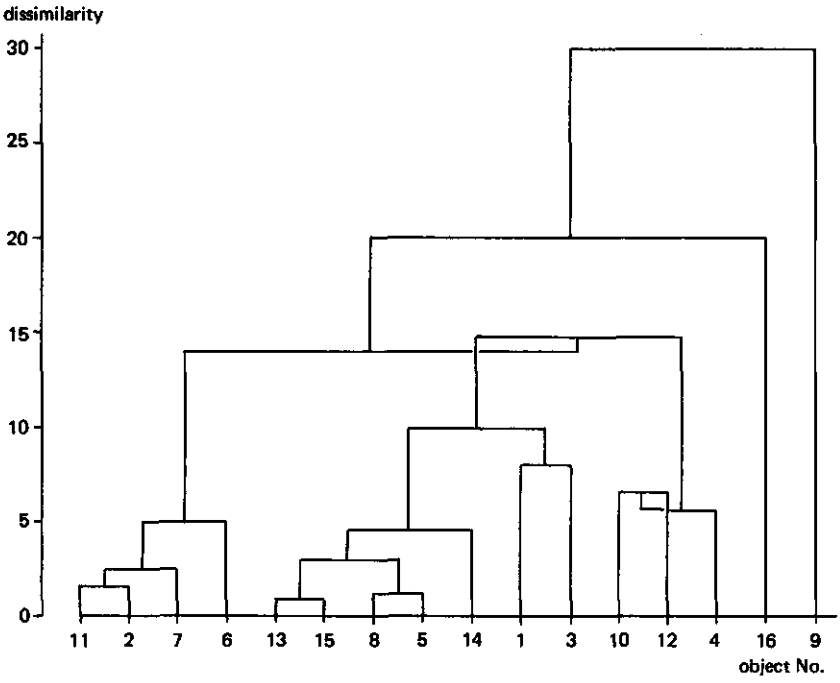


Fig. 2. Fictive dendogram for 16 objects.

as the similarity between the subsets that it connects. The measure of similarity or dissimilarity depends on the actual method.

Certain forms of dendrograms are usually considered, largely intuitively, more favourable than others. Williams et al. (1966) formalized this method of assessment by defining the following criteria.

Chaining. The phenomenon of chaining occurs where single objects (for instance No 7 and 6 or No 16 and 9 in Fig. 2) must be added repeatedly to an ever growing subset in order to obtain partitions of higher level. If chaining is abundant then the dendrogram will show unbalanced partitions at the various levels, which is usually considered undesirable.

Here we consider only the case if the transition from a partition at level i to the one at level $(i + 1)$ requires the amalgamation of only *two* classes. The absolute value of the difference in number of objects in these two classes is denoted by δ_i. Williams et al. (1966) defined thus the following coefficient of chaining:

$$C = \frac{2 \sum_{i=1}^{n-1} \delta_i}{(n-1)(n-2)},$$

where n is the total number of objects. C varies between zero for balanced divisions throughout the dendrogram and unity for complete chaining. Its value for the example in Fig. 2 is 0.43.

Number of reversals. There are no reversals if the similarity between two subsets to be fused in a dendrogram is defined such that it is a monotone function of the partition level. If this monotonicity is not satisfied then reversals do occur, as for instance at the fusion of object No 4 with No 10 and 12 in Fig. 2. The authors consider reversals unfavourable because they hinder unambiguous interpretation of the dendrogram.

Stratification. Williams et al. (1966) considered the distribution of the values at fusions over the range of the coefficient, and suggested that ideally this is such that a relatively large proportion of that range is covered by, say, the last 20% of the fusions. For instance, in Fig. 2 that proportion is 0.5.

Descriptive accuracy. Instead of the form of the dendrogram, another type of criteria considers its accuracy.

A dendogram results usually from the analysis of a triangular matrix S, of all $\frac{1}{2}(n-1)(n-2)$ similarities, s_{ij}, between objects i and j, as calculated from the data (3.2.1). It is simple because it represents only $(n-1)$ similarities, notably those between the subsets which it connects. To establish the accuracy with which S is represented by a dendrogram, all inter-object similarities, s^*_{ij}, will be read from that dendrogram as the value of the similarity coefficient between the subsets to which the objects belong. For instance, from the dendogram of Fig. 2 is read: $s^*_{13,15} = 1, s^*_{13,3} = 10, s^*_{13,16} = 20$, etc. A new matrix, S^*, is thus formed. The more similar the matrix S^* is to S, the more accurate the representation by the dendrogram.

Various measures have been proposed for the deviation of S^* from S. The oldest and still most popular one is the product-moment correlation coefficient, $r(S,S^*)$, in this context introduced by Sokal & Rohlf (1962). They referred to it as cophenetic correlation coefficient. Of course, r may also be used as a measure for the difference between two dendrograms for the set of objects. Williams & Clifford (1971) decided not to use

metric information from a dendrogram and instead proposed an order statistic, analogous to r. Hartigan (1967) preferred a weighted sum of squared differences between the s_{ij} and s^*_{ij}.

2.6.3.2 Criteria for non-hierarchical classifications

Many alternative criteria are also possible for assessment of non-hierarchical classifications. The most prominent type of criterion uses the pooled sample-scatter matrix within classes, W, and the overall sample-scatter matrix, T. (T equals the matrix $(X-X_N)'(X-X_N)$, where $(X-X_N)$ is the data matrix reduced by the column means.)

Three alternatives, discussed by Demirmen (1969), are mentioned here. They will be discussed in more detail in 3.4.1.

a) $\mathrm{tr}(W)$

This measure has a simple geometrical interpretation: $\mathrm{tr}(W)/n$ is the mean squared Euclidean distance between each object and the centroid of the class to which it belongs. Of the three criteria $\mathrm{tr}(W)$ is most frequently applied; it was adopted for the present experiments too.

b) $\det(W)/\det(T)$

This quantity u, sometimes denoted by Λ, is Wilks's (1932) test statistic for testing equality of expected class centroids. Webster (1971) proposed it for assessment of soil classifications. As $\det(T)$ is constant for a given set of data, minimizing u is equivalent to minimizing $\det(W)$.

c) $\mathrm{tr}(W^{-1}B)$

B is defined by the identity $T = W + B$. This is Hotelling's (1931) criterion, used as an alternative test statistic for the same purpose as that of Wilks.

2.6.4 Discussion

In the preceding sections it has been shown why an established classification should be assessed, that a subjective approach to this is problematic, and how this could be made objective. However, also the latter is questionable; as explained below, a definitive solution is not available.

A numerical classification is the result of collecting and preprocessing data and the classification method used. Each of these may in principle be harmful for the result, but let us concentrate upon the classification method. Here again, there may be different detrimental factors.

Firstly, the principle of the method may be inappropriate in view of the purpose of the classification. One may think here of wrong decisions concerning the major choices discussed in 2.4, for instance overlapping *versus* disjoint classes, optimizing homegeneity *versus* fitting distributions, and also of more detailed issues, like the actual definition of homogeneity.

Secondly, although the principle may be sound, a completely satisfactory numerical procedure for application may not be available. Furthermore, when using a computer program for classification, the specification of user-parameters may be inappropriate, thus adding to the common type of numerical errors.

The second class of problems seems less difficult to overcome. It is largely open to a

systematic, and possibly even partly mathematical treatment. This is not so with the choice of the principle itself. That is made through a process of formalization, in which general intuitive notions and considerations about the purpose of the classification have to be translated into mathematical form. Because the purpose of a soil classification can in general only be specified more or less vaguely, any formalization implies inevitably uncertainty *a priori*. This difficulty cannot be evaded by application *a posteriori* of quality criteria because they suffer fundamentally from this same uncertainty. The one cannot compensate the other; even if they seem to do so, it would prove nothing, and if they fail to do so, it is impossible to spot the culprit. Apart from that, optimization should be tried through the method itself. The direct confrontation with any criterion might reveal undesirable features, possibly serious enough to discourage further use. This will be illustrated by the following examples.

Farris (1969) devised an agglomerative algorithm for stepwise maximization of the cophenetic correlation coefficient. This criterion had been used frequently, although until then only after the construction of a classification. Preliminary analysis already showed that consequent maximization in general would lead to dendrograms with reversals. Furthermore, the procedure entails least-squares clustering, the similarities however not being evaluated from the original variables but from the columns of S. Clearly, two objects having the same similarities to the other objects may differ greatly with respect to the original variables. So if 'compact' clusters are desired, $r(S,S^*)$ can lead to inferior solutions. Farris thus fell back to the basic question whether the purpose should be description of S, or description of the objects. This can only be answered by the users of the classification and not by mathematics.

Another example is found among the scatter criteria. Here again there is no compelling reason why one should be preferred *a priori* above the other. However, once it is inferred from the purpose that compact classes are needed, it can be argued that $\det(W)$ is less

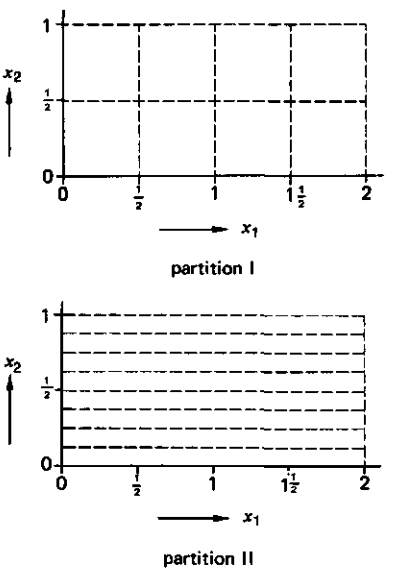


Fig. 3. Two partitions of the same bivariate uniform distribution.

appropriate than tr(W). For instance, det(W) fails to indicate Partition I of the bivariate uniform distribution in Fig. 3 as superior to Partition II, while tr(W) does.

In conclusion, the only remaining option for the final assessment of a soil classification is in the intended application. This admittedly takes a long time. Note for instance that realistic conclusions about principle are only possible if both its application and the input data are right. However, tedious and difficult to systematize as this proceeding may be, it may stimulate thinking about more explicit specification of the purpose, the first step towards better model building.

3 Choosing a numerical method for soil classification

3.1 Introduction

The specific problem in numerical classification is how to choose the most suitable method from the overwhelming number of alternatives. The large variety of methods already proposed in the literature certainly does not exhaust the options. Several reviews have been made (e.g. Bock 1974; Cormack 1971; Sneath & Sokal 1973). But the choice of methods has been discussed only in general terms, or for application in fields other than soil science.

The purpose of this chapter is to elucidate the problem of choosing a numerical method of soil classification, rather than to give another review of methods. Several existing methods will only be mentioned to exemplify the options. Other methods will receive more attention, according to their relevance for soil classification. For reasons given in 2.4.3, the discussion will be limited to methods leading to disjoint classes.

In general, there seem to be three approaches in selecting or devising a classification procedure. By the first approach one chooses a procedure directly, guided by intuitive notions about classification, and possibly simulating a human classificatory process. Although concepts from other fields like statistics or information theory may be borrowed and built in, there is no explicit and unambiguous statement as to which requirements a classification procedure or resulting classifications should fulfil. Indeed, although elaborate computations are made, as mentioned by van Emden (1971, p. 35), the problem for which the outcome is intended to be a solution, is not defined. Given the input, the outcome is not determined by anything other than the actual method or algorithm itself. This will be referred to as the *heuristic approach*. Much work along this line was done by Lance & Williams (1966) and Sneath & Sokal (1973). The problem here is of course that there is little on which the choice of a method can be based.

Secondly, one could formulate beforehand one or more *mathematical requirements* to be satisfied by a method. Methods not satisfying those requirements are not further considered.

The third approach resembles the second one in that requirements are explicitly set out beforehand. But they concern the resulting classification instead of the method itself. The main condition is that some previously specified *objective function* is optimized, possibly subject to one or more side conditions. In other words, classification problems are now approached through mathematical programming.

Following the second or third approach, the major problem is to define mathematical requirements or an objective function such that the purpose of the classification is reflected as closely as possible. About this very issue there is often much uncertainty at the present stage of development. Therefore we feel that these two strategies need not, for the moment, lead to better classifications than the heuristic approach.

The present distinction regards only the way a method is chosen. This is the reason

why there is some overlap in the methods actually resulting from the three approaches. For instance, both heuristic and optimization methods may aim at homogeneity of classes or separation of populations. They may share common elements, such as a similarity coefficient, or after all appear to satisfy mathematical requirements. The distinction seems, nevertheless, to provide a useful framework for discussion.

3.2 Heuristic approach

The heuristic approach formalizes preconceptions about classification in devising a procedure which one expects to produce satisfactory results. It has led to a large variety of methods, whose relative merits are largely unknown.

Many methods of this category proceed in two stages: firstly similarities between objects are calculated, and secondly a classification is established through analysis of the similarities. Definitions of similarity and methods of analysis have largely been developed separately and will here be treated likewise.

3.2.1 Similarity coefficients

Most of the similarity coefficients used fairly frequently will be discussed in the sequel. In selecting a similarity coefficient, the number of candidates is usually reduced because several coefficients may be not defined for the type of variables in hand, or may give undesirable results in chosen instances.

3.2.1.1 Correlation coefficient

The product-moment correlation coefficient has had a long-continued and frequent use in both psychology and ecology, where in Q-type studies it has been applied to pairs of objects instead of pairs of variables. Michener & Sokal (1957) were the first to use it specifically in numerical taxonomy. Between objects a and b it is computed as:

$$r_{ab} = \frac{\sum_{j=1}^{m} (x_{aj} - \bar{x}_a)(x_{bj} - \bar{x}_b)}{\sqrt{\sum_{j=1}^{m} (x_{aj} - \bar{x}_a)^2 \sum_{j=1}^{m} (x_{bj} - \bar{x}_b)^2}},$$

where: x_{aj} = value of variable j for object a
$\bar{x}_a$ = mean of all values for object a
m = total number of variables

This coefficient is now generally out of favour for taxonomic use, basically because it is more clearly realized that its properties are often undesirable. Imagine three objects a, b and c with 'measurement profiles' as in Fig. 4. Now $r_{ab} = r_{ac} = -1$ and $r_{bc} = 1$. The common interpretation of these values as minimum and maximum similarity, respectively, is not in agreement with what the picture shows. This illustrates that r is only sensitive for differences in shape of the profiles, their levels are not taken into account. One speaks of a 'shape coefficient' for short. For a biological classification this might be appropriate, for instance if organisms of different age should be compared, and differences in size are regarded as immaterial. For soil objects, however, differences in 'size' (level of the mea-

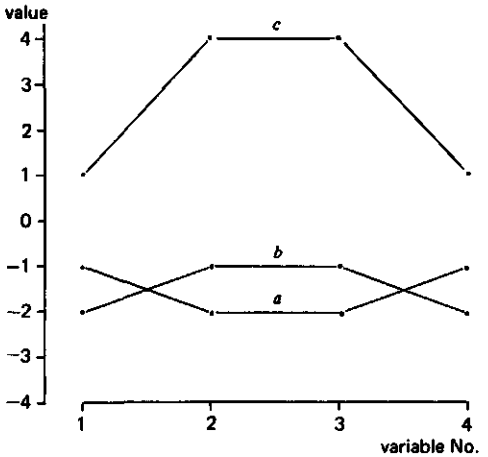


Fig. 4. Diagram with three fictive measurement profiles.

surement profile) are likely to be relevant, in which case another type of coefficient is required.

More information about the use of r as a similarity coefficient is given, for instance by Cronbach & Gleser (1953) and Eades (1965).

3.2.1.2 Association coefficients

Association coefficients are only defined for dichotomous variables; for continuous variables dichotomy is, unfortunately so, required. They have in common that the basic arrangement of data for their computation is a 2 x 2 table as below.

object i \ object j	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

Let the variables be coded 0, 1. Then, for instance, b is the number of variables for which 1 has been recorded for object i and 0 for object j. All coefficients of association are functions of three or four entries of this table, such that it increases with rising proportion of matched values (a and d). Many functions of that kind might be devised, in addition to the large number already been used in numerical classification. Only a few of the best known are mentioned.

Coefficient of Jaccard: $S_J = a/(a+b+c)$

Simple matching coefficient: $S_{SM} = (a+d)/(a+b+c+d)$

Phi coefficient: $S_Q = (ad-bc)/\{(a+b)(a+c)(c+d)(b+d)\}^{\frac{1}{2}}$

S_Q is the product-moment correlation coefficient for data coded 0, 1.

One of the issues in the choice of a coefficient is whether negative matches (d) should be taken into account, i.e. whether absence of a feature in two objects should be taken as evidence for similarity between them (as with S_{SM} and S_Q). For soil classification, there seems no reason for disregarding negative matches.

Association coefficients are in general less suitable for soil classification, because multi-valued and continuous variables cannot be handled adequately and variables cannot be weighted deliberately. One reason why they were not excluded from this review is that they form a starting point for generalization towards coefficients that are less restrictive. Early examples are the coefficients of Rogers & Tanimoto (1960) and that of Smirnov (1960). Here only a recent proposal of Gower (1971) for a general coefficient of similarity will be discussed.

3.2.1.3 General similarity coefficient of Gower (1971)

By the general similarity coefficient of Gower, S_G, weighting as well as simultaneous handling of dichotomous, multi-valued and continuous variables is possible.[4]

$$S_G = \frac{\sum_{j=1}^{m} s_{abj} w_j}{\sum_{j=1}^{m} \delta_{abj} w_j}$$

where: s_{abj} = contribution of variable j to the similarity between objects a and b,
w_j = weighting factor for variable j,
δ_{abj} = 1 if comparison of values of objects a and b for variable j is valid, otherwise 0,
m = total number of variables.

The contributions s_{abj} are assigned as follows:

a. For nominal variables (both with two and many values), $s_{abj} = 1$ if the two objects a and b agree in variable j and $s_{abj} = 0$ if they differ.

b. If, for presence-absence data, negative matches are not considered significant, $\delta_{abj} = 0$ and s_{abj} is unknown but conventionally set at 0. Otherwise as (a): $s_{abj} = 0$ or 1 for unmatched or matched positive values respectively.

c. For metrical variables (both multi-valued and continuous), $s_{abj} = 1 - |x_{aj} - x_{bj}|/R_j$, where x_{aj} is the value of variable j for object a and R_j is the range of variable j in either the population or the sample. In the latter case, this formula ascertains that, as with (a) and (b), s_{abj} ranges from 0 to 1, but intermediate values are now possible.

Two of the association coefficients mentioned before are special cases of S_G. If there are only dichotomous variables and these were equally weighted and treated as under (a), S_G would be identical to the 'simple matching coefficient'. If they were all treated as under (b) that would amount to using the coefficient of Jaccard. Furthermore, if all variables were to be treated as under (c), the formula leads to the complement of the 'mean character difference', which will be discussed among measures of distance. Gower (1971)

4. The same procedure for evaluating similarity was earlier applied to soil profiles by Rayner (1966) and Muir et al. (1970).

suggested a further generalization by introducing weight as a function of the two values involved, $w_j(x_{aj},x_{bj})$, instead of w_j (constant factor for a given variable). Effects of transforming variables can then be simulated. Gower has shown that if there are no missing values and $w_j \geqslant 0$, any matrix of his coefficients is positive semi-definite, so that methods operating on that type of matrices can be applied. For instance, one could determine the coordinates of points in Euclidean space with mutual distances proportional to $(1-S_G)^{\frac{1}{2}}$ for each pair of objects. The flexibility offered by this coefficient makes it more suitable for soil classification than association coefficients. However, transforming and partitioning are preferably not intermingled but kept apart, so that classification methods may remain limited to partitioning. Such methods are presumably easier to investigate and to compare mutually. Furthermore, from the viewpoint of data handling, it may be more efficient if the same transformed data can be used as input for various programs.

3.2.1.4 Canberra metric

The Canberra metric has been defined by Lance and Williams (1967a) as follows:

$$D_c(a,b) = \sum_{j=1}^{m} |x_{aj} - x_{bj}|/(x_{aj}+x_{bj}),$$

where, as before: x_{aj} = value of variable j for object a
m = total number of variables.

It has been employed in soil classification by Moore & Russel (1967), Campbell et al. (1970), Cuanalo & Webster (1970) and Webster & Burrough (1972a).

The coefficient is defined for variables with non-negative values only. It can be shown to be a metric (Lance & Williams, 1967a), although unfortunately it is sometimes referred to as the 'non-metric coefficient'. The denominator in the formula makes the coefficient dependent on the position of origin in hyperspace, and scale independent. The author regards both as undesirable.

For position of origin, interval variables are often involved in soil data, so that the origin will be arbitrarily sited. If so, the coefficient may lead to undesirable results. Imagine for instance three objects a, b and c for which 0, 1 and 2 has been recorded for 'dry', 'moist' and 'wet' respectivity. Then $D_c(a,b) = D_c(a,c) = 1$ and $D_c(b,c) = 1/3$. The same states could equally well have been coded 1, 2 and 3, resulting in $D_c(a,b) = 1/3$, $D_c(a,c) = 1/2$ and $D_c(b,c) = 1/5$. Note in addition that if, for each variable, any one value of the pair (x_{aj}, x_{bj}) is zero, a and b will have a fixed distance (m), irrespective of the magnitude of the non-zero values.

Scale independence is undesirable in soil classification when the user wants to set different emphasis on different soil properties. Scale independence would instead lead to weighing in an unpredictable manner.

3.2.1.5 Coefficient of Bray and Curtis

A measure has been suggested by Bray & Curtis (1957) for use in quantitative ecological studies. They referred to it as 'Index of Similarity', defined as:

$$S_{BC}(a,b) = \frac{2\sum_{j=1}^{m} \min(x_{aj}, x_{bj})}{\sum_{j=1}^{m} x_{aj} + \sum_{j=1}^{m} x_{bj}},$$

which can be rewritten as:

$$S_{BC}(a,b) = 1 - \frac{\sum_{j=1}^{m} |x_{aj} - x_{bj}|}{\sum_{j=1}^{m} (x_{aj} + x_{bj})}.$$

If the variables are fractions, summing up to 1 for each object, the equation reduces to:

$$S_{BC}(a,b) = \sum_{j=1}^{m} \min(x_{aj}, x_{bj}).$$

Hole & Hironaka (1960), after scaling each variable between 0 and 100, applied the coefficient to soil profiles. Application to soil data was later continued by Bidwell & Hole (1964a), Bidwell et al. (1964) and Sarkar et al. (1966).

S_{BC} varies between 0 and 1, increasing with similarity of the objects. As the Canberra metric, S_{BC} is dependent on the position of the origin and only defined for non-negative values. Also the same undesirable behaviour at the origin exists here. Further, S_{BC} is scale dependent but not additive over variables, which makes control of the weight of variables difficult. For these reasons I consider this coefficient generally unsuitable for soil classification.

3.2.1.6 Minkowski metrics

Several measures of distance have been derived from the general definition of Minkowski metrics,

$$d_p(a,b) = \left\{ \sum_{j=1}^{m} |x_{aj} - x_{bj}|^p \right\}^{1/p},$$

where p determines the actual metric.

In particular, d_1/m is known as the 'mean character difference', and d_2 is the Euclidean distance. A possibly interesting limiting case is

$$\lim_{p \to \infty} d_p = \max\{ |x_{a1} - x_{b1}|, |x_{a2} - x_{b2}|, \ldots, |x_{am} - x_{bm}| \}.$$

Both 'mean character difference' and Euclidean distance have a long history of application in natural and human sciences and especially d_2 has been frequently employed in soil classification; they will be discussed under separate headings.

3.2.1.7 Mean character difference

The 'mean character difference',

$$\frac{1}{m} \sum_{j=1}^{m} |x_{aj} - x_{bj}|,$$

is also known as the Manhattan metric or 'city-block distance'; when x_{aj} and x_{bj} represent group means, it is Czeckanowski's 'durchschnittliche Differenz'. It has been applied to soil data by Moore & Russell (1967).

The measure d_1 is additive over variables, in contrast to other Minkowski metrics. This facilitates the computation of the average contribution (weight) of a variable. Another property of d_1 (and other d_p, except d_2) is that it is not invariant with rotation of the coordinate axes. This means that if a partition is desired through which the average d_1 within classes is minimized, the solution will generally differ for rotated and unrotated data, although the choice between them may be arbitrary. Another difference from d_2 is that with d_1 the set of points equally distant from two class representatives (boundary between two classes) is generally only piecewise linear instead of completely linear. The two last

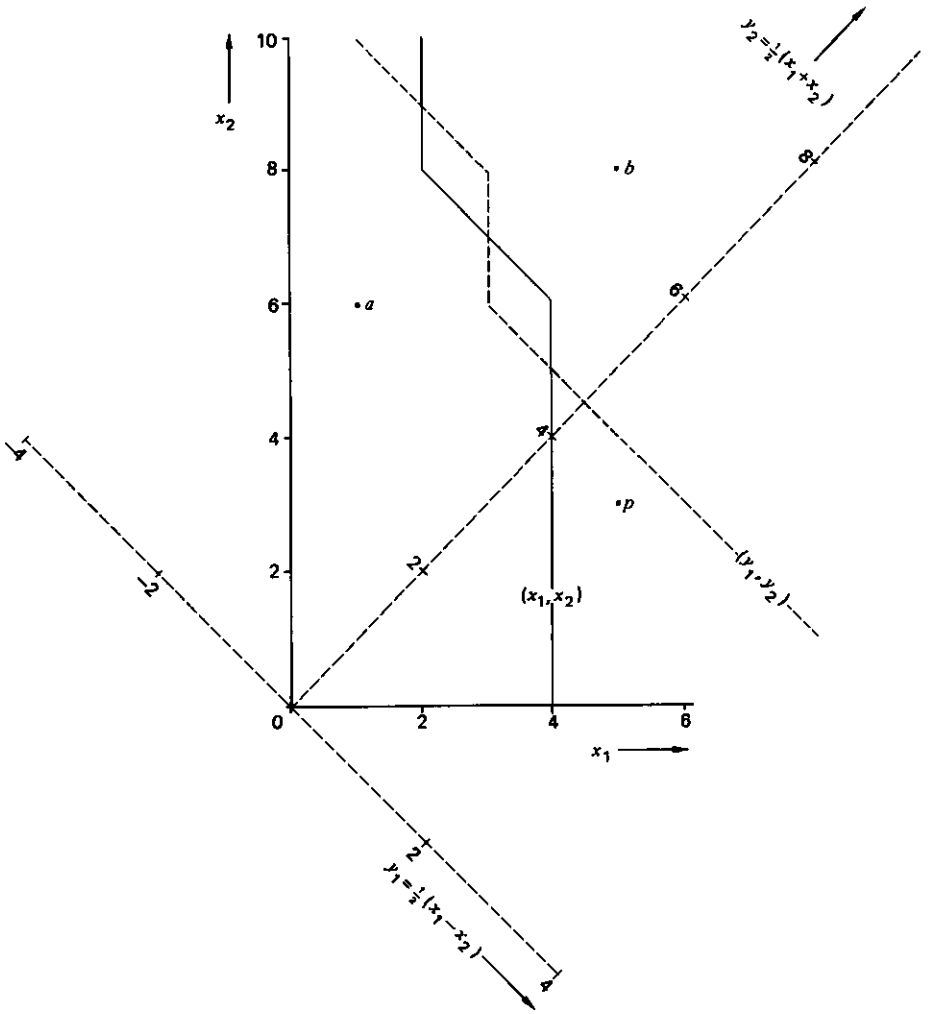


Fig. 5. Effect of rotation of axes on the boundary between two classes as implied by allocation according to smallest d_1.

mentioned features of d_1 are illustrated in Fig. 5. Indicated therein is the boundary (solid line) between two classes with representatives a and b, based on d_1. Object p belongs to class b. After transformation of the original variables x_1 and x_2 to $y_1 = \frac{1}{2}(x_1 - x_2)$ and $y_2 = \frac{1}{2}(x_1 + x_2)$, another boundary (dotted line) appears and now p belongs to class a.

3.2.1.8 Euclidean distance

The Euclidean distance was first proposed for numerical classification by Sokal (1961). It is defined as

$$d_2(a,b) = \left\{ \sum_{j=1}^{m} (x_{aj} - x_{bj})^2 \right\}^{\frac{1}{2}}.$$

It is related to the 'coefficient of racial likeness', developed by Pearson (1926), and is sometimes referred to as 'taxonomic distance' (Sneath & Sokal, 1973). Most soil scientists who until now have experimented with numerical classification have applied Euclidean distance or its transforms: $d_2/\sqrt{m}$, d_2^2 and d_2^2/m.

This measure will be discussed in more detail, not only because its calculation is the first part of several heuristic methods, but also because it has been included in objective functions. (Our experiments were concentrated in particular on the sum of d_2^2 between each object and the centroid of the class to which it is assigned.)

While d_1 and d_2^2 are additive over variables, d_2 is not. Further, d_2 is a metric, while d_2^2 is not. Additivity may be put to advantage in controlling weights. Since the average contribution of a variable to all $\binom{n}{2}$ inter-object distances d_2^2 (i.e. the effective weight in the sense of Burr; 2.3.1.2) is twice the variance of that variable, multiplication of a variable by a factor λ increases its average contribution by a factor λ^2.

Two other differences from d_1, mentioned before, are that the boundary between two classes on the basis of shortest distances to the centroids, is always completely linear as well as invariant under rotation of the coordinate axes. The implications of this are nor fully known. A better mathematical tractability, however, may probably be anticipated; this is exemplified by the fact that some interesting methods of analysis have been devised which explicitly make use of the linearity of class boundaries (3.4.3). More generally, working with d_2 at least potentially has the advantage of using the powerful mathematical tools based on the Euclidean metric, geometry or Euclidean spaces and least squares methods.

The distribution of d_2 depends on the distribution of the population within which the distance is measured. If the m variables are normally and independently distributed, each with variance σ^2, then d_2^2 is distributed as $2\sigma^2 \chi_m^2$.

Hence $\mathcal{E}(d_2^2/m) = 2\sigma^2$,

and $$\mathcal{E}(d_2/\sqrt{m}) = \sqrt{2\sigma^2/m}\,\mathcal{E}\sqrt{\chi_m^2} = \frac{2\sigma}{\sqrt{m}} \cdot \frac{\Gamma(\frac{1}{2}(m+1))}{\Gamma(\frac{1}{2}m)},$$

from which the variance can be calculated. ($\mathcal{E}$ denotes expectation.)
It would be misleading to term d_2 a 'size coefficient', in contrast to a 'shape coefficient' as the correlation coefficient. As pointed out by Cronbach & Gleser (1953), d_2^2 can be

decomposed into three components. Translated into vector notation the procedure is as follows.

Any vector a, representing an object in Euclidean space, can be decomposed into $a_N = \alpha(1,...,1)$ and $a_R = \lambda a^*$ in the residual space R, with $N \perp R$ and a^* normalized, so that $(a^*)^2 = 1$.

Hence we have: $a = a_N + \lambda a^*$

and similarly for a second vector: $b = b_N + \mu b^*$.

For the squared Euclidean distance between a and b we can thus write:

$$\begin{aligned} d_2^2(a,b) = (a-b)^2 &= (a_N - b_N + \lambda a^* - \mu b^*)^2 \\ &= (a_N - b_N)^2 + (\lambda a^* - \mu b^*)^2 \\ &= (a_N - b_N)^2 + \lambda^2 + \mu^2 - 2\lambda\mu(a^*,b^*) \end{aligned} \quad (1)$$

and also:

$$d_2^2(a^*,b^*) = (a^* - b^*)^2 = 2 - 2\,(a^*, b^*). \quad (2)$$

Inserting Equation 2 into Equation 1 yields:

$$\begin{aligned} d_2^2(a,b) &= (a_N - b_N)^2 + \lambda^2 + \mu^2 + \lambda\mu\{d_2^2(a^*,b^*) - 2\} \\ &= (a_N - b_N)^2 + (\lambda - \mu)^2 + \lambda\mu d_2^2(a^*,b^*). \end{aligned} \quad (3)$$

The first component, $(a_N - b_N)^2$, accounts for the difference between the average value of a and b, i.e. between the levels of the two measurement profiles. So it is true that $d_2^2(a, b)$ is sensitive to differences in 'size'. However, the second component accounts for differences in the scatter of values about their mean, and the third for differences in measurement profiles after adjusting for both level and scatter. Hence $d_2^2(a, b)$ is also sensitive to differences in 'shape'.

As mentioned in 2.3.1.2, the correlation coefficient r between two objects is equivalent to the Euclidean distance after standardizing the objects to zero level and unit scatter:

$$r_{ab} = (a^*, b^*) = 1 - \tfrac{1}{2} d_2^2(a^*, b^*). \qquad \text{cf. Equation 2.}$$

Equation 3 shows that such standardization implies loss of information about level and scatter. This is defensible only if that information is definitely considered irrelevant.

The situation is different if the sum of the values for two or more variables is by definition equal for each object. This is frequent in soil data, for an m-valued nominal variable represented by m binary variables, as well as fractions or percentages summing to a total of 100%. In both, the objects already belong to a plane orthogonal to the vector (1, ..., 1); the texture triangle is a well known example of such a plane. Quite straightforwardly, distances between points in that place can be calculated, but various other procedures for this type of data have been suggested; they are reviewed by Gower (1972). For instance, Edwards (1971) proposed square-root transformation of fractions, so that each object is projected onto a hypersphere with unit radius and its centre at the origin. He defined the distance between two objects a and b as the length of the shortest arc between a and b:

$$\arccos \sum_{j=1}^{m} \sqrt{x_{aj} x_{bj}}$$

or, as an approximation, the length of the chord between a and b:

$$\left\{ 2(1 - \sum_{j=1}^{m} \sqrt{x_{aj} x_{bj}}) \right\}^{\frac{1}{2}}$$

I will not go into the pros and cons of this and other procedures; the choice depends on the purpose of the classification. I would tentatively advise calculation of distances directly in the original plane. This is simpler and in practice the outcomes after projection are not likely to differ much (Krzanowski, 1971). This procedure cannot be substituted by arbitrary deletion of one of the composing variables as being 'redundant'. By deletion of variables, the distances in the original plane are transformed in a way only appropriate if differences in the deleted variable are judged completely irrelevant. However, the purpose of a classification will generally require a more balanced weighting.

One remark should be made on the use of d_2 with correlated variables. It is sometimes stated that in any study of similarity, uncorrelated variables should be employed. Cronbach & Gleser (1953) analysed the implications in detail and concluded that d_2 can be meaningfully interpreted also if the variables are correlated. I would agree. Any principal component based on the variables contributes to d_2^2 in proportion to the corresponding eigenvalue of the variance-covariance matrix between variables. For description, of which classification is a special case, this may be desirable. I see, for this type of analysis, no reason why all principal components should be given equal weight *a priori*, especially if one thinks of the last components, which often account for variance largely due to measurement error. The above, of course, re-emphasizes the necessity of careful selection and scaling of variables.

In conclusion, d_2 seems to fit the approach of numerical soil classification developed so far. Non-mathematicians can very easily interpret it geometrically, and it allows the user to specify weights in terms of the familiar variance. Squaring differences in calculating d_2, instead of taking absolute values in calculating d_1, amounts to giving more attention to a few large differences than to many minor differences. That might be attractive. Nevertheless, truly conclusive arguments about choosing *a priori* between d_1 and d_2 or other d_p, can in my view hardly be given; their relative suitabilities should ultimately be shown in practice.

3.2.1.9 Mahalanobis' generalized distance

The generalized distance of Mahalanobis is calculated from the formula:

$$D^2(a,b) = (a-b)' S^{-1} (a-b),$$

where S is the pooled variance-covariance matrix within groups. The measure was originally designed for testing equality of two expected group centroids, estimated by a and b (Hotelling, 1931). The formula can also be applied to any pair of individuals. If so, S may be taken equal to T/n or to the within-group variance-covariance matrix based on a classification given beforehand. For applications to soil data, see Hughes & Lindley (1955), van den Driessche & Maignien (1965) and Prusinkiewicz (1969). Allocation of objects to the centroid that is nearest in the sense of D^2 is equivalent to maximizing Hotelling's criterion $\mathrm{tr}(W^{-1}B)$, to be discussed in 3.4.1.2.

The important difference from d_2 is that in the definition of D the underlying princi-

pal components receive equal weight. The desirability of this was already questioned in the discussion of d_2. Indeed we envisage scaling of the primary variables on purpose, such that changing of any of these variables by one unit is expected to lead to an approximately constant change of a weighted sum of the secondary variables ultimately of interest. Moreover, pooling of scatter matrices within groups as well as the choice of a classification *a priori* may be problematic.

For these reasons, we expect D^2 to be generally unsuitable for use in numerical soil classification, without questioning its value as a test statistic. Another discussion of the usefulness of D^2 in relation to d_1 and d_2 is found in, for instance, Huizinga (1962).

3.2.2 Methods of analysis

An overwhelming number of strategies have been devised for selecting a partition of a given set of objects. This section deals with methods for accomplishing a partition without any explicit specification of mathematical requirements or objective function. These methods may or may not use a previously chosen similarity coefficient, such as discussed in 3.2.1.

We shall follow the usual dichotomy into hierarchical and non-hierarchical methods, leading, respectively, to hierarchical and one-level classifications. Of course, if a one-level classification is desired, hierarchical methods can in principle also be used, by choosing a partition at any level. Conversely, one could arrive at a hierarchical classification by repeated application of a non-hierarchical method to any subset obtained.

3.2.2.1 Hierarchical methods

Three different strategies are possible for establishing a hierarchical classification. Firstly, one could start with the partition at the lowest possible level, each subset containing exactly one object. The iteration step is fusion of two subsets, which renders a new partition with one class less than the preceding one. The final stage is the set containing all objects, or any earlier stage. Such methods are called *agglomerative*. Secondly, the other extreme start could be the complete set; the iteration step is splitting of a subset into two new subsets. Here the endpoint is the stage where all objects are apart, or any earlier stage. This type of method is called *divisive*. A third possibility is to transform the original distances between the objects by an iterative procedure, such that finally for every triplet of objects i, j and k the ultrametric inequality is satisfied: $d(i,j) \leqslant \max\{ d(i,k), d(j,k) \}$.
The partitions at the various levels are thus simultaneously determined.

Once a dendrogram has been produced, it may admittedly usefully summarize similarities, but for mere classification very often only one partition is finally used instead of the whole series. Hierarchical methods are nevertheless popular, probably because the way a partition is found is computationally efficient, rather than because there is a general need for hierarchical classifications. The computational efficiency is attributable to the fact that each iteration step must find a partition that has a hierarchical relation to the previous one, so that many alternatives can be ignored. Efficiency of such partition is generally at the expense of the quality that would be reached without constraints.

3.2.2.1.1 Agglomerative methods Partitions by agglomerative methods are usually polythetic; they may happen to be monothetic, but are never so by definition. Polythetic partitioning by agglomeration is apt to have an important computational advantage over that by division. This is because at each iteration step of an agglomerative method, the only task is to select the pair of subsets to be amalgamated. If the number of subsets at that stage is k, there are $\frac{1}{2}k(k-1)$ alternatives. However with a divisive method, one must decide which subset should next be divided, as well as how it should be divided. If that subset contains n objects, there are $(2^{n-1}-1)$ possible divisions.

All agglomerative methods start by calculating similarities, how ever defined, between all $\frac{1}{2}n(n-1)$ pairs of objects. The most similar pair is then selected to form a new subset. At any iteration step, the next selection should be prepared by calculating the similarities between the new subset and all subsets that remained. All $(n-1)$ steps together require $(n-1)^2$ such calculations. For each of these computations the original data could be used. Another possibility, computationally more attractive, is to calculate the new similarities from two or three similarities that are already known. Therefore the following is confined to this so-called *combinatorial* type of strategy. See, for instance, Orloci (1967a) and Rohlf (1970) for a non-combinatorial method.

Adopt as similarity coefficient a measure of distance, d, increasing with growing disparity of objects. Assume that subsets i and j were fused at some iteration step to form $k = \{i,j\}$, with n_i, n_j and $n_k = n_i + n_j$ objects, respectively. Then the distance d_{hk}, between any remaining subset h (with n_h objects) and the new subset k, is determined by d_{hi}, by d_{hj} and possibly by d_{ij}, n_i, n_j and n_n. As pointed out by Lance & Williams (1966), one may use for all known combinatorial methods a specific form of the general equation:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma \mid d_{hi} - d_{hj} \mid$$

where the parameters α_i, α_j, β and γ determine the method. This allows the following brief characterization.

Nearest-neighbour: $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = 0$; $\gamma = -\frac{1}{2}$.
This method has an extra computational advantage because, simply, $d_{hk} = \min \{ d_{hi}, d_{hj} \}$. This implies that among all inter-object distances between two subsets, the shortest one is taken as the distance between those subsets. It often results in long, straggling clusters.

Furthest-neighbour: $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = 0$; $\gamma = \frac{1}{2}$.
Here $d_{hk} = \max \{ d_{hi}, d_{hj} \}$. This method has the same computational advantage as Nearest-neighbour, but gives clusters that are more compact.

Median sorting: $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = -\frac{1}{4}$; $\gamma = 0$.
This method has been suggested by Gower (1967). It seems to have found little application. If d stands for d_2^2, then the method implies that at each iteration step the centre of k is defined as the midpoint of the line between the centres of i and j.

Weighted pair-group method: $\alpha_i = \alpha_j = \frac{1}{2}$; $\beta = \gamma = 0$.
Sokal and Michener (1958) introduced this method together with the next one; they preferred the present one on phylogenetic grounds.

Unweighted pair-group method: $\alpha_i = n_i/n_k$; $\alpha_j = n_j/n_k$; $\beta = \gamma = 0$.
The many applications of this method in recent years suggest that it has largely replaced the weighted version above. Some authors speak of 'group-average method' following Lance & Williams (1966). The reason for the adjective 'unweighted' is that in effect the distance between two subsets is the unweighted mean of all inter-object distances between those subsets.

Centroid sorting: $\alpha_i = n_i/n_k$; $\alpha_j = n_j/n_k$; $\beta = -\alpha_i\alpha_j$; $\gamma = 0$.
If combined with d_2^2 as a measure of distance, this choice of parameters ensures that a new subset is always represented by its centroid. The fact that resulting dendrograms may show reversals (2.6.3) is generally felt as a drawback of this method.

Flexible sorting: $\alpha_i = \alpha_j = \frac{1}{2}(1-\beta)$; $\gamma = 0$.
The generality of the equation above led Lance & Williams (1967b) to suggest that the user should be free to choose the parameters as suited for his particular classification problem. The restriction $\alpha_i + \alpha_j + \beta = 1$ can be shown to warrant the absence of reversals. Most commonly β is fixed at -0.25; this seems to result on average in a reasonable compromise between the tendencies to 'chaining' (2.6.3) and to form 'non-conformist groups'. Williams et al. (1971) attempted to rationalize further the choice of β.

Ward's method: $\alpha_i = \dfrac{n_h + n_i}{n_h + n_k}$; $\alpha_j = \dfrac{n_h + n_j}{n_h + n_k}$; $\beta = \dfrac{-n_h}{n_h + n_k}$; $\gamma = 0$.

As an early example of the mathematical programming approach, Ward (1963) proposed to optimize a prespecified objective function in classification. He implemented the idea in a non-combinatorial procedure to minimize the within-group sum of squares. It is also called the minimum variance method. Wishart (1969a) found the above combinatorial equivalent, which is the reason why the method is mentioned at this point. Some experiments are reported in Chapter 4.

The main emphasis in this study is on optimizing homogeneity (2.4). In some way or another, the above methods indeed aim at homogeneous classes, however with widely different rigour. In this respect, Nearest-neighbour is obviously at one (negative) extreme, and one finds Furthest-neighbour and Ward's method at the other.

We shall not go into the relative merits of the individual methods or the many issues involved in their application, such as the compatibility with the various coefficients of similarity. Some information on these aspects is given, for instance, by Burr (1970), Kuiper & Fisher (1975) and Lance & Williams (1967b).

3.2.2.1.2 Divisive methods At the beginning of 3.2.2.1.1, I explained why in principle a divisive method is computationally more laborious than an agglomerative one. This is clearly illustrated by the method of Edwards & Cavalli-Sforza (1965), which is just the divisive version of Ward's method. With the latter, a hundred objects or more can easily be handled. Edwards & Cavalli-Sforza warned that even with a modern computer the number of objects that can be handled by their method is limited to about 16.

On the other hand, the final number of classes is usually small compared to the number of objects. Thus fewer steps are needed with dividing than with fusing. This could

compensate to some degree the greater computational effort for a single step. Furthermore, compared with the (unknown) optimum partition, both agglomerative and divisive methods are only approximate because the decisions taken at successive steps are irrevocable. Nevertheless it seems reasonable to expect that the less steps made beforehand, the less a partition at a given level will deviate from the best.

To accommodate an acceptable number of objects, various procedures have been suggested that avoid the total enumeration of alternatives required by the method of Edwards & Cavalli-Sforza. This, however, introduces an additional error source which possibly cancels the advantage of dividing. For instance, one could use the most deviating object of a given subset as a 'condensation core' and allocate each object either to this or let it stay in a residual group (Macnaughton-Smith et al., 1964). Another option is to calculate the first principal component of a given subset, which is then split on the basis of the scores on that component (Fisher, 1958). Similar methods were outlined by Cox (1957) and Switzer (1970).

The remarks above tacitly imply the use of polythetic methods. A separate type of method, however, is based on monothetic dividing. The principle of such methods is that a subset is split according to the variable which, within that subset, is most closely associated with all other variables. The methods may differ in the way 'association' is defined.

A well-known procedure for dichotomous variables, for instance, is to calculate from the 2 x 2 tables the χ^2 - values for each pair of variables and then to split according to the variable with the greatest sum of χ^2-values ('association analysis'). In contrast to other methods, the computational effort is proportional to n and $\frac{1}{2}m(m-1)$, if n and m represent the number of objects and variables, respectively. If n is high and m is low, this may be advantageous. However, as pointed out in 2.4.5, the constraint that the class boundaries must be perpendicular to the coordinate axes is apt to cause unacceptable deviation from the desired homogeneity within the classes. Monothetic methods are therefore not further discussed; the reader is referred to Lance & Williams (1965), Macnaughton-Smith (1965) and Gower (1967).

3.2.2.2 Non-hierarchical methods

Several non-hierarchical methods are used for optimizing a specified objective function; these will be discussed in 3.4. Methods that do not so, have often the purpose of finding 'natural' clusters, i.e. distribution fitting (2.4). Some examples are described below.

Single linkage (Sneath, 1957): Two objects are said to be 'linked', if their mutual distance is shorter than a threshold specified by the user. The set is partitioned into subsets within which each object is linked with at least one co-member, and while between the subsets no links exist. The partitions that result from the choice of any threshold are just those obtained from the Nearest-neighbour method.

An important drawback of this method is that two distinct clusters may be fused by 'noise' points in a saddle region between them.

Mode analysis (Wishart, 1969c): This method was suggested as an improvement of single linkage, especially to reduce the effect of 'noise' points. The algorithm is as follows:

a. Estimate the (probability) density k_i at each point i by the number of other points lying within a threshold distance r from i.
b. Remove the points with small density, i.e. those for which k_i is less than a threshold density k.
c. Cluster the remaining points ($k_i \geqslant k$) by single linkage.
d. Allocate, according to a suitable rule, each point with small density to a cluster, e.g. to the cluster that contains the point that is closest.
Method of Schnell (1964): In the first stage a density function is estimated from the data. This function is the sum of individual density functions around each object. The shape of these functions is specified by the user, e.g. multivariate normal with independent components, each with variance σ^2. At the second stage, each object will be moved stepwise to a local maximum of the density function by the method of steepest descent. Objects arriving at the same maximum together form a subset. The number of resulting subsets depends on σ. One practical problem here is that the required computer time is soon prohibitive.

Many other methods of the present type were devised, for instance by Forgy (1963), Bonner (1964), Goodall (1966), Carmichael et al. (1968), Flake & Turner (1968) and Boon van Ostade (1969). Reviews by Ball (1966) and Spence & Taylor (1970) discuss many of them.

3.3 Imposing mathematical requirements

One way to systematize the choice of a classification method is to impose mathematical requirements that a method should meet. It may happen that, fortunately, only one method satisfies a given set of requirements. In general, however, this approach merely results in the exclusion of evidently bad methods from further consideration.

The principle was applied in part in earlier papers (e.g. Rubin, 1967), but Jardine & Sibson (1968) followed it more consistently. The authors proposed seven requirements and then found that Nearest-neighbour was the only method which satisfied those requirements. Williams et al. (1971), however, made strong objections to these requirements after out less satisfactory experiences with Nearest-neighbour.

A more liberal line was taken by Fisher & van Ness (1971). They tested some well-known methods against a series of requirements. The results were summarized in an 'admissibility' table. Once a subset of the requirements has been chosen in accordance with a particular problem, one can read the admissible method(s) directly from this table. A part of Fisher & van Ness's table is presented as illustration (Table 1).

'Minimum least squares with fixed k' means a method that finds the partition with minimum sum of squares within subsets, the number of subsets being fixed (3.4.1.1) According to Fisher & van Ness (1971), a classification method satisfies the requirements of:
'convexity', if it always leads to a partition into subsets whose convex hulls do not intersect;
'(k-group) well-structuredness', if it leads to a partition with all within-group distances smaller than all between-group distances, whenever such a partition exists;

Table 1. Admissibility table after Fisher & Van Ness (1971). See text for explanation.

Clustering procedure	Convexity	(k-group) well-structuredness	Point-proportion invariance
Nearest-neighbour	No	Yes	Yes
Furthest-neighbour	No	Yes	Yes
Minimum least squares with fixed k	Yes	No	No
Centroid	No	Yes	No

'point-proportion invariance', if duplicating one or more objects any number of times does not affect the separation of the subsets.

The problem of choice is thus shifted to other issues, and it seems to be difficult to define mathematical requirements that are really appropriate for a particular classification problem. The trickiness of choice is already evident by the argument of Jardine & Sibson (1971) with Williams et al. (1971). One difficulty lies in the intransparency of the relations between the requirements. Once any requirement is accepted, one might no longer be free with respect to others.

In conclusion, until more insight is gained into the applicability of the various requirements, imposing mathematical requirements is expected to offer a useful framework for analysing the behaviour of methods, rather than a ready tool for decision-making in practice.

3.4 Approach by objective functions

In Section 3.2, it is reported how methods have been designed by the heuristic approach, and in Section 2.6 how the resulting classifications possibly could be assessed by mathematical criteria. An alternative is to define a criterion or objective function beforehand. A procedure is then sought to optimize that function, possibly within the limits posed by one or more side conditions. In other words, a classification is found through mathematical programming. The difference from imposing mathematical requirements (3.3) is that now the criteria are directed towards the results of the method instead of the method itself.

Definition of an objective function and imposing side conditions, as well as various ways of finding the optimum, will be discussed successively in the sequel.

3.4.1 Objective functions

Naturally, an objective function should be chosen in accordance with the purpose of the classification. Thus one has to make the major choice already discussed in 2.4, between optimization of homogeneity and of separation populations. Discussions on the actual definition of an objective function in relation to these two purposes are followed by a brief review of some other targets suggested in the literature.

3.4.1.1 Functions related with homogeneity

It is repeated here that the rationale for defining homogeneous soil classes is the need for a most accurate estimate of the properties of any soil object given its class membership. The more accurate (and relevant) the information about a soil, the better the use of that soil can usually be planned.

For all members contained in one class the same estimate will be used: the class concept already mentioned in 2.1. Such a concept may be a specification of the class boundaries of properties. For the present discussion, however, it is convenient to assume that each class concept specifies a *class representative*: one real or hypothetic object meant to represent all the members of that class. The error of an estimate could then be measured by the distance, how ever defined, between the object x involved and the representative c_i of the class i to which it has been allocated, $d(x,c_i)$. The mean error, $d(x,c_i)$ averaged over all x, could also be interpreted as a general measure of homogeneity. The actual definition of such an objective function will be implied by the definition of distance, discussed already in 3.2.1. There it was concluded that the squared Euclidean distance d_2^2 seems to be suitable. Actually the mean $d_2^2(x,c_i)$, hereafter designated E_c, is a frequently adopted objective function. If for a given partition the c_i are optimal, i.e. at the centroids of the classes, then E_c can be shown equal to $\mathrm{tr}(W)/n$, where W is the pooled within-group scatter matrix (2.6.3) and n is the number of objects.

In later sections, we shall return to the use of E_c. This objective function is compatible with a natural measure of accuracy of soil maps. Imagine a map of which the legend specifies a number of class representatives c_i. Such a map implies a model of the soil as a finite set of disjoint internally homogeneous areas. For each variable, any c_i given an estimate for any point in those areas which are allocated to class i. The usual measure of in accuracy, E_t, of such a model is the mean squared error, summed over the variables:

$$E_t = \sum_{j=1}^{m} \& (x_j - c_j)^2 ,$$

where x_j is the actual value of variable j at a random point x and c_j is the estimated value of variable j for the area to which x belongs. (The expectation & has been taken over the whole map.)

If the classes had been optimally delineated (with 100% purity), i.e. such that each point in the area is allocated to the class with the nearest representative, then E_t would equal E_c. Through inaccuracies in delineation, however, E_t will generally be greater than E_c. The additional error $E_d = E_t - E_c$, might be a more powerful measure of the inaccuracy with which given classes have been delineated than the customary 'purity'. The present homogeneity function may thus be interpreted as the contribution from classifying to the residual error not explained by a soil map.

Other objective functions related to homogeneity have been proposed by Gower (1974) and Hartigan (1975) who, for binary variables, maximized the number of values of the objects correctly predicted by their associated representatives. Scott (1969) minimized the average $d_2(x,c_i)$ instead of $d_2^2(x,c_i)$, and for computational convenience he sited the c_i at the class centroids, although for this measure of distance that will generally not lead to the optimum partition.

3.4.1.2 Functions for separating populations

In 3.2.2.2, two heuristic methods aiming at separation of population were briefly outlined. Only recently has a general mathematical framework been constructed to treat the problem of separating populations as one of mathematical programming. Scott & Symons (1971) considered maximum likelihood estimation of the parameters of a given number of multivariate normally distributed populations and of an allocation vector γ, of which the ith element indicates the population from which the ith object arose. Sclove (1973) provided a mathematical formulation for the more general case where the distributions belong to any parametric family. He worked this out for multinormal distributions (with equal as well as different covariance matrices) and for distributions underlying 2^m tables.

It turned out that maximum likelihood estimation of γ corresponds to optimization of objective functions that are typical for the assumed nature of the distributions and that possibly may depend on information *a priori*. The results of Scott & Symons (1971) for two situations are worth mentioning here.

Firstly, if the populations are multivariate normal with covariance matrices all equal to the identity matrix I, then maximum likelihood estimation of γ amounts to finding the partition which minimizes $\mathrm{tr}(W)$, the objective function discussed in the preceding section. Thus under these circumstances the purposes of separating populations and optimizing homogeneity lead to the same objective function. The second situation is as above, but the covariance matrices are now equal to an unknown matrix S. The appropriate objective function can be shown to be $\det(W)$.

Difficulties in using $\det(W)$ may be summarized as follows. Generally $\det(W)$ would not lead to compact classes, and the user cannot weight the variables deliberately (2.6.4.). But also from the point of view of population separation, $\det(W)$ is not very satisfactory. The assumptions of normality and equal covariance matrices on which its use is based, will often be unrealistic for soils. A model with different covariance matrices (Scott & Symons, 1971) might then prove more useful. But now minimization is computationally even more laborious than it is with $\det(W)$. Scott & Symons (1971) warned further that if one of two populations is noticeably more weakly represented in the mixture than the other, the maximum likelihood estimation of γ has the tendency to degenerate to an even split of the mixture, especially if the separation of the populations is only moderate. Application of $\det(W)$ as an objective function requires that any new object should be allocated to a class on the basis of shortest Mahalanobis distance, which is hardly practicable in the field. Application of $\mathrm{tr}(W)$ requires subsequent allocation on the basis of shortest Euclidean distance.

Friedman & Rubin (1967) found that clustering Fisher's 150 Iris records with $\det(W)$ as objective function gave better separation of the Iris species, than clustering by $\mathrm{tr}(W)$. Notice, however, that this type of test does not allow for conclusions about the relative suitability for constructing homogeneous classes. Other papers dealing with the use of $\det(W)$ in cluster analysis are by Demirmen (1969), Marriott (1971) and Webster (1971); Bock (1974) and Day (1969) discussed population separation in general.

3.4.1.3 Other objectives and objective functions

Beside maximizing homogeneity and separation of populations, there are several other classification purposes leading to objective functions. For instance, Boulton & Wallace (1970) proposed partition of a set of objects such that after Shannon-Fano coding of the original data the complete code message would be as short as possible. Other measures used in information theory were suggested by Watanabe (1965), Orloci (1972) and van Emden (1972).

Another objective that has received much attention is to minimize the distortion of distances. The original distances between the objects are distorted when, after classification, the objects are shifted to class representatives or the mutual distances are represented by a dendrogram. Various measures of distortion could be defined. The cophenetic correlation coefficient is a well known example, applicable to hierarchical classifications (2.6.3 and 2.6.4). A special version for non-hierarchical classifications has been proposed and studied by Koontz & Fukunaga (1971).

General discussions on this type of objective function are found in, for instance, Lerman (1970), Anderson (1971) and Gower (1972). The latter two authors applied the principle also to ordination.

3.4.2 Side conditions

Frequently an objective function is defined such that not all aspects of the goodness or usefulness of a solution are taken into account. One or more side conditions are then added to exclude 'inadmissible' solutions. A solution is thus sought by which the objective function has its extreme value *and* which satisfies the side conditions. There is a certain interchangeability. Considerations of usefulness that are taken into account by side conditions could in principle also be embodied in an objective function, and *vice versa*.

The preceding already implies that restrictions to be imposed by side conditions may be of varying nature. For instance, in addition to minimization of an overall measure of heterogeneity, like $\mathrm{tr}(W)$, one may require the heterogeneity of each individual class to be less than a specified maximum, or the number of objects in each class to be within a given range. Such side-conditions have been implemented by, for instance, Ball (1966) and Crawford & Wishart (1968).

Special attention will be paid to two important types of restrictions. The first is concerned with the number of classes and the second with the contiguity of the classes in geographical sense.

3.4.2.1 Number of classes

An objective function may either have an explicitly built-in number of classes, or may yield this number on application to data. In practice, however, the number of classes will often be specified *a priori* as a side condition.

If separation of populations is the aim, the number of classes would ideally equal the number of populations. Determination of the latter, however, is usually part of the problem. One approach to this question is to rephrase it as a problem of testing. Never-

theless it still seems to be largely unsolved. Friedman & Rubin (1967), Scott & Symons (1971) and Sclove (1973) commented on the issue; Engleman & Hartigan (1969) did empirical research in the univariate situation. Marriott (1971) reasoned that, while minimizing $\det(W)$ for partitions with increasing number of classes k, a marked fall of $k^2 \det(W)$ below any obvious trend may indicate an optimal k. Webster & Burrough (1972a) illustrated this approach with soil data.

In minimizing heterogeneity of classes, expressed, for instance, as $\mathrm{tr}(W)$, some restriction on k is necessary. Otherwise one ends up with as many classes as objects. Then no simplification of the data would have been attained at all. Too many classes would make a classification intractable. Decreasing k generally facilitates the use of the classification but increases heterogeneity. Thus some compromise must be made between homogeneity and tractability.

In the absence of definite external considerations, it is intuitively appealing to choose k such that increase of k results in relatively little gain in homogeneity (4.3.4). Landwehr (1972) used the jack-knife method to study the variance of parameters resulting from cluster analysis of a sample set of objects. Rather than the maximum mentioned above, a useful upper bound on k could perhaps be indicated by setting bounds to the size of confidence sets. But more work has to be done.

Of course, when the classification is used for soil survey, then k specifically affects the intricacy of the map. Human perception of maps is now being analysed by cartographers and psychologists, but this analysis needs further development. At present, the influence of k on tractability in various circumstances seems unspecifiable with an exactitude sufficient to allow any realistic formalization of this aspect in an optimization model. For the time being, one must content oneself with determining k in the conventional way, by trial and error, guided by experience.

3.4.2.2 Geographic fragmentation of the classes

By numerical classification as discussed so far, only the position of objects in variable space is taken into account, not that in geographic space. However, since a main purpose of soil classifications is to provide a basis for soil survey, the geographical aspect is essential. When this is not taken into consideration, one might end up with classes of which the occurrences form too intricate a pattern to be revealed by a routine survey or to be represented on a map in readable form. When the intricacy of a map would hinder information retrieval, it may be efficient to sacrifice some class homogeneity for a more readable map. This could also reduce effort in map compilation. Here the same kind of compromise should be found as in choosing the number of classes (3.4.2.1). So far, two strategies have been proposed to exploit data on geographic position; the first by setting a side condition, the second by building it into the objective function.

Frequently suggested by geographers but quite drastic is the construction of classes such that each one only occurs in a single contiguous part of the area to be mapped. This approach is known as *regionalizing*. Fragmentation of classes is entirely avoided, but the area in which a class occurs could still have intricate boundaries. Furthermore, such a rigid side-condition may seriously affect class homogeneity, in that there is no reason why similar soils should be adjacent. An additional difficulty is that one must establish which pairs of objects are contiguous and which are not. However if the objects are points in

two or three dimensions, such as soil profiles, any definition of contiguity is apt to be somewhat arbitrary. Establishing contiguity may be laborious if there are many irregularly distributed objects. Examples of this strategy are given by Spence (1968), Gabriel & Sokal (1969) and Dale et al. (1971). Webster & Burrough (1972b) applied it to soil data.

Another strategy consists of combining distances in variable and geographic space into a single measure of disparity. This is more flexible; in principle the intricacy of boundaries is reduced and some fragmentation of the classes is allowed for. Bunge (1966) proposed a simple version of this strategy, by computing a linear combination of both distances. This might also be attained by simply adding the (suitably weighted) geographic coordinates to the set of variables. Webster & Burrough (1972b) rightly pointed out that this method has the tendency to place widely separated but otherwise similar objects in different classes. They remedied this by reducing influence of geographic distance if that distance were large. To that end, they computed the combined distance D^*_{ij} between objects i and j by e.g.:

$$D^*_{ij} = \frac{D_{ij} + \frac{w}{d^2_{ij}} \cdot \frac{d_{ij}}{d_{max}}}{1 + \frac{w}{d^2_{ij}}}$$

or $D^*_{ij} = D_{ij}\,(1 - e^{-d_{ij}/w})$,

where: D_{ij} = distance in multivariate space between objects i and j,
d_{ij} = geographic distance between objects i and j,
d_{max} = geographic distance between the most distant pair of objects,
w = weighting factor.

They applied these 'smoothing' formulae on data from 84 soil profiles in a 100 m x 100 m grid and the Canberra metric (3.2.1.4) as D_{ij}. By the unweighted pair-group method (3.2.2.1.1), they found less fragmentary classes occurring in more compact areas and with hardly larger within-class variation. The above formulae still have the weakness that with geographically close, but otherwise dissimilar objects an extremely large dissimilarity may be required for allocation to different classes. This becomes acute for an irregular grid and discontinuously varying properties. Other procedures were proposed by Haralick & Kelly (1969), Taylor (1969), Dale et al. (1971) and Monmonier (1972).

The state of the art seems to be one of a collection of methods which at best are suitable for special situations only, and are based on no more than our superficial knowledge of the psychology of map perception. Methods that simultaneously employ information about properties and geographic position deserve further development. For the time being it seems sensible to use only the properties in classifying and, if necessary, to generalize the map afterwards.

Other discussions on this topic are found in, for instance, Berry & Marble (1968), Johnston (1970) and Spence & Taylor (1970).

3.4.3 Optimization

The search for the classification for which an objective function reaches its extreme value, possibly after setting side conditions according to the purpose, is a strictly mathe-

matical problem. It usually entails a difficult type of non-linear programming problem. Several procedures are reviewed below, firstly those leading with certainty to the solution, next the ones that try to approach it.

3.4.3.1 Exact solution

The analytic solution for minimum variance partition of a univariate normally distributed population has been given by Cox (1957). In general an analytic solution is not feasible.

A straightforward procedure to find the best partition of a sample set of objects is to generate all admissible partitions in turn and to evaluate the objective function for each of them. Unfortunately, as the number of objects grows, the number of possible partitions becomes very soon inhibitive. More specifically, the number of possible partitions of n objects into k non-empty classes is, according to Feller (1950):

$$P(n,k) = \frac{1}{k!} \sum_{i=0}^{k-1} (-1)^i \binom{k}{i} (k-i)^n$$

and for $n >> k$, approximately: $P(n,k) \approx k^n/k!$.
Thus, for instance, $P(10,5) = 42\,525$, $P(20,5) \approx 75 \times 10^{10}$ and $P(50,5) \approx 10^{33}$.

Enumeration (consideration one by one) of all possible solutions was applied, for instance, by Edwards & Cavalli-Sforza (1965) for minimizing tr(W) at each successive division in a hierarchical analysis. They found that a set of only 16 objects was about the largest to be analysed within reasonable computer time. But a sufficiently accurate representation of variations and co-variations of a series of soil properties will often require a sample size of hundred or more.

Thus, regardless of the adopted objective function and the state of computer technology, some form of directed search will nearly always be necessary. Methods for this must be derived from the nature of the objective function. In the case of tr(W) the following possibilities exist.

Firstly, Fisher (1958) deduced from the fact that tr(W) is additive over subsets, the following suboptimization theorem:

Let $A_1 : A_2$ denote a partition of set A into two disjoint subsets A_1 and A_2 and let P_i^* denote a least squares partition[5] of A_i into k_i subsets ($i = 1,2$). Then, among the class of subpartitions of $A_1 : A_2$ employing k_i subsets of A_i ($i = 1,2$) a least squares subpartition is $P_1^* : P_2^*$.

In other words, the search of a least-squares partition of subset A_1 may be carried out independently from that of subset A_2. Application of this theorem may considerably decrease computer time. Fisher (1958) investigated the problem of partitioning a time series of 96 values into 10 contiguous classes (i.e. periods). He reported that the best admissible partitions for all k from 1 to 10 were actually obtained in 3 minutes, while without applying the theorem more than 280 years would have been required for $k = 10$ only.

The suboptimization theorem was also used by Jensen (1969), who treated the search for least squares partitions as a problem of dynamic programming. He analysed the

5. That is, a best partition under tr(W).

number of times that contributions from subsets to tr(W) have to be calculated. A set of 20 objects to be partitioned into 5 classes for instance would require $5P(20,5) \approx 5 \times 75 \times 10^{10}$ such calculations when proceeding by enumeration (see above). Although Jensen's dynamic programming method would reduce this number to a thousandth, the remaining amount of computation is clearly prohibitive for so small a problem as this. The reason why suboptimization seemed promising in Fisher's example apparently is the imposed contiguity constraint.

The second possibility to direct the search arises from the geometric consideration that each pair of subsets in a partition with minimum tr(W) must be separated by a hyperplane (Gower, 1967). Any solution which does not satisfy this condition of 'convexity' can be ignored *a priori*. The number of distinct dichotomies of a set of n points in m dimensions induced by hyperplanes is in general:

$$P_m(n) = \sum_{i=1}^{m} \binom{n-1}{i},$$

which for $n-1 \leqslant m$ is interpreted as 2^{n-1} (Harding, 1967). For large n and small m, this is approximately equal to $n^m/m!$. In a two-dimensional space, all such dichotomies could be generated in a systematic way by an algorithm of Harris et al. (1972), who also outlined a strategy for higher-dimensional spaces. As indicated above, the condition of convexity also applies for more than two classes. However, I do not know of any efficient procedure for exploiting this, apart from the one-dimensional case, for which the number of partitions still to be considered is reduced to $\binom{n-1}{k-1}$ (Dagnelie, 1966).

Neither the suboptimization theorem nor the convexity condition is sufficiently powerful to allow exact solutions of problems of the size usually involved in soil classification.

3.4.3.2 Approximate solution

The excessive number of options for testing led some authors to restrict the search to a manageable *sample* of partitions. It is then hoped that the best or near best partitions will be included in the sample. The chance that this will happen might be increased by raising the sample size or by sampling only some more promising sub-class of partitions. The disappointing results of Fortier & Solomon (1964) with this approach were attributed to a long tail in the distribution of the objective function. See Dagnelie (1966) and Switzer (1970) for other examples of this method.

Quite different from sampling is the wide class of *relocation methods*. Here an initial partition is taken as a starting point, and is improved stepwise.

Most relocation methods proposed so far aim at minimizing E_c, the mean squared Euclidean distance of an object to the representative of its class (3.4.1.1). Obviously, with a series of fixed class representatives, E_c is minimized by allocating each object to the class with the nearest representative. Inversely, with a given partition, E_c is minimized by taking the centroids as representatives of the classes (then E_c equals tr(W)/n). But the problem is that both optimum partition and optimum class representatives are unknown; they must be found simultaneously. Relocation methods try to solve this by *relaxation*; the errors resulting from an initial approximation are considered as constraints to be relaxed.

To begin with, one determines the set of class representatives, C_1, that is optimum for a given initial partition P_1. This partition should then in general be altered to render it optimum for C_1: this results in P_2. Now C_1 is no longer optimum: one obtains a new set of centroids, C_2, which may give rise to a new partition, P_3, and so on. Thus if P_i and C_i denote the partition and the class representatives, respectively, in stage i, the iteration will proceed as:

$$P_1 \rightarrow C_1 \rightarrow P_2 \rightarrow C_2 \rightarrow \ldots \rightarrow P_i \rightarrow C_i \rightarrow P_{i+1} \rightarrow C_{i+1} \rightarrow \ldots$$

In this basic form, the principle has been applied by, for instance, Jancey (1966). The process may be stopped according to various rules, for instance when the number of iterations exceeds a given limit or the rate of improvement is insufficiently large, or when iteration does not change the partition any further.[6] Even in the latter case there is no guarantee that tr(W) will reach the global minimum; counter-examples are available. Dependent on the starting-point (P_1) the process may end up in some local minimum. This is the major drawback of relocation methods. Various sophistications of the basic principle have been proposed that try to avoid local minima and to speed up the process. Some of these are outlined below.

Acceleration has been suggested by Jancey (1966). In the present context it means that, on arrival at C_i, one tries to bypass one or more iteration steps by extrapolating towards C_{i+j} ($j > 1$) on the basis of the preceding centroid positions, assuming that these positions converge to a limiting position. So far this method does not seem to have found any application.

Another option, now frequently employed, is to adapt the position of the representative of any class immediately after its content has been changed, rather than to postpone re-calculation of the centroids until every object has been relocated. This admittedly introduces greater computational effort for each iteration step. However, some objects may now be transferred to other classes one step earlier than otherwise, because they will be compared with recent centroids rather then with backward class representatives. The final solution now depends not only on the starting-point but also on the order in which the objects will be relocated. Immediate re-calculation of centroids has been applied by, for instance, MacQueen (1966), who however relocated each object only once.

The procedure of immediate re-calculation provides further options. If any object is simply allocated to the nearest current centroid at any stage, such allocation will be conservative in that it is positively biased towards the centroid that represents the object at that stage. If the object is not allowed to contribute to the calculation of the centroid of its class, its distance from that centroid tends to increase. Furthermore, the centroid of any other class would have been nearer if that object were allocated to it beforehand.

The above can be introduced by allocating any object such that the resulting tr(W) is minimized. The decrease in tr(W), resulting from switching object t from its current class P to any other class Q, can be shown to equal

---

6. The finiteness of the present iterative procedures is guaranteed by the fact that there exist only a finite number of distinct partitions of a finite set of objects, and that only operations are volved which decrease the objective function. See Needham (1966) for a formal proof.

$$\frac{n_P}{n_P - 1} \cdot d_2^2(t,p) - \frac{n_Q}{n_Q + 1} \cdot d_2^2(t,q).$$

where n_P and n_Q denote the number of objects in P and Q, respectively, and p and q denote their centroids before the switch. This refinement is available in the program of Wishart (1969b) and referred to as 'removal'. It was adopted as a fixed procedure for the experiments reported in Chapter 4.

One could also consider a switch of several objects together. If any subset T of class P, with n_T objects and centroid t, would be switched to class Q, then tr(W) would decrease by (Howard, 1966):

$$\frac{n_P}{n_P - n_T} \cdot d_2^2(t,p) - \frac{n_Q}{n_Q + n_T} \cdot d_2^2(t,q)$$

At equilibrium after iterative allocation of single objects, it may still be possible to reduce tr(W) further by switches of this more general type. In other words, more local minima might be avoided. But, of course, now any switch has to be selected from a much larger number of alternatives. Furthermore, the fact that a solution cannot be further improved by the present type of switches is not sufficient condition for optimality. The process could avoid local minima better if a more general type of switch were considered, that of transferring any set T from class P to class Q simultaneously with any set U from Q to P. But then the number of alternatives would increase considerably, although reaching the optimal partition would not be guaranteed if there were more than two classes.

When no further improvement can be achieved by switches of single objects, one could, instead of searching systematically, try some haphazard reshuffling of the objects to get the process going again. This has been done by, for instance, Forgy (1965) and Rubin (1967).

One could perhaps affect the final solution favourably by starting at a well chosen initial partition, rather than just any one. This, in addition, could also reduce the number of iterations required. Favourable initial partitions may be obtained by Ward's method (3.2.2.1.1), because this minimizes tr(W) at each fusion. The initial partition could be read from the dendrogram, at the level corresponding with the desired number of classes. When the number of objects is too high for hierarchical classification one could apply Ward's method to a random subset, calculate the centroids of the resulting classes, and use those centroids as a starting point for iteration with the entire set. See Chapter 4 for examples of this procedure.

Wishart (1970) studied the effect of different starting-points on the final partition, using artificial bivariate data. See 4.3.2 for similar trials with observed multivariate data.

Risk of local minima has been investigated by Harris et al. (1972) for two sets of objects. The one set dealt with 97 species of bees as described by Michener & Sokal (1957). The other was taken from the classical data of Fisher (1936) on 3 species of iris with 50 objects each. In both cases the prescribed number of classes was two, and only the scores on the first two principal components were used. The authors concluded from their analysis that for the irises, there was only one minimum of tr(W); this global minimum would be found whatever the starting-point for iteration. For the bees, there were two local minima; with an initial convex partition chosen at random, the probability

of reaching the global minimum is 0.62.

Ball (1966) developed a program for minimizing tr(W) in combination with side conditions on the size and heterogeneity of individual classes and on their separation. The number of classes is not kept constant during this process.

Relocation methods may also be used to optimize objective functions other than tr(W). For instance, Koontz & Fukunaga (1971) used relocation in preserving distance while Sclove (1973) discussed it in considering separation of populations. A general review of relocation methods can be found in, for instance, Bolshev (1969).

4 Experiments with a classification procedure

Based on the considerations in Chapters 2 and 3, a classification procedure has been devised that could be introduced in soil survey. In 4.1, the procedure is outlined and its rationale given. In 4.2, it is described in detail and illustrated by an application. In 4.3, investigations on various particular elements of the procedure are reported.

4.1 Outline and rationale of the procedure

The procedure discussed in this chapter is intended to be an aid in constructing suitable classifications for detailed and semi-detailed soil surveys as done now in the Netherlands. These surveys are mainly in Holocene clay and peat areas, and in Pleistocene sand and loam areas. They support the planning of land-use, agricultural or non-agricultural.

Soil survey for planning land-use generally tries to provide, at lowest costs, the most valuable soil information about a given area. Apart from the efficiency with which a user can retrieve data from the map and the memoir, the value of the information depends on the relevance and the precision of the data presented, and on the errors included. The precision and relevance, respectively, depend on the number of mapping units and the properties in terms of which they are described. I believe that at present neither of these two factors, nor the mentioned efficiency of data retrieval can be satisfactorily included in a comprehensive model for optimization (see also 2.2.2 and 3.4.2). The role of numerical classification in the present procedure will thus concentrate on minimization of error.

In 3.4.1.1, an overall measure of error, denoted by E_t, was suggested for soil maps. It assumed that the survey would be based on a classification specified by class representatives, i.e. for each class a 'central' value is given for every variable, and that mapping units would be delineated that correspond to these class representatives. The best delineation with respect to E_t would be that by which each profile in the area was allocated to the most similar class representative (i.e. at shortest Euclidean distance in the space with the variables as mutually perpendicular axes). In that case, E_t equals E_c: the contribution due to the heterogeneity of the classes. In principle, E_c can be minimized by constructing classes as homogeneous as possible. In practice, however, there will always be an additional contribution, E_d, due to the fact that delineation is accompanied by inaccuracies and cartographic generalizations. This makes construction of optimum classifications for soil survey considerably more complicated. Accuracy of delineation may not only depend on the survey method, but also on the classification used. The advantage of more homogeneity of the classes could even be overruled by worse accuracy of delineation.

In the present procedure, homogeneity of classes is kept as the main criterion, but with certain restrictions in favour of accuracy of delineation. (The principle of optimizing homogeneity in classification was discussed in 2.4.2, along with that of distribution fitting.)

With respect to class homogeneity, it is recalled from 3.4.1.1 that, for a given partition of a set of profiles, E_c is minimum if the centroids of the classes are chosen as representatives. If so, E_c is identical to $\mathrm{tr}(W)/n$, the objective function argued in 3.4.1.1 from a more general viewpoint. (The number of objects, n, is a constant, and so is immaterial for minimization.) This function can be minimized by Ward's method (3.2.2.1.1) or by relocation algorithms (3.4.3.2). That will lead to the disjoint, polythetic type of classes, discussed in 2.4.3 and 2.4.5.

For delineation of classes, one would expect adapted allocation of profile observations often to be more efficient than random allocation (2.2.2). Therefore the classification procedure should not inhibit survey with adapted allocation. So some practical requirements were added to the theoretical considerations above. These requirements were:

a. The classification should be available at an early stage of the field work. It can therefore only be used on a limited set of profile observations made before the actual mapping.

b. This set should be sufficiently representative, so that the classification may remain unchanged throughout the survey. But if some adaptation would prove necessary, then that should be feasible without undue effort.

c. New profiles must be identified on the spot. The classification should thus permit simple and quick identification (2.5).

d. To keep track of soil boundaries in the field, differences between the classes should be easy to memorize.

Suppose that for a given survey, 100 variables are involved in profile descriptions and 30 classes are needed for a sufficiently accurate representation. This is not an unrealistic example. But if, straightforwardly, 30 classes be constructed with centroids consisting of 100 values each, then there are likely to be serious shortcomings on points b, c and d above. Generally, allocation of new profiles to the centroid at shortest Euclidean distance would hardly be possible in the field. Furthermore, identification had to be completely revised for each profile involved, if classes be adapted during the survey. Also differences between classes would be difficult to memorize.

The strategy followed in this study is to split the problem into manageable parts. The procedure is schematically represented in Fig. 6. At the first stage a series of classifications is constructed, each based on a subset of the variables only. This series is combined into an *interim* classification. If there are too many interim classes all to be mapped individually, then at a second stage of the procedure, they are fused to larger classes. The result will be referred to as the *final* classification. Thus the interim and final classification respectively constitute a lower and a higher level of a hierarchical classification, a type discussed in 2.4.4. At both stages, $\mathrm{tr}(W)$ is minimized, though with constraints. The procedure can be outlined as follows.

First stage. As a preliminary, the set of variables is divided into subsets, and on the basis of each subset, a classification is constructed by relocation. The numbers of classes are chosen by the user. In the following, these classifications will be referred to as *special* classifications. The interim classification is then obtained as a combination of the special classifications, i.e. all soil profiles allocated to the same special classes together constitute one interim class.

If each variable represents the value of a given property (e.g. the percentage of clay) in

a given depth interval of any soil profile, then the variables may be grouped by property or by depth interval.

If the latter, each special classification regards one interval, and the interim classes are characterized by a unique sequence of types of layers. Norris & Dale (1971) followed this strategy, though in combination with other cluster methods. Separate classifications of layers would be unsatisfactory in our case, where the profiles are divided into 20 depth intervals. Sequences of 20 layer types would be too difficult to handle by the users (in the first place the surveyor).

If the number of properties is not too large, grouping the variables by property yields a more manageable classification. This type of grouping is chosen in the present study. Each special classification is thus of a single property, and the classified objects are profiles of that property only, as (could be) observed in vertical direction. Such a profile will be referred to as a *depth profile* (as for clay content), in contrast to the soil profile, which is the combination of depth profiles for the respective properties. The centroid, or more generally, the class representative of a class of depth profiles will be termed a *central depth-profile*.

Classifications of depth profiles for *clay* and *carbonate* have long been used by the Netherlands Soil Survey Institute (Bodemkaart van Nederland, 1:50 000, 1964). Study of soils by depth profiles, in addition to horizons, has been argued by Bennema (1974) for general reasons.

The advantage of constructing the interim classification from special classifications lies in economy of class definition. The representative of each interim class is the union of certain central depth-profiles from a common collection. If there be m_j intervals and k_j classes for the jth property then, with l properties, the whole collection of central depth-profiles will consist of $\sum_{j=1}^{l} m_j k_j$ average values. That may be far less than the mk values involved when the set of soil profiles would have been partitioned into k classes on the basis of all m variables simultaneously.

Second stage. The number of interim classes equals the number of different combinations of central depth-profiles occurring in the given sample. This number depends on the number of central depth-profiles chosen for each property, and on statistical dependence between the properties; it is not known beforehand. If too many classes arise all to be mapped individually, their number can be reduced by fusion of classes at the second stage. Only the final classes so constructed need to be mapped.

The fusions may be chosen by Ward's method in order to minimize $\mathrm{tr}(W)$ as at the previous stage. Alternatively, the principle of minimizing $\mathrm{tr}(W)$ could be combined with considerations regarding the geographic contiguity of the classes (3.4.2.2) or their relationships to the landscape.

As indicated in Fig. 6, weighting of the variables will be at both stages of the procedure: first the intervals are weighted, then the properties.

One advantage of the present procedure is that identification of new profiles is easier, hopefully easy enough to be acceptable as part of the field work. When a soil profile is to be identified, the central depth-profile at shortest Euclidean distance is selected for each property successively. If the number of central depth-profiles is not too high, this is relatively easy, especially if the central depth-profiles are represented by diagrams, as in

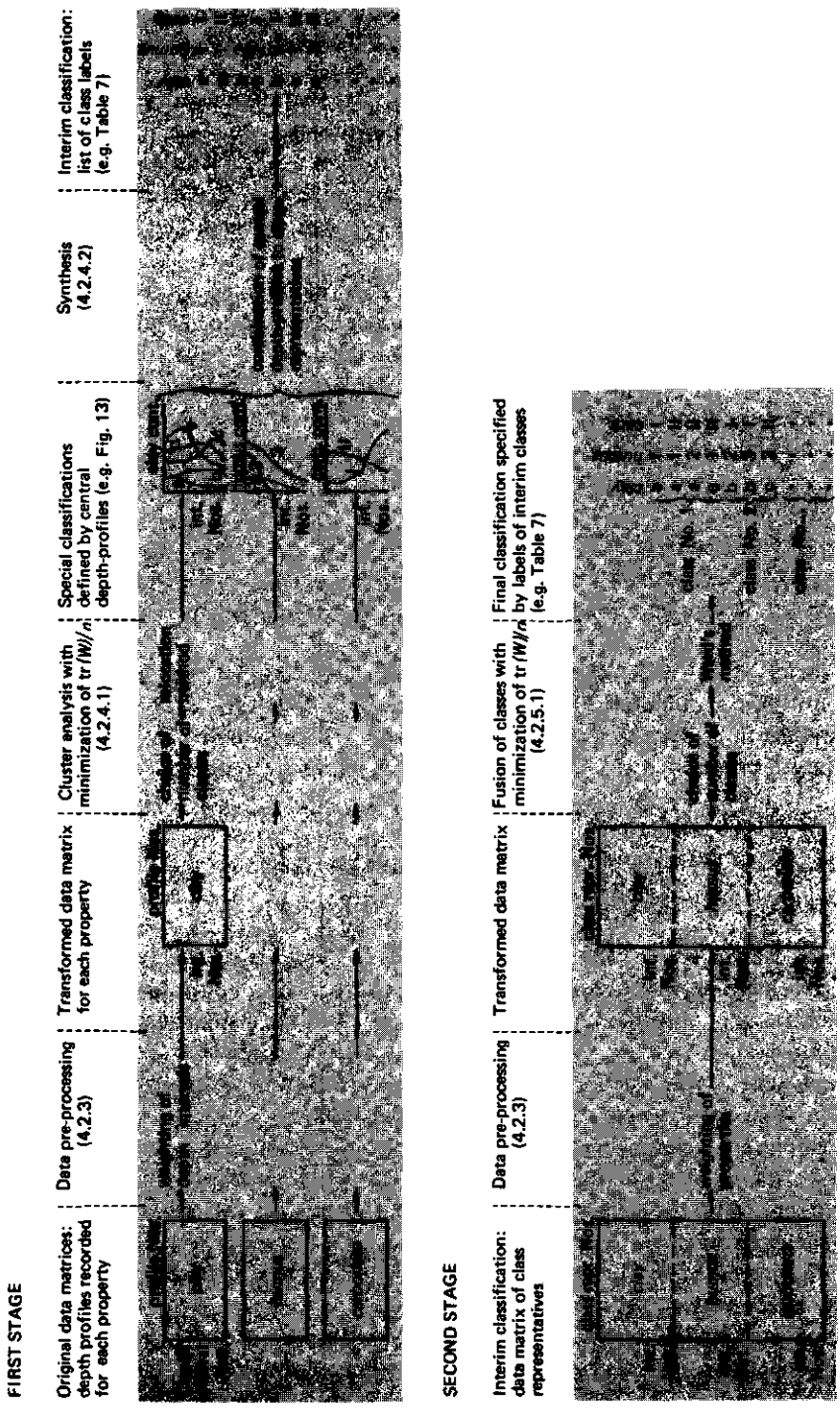


Fig. 6. Scheme of the classification procedure. (For simplicity only three of the soil properties used in Section 4.2 are included.)

Fig. 13. The identification can be recorded as a string of characters on the field-map. The jth character in such a string would then refer to the central depth-profile to which the observed depth-profile for the jth property has been allocated. Any such character string is the label of an interim class and, given a particular one, the corresponding final class can be read from a table or key specifying the grouping as determined at the second stage.[7]

A further advantage is that earlier identifications to interim classes need not be revised when the final classification is adapted during the survey. Also, differences between classes are easier to memorize, because one need not consider all variables at once.

4.2 Detailed description and application

In this section, the various steps of the procedure are discussed in detail and exemplified with an application. Since the procedure is intended for routine soil survey, it has been applied to data as usually collected in such surveys.

Use was made of the fact that the complete set of data from one of the surveys by the Netherlands Soil Survey Institute (Kamping & Rutten, 1969) had already been put onto punch cards for other purposes. This data set was used for the experiments in the present and the next section.

The classes constructed were not delineated in the field because this would have required a new survey. So the classification was not finally assessed (2.6), but the geographic distribution of the classes was studied from the data already collected. Comparison of the maps from the experimental classification with those from the conventional was pointless, because the weights assigned to the variables differed considerably between the two classifications (4.3.4).

4.2.1 Purpose of the classification

The survey for which a classification was required was typical of the commissioned surveys on scale 1 : 25 000 in marine clay areas in the Netherlands. The survey was in support of a reallotment project in the area described below. Such a project is comprehensive in that it is intended to improve physical conditions, in a general sense, for agricultural production by the individual farmers. Also non-agricultural land-use (such as recreation projects) may be included in the plans. Important elements of a reallotment project are usually: improvement of drainage and permanent soil conditions related to structure and micro-relief, more efficient parcelling, access roads and site of buildings.

Soil data are used in many aspects of planning. For instance, they may indicate:

- how and where drainage and soil structure could be improved,
- how levelling should be done,
- optimum changes in the plan of roads and watercourses,
- estimation of production capacities.

The use of soil data in reallotment projects has been analysed in detail by Naarding (1970).

---

7. If the combination at hand be not included in the pilot sample, neither will it be represented in the key. The latter can as yet be completed by allocating such a combination to the class containing the combination at smallest Euclidean distance.

4.2.2 Data collection

4.2.2.1 Test area

The survey of Kamping & Rutten covered a reallotment area of 81 km^2 around the villages of Stedum and Loppersum, in the north-eastern part of the Province of Groningen (Fig. 7).

Geologically the test area consists of young marine clay. Three types of landscape were distinguished by Kamping & Rutten (1969). The oldest type originated about 900 BC from tidal marsh with banks and basins. It forms a wide belt from east to west through the area. This region was settled before the Christian era. To escape periodic floods, the earliest farmers raised dwelling mounds ('terpen' in Dutch) on which to build their villages or single farms. During the last centuries, some mounds were removed, the earth being used for fertilizing elsewhere.

In a period of marine incursions from about 250 until 500 AD the sea penetrated through the banks into the southern hinterland. The landscape that developed there is in general flat and slightly lower than the one mentioned above. The soil in the south-western part of the area has a typical sticky structure, associated with conditions of deposition at that time. Several fields have been quarried for brick-making.

The northern part of the area was formed mainly during the Middle Ages when, bit by bit, the coastal marsh in the former shallow bay Fivel was being embanked.

Traditional crops are sugar-beet, potato, wheat and barley. Grassland is an increasing land-use.

Kamping & Rutten produced a soil map on scale 1 : 25 000. The legend was constructed by subdividing some subgroups of the Dutch soil classification (de Bakker & Schelling, 1966). Nearly all soils fell into the subgroup of Polder vague soils. In the US system (Soil Survey Staff, 1975) many would be identified as fine-silty, mixed (calcareous), mesic Typic Fluvaquent, other as fine-clayey, illitic (non-acid), mesic Typic Fluvaquent, sometimes slightly intergrading to a Natraqualf. Plaggepts occur on dwelling mounds.

4.2.2.2 Profiles

Soil profiles were described by augering to a depth of 2 m.

The description of any such profile hereafter represents an object. A 'free survey' system was followed, so the siting of borings was adapted instead of random. As it happened, the borings proved finally to be evenly distributed over the area. In total, 2212 borings were made, giving an average density of about 27 per square kilometer.

4.2.2.3 Variables

Six properties were observed throughout the whole depth of each boring. Because the depths of transitions had been rounded off to 5 cm, the borings can be divided into 40 intervals of 5 cm, for each of which 6 values are available: one value for each property. So initially there were 6 x 40 = 240 variables.

In the following the properties are successively discussed. In this section and the next,

they will be designated by the italicized names given below.

Clay and *humus*. Contents of clay and humus were estimated visually and by finger test, supported by reference samples. Clay is here defined as the fraction of mineral particles smaller than 0.002 mm in equivalent diameter.

By humus is meant the organic matter in the upper part of the soil.

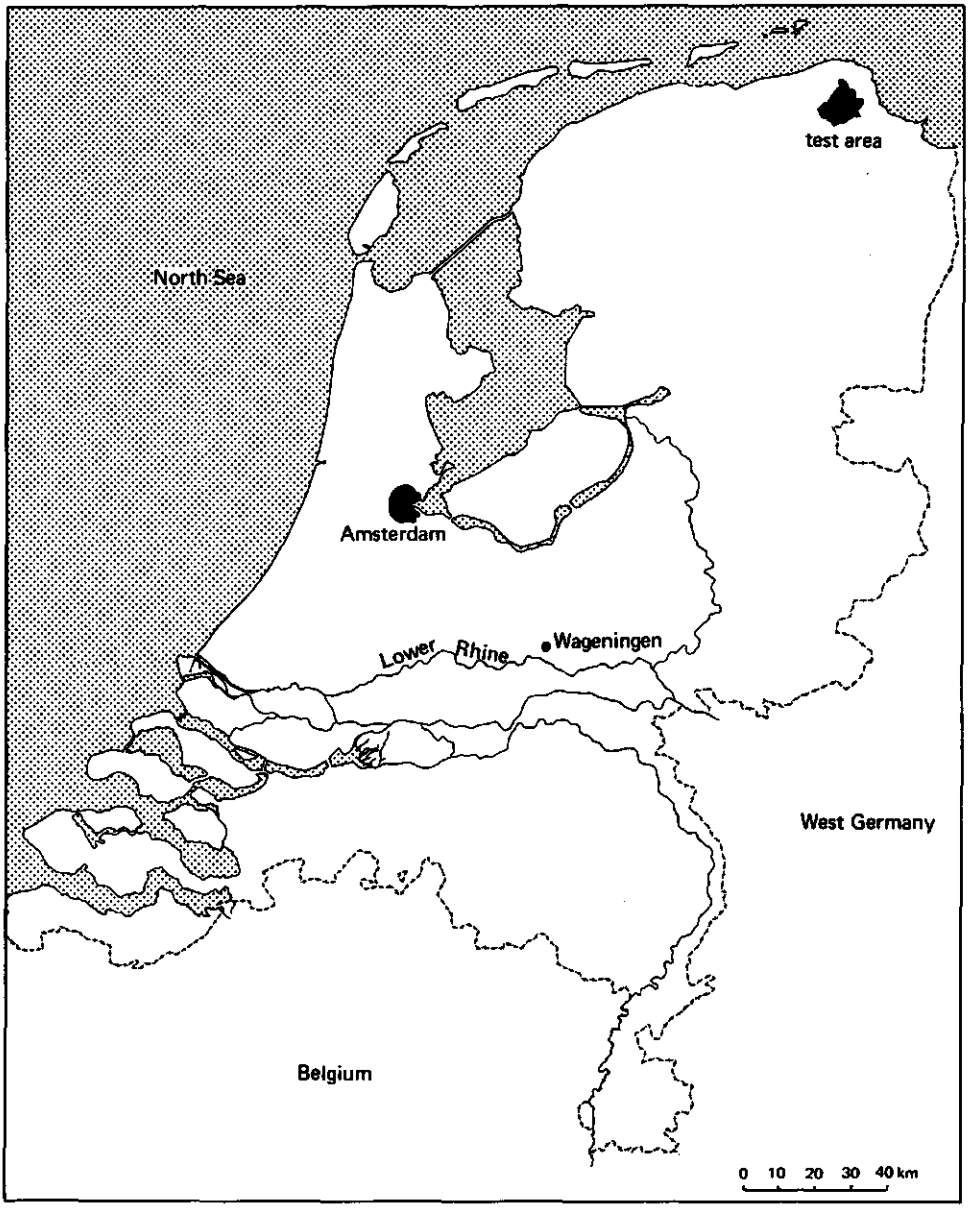


Fig. 7. Location of the test area.

Clay and humus contents were recorded as percentages by mass of the mineral constituents and of total soil material, respectively. Their ranges of variation differed considerably: for clay this was between 3 and 54%, and for humus was between 0 and 10%.

Carbonate. A rough indication of the amount of carbonates was obtained by observing the reaction with 10% HCl. Originally, 'no', 'weak' and 'strong' effervescence was recorded, as –, + and ++, respectively. So for each interval, a 3-valued ordinal variable (2.3.1.1) was created. Comparison with laboratory data showed that the carbonate contents corresponding to the reactions mentioned above are roughly proportional to the values 0, 1 and 2, respectively. The codes were transformed to these values, so that in the following the variables could be treated as metrical.

Ripening. Physical ripening was assessed by squeezing the material by hand. The three values originally recorded were: 'fully ripened', 'half ripened' and 'not ripened'. It was considered that in this case the difference in ripening between 'fully' and 'half' was roughly of the same importance as the difference between 'half' and 'not'. Therefore, as for *carbonate*, the values were transformed into 0, 1 and 2, respectively, again resulting in 3-valued metrical variables.

Knip. A special structure sometimes occurs underneath the top of alluvial soils. In Dutch, it is termed 'knip', which I leave untranslated. It is associated with relatively strong swelling and shrinking. As a consequence the profile has some unfavourable properties from an agricultural point of view. Soils with finer textures become impermeable in the wet season, hence are often saturated with water. In the dry season, these soils break up into hard, blocky or prismatic fragments, and capillary water supply is easily interrupted as a result of high compactness. The phenomenon is less distinct in soils with coarser textures. Its absence and presence has been recorded with 0 and 1, respectively. Because actually many intermediate forms are possible between 'knip' and 'no knip', these variables can be considered as conceptually continuous, like *carbonate* and *ripening*.

Peat. Some subsoils in the area contain peat or peaty material. The amount of organic matter was estimated in five classes:

- no organic matter visible,
- organic matter visible, but less than required for qualification 'peaty clay',
- peaty clay,
- clayey peat,
- peat.

The latter three classes are as defined by de Bakker & Schelling (1966).

Although both *peat* and *humus* data refer to the amount of organic matter, it was decided to maintain the distinction between the two throughout analysis. In the test area, humus and peat are genetically and morphologically different components with different physical and chemical properties. They could always be estimated independently from each other, because they occur in different parts of the profile.

The five classes were originally denoted by alpha-numeric codes. Because the average percentages of organic matter within the classes may be expected to be roughly equidistant, the values have been transformed into 0, 1, 2, 3 and 4, respectively.

4.2.3 Data pre-processing

4.2.3.1 Reduction of the number of profiles

As argued in 4.1, the classification procedure should be applied to a limited number of profiles. A reconnaisance by some 600 borings would be about the maximum acceptable for surveys like the present. To incorporate this restriction, a sample, designated Set B, of 600 borings was drawn from the complete set, A, of 2212 borings. To ensure a fairly even spread of the samples, the area was stratified into 49 sub-areas. Borings were then drawn, at random, in numbers proportional to those in each sub-area. The choice of sample size will be further discussed in 4.3.1.

4.2.3.2 Reduction of the number of variables

The number of variables was reduced by two measures. Firstly, it was felt that doubling the width of the intervals from 5 to 10 cm and taking the average value for each would not distort the information seriously. Information would only be distorted if where transitions were recorded within the new 10-cm intervals. As the total number of recorded transitions per profile is already low (usually less than 5), distortion only occurs infrequently. Furthermore, recording the depths of transitions in multiples of 5 cm must often have gone beyond the actual accuracy of observation. This first measure reduced the total number of variables from 240 to 120 (= 20 x 6).

Secondly, on examination several variables turned out to be constant within Set B. As such variables do not contribute to Euclidean distances, they were excluded from analysis. (During survey, however, one should check how far variation does occur in other samples from a population.) Eighty-eight variables remained; they are represented in Table 2.

To complement the cluster analyses, principal component analysis was applied to the *clay* data of Set B. The matrix of correlation coefficients used for this analysis is given in Table 3. Table 4 shows that the first four components together explain most (92%) of the variance. The first component is roughly proportional to the clay percentage averaged over the intervals. The second, third and fourth component, respectively, correspond to the linear, quadratic and cubic trend with depth. Fig. 8 shows a random subsample of 300 *clay* depth-profiles from Set B, projected onto the first and second component. In this

Table 2. Variables included in the analyses.

Property	Values	Depth interval (downward along profile)	Number of variables
Clay	0-100%	1-20	20
Humus	0-100%	1-12	12
Carbonate	0, 1, 2	1-20	20
Ripening	0, 1, 2	6-20	15
Knip	0, 1	1-8	8
Peat	0, 1, 2, 3, 4	8-20	13
Total number of variables			88

Table 3. Correlations between percentage clay in different depth-intervals, calculated from a Set B of 600 profiles.

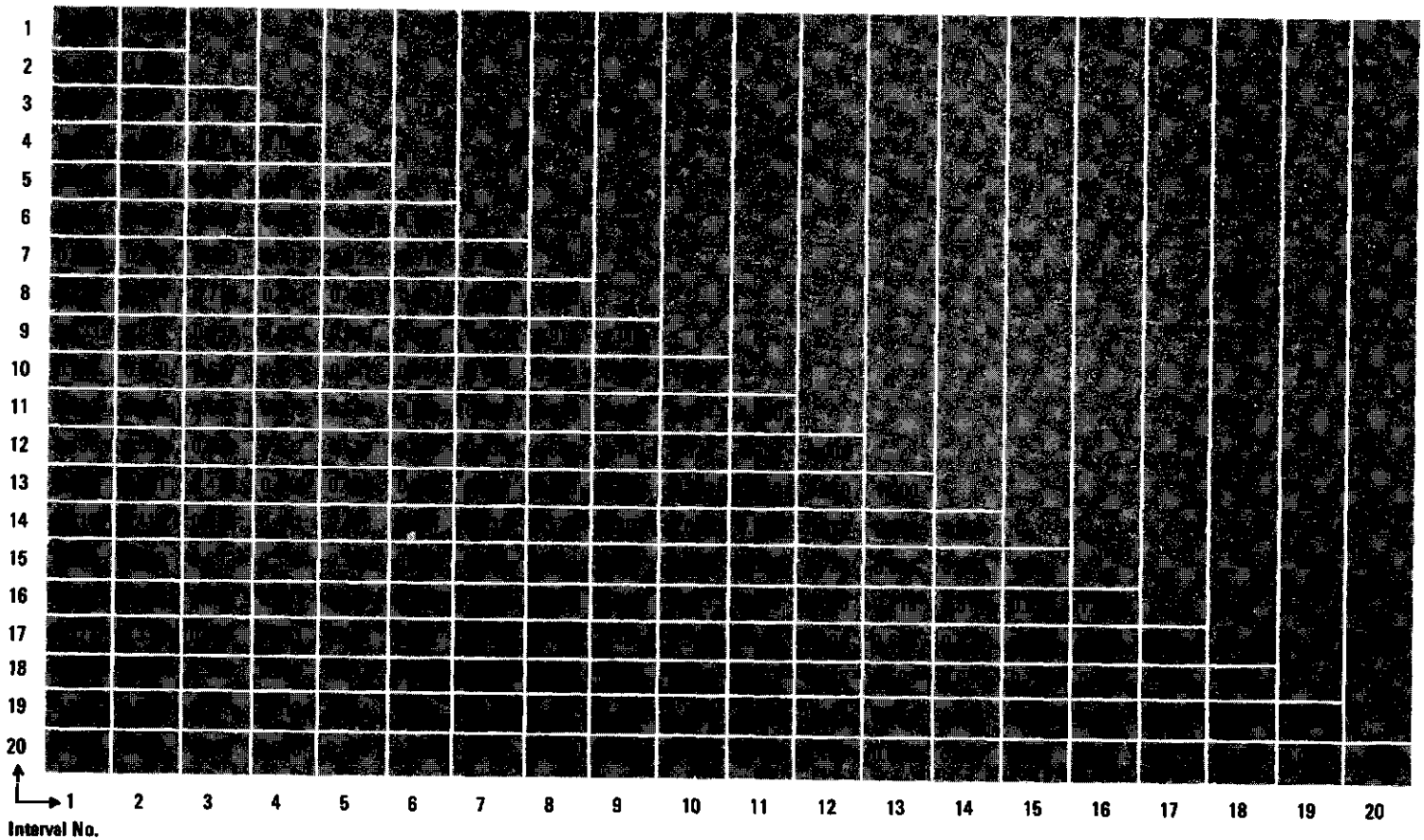


Table 4. Principal component analysis applied to standardized percentage clay in 20 depth intervals of 600 profiles (Set B).

Component	1st	2nd	3rd	4th
Eigenvalue	12.61	3.64	1.42	0.75
Explained variance (%)	63.1	18.2	7.1	3.7
Cumulative variance (%)	63.1	81.3	88.4	92.1
Interval 1	0.220	−0.262	0.195	−0.210
Interval 2	0.220	−0.267	0.190	−0.204
Interval 3	0.223	−0.274	0.164	−0.193
Interval 4	0.224	−0.279	0.148	−0.153
Interval 5	0.226	−0.265	0.147	−0.084
Interval 6	0.234	−0.226	0.073	0.100
Interval 7	0.242	−0.169	−0.036	0.255
Interval 8	0.246	−0.115	−0.103	0.349
Interval 9	0.243	−0.068	−0.199	0.387
Interval 10	0.245	−0.022	−0.258	0.284
Interval 11	0.239	0.043	−0.338	0.069
Interval 12	0.236	0.084	−0.343	−0.080
Interval 13	0.233	0.141	−0.296	−0.235
Interval 14	0.224	0.206	−0.203	−0.344
Interval 15	0.220	0.237	−0.096	−0.316
Interval 16	0.211	0.273	0.067	−0.175
Interval 17	0.202	0.299	0.190	−0.035
Interval 18	0.196	0.302	0.281	0.099
Interval 19	0.191	0.278	0.347	0.217
Interval 20	0.185	0.283	0.350	0.216

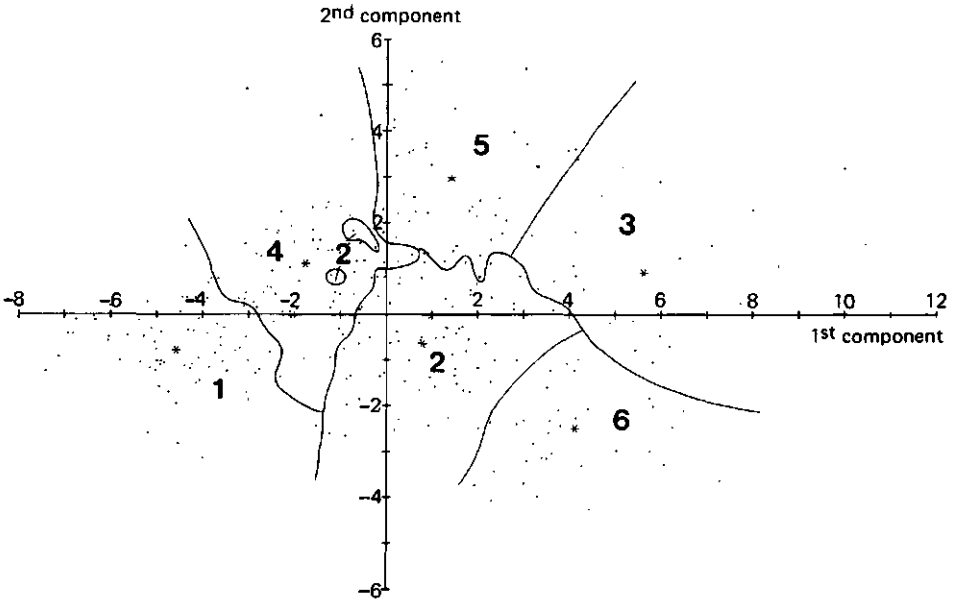


Fig. 8. Scatter of *clay* depth-profiles projected onto the first two principal components (specified in Table 4) and allocated to the centroids in Fig. 12 (with projections marked*).

scatter diagram, the depth profiles are delineated according to their later classification by the relocation method (4.2.4). Natural clusters did not appear from the diagram.

There was clearly much redundancy in the *clay* data, due to correlations between the intervals. The same may be expected for the other properties. However, as mentioned in 2.3.2.1, defining the classes in terms of principal components would hamper identification in the field. So the classification was based on the original variables. The possibility of reduction by deleting a number of depth intervals will be discussed in 4.3.4.

4.2.3.3 Weighting of the variables

The problem of weighting was discussed in general terms in 2.3.1.2. The classification procedure being chosen, the following may be added.

It is recalled from 3.2.1 that with $\mathrm{tr}(W)/n$ as objective function, the effective weight in the sense of Burr (1968) (i.e. the average contribution of a variable to all $\binom{n}{2}$ inter-object distances) equals twice the overall variance of that variable. Weighting will therefore be further discussed in terms of the more usual variance.

The relocation method minimizes $\mathrm{tr}(W)/n$, i.e. the sum of the pooled within-class variances. But apart from the sum, also the distribution of these variances over the variables will affect the usefulness of a classification. By choice of weights, the user can influence the distribution. Multiplication of a variable by a scale factor λ multiplies the overall variance by λ^2, but the effect on the within-class variance is more complicated. This depends on the number of classes and on the multivariate frequency distribution, as will be discussed in 4.3.4.

For an optimum choice of weights several aspects must be considered. If assessment factors (e.g. moisture supply, bearing capacity) are to be estimated for each mapping unit, then one would like to minimize the error of such estimates by suitable weighting of the variables on which the classification is to be based. For that purpose, the classes could be made as homogeneous as possible with respect to the assessment factors, by weighting the variables in proportion to their influence on the factors. In addition one could increase the accuracy of delineation in the field, by giving higher weights to variables that can be observed more accurately and that have closer correlation with visible landscape features.

An entirely quantitative approach to weighting encounters many difficulties (e.g. lack of data and non-linearity of relationships). Therefore approximations must suffice for the moment. One way would be to question a panel of experienced soil scientists provided with a specification of the purpose and the method of the survey. (In conventional classification, the process of weighting is implicit, whereas here the soil scientists would be asked to express their opinions in a quantitative form.) This has not been undertaken in the present study. Instead I confined myself to a provisional, rather arbitrary choice of weights, and in a separate study I concentrated on the effects of differences in weighting on the classification (see 4.3.4).

Weights are to be chosen at both stages of the classification procedure. At the first stage, the *depth intervals* are weighted before constructing the special classification. This can be done separately for each property. In the present soil survey, there was no immediate evidence that features of the subsoil, for instance, of interest for drainage, were markedly less important than those of the topsoil. Therefore, for every property the scale factors were set at 1 for each interval. In other words, the special classifications were

based on the original data. The overall variances of the original variables in Set B are presented in Fig. 17.

At the second stage, the *properties* should be weighted, because the interim classes to be fused are defined by all properties simultaneously. The weights of the intervals are already fixed, so weighting of the properties is confined to application of a common scale-factor to all data on the same property. These factors should be selected in relation to the number of classes in the special classifications. The homogeneity of the final classes (i.e. classes of the final classification) for a given property, depends on the homogeneity of the interim classes and on how these are fused by the classification method. The homogeneity of the interim classes is controlled by the choice of the number of special classes for that property. Fusion of the interim classes is controlled by the choice of scale factors.

In the present application, six classes were constructed for *clay*, three for *humus* and *carbonate*, and two for *ripening*, *knip* and *peat*. This choice seemed not unreasonable in view of the relative importance of the properties, their range of variation in the area, and the accuracy of observation. Because the importance of the properties was assumed to be sufficiently accounted for already in these numbers of classes, the scale factors were chosen such that the overall variance, calculated from the central depth-profiles and summed over intervals, was equal for each property.

4.2.4 Classification at the first stage of the procedure

At this stage, a special classification is separately constructed for each property. The special classifications are then combined into an interim classification.

4.2.4.1 Special classifications

A special classification for *clay* was constructed by iterative relocation of the 600 *clay* depth-profiles of Set B. The same was done for the other properties. Relocation methods are dealt with in 3.4.3.2. The following method was selected.

Step 1. Choose the number of classes.

Step 2. Choose an initial partition of the set of depth profiles and calculate the centroids.

Step 3. Relocate successively each depth profile, such that tr(W) is minimized. The decrease in tr(W) by switching a depth profile t from its present class P to any other class Q equals:

$$\frac{n_P}{n_P-1} \cdot d_2^2(t,p) - \frac{n_Q}{n_Q+1} \cdot d_2^2(t,q),$$

where n_P and n_Q denote the number of depth profiles in P and Q, respectively, before the switch, while p and q denote the centroids of P and Q. Whenever a depth profile is switched, the two centroids involved are immediately recalculated. (This makes the result dependent on order, as will be discussed in 4.3.3. In the present example the depth profiles were relocated in random order.)

Step 4. Iterate Step 3 until no switches occur during the previous iteration.

This relocation algorithm can easily be generalized to incomplete data sets. This is achieved by calculating changes in tr(W) not by Euclidean distances, but by summation of the contribution of the individual variables. The decrease in the pooled within-class variance by relocating object t from class P to class Q is for variable j:

$$\frac{1}{n_j}\left\{\frac{n_{Pj}}{n_{Pj}-1}\cdot(t_j-p_j)^2-\frac{n_{Qj}}{n_{Qj}+1}\cdot(t_j-q_j)^2\right\}$$

if t_j (the value of variable j for object t) is known, and 0 if the value is missing. Here
n_j: total number of known values of variable j,
n_{Pj}: number of known values of variable j in class P,
p_j: mean value of variable j in class P.
I used this generalized algorithm for classification of texture depth-profiles in sandy areas, to exclude genetically different parent materials (e.g. peat and boulder clay) from the analysis.

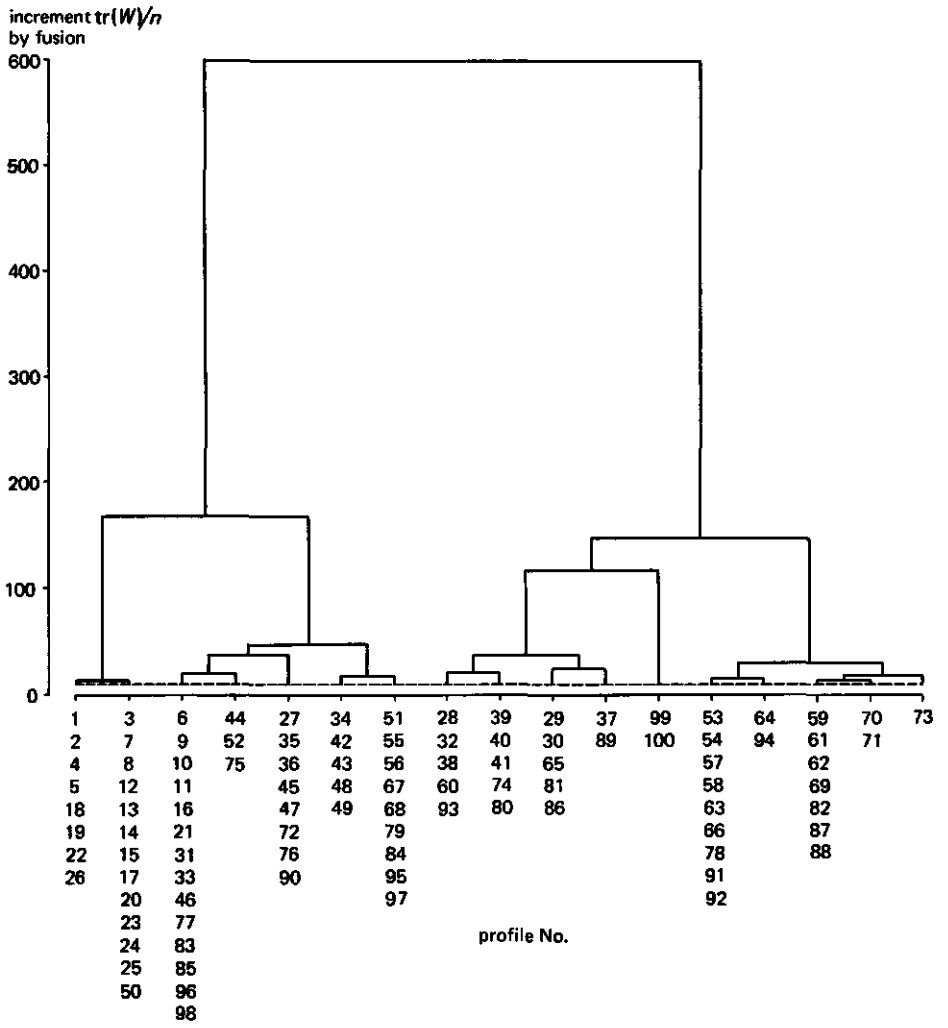


Fig. 9. Dendogram by Ward's method, applied to Set C of 100 *clay* depth-profiles.

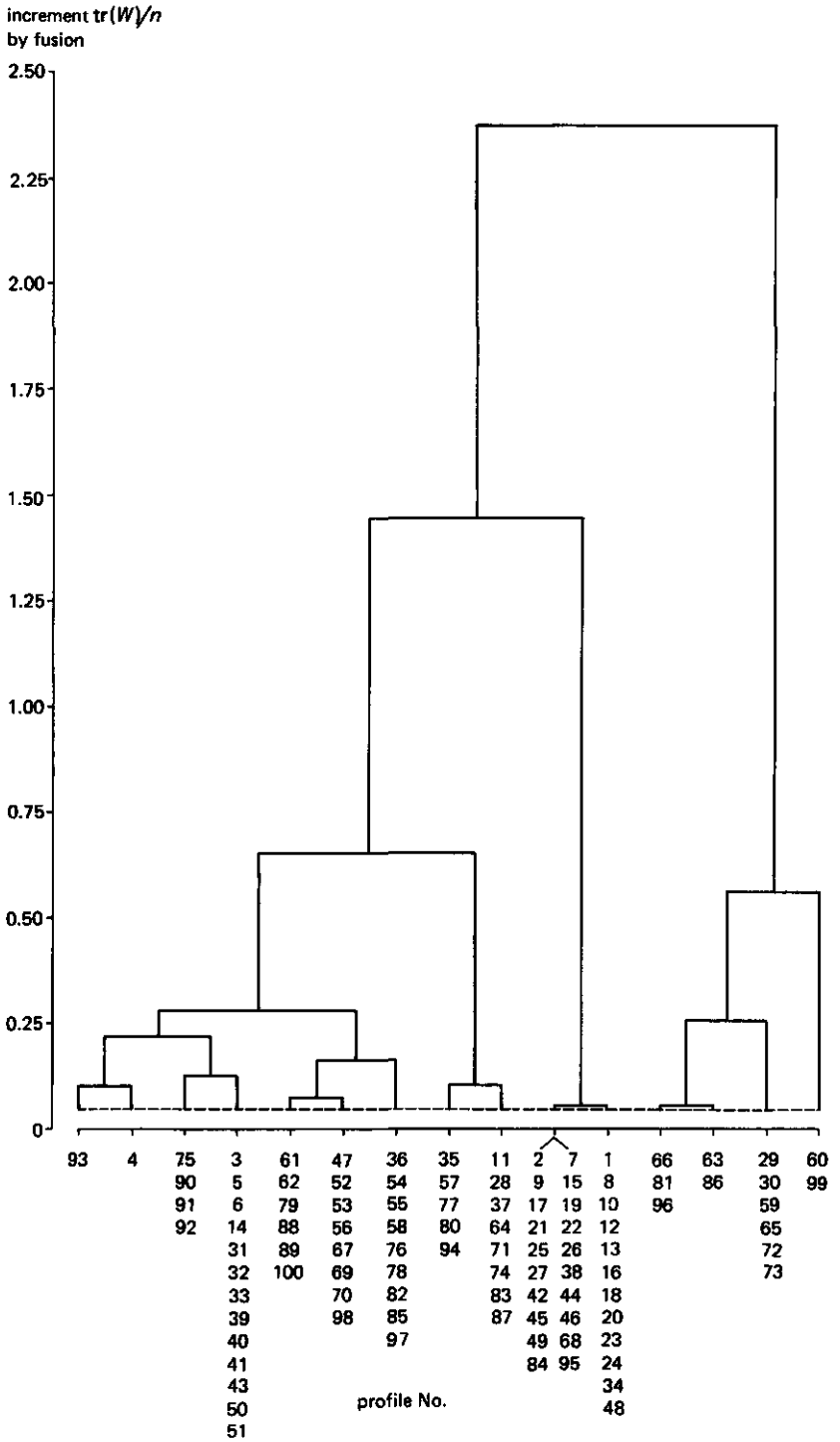


Fig. 10. Dendogram by Ward's method, applied to Set C of 100 *carbonate* depth-profiles.

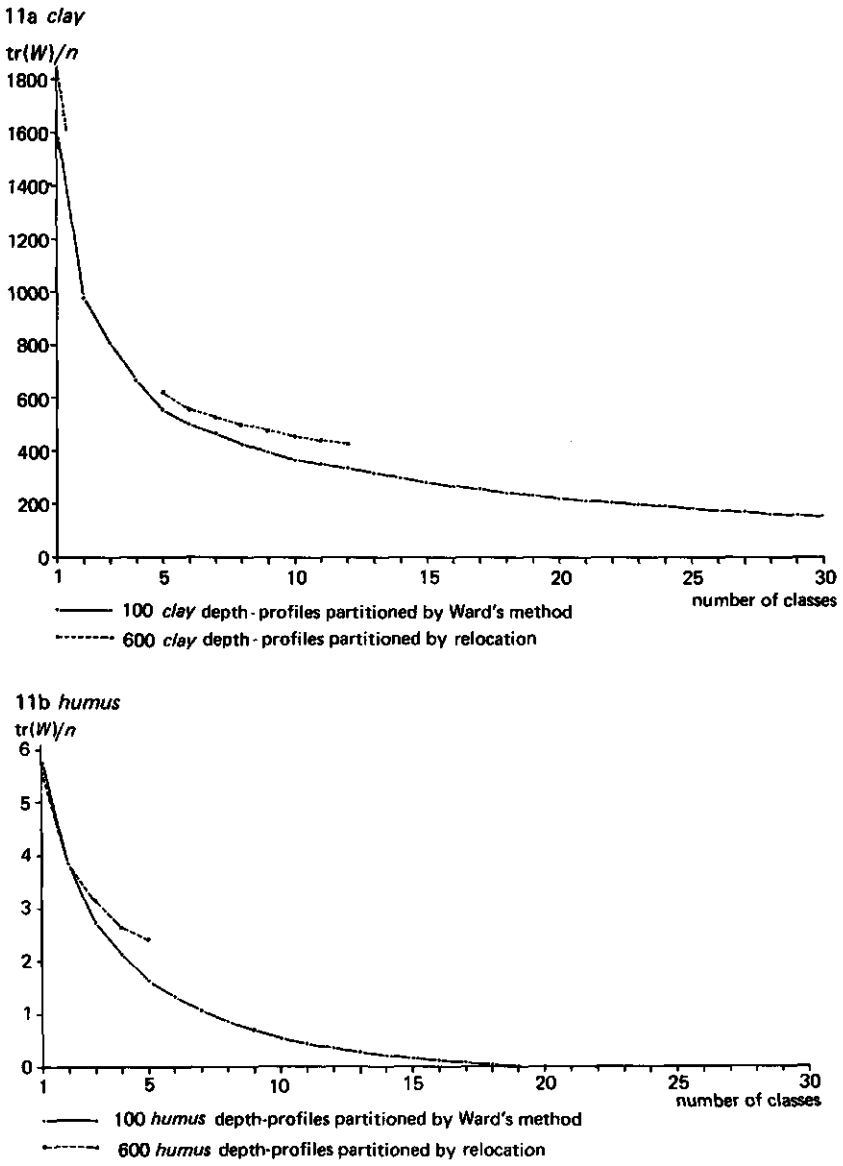
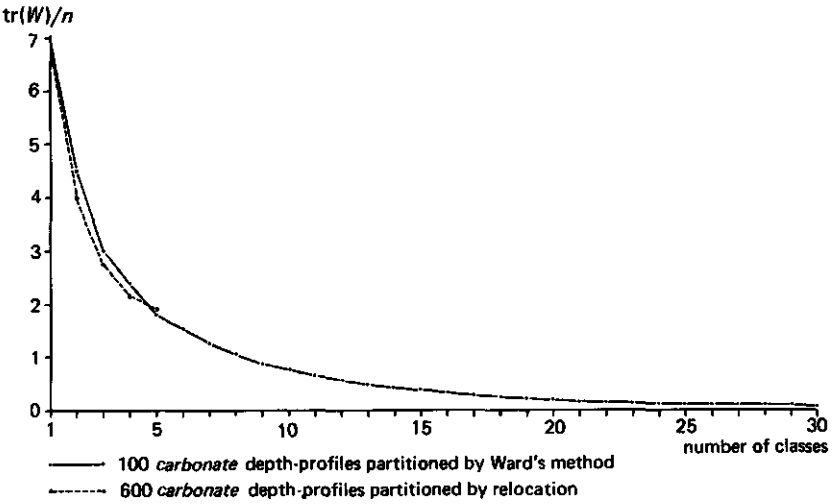


Fig. 11. Relation between the number of special classes and $tr(W)/n$

The choice of the number of classes and the initial partition may be supported, in principle, by application of cluster analysis to a subsample of depth profiles.[8] In the present application I previously applied Ward's method to a subsample of 100 depth profiles, drawn at random from Set B and designated Set C. Ward's method, summarized in 3.2.2.1.1, starts from a matrix of squared Euclidean distances between the objects, and

8. The value of such preliminary analysis should not be over-estimated; see 4.3.2 and 4.3.4.

11c *carbonate*



progressively fuses classes such that tr(W) is minimum. A sample size of 100 seemed a reasonable compromise between representation of the data set and computational effort, which increases with the square of the number of objects.

The results of Ward's method were recorded in the form of dendrograms. Those for *clay* and *carbonate* are given as examples in Fig. 9 and 10. From each of these dendrograms a graph was constructed to show the relation between tr(W)/n and the number of classes. The graphs for *clay*, *humus* and *carbonate* are represented by solid lines in Fig. 11a, b and c, respectively. They, of course, only approximate the relations which would be obtained for the complete Set B without imposing a hierarchical structure (4.3.4).

The effect on the homogeneity of the classes, as shown by the mentioned graphs, was one of the factors considered in choosing the number of classes. The more important a property is for the purpose of the survey, the more homogeneity will be required for that property. However, it would be pointless to construct many classes for a property which, in the field, is only roughly assessed. Furthermore, if the special classifications have too many classes, an excessive number of interim classes may result. The effect of the number of special classes on the final result will be discussed in 4.3.4. In the present application, the following choice was made: 6 classes for *clay*, 3 classes for *humus* and *carbonate*, and 2 classes for *ripening*, *knip* and *peat*.

To create an initial partition for relocation, the partition of Set C with the chosen number of classes was read from the dendrogram and the centroids of the classes were calculated. The remaining depth-profiles (Set B minus C) were then allocated to the class with the nearest centroid. The effect of the initial partition on the result of relocation will be discussed in 4.3.2.

For *clay*, for instance, relocation resulted in a final partition of the 600 *clay* depth-profiles of Set B. The centroid of each of the six classes consists of the class averages for *clay* in each of the 20 depth intervals. These centroids are presented, with a code for reference, in Fig. 12.

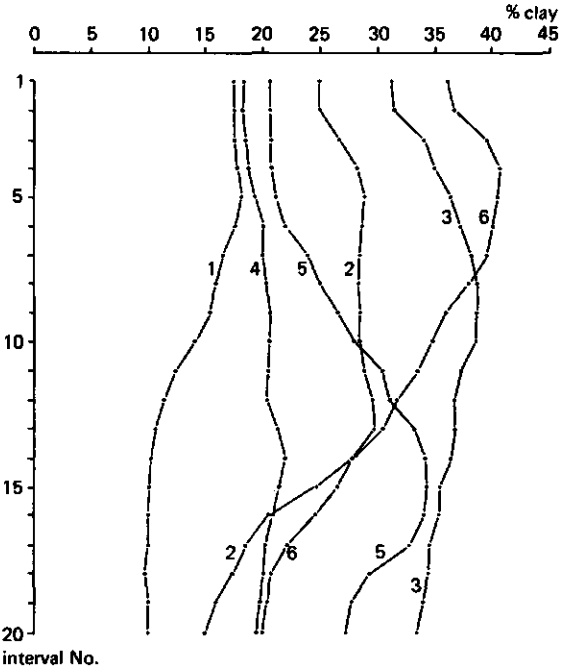


Fig. 12. Six centroids resulting from relocation with Set B of 600 *clay* depth-profiles.

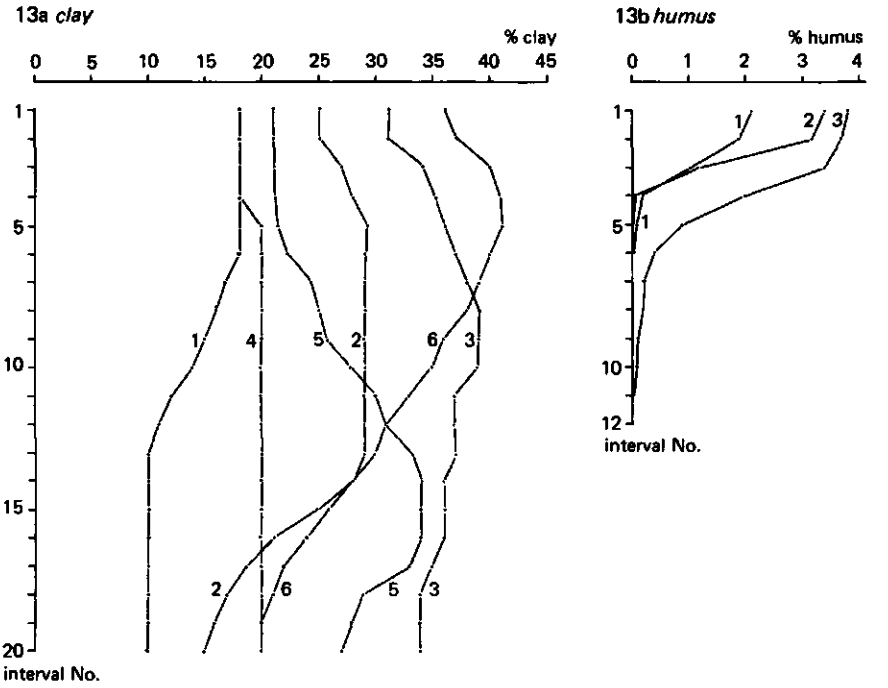


Fig. 13. Central depth-profiles resulting from separate relocations with 600 depth-profiles for each property (Set B).

With diagrams like Fig. 12, a trained surveyor should often be able directly to select the centroid at shortest Euclidean distance from an observed depth-profile. But calculating this has to be fairly easy in case of doubt. For that reason, the values of the centroids were rounded off. Identification may also be facilitated by smoothing, i.e. by removing small, insignificant deviations from an obvious trend in a centroid. By rounding and smoothing, the central depth-profiles no longer coincide with the centroids, hence the objective function E_c is no longer equal to $\mathrm{tr}(W)/n$. A moderate increase in E_c will usually result from these simplifications.

The rounded and smoothed central depth-profiles for *clay* and the other properties are presented in Fig. 13. Except for *clay*, these diagrams are nearly identical to those of the centroids which, therefore, are not presented here. After allocation to the central depth-profiles the observed depth-profiles were distributed over the classes as indicated in Table 5. The pooled within-class variances in each of the intervals are presented in Fig. 17. Values of E_c for the special classifications are given in Table 10, those of $\mathrm{tr}(W)/n$ are underlined in Table 17.

Some information about the speed of the relocation process is given in Table 6, with the number of iterations required (including the last iteration without switches), and the total number of switches during the process. The cluster analyses required 12-50 seconds of computer time, using Wishart's CLUSTAN 1A package, with an IBM 360.

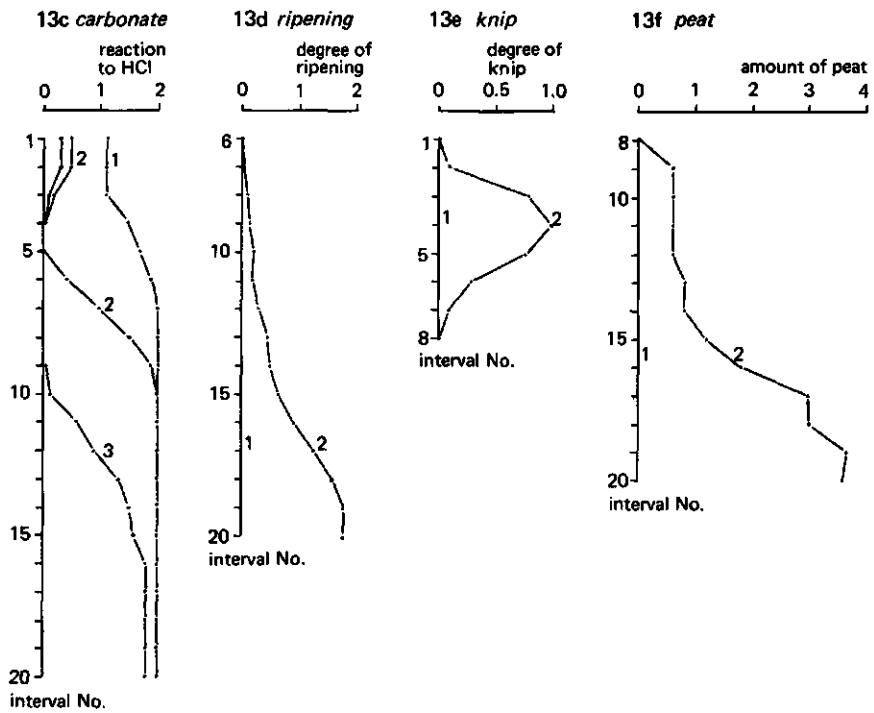


Table 5. Frequencies of the special classes, after allocation of Profile Set B to the central depth-profiles in Fig. 13.

	Class No						Total
	1	2	3	4	5	6	
Clay	140	141	55	109	78	77	600
Humus	294	215	91				600
Carbonate	255	275	70				600
Ripening	508	92					600
Knip	479	121					600
Peat	595	5					600

Table 6. Computational effort in relocation.

	Number of	
	iterations	switches
Clay	6	113
Humus	4	53
Carbonate	2	17
Ripening	2	6
Knip	2	2
Peat	1	0

4.2.4.2 Synthesis; interim classification

The interim classification was obtained by combining the special classifications: all soil profiles belonging to the same special classes for the respective properties together constituted one interim class. The representative of an interim class was formed by the union of the corresponding central depth-profiles. In the present application, there were 6 x 3 x 3 x 2 x 2 x 2 = 432 different combinations of special classes. Actually, only 108 of these combinations arose from Set B. These combinations and their frequencies are listed in Table 7.

4.2.5 Classification at the second stage of the procedure

4.2.5.1 Method

It would be neither feasible nor necessary to map all 108 interim classes individually. Several interim classes were therefore lumped to obtain a final classification. Ward's method of cluster analysis was applied to find the best fusions, i.e. those with least increase of within-class variances. This time, however, it did not need to be confined to a subsample, because the complete set of 108 classes could be reasonably handled. Also, the number of fusions was relatively low, so that deviation from optimum by constraining to a hierarchy was not expected to be serious, and the partition was not further optimized by relocation.

Another difference from the previous stage is that amalgamation started from class representatives, to be weighted with class frequencies, instead of individual depth-profiles.

Table 7. Final classification defined by combinations of the central depth-profiles in Fig. 13.

Class No	a[1]	b[1]	c[1]	d[1]	e[1]	f[1]	Frequency
1	4	1	2	1	1	1	18
2	2	1	2	1	1	1	23
	3	1	2	1	1	1	2
	6	1	2	1	1	1	3
3	5	1	2	1	1	1	15
4	1	2	2	1	1	1	1
	4	2	2	1	1	1	10
5	2	2	2	1	1	1	22
	5	2	2	1	1	1	7
6	3	2	2	1	1	1	5
	6	2	2	1	1	1	13
7	1	3	2	1	1	1	3
	4	3	2	1	1	1	4
8	2	3	2	1	1	1	9
	3	3	2	1	1	1	1
	5	3	2	1	1	1	6
	6	3	2	1	1	1	1
9	1	1	1	1	1	1	55
10	4	1	1	1	1	1	39
11	5	1	1	1	1	1	21
12	2	1	1	1	1	1	21
	3	1	1	1	1	1	2
	6	1	1	1	1	1	1
	3	2	1	1	1	1	1
	6	2	1	1	1	1	1
13	1	2	1	1	1	1	36
14	2	2	1	1	1	1	8
	4	2	1	1	1	1	14
	5	2	1	1	1	1	4
15	1	3	1	1	1	1	35
16	2	3	1	1	1	1	6
	4	3	1	1	1	1	6
	5	3	1	1	1	1	3
17	1	1	3	1	1	1	4
	4	1	3	1	1	1	3
18	2	1	3	1	1	1	3
	5	1	3	1	1	1	3
	6	1	3	1	1	1	1
	2	2	3	1	1	1	2
	3	2	3	1	1	1	3
	5	2	3	1	1	1	2
	6	2	3	1	1	1	2
19	4	3	3	1	1	1	2
	5	3	3	1	1	1	2
20	3	1	2	2	1	1	3
	6	1	2	2	1	1	4
	3	2	2	2	1	1	1
	5	2	2	2	1	1	1
	6	2	2	2	1	1	3
21	2	3	2	2	1	1	1
	3	3	2	2	1	1	3
	4	3	2	2	1	1	2
	5	3	2	2	1	1	1
	5	3	3	2	1	1	1
22	1	1	2	2	1	1	1
	4	1	2	2	1	1	1
	1	1	1	2	1	1	3
	4	1	1	2	1	1	1
	4	2	2	2	1	1	2
23	1	3	1	2	1	1	2
	4	3	1	2	1	1	1
24	2	2	1	2	1	1	3
	3	2	1	2	1	1	2
	4	2	1	2	1	1	1
	5	2	1	2	1	1	2
	3	1	1	2	1	1	1
25	3	1	3	2	1	1	2
	5	1	3	2	1	1	1
	6	1	3	2	1	1	1
	3	2	3	2	1	1	5
	5	2	3	2	1	1	2
26	2	1	2	1	2	1	22
	4	1	2	1	2	1	3
	5	1	2	1	2	1	2
27	3	1	2	1	2	1	2
	6	1	2	1	2	1	8
28	2	2	2	1	2	1	14
	4	2	2	1	2	1	1
	2	3	2	1	2	1	1
29	3	2	2	1	2	1	5
	6	2	2	1	2	1	13
30	2	1	1	1	2	1	1
	6	1	1	1	2	1	1
	2	2	1	1	2	1	1
31	3	1	3	1	2	1	2
	6	1	3	1	2	1	1
	2	2	3	1	2	1	1
	3	2	3	1	2	1	3
	5	2	3	1	2	1	2
	6	2	3	1	2	1	2
32	2	1	2	2	2	1	1
	3	1	2	2	2	1	2
	6	1	2	2	2	1	8
	2	2	2	2	2	1	2
	3	2	2	2	2	1	1
	6	2	2	2	2	1	3
	6	2	1	2	2	1	1
33	3	1	3	2	2	1	1
	4	1	3	2	2	1	1
	6	1	3	2	2	1	5
	3	2	3	2	2	1	3
	5	2	3	2	2	1	2
	6	2	3	2	2	1	5
	3	3	3	2	2	1	1
34	3	2	1	2	1	2	1
	5	2	1	2	1	2	1
	3	1	2	2	1	2	1
35	3	2	3	2	1	2	2

1. a = *Clay;* b = *Humus;* c = *Carbonate;* d = *Ripening;* e = *Knip;* f = *Peat.*

The usual recurrence formula (3.2.2.1.1) for calculating new increments of the sum of squares within classes could still be used, provided that a matrix of correct initial increments was the starting point. Squared Euclidean distances can normally be used for this, but because in this application the objects to be fused represented different numbers of profiles, the initial increments were calculated according to:

$$\frac{n_i n_j}{n_i+n_j} \cdot d_2^2(i,j),$$

where n_i and n_j are the number of profiles in classes i and j, respectively, and $d_2^2(i,j)$ is the squared Euclidean distance between the class representatives of i and j. Section 4.2.3.3 states how the properties were weighted before cluster analysis.

4.2.5.2 Final classification

The dendrogram resulting from Ward's method is presented in Fig. 14. A graph indicating the relation between the number of classes and tr(W)/n was derived from this

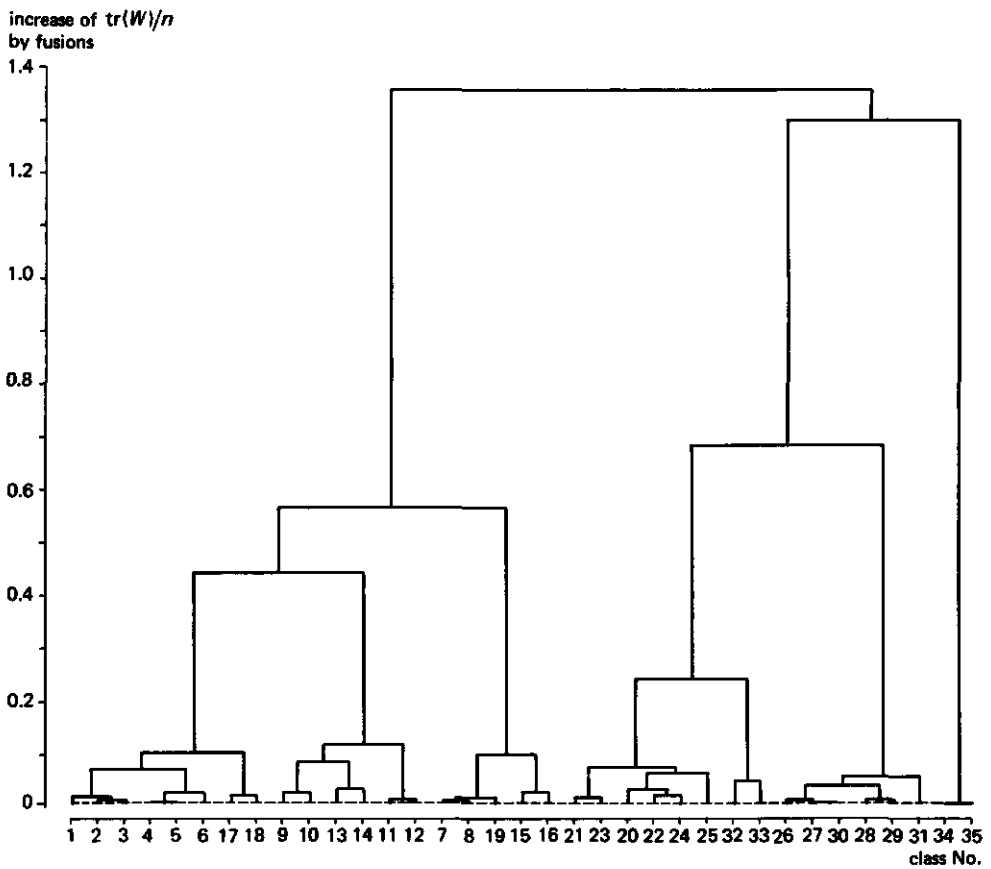


Fig. 14. Dendogram by Ward's method, applied to 108 combinations of central depth-profiles with previous standardization of the properties (beginning with the 35 classes specified in Table 7).

dendrogram (Fig. 15). A total of 35 classes seemed a reasonable compromise between the demands for homogeneity and manageability; it is a usual number for the present type of surveys. The choice of the number of classes is dealt with in 4.3.4.

The classification with 35 classes was read from the dendrogram; it is specified in Table 7. Values of $\mathrm{tr}(W)/n$ were, separately for each property, calculated from the untransformed data of Set B. These values are presented in Table 17.

4.2.6 Geographic distribution of the classes

A definitive test of the classification in the field was beyond the scope of this study. Using the same data as before, however, the geographical distribution of the classes in the test-area could be examined. Two series of maps were produced by computer.

Profile location maps with class-labels (point maps) were separately made for the six special classifications, so that one map was obtained for each property. To make them, all profiles in the test area (Set A) were allocated to the central depth-profiles depicted in Fig. 13 on the basis of shortest Euclidean distance. The corresponding labels were plotted on the profile locations (Map 1-6).

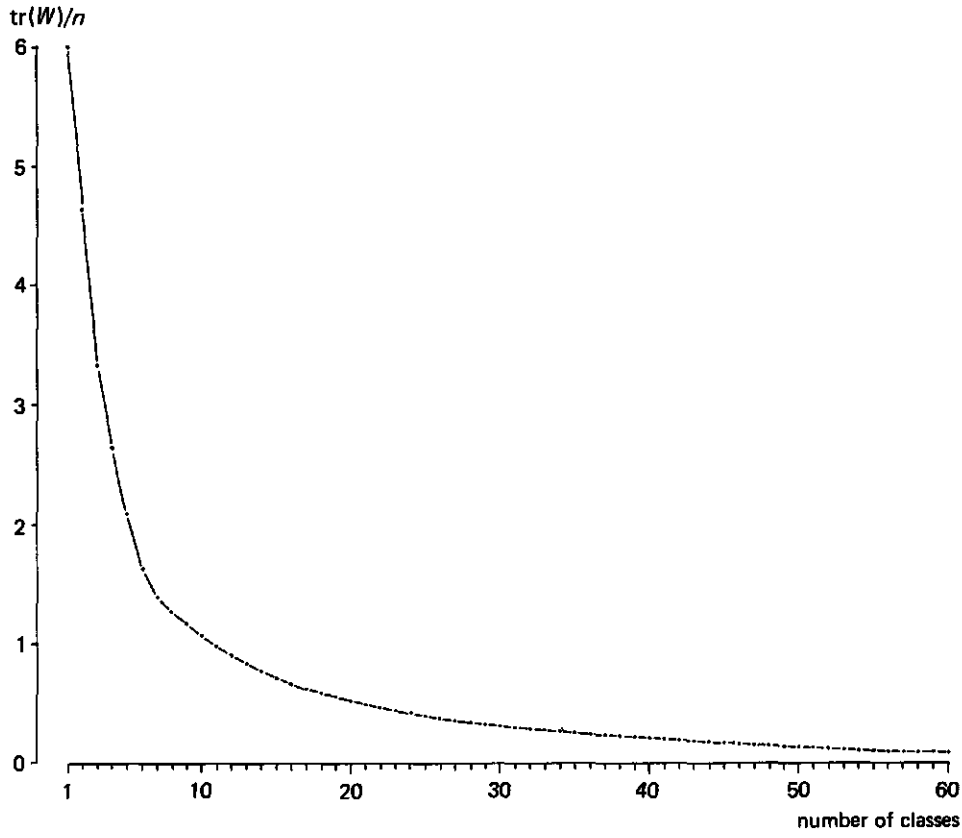


Fig. 15. Relation between the number of final classes and $\mathrm{tr}(W)/n$, derived from the dendogram in Fig. 14.

The point maps reflect the class occurrences only partially. For a more complete image, occurrences have been delineated. Manual delineation would inevitably have introduced a degree of subjectivity undesirable in this test. Existing programs for contouring were not applicable because they operate only with a single quantitative variable, and the present classes, being polythetically defined, could not be delineated by superposition of contour maps. For that reason, the following automated mapping procedure was devised and applied to the data.

First a rectangular sub-area of 1.875 km x 3.750 km was selected in the test area, where a diverse pattern was expected. The location of this sub-area is indicated on the point maps. Using the original data from all 195 borings in this sub-area and a computer program SYMAP (Shepard, 1969), interpolations were carried out for each variable separately. This resulted in an estimated value for every variable in each of 118 x 141 rectangles into which the sub-area had been divided. According to these values, the rectangles were allocated to the central depth-profiles and to the final classes specified in Table 7. The maps, No 7-12, were finally produced by a line-printer, the class labels being printed at positions corresponding to the rectangles. In the present example, the maps are printed in colour. The colour separates were obtained by photographic reduction (4x) of a separate line-printer map of each class or combination of classes, depending on the chosen colour mixtures. An account of this procedure is presented elsewhere (de Gruijter & Bie, 1975).

4.3 Some technical aspects of the procedure

The experiments reported in this section were carried out to obtain some global information about the effect of certain decisions on the results of the classification procedure. Those decisions concern the number of objects (sample size), the initial partition from which relocation is to start, the order in which the objects are relocated, and the number of classes and weights.

4.3.1 Sample size

Problem. The requirement that the classification must be available at an early stage of fieldwork implies that its calculation should be based on a profile sample of limited size. But when the sample is too small, it may misrepresent the population of profiles in the area concerned. The resulting classification may then deviate too much from the (unknown) optimum classification of the population. Statistical theory does not provide a direct answer about the sample size required for cluster analysis in various situations.

Method The relocation algorithm as described in 4.2.4.1 was applied to three samples, consisting of 100, 600 and 999 profiles. Centroids and values of objective function were compared.

The first two samples were, respectively, identical with Set C and Set B (4.2), while the third was obtained by adding 399 profiles to Set B, selected at random from Set A minus Set B. Fifteen classes were constructed in each of the three analyses, using the clay percentages recorded for the twenty intervals.

An initial partition for relocation with the 100 profiles was read from the dendrogram in Fig. 9. After relocation, fifteen centroids resulted. The remaining 500 profiles in Set B

were successively allocated to these centroids, the latter being immediately recalculated after each allocation. Relocation was then continued with the 600 profiles. The centroids resulting from this were similarly used as a starting point for relocation with the 999 profiles.

Results. Three series of fifteen centroids resulted from relocation with 100, 600 and 999 profiles. The differences between each pair of series were evaluated as follows. Each centroid of the one series was compared with the most similar centroid of the other series. More precisely; the pairs of centroids to be compared were chosen such that the sum of the squared Euclidean distances between them was minimum. For each pair of centroids, the absolute value of the differences in percentage of clay were calculated for the twenty intervals. Summary statistics of the $15 \times 20 = 300$ differences thus assessed, are given in Table 8.

The values of the objective function after additions of new profiles and relocations are presented in Table 9. A similar comparison of function values was made in the case of the special classifications described in 4.2. In Table 10, two values of E_c are given for each property: the mean squared Euclidean distance of 600 profiles (Set B) to their respective class representatives, depicted in Fig. 13, and the same for all 2212 available profiles (Set A).

Conclusions. Table 8 shows that increasing the sample size from 100 to 600 profiles considerably affected the final centroids. A further increase with about 400 profiles, however, hardly changed the centroids. This suggests that a least squares partition of the test area into fifteen classes defined on *clay*, could be well approximated using the centroids based on 600 profiles. This is supported by Table 9, indicating that $\mathrm{tr}(W)/n$ hardly changed when new profiles were added to the sample of 600.

Table 8. Differences between three series of fifteen *clay* centroids, resulting from relocation with 100, 600 and 999 *clay* depth-profiles.

Sample size	Mean difference	Mean squared difference	Minimum difference	Maximum difference
100 or 600	3.37	17.48	0.01	12.10
100 or 999	3.46	20.19	0.00	15.55
600 or 999	0.81	1.21	0.00	3.57

Table 9. Change of $\mathrm{tr}(W)/n$ for fifteen classes of *clay* depth-profiles, by allocation of more depth-profiles and relocation.

	Step 1 Partition by Ward's method	Step 2 Partition optimized by relocation	Step 3 500 new profiles allocated	Step 4 Partition optimized by relocation	Step 5 399 new profiles allocated	Step 6 Partition optimized by relocation
Sample size	100	100	600	600	999	999
$\mathrm{tr}(T)/n$	1585.2	1585.2	1831.2	1831.2	1745.6	1745.6
$\mathrm{tr}(W)/n$	284.6	273.3	412.3	390.0	388.0	385.6

The same sample of 600 profiles seems also to be sufficient for constructing six classes on *clay*, and for the classifications on *carbonate, ripening* and *knip* (Table 10). However, the sample was probably less adequate for the classifications on *humus* and *peat*, since E_c markedly increased when new profiles were allocated. The dwelling mounds in the test area constitute a small minority of profiles, which strongly differ from the majority in humus contents. Such minorities could be easily misrepresented in a sample like the present. A similar explanation applies to peat; the subsoil of only a small part of the test area consists of peat or peaty clay.

These conclusions emphasize that numerical classifications should not be applied blindly in soil survey. A control procedure is necessary, so that the classification is adapted when class representatives deviate too much from profiles observed during the survey.

4.3.2 Initial partition

Problem. The relocation process as specified in 4.2.4.1 may stop at some local minimum of $\mathrm{tr}(W)/n$. Different solutions may thus result from the same data, partly depending on the initial partition. The initial partitions for the relocations discussed in 4.2.4.1 were obtained using Ward's method, hence were already optimized to some extent for $\mathrm{tr}(W)/n$. It is not known how far such a starting point has a more favourable effect on relocation than does random choice.

Method. Relocation was done twice with the same data, starting from two different initial partitions.

In both procedures, 600 *clay* depth-profiles (Set B; see 4.2) were partitioned into 12 classes, relocating the profiles in the same order. One initial partition was created by randomly allocating 50 profiles to each of 12 classes. For the second initial partition, Ward's method was applied to a subset of 100 profiles (Set C). The partition into 12 classes was read from the dendrogram (Fig. 9) and the centroids were calculated. The complete set of 600 profiles was then allocated to these centroids, such that $\mathrm{tr}(W)/n$ was minimized by each individual allocation.

Table 10. Values of the objective function, E_c, after allocation of different numbers of profiles to the central depth-profiles in Fig. 13.

Property	Step 1 (600 samples)[1]		Step 2 (2212 samples)[2]	
	$\mathrm{tr}(T)/n$	E_c	$\mathrm{tr}(T)/n$	E_c
Clay	1831.2	568.4	1793.0	568.1
Humus	5.577	3.169	6.102	3.758
Carbonate	6.926	2.784	6.861	2.850
Ripening	2.4540	0.8338	2.4237	0.8343
Knip	0.5194	0.1591	0.5242	0.1591
Peat	0.5452	0.1182	0.5106	0.1494

1. Partition optimized by relocation.
2. 1612 new profiles allocated.

Results. The value of $\mathrm{tr}(W)/n$ after relocation were as follows:
initial partition selected at random : $\mathrm{tr}(W)/n = 21.84$
initial partition selected by Ward's method : $\mathrm{tr}(W)/n = 21.41$
The centroids that finally resulted from the initial partition selected by Ward's method are depicted in Fig. 16. These centroids were pair-wise compared with those resulting from the randomly selected initial partition. For each pair of centroids, chosen such that the sum of the squared Euclidean distances between them was minimum, the differences in clay content were calculated. Squares and absolute values of these differences were averaged over the 20 intervals; the results are presented in Table 11. The differences ranged between 0 and 17% clay.

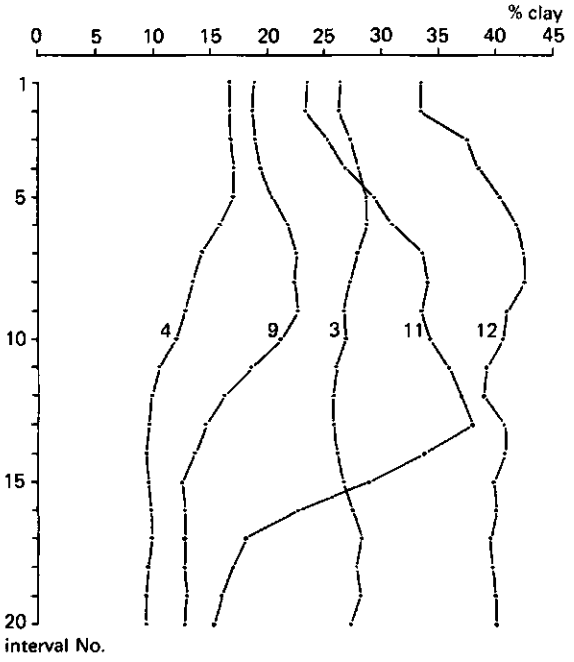
The relocation starting from the random initial partition required 13 iterations with a total of 1006 switches, and 8.67 min of computer time (IBM 360). Preparation of the initial partition by Ward's method took 5.86 min. Relocation then required 9 iterations with a total of 200 switches, and 2.15 min.

Conclusions. In this test, the initial partition with lower $\mathrm{tr}(W)/n$ has also given a final partition with lower $\mathrm{tr}(W)/n$. The difference in $\mathrm{tr}(W)/n$, however, is too small to be of practical interest. An initial partition obtained by application of Ward's method to part of the data apparently need not be clearly advantageous in terms of minimization. This supports a similar conclusion of Wishart (1970), who even obtained better partitions from 'worse' starting points.

Although computational effort in relocation was much reduced by starting from an initial partition derived from a subsample, only a slight net reduction in computer time remains when the time used for preparing the initial partition is taken into account. Such an initial partition could be more profitable when relocation is to be carried out with a larger data set than the present one. Apart from the initial partition, preliminary cluster analysis of a subsample could help in deciding a range for the number of classes and for the weights of the variables.

Table 11. Differences between *clay* centroids, due to difference in initial partition.

Centroid pair No	Mean absolute difference	Mean squared difference
1	0.82	0.90
2	2.27	7.60
3	3.09	19.86
4	0.61	0.52
5	2.74	10.88
6	9.29	104.87
7	2.55	9.12
8	2.42	8.44
9	1.57	2.91
10	0.87	1.37
11	0.72	0.84
12	3.08	18.19
1-12	2.50	15.46



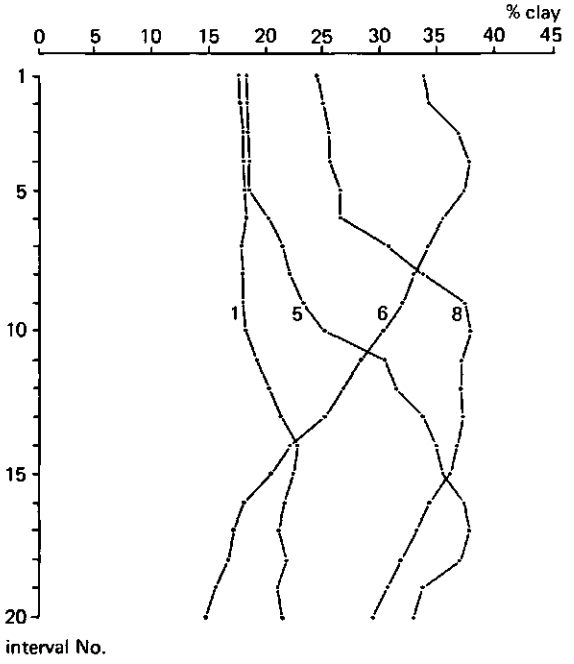
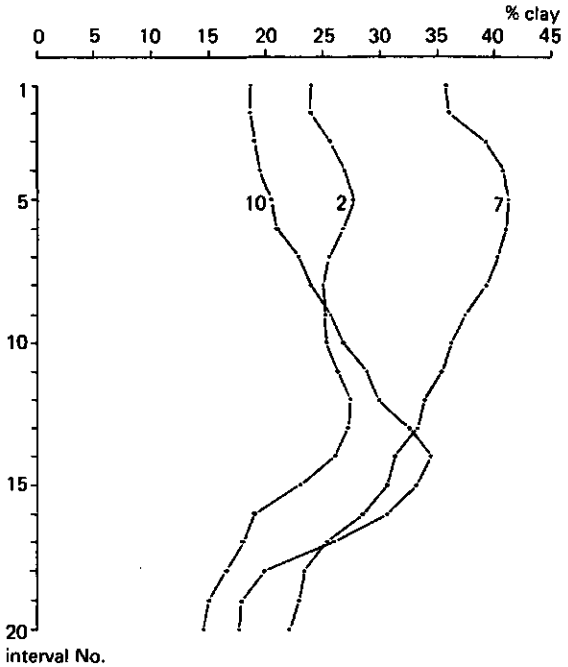


Fig. 16. Twelve centroids resulting from relocation with Set B of 600 *clay* depth-profiles, starting from an initial partition obtained by Ward's method.



Wishart (1970) noted that his sample from a population with a standardized bivariate normal distribution could be dichotomized in various ways, each however with nearly the same values of tr(W)/n. The two solutions found in the present test also have almost the same tr(W)/n value; the centroids of the one solution are all fairly similar to those of the other solution, except centroids No 6 (Table 11).

4.3.3 Order dependence of the solution

Problem. The relocation algorithm used in my experiments prescribes recalculation of the centroids as soon as any object has been transferred from one class to another. The order in which the objects are relocated will therefore affect the process and generally also the final result, when local minima exist. Lack of insight into the consistency of results at present impedes reliable interpretation.

Method. Relocation was with the same data and the same starting point, the objects being placed however in different orders.

For this test 600 *clay* depth-profiles (Set B) were used. A subsample of 100 profiles (Set C) was first partitioned by Ward's method into fifteen classes (Fig. 9). The centroids were calculated and the remaining 500 profiles allocated in random order designated I. The entire set of 600 profiles was then relocated until a stable partition was reached. This was repeated with another random order of profiles (II) and with an order (III) that was the reverse of Order II.

Results. The values of tr(W)/n after relocation were as follows:

Profile Order I : tr(W)/n = 19.61
Profile Order II : tr(W)/n = 19.83
Profile Order III: tr(W)/n = 19.50

The three series of fifteen centroids were mutually compared in the same way as in 4.3.1 and 4.3.2. For each pair of series, the pairs of centroids to be compared were chosen such that the sum of the squared Euclidean distances between them was minimum. As first evidence of the close similarity between the three series, it turned out that the centroids of the second and third series closest to any given centroid of the first series, were also closest to each other.

Mean difference, mean squared difference and maximum difference in clay content between each two series of centroids are presented in Table 12. The minimum difference between each pair of series was zero.

Conclusions. In this instance, different solutions were obtained when the objects were relocated in different order. However, the differences in $\mathrm{tr}(W)/n$ were negligible. The same holds for the series of centroids as a whole. When averaged over centroids and intervals, the difference in clay content does not exceed the usual error of field estimates. Some particular centroids (No 3, 4, 6, 8 and 12) were, however, slightly unstable.

4.3.4 Choice of the number of classes and weights

4.3.4.1 Problem

As argued in 2.3.1.2 and 3.4.2.1, formal rationalization of the choice of the number of classes and weighting of variables is still beyond reach. In the present classification procedure, the choice of the number of classes and of weights has therefore been left to the user, instead of being built in a more comprehensive model. The choices could be made in a dialogue between pedologist and statistician, possibly supported by a trial-and-error

Table 12. Differences between *clay* centroids, due to difference in order of relocation.

Centroid pair No	Order I or II			Order I or III			Order II or III		
	Mean abs. difference	Mean sq. difference	Max. diff.	Mean abs. difference	Mean sq. difference	Max. diff.	Mean abs. difference	Mean sq. difference	Max. diff.
1	0.93	1.45		0.29	0.19		1.19	1.82	
2	1.39	2.69		2.77	9.09		1.72	3.26	
3	3.04	16.26		0.51	0.31		3.16	16.90	
4	2.80	8.60		2.14	5.30		4.59	24.96	
5	0.48	0.45		2.59	11.43		2.64	11.65	
6	3.72	19.39		0.18	0.05		3.84	20.37	
7	0.49	0.34		0.66	0.79		0.27	0.14	
8	1.36	2.61		3.08	12.26		2.18	7.79	
9	1.41	2.41		1.55	2.75		0.36	0.28	
10	0.40	0.32		0.22	0.07		0.47	0.36	
11	0.00	0.00		0.44	0.25		0.44	0.25	
12	3.35	26.16		0.56	0.46		3.66	31.62	
13	1.57	4.57		0.57	0.50		2.06	7.52	
14	0.27	0.15		0.28	0.10		0.51	0.39	
15	0.12	0.02		0.35	0.16		0.36	0.20	
1-15	1.42	5.69	8.87	1.08	2.91	6.56	1.83	8.50	8.98

procedure. Insight is therefore needed in how the classes are affected by choice of weights and the number of classes.

The effect on $\mathrm{tr}(W)/n$ of increasing the number of classes can be qualitatively predicted: the optimum partition of a given set into $(k + 1)$ classes will have a lower $\mathrm{tr}(W)/n$, than the optimum partition into k classes, as long as $(k + 1)$ does not exceed the total number of different objects. At the same time, however, the manageability of the classification is likely to decrease, both for surveyors and for users of the soil map. A compromise should be based, among other things, on information about the relation between $\mathrm{tr}(W)/n$ and k. In the application of 4.2, this information was obtained by applying Ward's method to a subsample (Fig. 9 and 10). A practical question is whether such information is sufficiently accurate or too distorted by application of an agglomerative method to a subsample.

Minimization of $\mathrm{tr}(W)/n$ tends towards compact, spherical subsets in multivariate space, hence a uniform distribution of pooled within-class variances over the variables. If smaller within-class variances are required for more important variables than for others, then that can in principle be achieved by enlarging the scales of the more important variables before partitioning the set. When roughly spherical subsets result in the multivariate space thus transformed, these subsets will be flattened after retransformation to the original space. Thus the within-class variance of an original variable will tend to be smaller accordingly as its scale is previously enlarged.

The actual distribution of the objects in multivariate space determines how closely the subsets approximate a spherical shape. The effect of altering scale on within-class variances will therefore depend on this distribution. The effect also depends on the number of classes. So the number of classes and weighting should be considered in association with one other.

Quantitative prediction of the effect of scale alteration is only feasible for populations with a simple parametric distribution. Little is known about the effect on classifications based on samples from real populations of soil profiles. For these reasons a general exploration seemed profitable.

4.3.4.2 Method

(a) Special classification with different numbers of classes were calculated, using the same weights for the intervals. (b) Conversely, partitions with the same number of classes were calculated using different weights. In both cases, the effect on the pooled within-class variances was assessed. (c) Variances within the mapping units of the existing soil map (Kamping & Rutten, 1969) were calculated for comparison. (d) The interim classes were fused to different numbers of final classes, keeping the weights of the properties constant. (e) Finally, different weightings were applied to the properties.

4.3.4.3 Results and conclusions

(a) Number of special classes As described in 4.3.2, 600 *clay* depth-profiles (Set B) were partitioned into 12 classes by relocation with a random starting point. This partition was used for the present test as follows.

The number of classes was first reduced to 11 by fusing two classes, chosen such that

tr(W)/n increased as little as possible. Objects were then reallocated again. Two of the final classes were fused as before, to obtain an initial partition for relocation with 10 classes. This was repeated until the number of classes was reduced to 5 (Fig. 18). The tr(W)/n of the final partitions can be compared with those resulting from Ward's method and 100 profiles. Both series of values are presented graphically in Fig. 11a. The same procedure as above was applied to the depth profiles of the other properties, except that other num-

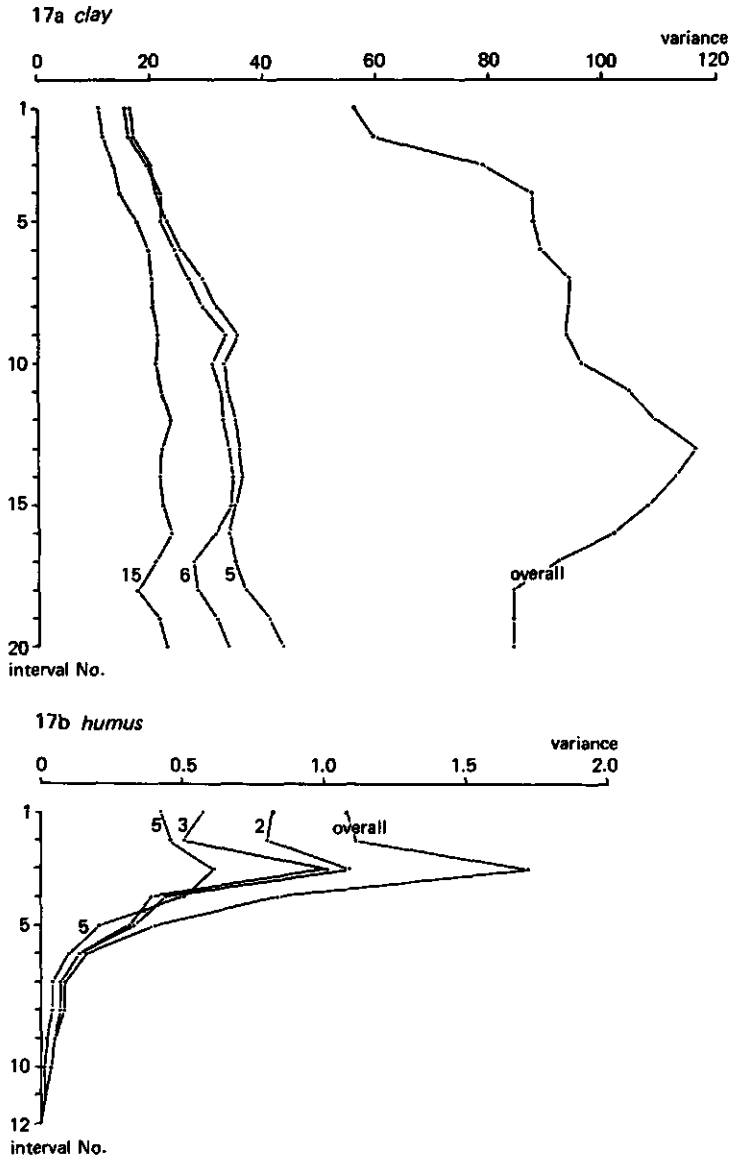
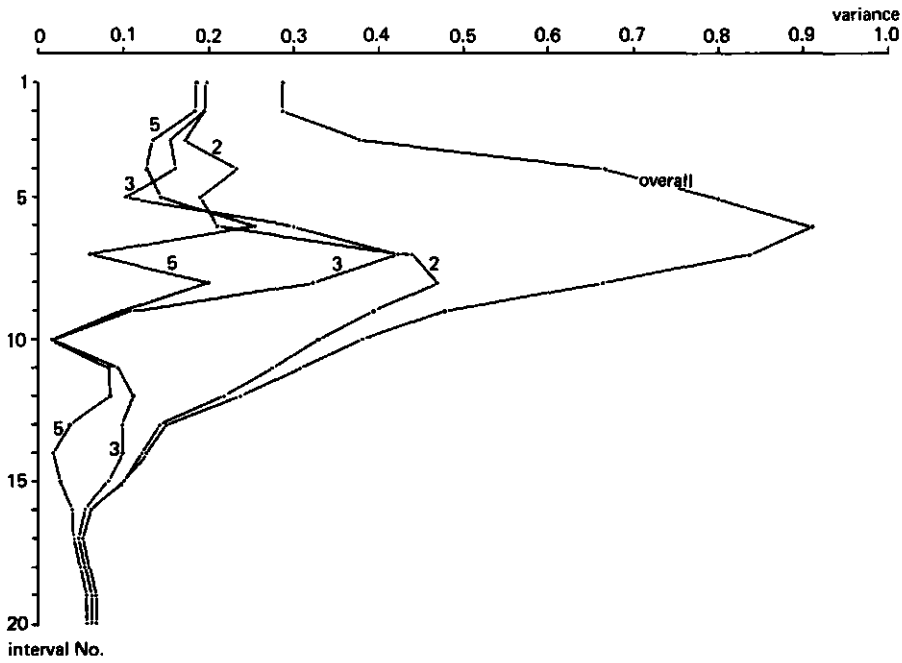


Fig. 17. Overall variances and variances within different numbers of special classes, constructed by relocation with 600 depth profiles for each property (Set B).

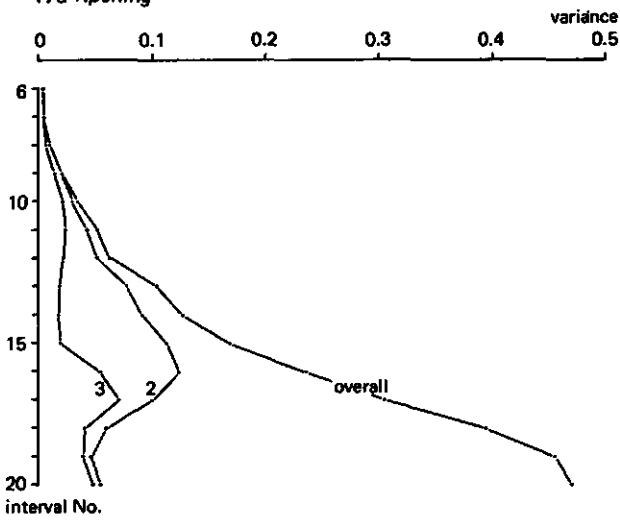
bers of classes were chosen. The tr(W)/n values for *humus* and *carbonate* are presented in Fig. 11b and c. The results obtained for the other properties are similar. The pooled within-class variances for most of the partitions are presented in Fig. 17.

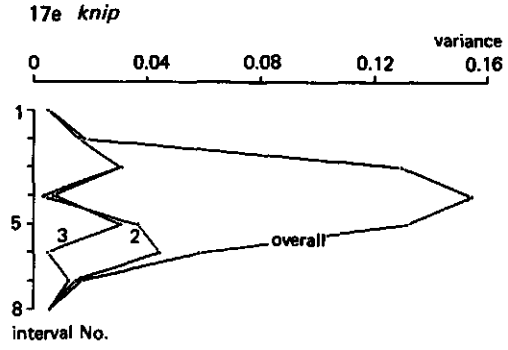
Decreasing the number of central depth-profiles increases the within-class variances, but will generally decrease the number of different combinations of such profiles. Some information on this is given in Table 13.

17c *carbonate*



17d *ripening*





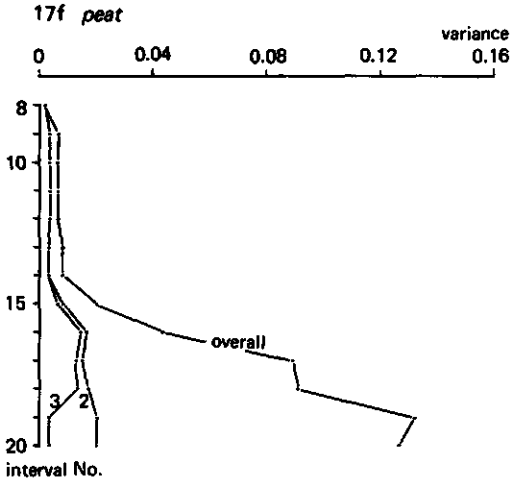


Table 13. Effect of the numbers of central depth-profiles on the number of combinations occurring in two sets of soil profiles.

Number of central depth-profiles for						Number of combinations of central depth-profiles occurring in	
clay	*humus*	*carb.*	*ripe.*	*knip*	*peat*	Set B (600 soil prof.)	Set A (2212 soil prof.)
12	5	5	3	3	4	249	497
10	5	4	3	3	3	223	447
10	5	4	2	2	2	198	·
8	5	4	2	2	2	176	·
8	3	3	2	2	2	130	·
6	3	3	2	2	2	108	157
5	3	3	2	2	2	92	·

The effect on $\mathrm{tr}(W)/n$ is one of the criteria in choosing the number of central depth-profiles for a given property. Fig. 11 shows that Ward's method applied to 100 profiles gives $\mathrm{tr}(W)/n$ values which, for the same number of central depth-profiles, may markedly differ from those obtained by relocation with 600 profiles. The effect on $\mathrm{tr}(W)/n$ is particularly marked within the range of numbers of practical interest for soil survey. In applications, it will therefore be better to find out this effect directly, using all available profiles for relocation with different numbers of classes.

Fig. 11 does not show clear discontinuities in the graphs of $\mathrm{tr}(W)/n$ against the number of classes. This indicates that in the present case there is no 'natural' clustering or a number of classes corresponding with that.

For *clay*, this is supported by the scatter diagram (Fig. 8). Rather than a natural number of classes, graphs as in Fig. 11 could indicate the number of classes below which the heterogeneity would be too high for the intended purpose.

The main reason to keep the number of central depth-profiles low is that otherwise the identification of profiles and delineation of classes in the field would cost too much time. Not only the allocation to the central depth-profiles would be more tedious. Also the number of different combinations would increase considerably (Table 13), and for display on one map these combinations would have to be fused again into a manageable number of classes. Much of the detail achieved in the first stage would then be lost. Further, most classes of the final classification would contain many combinations and would be difficult to handle. More combinations would not be represented in the pilot sample (Table 13), hence would have to be allocated to the classes *ad hoc* during the field work. For these reasons, the numbers of central depth-profiles chosen for the application in 4.2 seem a reasonable compromise. As mentioned before, the number of classes (35) of the final classification in this application was a number usual for similar surveys in the Netherlands.

Figures 12, 16 and 18 (see separate overlays) illustrate that some classes may remain almost unaltered when the number of classes is changed. For instance, after reduction of the number of central depth-profiles for *clay* from 6 to 5, and of the new classes (No 3) was nearly identical to the union of two of the old classes (No 3 and 6). Therefore, when for surveys at different scales, different numbers of central depth-profiles are required for the same property, constraining to a hierarchical system apparently would not always lead to an important loss of homogeneity as compared with unconstrained optimization.

Figure 17 bears out that when the number of central depth-profiles for a given property is increased, the within-class variances in different intervals become more equal. Geometrically, the more classes the smaller and the more spherical they are.

(b) Weighting of depth intervals First standardization to equal overall variance was tried to see which differences in within-class variances are due to statistical dependence between the intervals. Further, extreme scale alterations were tried by assigning zero weights to a number of intervals.

Using the final partitions of the *clay* depth-profiles with 15, 10 and 5 classes as starting points, relocation was repeated after standardizing the data. Standardization was carried out by multiplying the percentages clay by factors such that the overall variance in each interval equalled 100. This value is an arbitrary one, but it does not affect the final partitions. After retransformation to the original scales, the central depth-profiles

and within-class variances were calculated and compared with those of the partitions based on original data. Five central depth-profiles, calculated from original data, are depicted in Fig. 18. They hardly differ from those based on standardized data: averaged over intervals and classes, the absolute value of the difference in clay percentage was 0.5. The same is true for both sets of 10 and 15 central profiles: here the mean difference in clay percentage is 0.6 and 0.5 respectively.

Comparing the smallest pooled within-class variance in the twenty intervals with the highest one, gives an indication how far the classes approach a sphere. Minimum and maximum within-class variances are presented in Table 14. For both types of partitions, the pooled within-class variance in each interval is represented in Fig. 19.

The sample of *clay* depth-profiles (Set B) was also used for tests with more drastic changes of the interval weights than implied by standardization. The scales were altered to wide extremes by transforming the percentages clay of selected intervals to zero variance, which was effectuated by deleting the data of those intervals during relocation. Two experiments of this type were done.

Firstly, relocation was applied to the original clay percentages in only the upper 12 intervals. This corresponds with the usual depth (1.2 m) of augerings in Dutch soil surveys. An initial partition was obtained by Ward's method on the subsample of 100 profiles (Set C), using only the upper 12 intervals. Twelve classes were constructed. The corresponding central depth-profiles are presented in Fig. 20. They can be compared with those based on all 20 intervals (Fig. 16). The pooled within-class variances are presented in Fig. 21, together with those from the partition based on all 20 intervals (4.3.2).

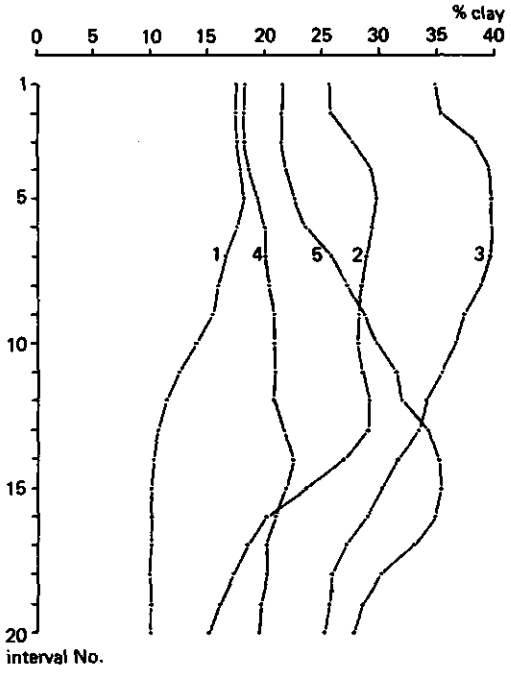
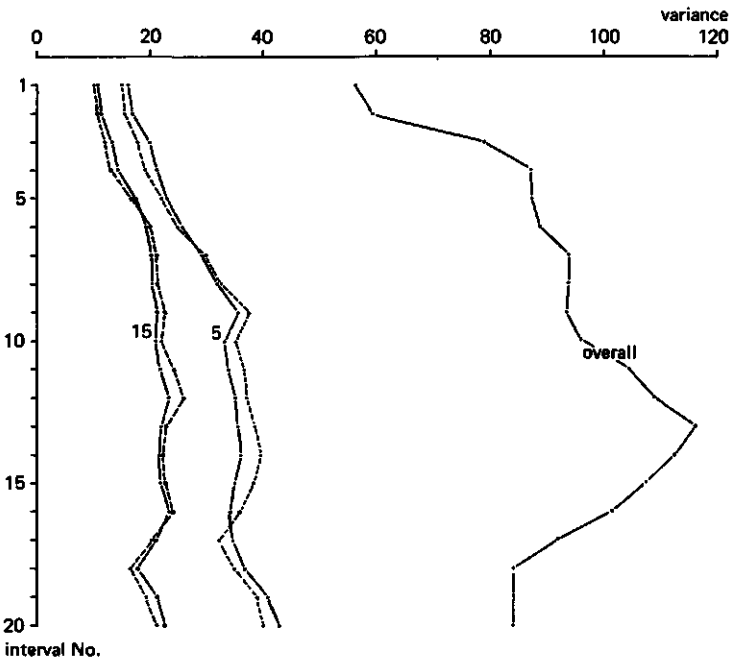


Fig. 18. Five centroids resulting from relocation with Set B of 600 *clay* depth-profiles, using untransformed data.

Table 14. Effect of standardization and number of classes on minimum and maximum value of the pooled within-class variances for partitions of 600 *clay* depth-profiles (Set B).

	Partitions and variances calculated from		
	original data	standardized data	standardized and original data, resp.
Minimum variance within 5 classes	16.33	21.88	15.07
Maximum variance within 5 classes	43.11	47.84	40.24
Minimum variance within 10 classes	12.66	16.00	10.94
Maximum variance within 10 classes	31.76	35.87	30.17
Minimum variance within 15 classes	10.85	15.09	10.23
Maximum variance within 15 classes	23.48	25.04	25.71



Relocation with original data (cf. Fig. 17a)

Relocation preceded by standardization to equal overall variance in each interval. Within-class variances calculated from the original data

Fig. 19. Effect of previous standardization on variances within 15 and 5 classes, constructed by relocation with Set B of 600 *clay* depth-profiles.

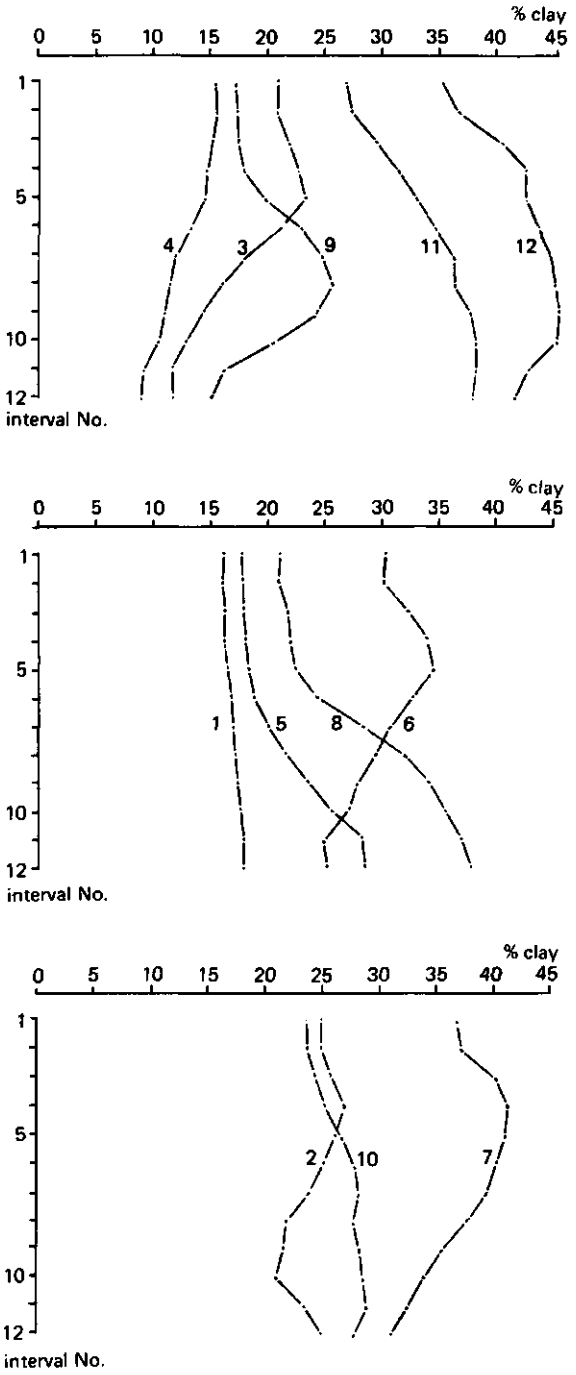


Fig. 20. Twelve centroids resulting from relocation with Set B of 600 *clay* depth-profiles to 1.2 m.

Secondly, relocation was aplied to the percentages clay in depth interval No 1 only. The initial partition was chosen such that it had twelve classes with equidistant limits. The same was done using the data successively of interval No 10 and 20. Minimum and maximum percentages clay in the final classes are given in Table 15; the pooled within-class variances in each of the 20 intervals are presented in Fig. 21. For a given interval, the within-class variance will mainly depend on the overall variance and on correlation with the interval on which the partition is based. Product-moment correlation coefficients between the intervals, calculated from the set (B) of 600 *clay* depth-profiles, are presented in Table 3.

It was concluded earlier in this section that, geometrically, the more the classes, the smaller and the more spherical they are. In addition to this, there is a tendency for

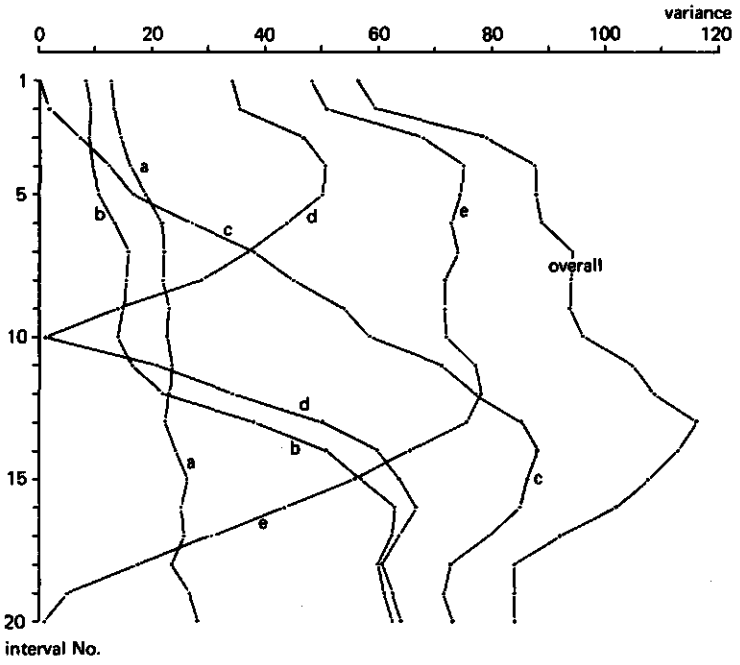


Fig. 21. Overall variances and variances within twelve classes constructed by relocation with 600 *clay* depth-profiles (Set B), using the data of interval No 1-20 (Line a), No 1-12 (Line b), No 1 (Line c), No 10 (Line d) and No 20 (Line e).

Table 15. *Minimum and maximum percentage clay in classes resulting from relocation with clay percentage (Set B) in single depth-intervals.*

Partition based on	Minimum clay % in Class 1	Maximum clay % in Class											
		1	2	3	4	5	6	7	8	9	10	11	12
Depth interval 1	10	14	16	18	20	23	25	27	29	31	34	38	42
Depth interval 10	4	10	13	15	18	23	24	28	31	34	38	43	52
Depth interval 20	3	8	12	16	18	20	23	25	28	30	34	39	54

within-class variances to be larger in intervals with larger overall variance, but Fig. 17 shows many exceptions to this. When the percentages of clay were previously standardized so that overall variances were equal, the within-class variances still differed considerably among intervals (Table 14). This must be due to the fact that the percentages clay in different intervals were not statistically independent (Table 3, correlations). Standardization hardly affected the central depth-profiles for *clay* when retransformed to the original scales. Fig. 19 indicates that within-class variances slightly increased in depth intervals with an original overall variance above the average (No 7-16), and slightly decreased in the other. But the effect was not proportional to the scale alterations and hardly of practical importance.

In contrast to standardization, assigning zero weights to intervals affected the results considerably. The central depth-profiles for *clay* based on the upper 12 intervals (Fig. 20) were only roughly similar to those based on all 20 intervals (Fig. 16, see separate overlays). Within-class variances in the upper 12 intervals decreased by about 30% at an average cost of doubling in the lower intervals (Fig. 21, Line a compared with b). The variances were even more drastically changed when partitions were based on only one interval (Fig. 21, Line c, d and e). With the high correlations between intervals (Table 3), the within-class variances only gradually increased with difference in depth from the intervals on which the partitions were based.

It would be attractive to use these correlations and to reduce the number of intervals by which the central depth-profiles are defined. As long as these intervals are not too far apart and the correlations are high enough, the increase in within-class variance in intermediate intervals may be acceptable. This would clearly facilitate identification, because comparisons are confined to fewer data. Furthermore, weighting with scale factors could be simulated by wider or closer spacing, according to the relative importance attached to the property at different depths. Avoiding scale factors would further facilitate identification because the usual scale, for instance percentage of clay, is preserved in each interval.

(c) Comparison with existing mapping units It seemed interesting to compare the variances within classes resulting from relocation with those within the mapping units actually defined for the survey of the test area. Pooled within-class variances were therefore calculated from the complete set of profiles (Set A), partitioned by

1. allocation to the central depth-profiles for each property (Fig. 13), as described in Section 4.2.4,
2. delineations on the available soil map of the test area at scale 1 : 25 000 (Kamping & Rutten, 1969).

To allow a comparison with the special classifications, the mapping units were grouped separately for each property. So mapping units with a common definition for *clay* were grouped, and likewise for the other properties. The definitions of these classes of mapping units are given in Table 16. Variances within the map delineations of these classes (thus including cartographic impurities) and within the special classes as we obtained them, are both presented in Fig. 22.

With some provisos the variances within the mapping units of the available conventional soil map (Fig. 22, Line m) may be compared to those within the experimental classes of depth profiles as we obtained them (Fig. 22, Line e). These limitations are:

Table 16. Definitions of mapping units used in the survey by Kamping & Rutten (1969), classified on the basis of single properties.

Property	Class of mapping units	Definition		
Clay		Average clay % in upper 25 cm		Clay % between 25 and 80 cm depth
	No 1	> 8%	and ⩽ 12%	If > 17.5%, then through a depth range < 15 cm
	No 2	> 12%	and ⩽ 17.5%	If > 25%, then through a depth range < 15 cm
	No 3	> 12%	and ⩽ 17.5%	> 25% through a depth range ⩾ 15 cm
	No 4	> 12%	and ⩽ 17.5%	No differentiation
	No 5	> 17.5%	and ⩽ 25%	> 35% through a depth range ⩾ 15 cm
	No 6	> 17.5%	and ⩽ 25%	⩽ 12% through a depth range ⩾ 15 cm
	No 7	> 17.5%	and ⩽ 25%	If ⩽ 12% or > 35%, then through a depth range < 15 cm
	No 8	> 17.5%	and ⩽ 25%	No differentiation
	No 9	> 25%	and ⩽ 35%	⩽ 17.5% through a depth range ⩾ 15 cm
	No 10	> 25%	and ⩽ 35%	⩽ 35%; if ⩽ 17.5%, then through a depth range < 15 cm
	No 11	> 25%	and ⩽ 35%	⩽ 35% from 25 to 40 cm depth; > 35% at a depth between 40 and 80 cm; if ⩽ 17.5%, then through a depth range < 15 cm
	No 12	> 25%	and ⩽ 35%	> 35% at a depth between 25 and 40 cm; if ⩽ 17.5%, then through a depth range < 15 cm
	No 13	> 25%	and ⩽ 35%	No differentiation
	No 14	> 35%		⩽ 25% through a depth range ⩾ 15 cm
	No 15	> 35%		If < 25%, then through a depth range < 15 cm
Humus	No 1	Soils of dwelling mounds, identified on the basis of their elevated position and thick (> 40 cm) antropic epipedon with phosphate spots, potsherds and dark colours, usually associated with more than 1% humus		
	No 2	Average humus % in upper 25 cm ⩾ $2.5(1+L/100)$, where L denotes clay % of the mineral part		
	No 3	Average humus % in upper 25 cm < $2.5(1+L/100)$		
Carbonate	No 1	Reaction to 10% HCl at least weak from 0 to 60 cm, and strong from 60 to 80 cm depth		
	No 2	Reaction at most weak from 0 to 25 cm, and no reaction from 25 to 50 cm depth		
Ripening	No 1	The subsoil is at most half-ripened within 80 cm depth		
	No 2	Ripened or nearly ripened to a depth of 80 cm		
Knip	No 1	Soils without 'knip'		
	No 2	Soils with 'knip'		
Peat		No differentiation		

firstly, the variances relate to different numbers of classes. Secondly, sampling was purposive instead of random and not independent of delineation on the map. Thirdly, the mapping units include cartographic impurities whereas the experimental classes do not. It can nevertheless be concluded that the variances within the experimental classes were, on average, lower than within the conventional mapping units. For several depths and prop-

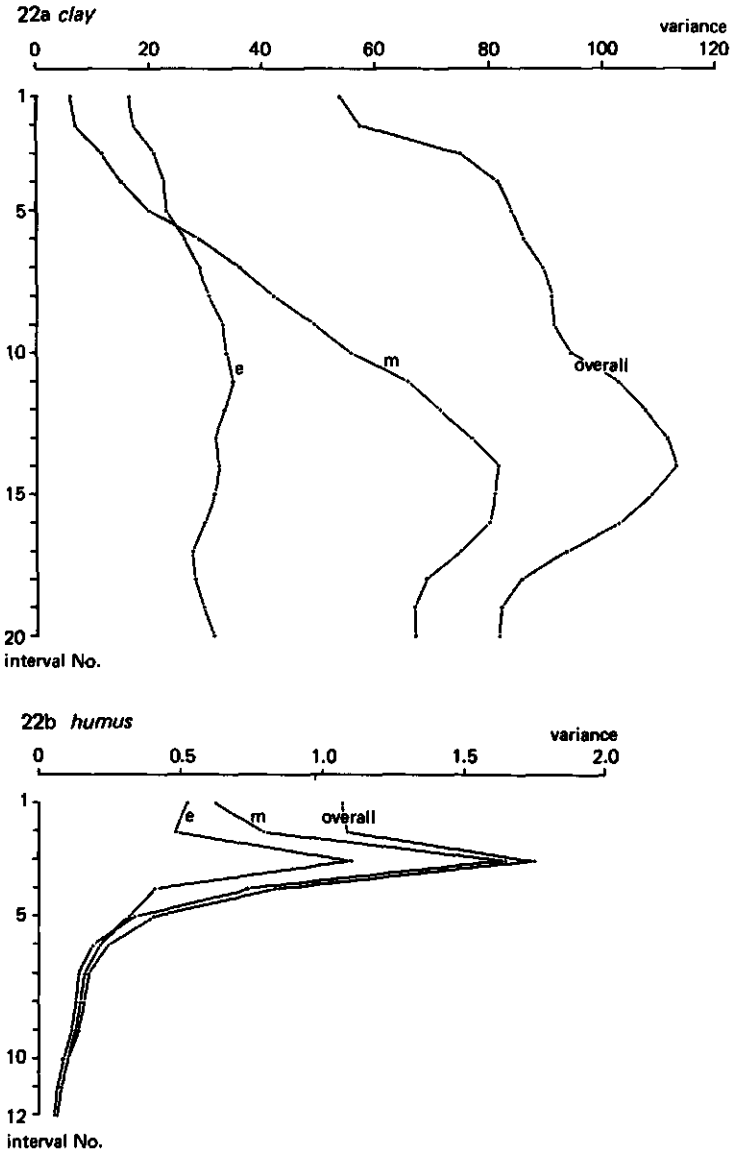
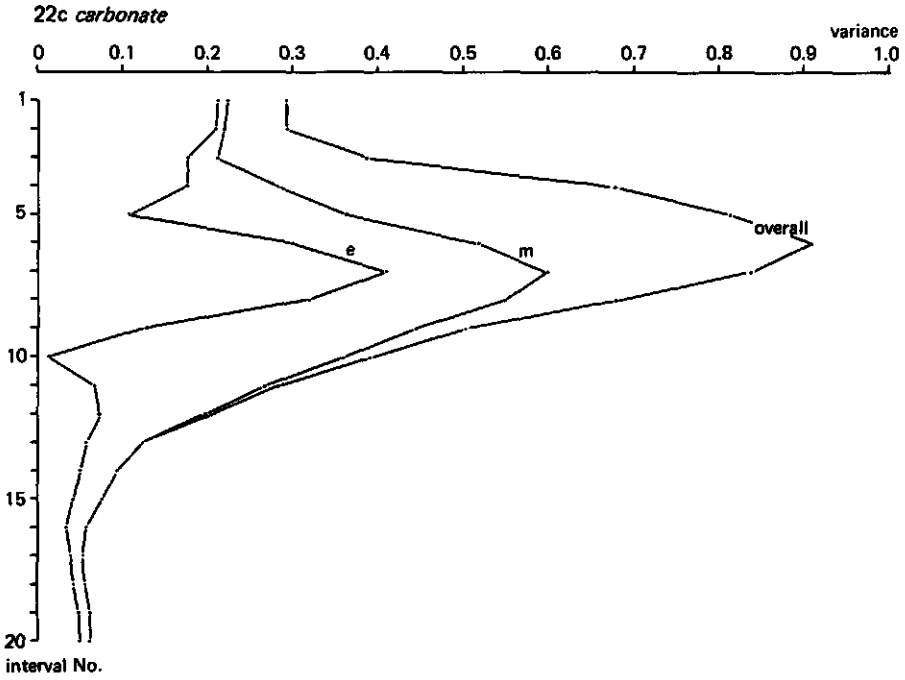
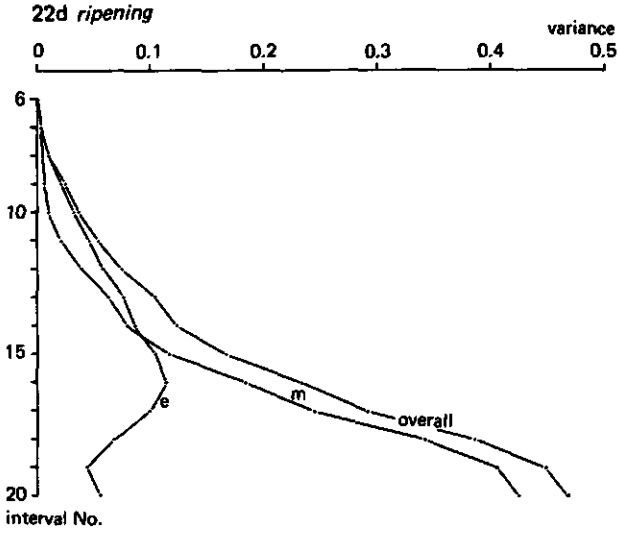
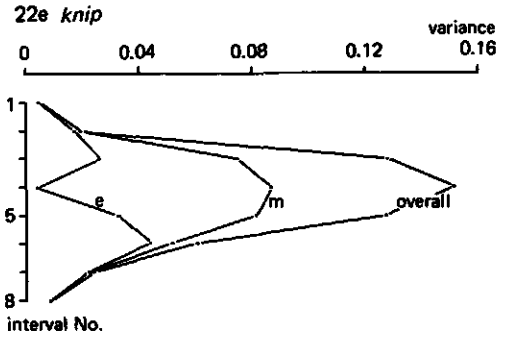


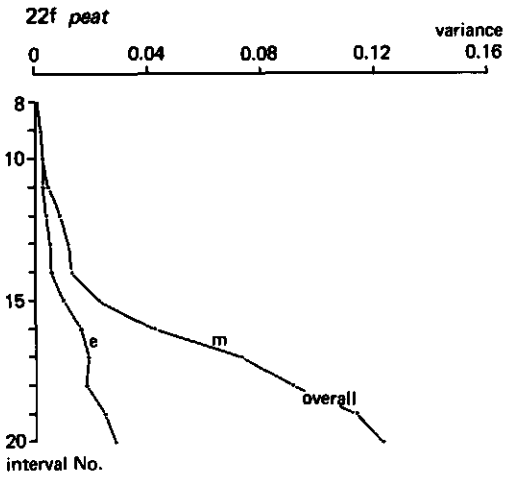
Fig. 22. Overall variances and within-class variances calculated from 2212 soil profiles (Set A), partitioned by:
(e) allocation to the central depth-profiles of Fig. 13,
(m) delineations on a soil map (scale 1 : 25 000), according to the mapping units defined in Table 16.

erties, the difference in homogeneity seems to be of practical importance. For instance, the conventional mapping units were highly homogeneous for *clay* in the topsoil, but the situation rapidly worsens with depth. No useful predictions could be made from the soil map about *clay* at depths of 1.2 m (usual for tile drains) or more, although the map was based on augerings to 2 m depth, because the definitions of the mapping units depend









only on the clay contents in the upper 0.8 m and lay special emphasis on the topmost 0.25 m (Table 16).

The situation for *humus* and *carbonate* is similar: the mapping units are too heterogeneous to provide useful predictions of amounts below the Ap-horizon. Although the mapping units strictly separate 'knip' soils from soils without 'knip', the variances within mapping units are relatively high for *knip*, because of both cartographic impurities and variations in the depth at which 'knip' occurs. The borings provided information about the presence of peat or peaty material in the subsoil, but this was not displayed on the map.

(d) Number of final classes The results so far concern the effect of the number of classes and of the weights at the first stage of the classification procedure. We will now concentrate on the second stage. Before classifying the 108 combinations of central depth-profiles into 35 classes, the properties were standardized to unit sum of overall variances (4.2.3.3).

Figure 15 indicates how $\mathrm{tr}(W)/n$ increases when the number of classes is reduced by Ward's method. The values of $\mathrm{tr}(W)/n$ presented there were calculated from transformed

data: when fusing finally results in one class, tr(W)/n coincides with tr(T)/n and equals 6, through a contribution of 1 from each property.

Pooled within-class variances were calculated from the original data of 600 profiles (Set B), after allocation to the 35 classes mentioned above. The same was done after similar allocations to 30 and 25 classes. The variances within 35 classes are represented in

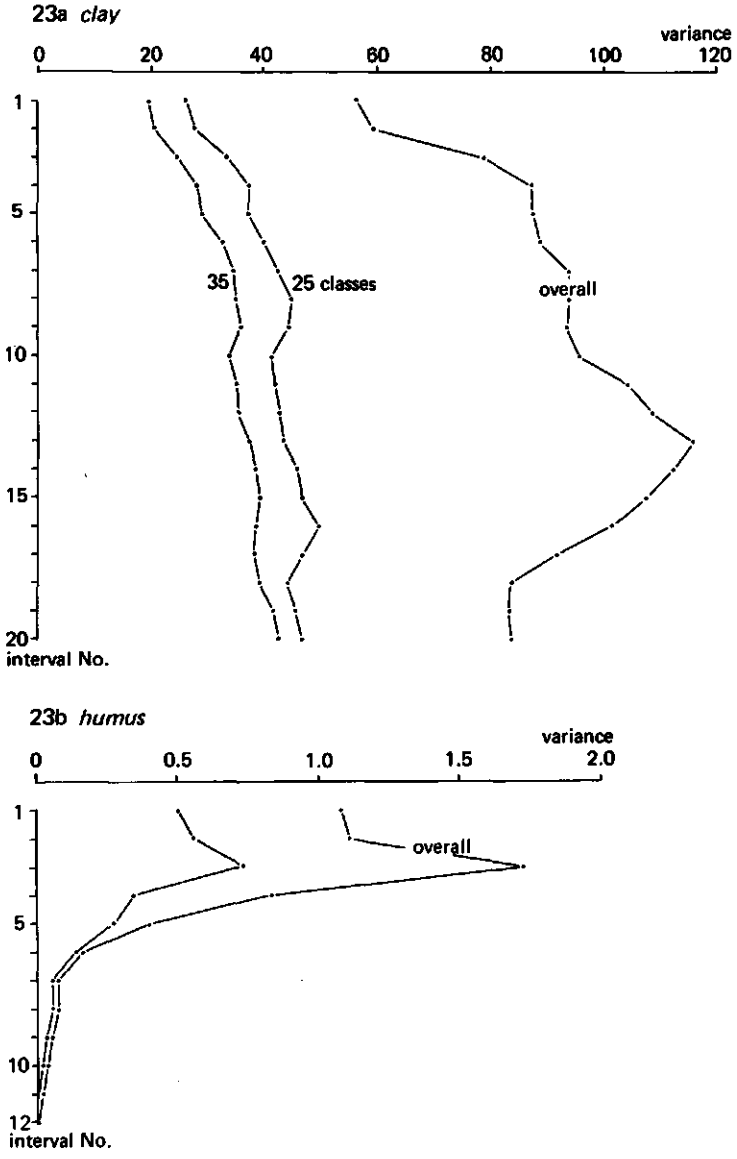
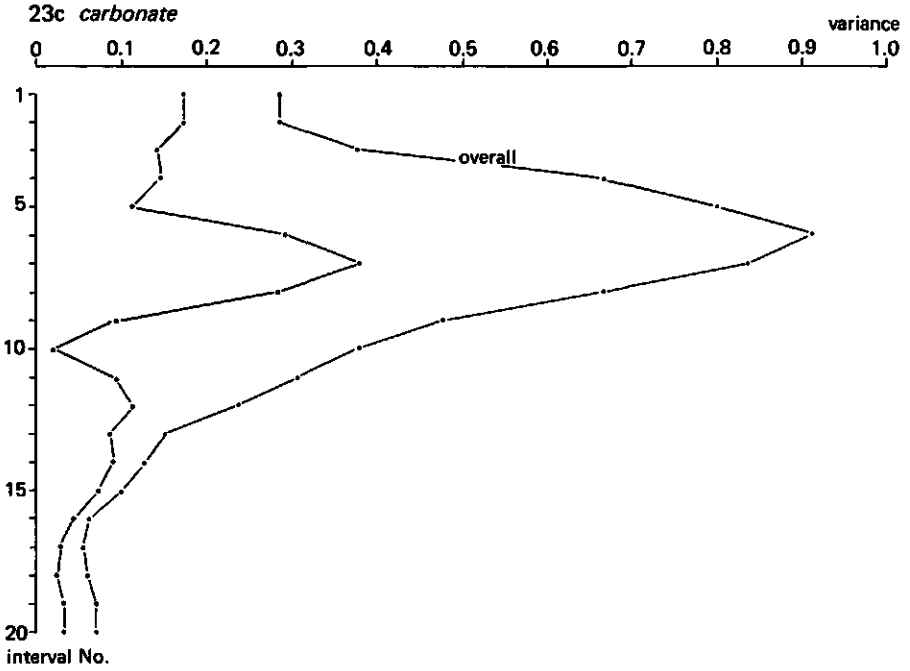
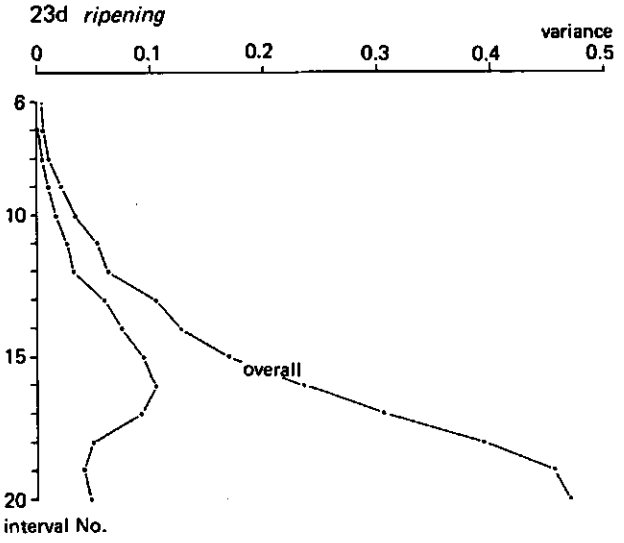
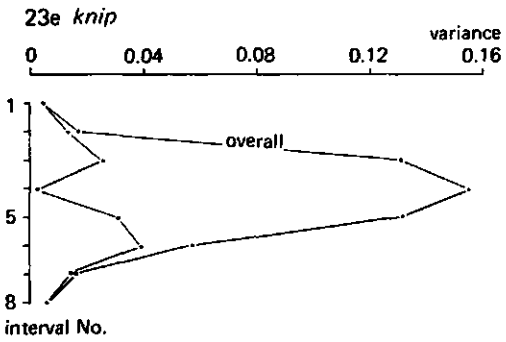


Fig. 23. Overall variances and within-class variances calculated from 600 soil profiles (Set B), allocated to the 35 final classes (for *clay* also 25 classes) specified in Table 7.

Fig. 23 for all intervals and properties. The variances within 30 and 25 classes have much the same distribution over intervals as those within 35 classes; they are not represented in Fig. 23 except for *clay* within 25 classes, where the differences are largest. The sum of the within-class variances over depth intervals is given as tr(W)/n in Table 17, for each of the three classifications and each property.







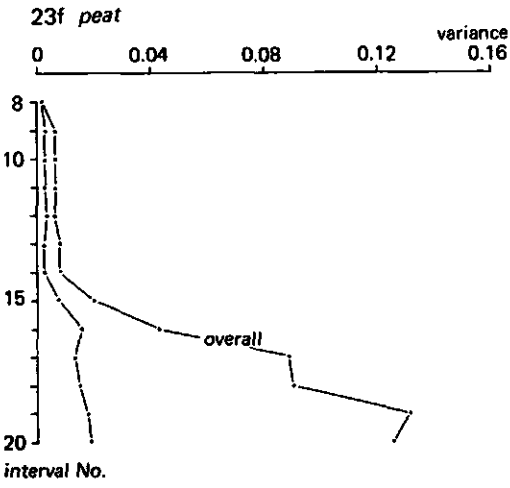


Table 17. Values of tr(T)/n and tr(W)/n for each property, calculated from the original data of 600 soil profiles (Set B), after allocation to the special classifications and to final classifications with different numbers of classes.

	Number of classes	*Clay*	*Humus*	*Carbonate*	*Ripening*	*Knip*	*Peat*
tr(W)/n after allocation to special classifications							
clay	6	563.2	5.106	5.376	2.022	0.4160	0.5135
humus	3	1750.0	3.164	6.784	2.421	0.5016	0.5395
carbonate	3	1379.8	5.326	2.774	2.255	0.4432	0.5382
ripening	2	1629.6	5.461	6.474	0.830	0.5040	0.5135
knip	2	1610.5	5.284	6.266	2.393	0.1584	0.5447
peat	2	1796.6	5.534	6.770	2.207	0.5184	0.1183
tr(W)/n after allocation to final classifications							
	35	683.1	2.746	2.434	0.659	0.1400	0.1053
	30	726.2	2.777	2.520	0.662	0.1440	0.1131
	25	831.6	2.813	2.596	0.669	0.1464	0.1131
tr(T)/n		1831.2	5.577	6.926	2.454	0.5194	0.5452

The graph of the number of final classes against $\mathrm{tr}(W)/n$ (Fig. 15) does not show clear discontinuities, indicating that there is no 'natural' number of classes in the present sample (Set B).

Reduction of the number of final classes from 35 to 25 by Ward's method, with previous standardization of the properties, unequally affected the pooled within-class variances for the individual variables. The distribution of these variances over the depth intervals remained approximately unaltered (Fig. 23), but for *clay* they increased by more than 20% on average, while for the other properties they increased by less than 10% on average (Table 17). Reduction in the number of final classes apparently is achieved mainly at the expense of homogeneity for *clay*. The dendrogram in Fig. 14 and Table 7 even show that only from the level with six classes, the classes begin to contain more than one central depth-profile for *ripening*, *knip* or *peat*. This is, of course, due to the chosen weighting of the properties.

(e) Weighting of properties In addition to standardization, zero weights were applied to the properties, in the same way as with the depth intervals. The special classifications obtained at the first stage could be directly used for this, because each of these classifications is based on only one property, implying zero weight for all the others. Pooled within-class variances were calculated from the Set B of 600 profiles, for each of the special classifications and each property. Their sums over depth intervals are presented as $\mathrm{tr}(W)/n$ in Table 17.

As mentioned before, the effect of weighting the depth intervals on the special classifications is partly determined by statistical dependence between the intervals. Similarly, the effect of weighting the properties on the final classification will be partly determined by statistical dependence between the properties. Therefore, this dependence was studied, using the contingency tables (Table 18) that resulted from allocation of the soil profiles of Set A to the special classifications specified in 4.2.4.1. On χ^2-test with confidence level 0.95, the hypothesis of independence was rejected for all 15 pairs of properties, except for the pair *knip* – *peat*.

The construction of 35 soil classes at the second stage of the application in 4.2 was preceded by standardization of the properties to equal sum of overall variance. Nevertheless this part of the procedure unequally affected the within-class variances calculated from the original data, when compared with the special classifications. The distribution of these variances over depth intervals remained roughly the same (Fig. 17 compared with 23, see separate overlays), but there was an average increase of about 20% for *clay* and an average decrease between 10 and 20% for the other properties (Table 17). The same effect appears from Table 7, where most of the classes contain two or more central depth-profiles for *clay* but only one for the other properties. Reduction to 35 classes at the second stage was apparently achieved mainly at the expense of homogeneity for *clay*.

Combining the special classifications into the interim classification generally will have reduced the within-class variances, due to increase in the number of classes and to statistical dependence between the properties (Table 18). Fusion of classes at the second stage increased the within-class variances but, except for *clay*, this apparently did not outweigh the previous reductions.

Both for standardization of the properties and for minimization of $\mathrm{tr}(W)/n$ by Ward's

Table 18. Contingency table for each pair of properties, after allocation of 2212 soil profiles (set A) to the central depth-profiles in Fig. 15. Frequencies expected under the hypothesis of statistical independence between the properties[1] are between brackets

	Peat		*Knip*		*Ripening*		*Carbonate*			*Humus*			Total
	1	2	1	2	1	2	1	2	3	1	2	3	
Clay													
1	524	0	517	7	507	17	481	33	10	236	185	103	524
	(519)	(4.7)	(423)	(101)	(448)	(76)	(241)	(220)	(63)	(250)	(203)	(71)	
2	459	0	345	114	432	27	135	295	29	234	169	56	459
	(455)	(4.1)	(370)	(89)	(393)	(66)	(211)	(193)	(55)	(219)	(178)	(62)	
3	197	13	128	82	110	100	37	91	82	63	124	23	210
	(208)	(1.9)	(169)	(41)	(180)	(30)	(97)	(88)	(25)	(100)	(82)	(28)	
4	426	1	406	21	383	44	225	166	36	251	109	67	427
	(423)	(3.9)	(345)	(82)	(365)	(62)	(196)	(179)	(52)	(204)	(166)	(57)	
5	302	6	281	27	263	45	121	132	55	165	104	39	308
	(305)	(2.8)	(249)	(59)	(263)	(45)	(142)	(129)	(37)	(147)	(120)	(41)	
6	284	0	108	176	197	87	19	211	54	107	168	9	284
	(281)	(2.6)	(229)	(55)	(243)	(41)	(131)	(119)	(34)	(136)	(110)	(38)	
Humus													
1	1052	4	854	202	943	113	528	426	102				1056
	(1047)	(9.5)	(852)	(204)	(903)	(153)	(486)	(443)	(127)				
2	845	14	643	216	694	165	327	403	129				859
	(851)	(7.8)	(693)	(166)	(735)	(124)	(395)	(360)	(103)				
3	295	2	288	9	255	42	163	99	35				297
	(294)	(2.7)	(240)	(57)	(254)	(43)	(137)	(125)	(36)				
Carbonate													
1	1006	12	1008	10	924	94							1018
	(1009)	(9.2)	(821)	(197)	(871)	(147)							
2	925	3	617	311	799	129							928
	(920)	(8.4)	(749)	(179)	(794)	(134)							
3	261	5	160	106	169	97							266
	(263)	(2.4)	(215)	(51)	(227)	(39)							
Ripening													
1	1892	0	1575	317									1892
	(1875)	(17.1)	(1527)	(365)									
2	300	20	210	110									320
	(317)	(2.9)	(258)	(62)									
Knip													
1	1766	19											1785
	(1769)	(16.1)											
2	426	1											427
	(423)	(3.9)											
total	2192	20											2212

1. This hypothesis is rejected by χ^2-test (confidence level 0.95) with all 15 pairs, except for the pair *knip-peat*.

method, the combinations of central depth-profiles were taken as representatives of the interim classes. These combinations would have differed somewhat from the proper centroids of the classes. Using the centroids for standardization and minimization would be computationally more laborious but the variances within the final classes would be easier predictable and probably lower on average.

Table 18 indicates that statistical dependence exists between the properties. As opposed to the intervals, however, this could not be employed to reduce the number of properties. Table 17 shows that if any of the properties is not used to define the classification, the within-class variances increase to an extent that hampers useful predictions about that property.

5 Conclusion

The tests reported in 4.2 and 4.3 do not suggest that the principle used in the classification procedure – minimization of $\mathrm{tr}(W)/n$ in two stages – should be revised. In 4.2, the method was applied but the available data do not allow a complete assessment of the method. Whilst the approach by central depth-profiles facilitates field identification of the soil profiles, the accuracy of mapping the classes and the effort and costs involved, have yet to be assessed. Further research will thus investigate whether the mapping units resulting from delineation are sufficiently homogeneous and, if not, how this can be improved by adaptation of the classification procedure. One possibility would be to try to achieve closer correspondence between soil classes and visible landscape features by the choice and transformation of the variables for classification. It would also be interesting to see wether 'natural' soil classes, resulting from distribution fitting (2.4.2), have closer correspondence to geographical patterns of soil variation than classes with optimum homogeneity.

A further aspect that may be considered is the map image. It is pointless to compare the experimental soil map (Map 12) with the existing soil map of the area. The classifications used are different, for instance other weights have been employed (see Fig. 22), and the classes have been delineated differently. Seen in isolation, however, units in Map 12 look more fragmented than tolerated by normal standards of map legibility. This partly results from the choice of one of the most complex parts of the test area for this experiment, as can be seen from the point map of the special classes for *clay* (Map 1). Furthermore, no generalization has been attempted. If normal criteria of generalization were applied, a much less fragmented pattern would arise. The fragmentation experienced on the maps of the special classifications for the constituent properties (Maps 1-11) must be deemed acceptable.

One may ask whether optimization procedures may not be used in constructing soil classes which, when delineated, give a more legible map. As illustrated in 4.3.2 and 4.3.3, it may be possible to construct different classifications of which the classes have almost the same homogeneity. This implies a certain freedom of choice, which could be used in favour of classes with more geographical contiguity.

For this approach, one or more measures of map complexity and intricacy will be required, for instance the total number and size of mapped occurrences or the length of the soil boundaries. These quantities must be included in the objective function or in constraints, and automatically evaluated for alternative classifications. Geographical information is here required from a representative pilot sample of the area. How ever attractive, it seems difficult to implement such an approach. So the suggestion in 3.4.2.2 is to base numerical soil classification on soil properties only, followed by map generalization if required. A system of automated cartography may allow this generalization to take place interactively. The effect each change in the map may have on the homogeneity of the classes concerned, may then be automatically evaluated.

The following comments relate to the method of optimization used in Chapter 4. The results of 4.3.1 confirm the importance of having a sufficiently representative sample from the area for which a classification is desired. A classification based on a small sample may differ significantly from that derived from larger ones. For all six soil properties, a sample of 100 depth profiles proved insufficient for the construction of a special classification; for two properties (*humus* and *peat*) even 600 seemed inadequate. It is difficult to give general advise on sample size. The type of variability in the area, the number of classes desired and sampling design are factors to be considered. It may therefore be recommended to apply the optimization procedure on various subsamples, to gain information on the reliability of the result. It seems profitable, also for other purposes, to investigate whether the size of the sample may be reduced by a more efficient sampling design. Stratification by landscape features is one possibility.

For a given sample, the result of optimization may vary with the initial partition and with the order in which the profiles are relocated. This arises as the relocation process may terminate at a local minimum instead of the absolute minimum of the objective function. In the experiment reported here there were insignificant differences in the values achieved for the objective function. In general, however, the question remains open how far the quality of a final partition falls short of the best possible one. It is therefore important to aim further research at the reliability of the relocation method. The use of Ward's method to select the initial partition did not give significant improvements over a random starting point. For practical use, it is recommended to repeat the relocation method with different initial partitions.

Some local minima may be avoided when, in addition to one by one relocation of objects, two objects simultaneously exchange places in the partition (3.4.3.2). The GENSTAT program of the University of Edinburgh contains this facility.

Another way to reduce the risk of a local minimum could be to construct from the sample an estimated multivariate frequency distribution. From this distribution, hypothetical soil profiles could then be drawn at random and allocated to continuously recalculated centroids, until the centroids achieve acceptable stability. An advantage of this approach is that it allows closer study of convergence. It is possible to construct realistic test problems amenable to analytical solutions. Theory of stochastic processes may also be applied. From the analysis of MacQueen (1966), it follows that if for a given population no partition exists that is only locally optimum (i.e. optimally adapted to the corresponding centroids but less well than is possible with other centroids), then the process mentioned will converge on the absolute optimum. A relocation procedure applied to a sample from such a population may still yield a suboptimum partition.

The choice of the number of classes and the weights of the variables is one of the most important aspects of the classification procedure. It is also one of the most difficult aspects to decide on. To investigate the influence the number of classes and weights have on the final result, sensitivity analysis is recommended. Section 4.3.4 showed that reduction in the number of classes always led to a gradual increase in $\mathrm{tr}(W)/n$; there seems to be no 'natural' number of classes. Only large changes in weights had significant impact on the central depth-profiles for *clay*. When certain intervals were given zero weight by excluding them from the analysis, heterogeneity in such intervals increased only moderately as long as the intervals used were not too far apart.

This suggests that the correlations between intervals may be used to define the central depth-profiles on the basis of a limited number of intervals. Weighting may then be

approached by the selection of these intervals. If for a property only one or two transitions (from one value to another) are recorded, and these are similar in each profile, a further simplification is possible by using the depths of the transitions as variables for the classification. As appears from the results of Chapter 4, the properties *carbonate, ripening* and *knip* could have been treated this way.

Although there was statistical dependence also between properties, the elimination of some properties from the classification process did not allow useful predictions to be made of these properties from the ones employed. The classes were too heterogeneous.

More generally, the introduction of an objective function for classification will allow a more quantitative investigation of which properties may be displayed together on one map, and which may require separate maps. For instance, Fig. 21 suggests that in the test area clay content in the subsoil is probably best conveyed by a separate map.

The same argument applies to the question of the size of the area for which a classification must be optimized. In the present study, a classification was optimized for an area of 8100 ha only. To what extent is such a classification also suitable for other, comparable areas? A systematic analysis of advantages and disadvantages of a collection of local classifications rather than one general classification must have high priority in further research. The homogeneity of the classes and the corresponding mapping-units is one important aspect. A two-stage procedure as followed in this study would open the way for a compromise whereby locally adapted classifications are constructed from a general (e.g. national) system of central depth-profiles. But work on the classification of texture depth-profiles in Dutch coversand areas (to be reported elsewhere) has shown that central depth-profiles derived from one area may differ significantly from those originating from others.

Another principal question concerning classification relates to the efficiency and accuracy of making a separate map to answer a given single query about the soil, compared with a more general map directed to various queries. Often one has to choose between compiling a 'general purpose' map or a number of 'special purpose' maps. Is greater heterogeneity of general purpose mapping-units too high a price for less effort in delineation? A measure of map accuracy, and information on the efficiency of mapping different classifications in various circumstances, may bring a rational choice within reach.

Summary

Classifications facilitate communication. Soil scientists have always paid much attention to the construction and improvement of classifications. By the conventional approach to soil classification the classes are constructed, partly by intuition, on the basis of knowledge and theory about variation in soil properties. This is usually followed by repeated testing and adapting. Beginning in the early 1950s, numerical methods for classification found use in biological and social sciences. By fixed and specified procedures, the classification is computed from a set of data on the objects (individuals) to be classified. This allows more intensive and consistent use to be made of the available information. The methods are also suited to automation. An indirect advantage is that principles of classification are reassessed in detail, because the method has to be specified explicitly and unambiguously.

This study was induced by classification problems experienced in soil surveys by the Netherlands Soil Survey Institute, and by lack of clear indications in the literature which of the options in the vast array of numerical classification methods are most suitable to soil survey. Chapter 1 outlines these problems.

The search for solutions begins with an analysis of the general aspects of soil classification (Chapter 2): the purpose of the classification, the choice of the data on which it is to be based, the possible preliminary processing of these data (e.g. changes in scale), the type of classification desired, the assessment once a classification is constructed, and the allocation of soil profiles to existing classes ('identification').

The usefulness of a classification in soil survey depends on how far the classes can be geographically delineated at acceptable costs, and how far a legible map arises, with units sufficiently homogeneous in those soil properties in which the users of the map are interested. The usefulness therefore depends also on the survey method to be applied. A method often used is referred to as 'free survey' (*sensu* Steur, 1961). By this method the surveyor gradually builds up an image of the geographic distribution of the classes, observing soil profiles where he expects most information from them, instead of sampling at random or systematically. Under certain conditions, that method seems more efficient. So the classification procedure to be designed had to fit into this type of survey. The main practical consequences of this requirement are that the classification must be available at an early stage of the field work, and that soil profiles can be easily identified in the field. The latter implies that the classes are defined in terms of properties that can be assessed in the field, and that definitions of the classes are not too intricate.

We can represent soil profiles as points in a multidimensional space with soil variables (e.g. clay content at a given depth) as mutually perpendicular axes. The demand for homogeneity implies that the boundaries of the classes are so adapted to the point distribution that the scatter is minimal within the classes, i.e. the classes are as compact as possible. This does not imply that if 'natural' clusters of points occur in the space, they

could not be intersected by a class boundary. Such clusters may have an elongated shape and be too heterogeneous for one or more variables.

Chapter 3 deals with the choice from the many numerical methods that are possible. Three different approaches are discussed, together with methods to which they have led.

The first approach is heuristic: intuitive conceptions about the classification process are directly transposed into a numerical procedure. This approach dominated the first stage of development of numerical classification. Numerical methods were developed for a large variety of actual classification problems. In a given situation, these methods did or did not provide a satisfactory solution; their general merits remained unknown. Elements of different methods were sometimes joined in apparently inconsistent procedures.

The second approach is to impose mathematical requirements upon the methods. In principle this seems attractive. In practice, it might be difficult to define a logically consistent set of requirements that properly reflects the purpose of the classification.

By the third approach an objective function and, possibly, side conditions are derived from the purpose of the classification. Then one tries to find a classification that optimizes the objective function and satisfies the possible side conditions. This approach seems to be more suitable than the previous two. The reasons of possible shortcomings of a calculated classification can be better analysed.

Objective functions, side conditions and optimization methods are discussed in Section 3.4. As objective function was selected: the mean squared Euclidean distance between the profiles and the (hypothetic) representative of the class to which each profile is allocated. If the centroid of a class, consisting of the means of the respective variables, is used as representative, this function is identical to the pooled within-class variance summed over variables. The smaller the value of this function, the more homogeneous the classes will be on average. The desired number of classes can be introduced as a side condition. A relocation method was selected for minimization of the objective function: an initial solution was stepwise improved, continuously re-allocating the profiles to the classes. Iteration was stopped when no further improvement was possible in this way.

Chapter 4 indicates how the relocation method was built into a classification procedure aiming at both homogeneity of classes and manageability in the field. This procedure was applied to data collected during a survey of a young marine clay area near the northern coast of the Netherlands (Kamping & Rutten, 1969). The data set consisted of field estimates of 2212 soil profiles, divided into 20 intervals of 10 cm thickness. The following properties were recorded for each depth interval: the percentages of clay and humus, the amount of carbonates, the degree of physical ripening, the presence of 'knip' (4.2.2.3) and the amount of peat in the subsoil.

At the first stage of the procedure (Fig. 6), a special classification was calculated for each of the six properties. The objects to be classified were 'depth profiles' recorded for the property in question; each variable represented a value of that property in a given depth interval. Thus the representative of a class can be referred to as a 'central depth-profile' (e.g. for percentage clay; Fig. 13). It consisted of the centroid of that class, which was slightly rounded and smoothed to facilitate the identification of new depth-profiles in the field. A synthesis was then made by combining the six special classifications into one classification. Each possible combination of central depth-profiles, with one central

depth-profile for each property, formed the representative of a new class. As this resulted in too many classes to delineate each individually, the classes were fused at a second stage to a limited number of larger units. Here too, the sum of the pooled within-class variance was minimized.

Compared with direct calculation of a classification from all variables, the present procedure has the advantage that the identification of soil profiles is much easier and can be easily updated when the classification is altered during the survey.

The calculated classes were mapped using a computer-aided technique with output on a line-printer. This technique has more general application in that polythetically defined classes can be delineated with the original profile data.

With the classification procedure specified in Chapter 4, several choices have to be made. In general, each of these choices influences the final result. Tests provided information on these effects.

The classification must be based on a sample of soil profiles from the area for which it is intended. This sample should be kept as small as possible for technical and economic reasons. Too small a sample leaves the classification too much to chance. Class boundaries may be insufficiently aligned to the population as a whole. Experiments showed that in the test area, a sample of 100 profiles was too small for classifications based on single properties. A sample size of 600 proved sufficient, except perhaps for classifications based on humus and peat contents.

As the relocation process may reach only a local minimum instead of the absolute minimum of the objective function, the result generally also depends on the initial solution and the order in which the profiles are relocated. Thus somewhat different classifications resulted from different initial solutions and orders, but the differences between the values of the objective function were unimportant. So a preliminary analysis of a subsample to construct an initial solution had few advantages; a random initial solution gave classes with nearly the same homogeneity.

By the choice of the number of classes and the weights attached to the variables, one can purposefully influence the final result. At the first stage of the procedure, one chooses the number of classes and the weights of the depth intervals, for each of the special classifications. At the second stage, one chooses the number of classes of the final classification, and the weights of the properties. If the number of classes was increased, the sum of the pooled within-class variance generally decreased, while the distribution of these variances over the variables became more even. If a larger scale factor was chosen for a given variable, with constant scale factors for the others, the classes generally became more homogeneous for that variable, i.e. the pooled within-class variance of that variable decreased. The number of classes and of variables and their possible non-linear transformations and scale factors should be chosen such that for the resulting classes sufficiently accurate estimates can be made of the variables in which users of the map are interested. However, the effects on the final result are difficult to predict quantitatively; they depend on the frequency distribution within the sample from which the classification is to be calculated. In addition, the homogeneity of the final classes for a given property is governed not only by the scale factor but also by the number of classes chosen for that property at the first stage of the procedure. For these reasons a sensitivity

analysis is recommended, so that one can concentrate on the parameters with an important influence and harmonize the various choices to each other.

Only when drastic scale alterations were applied to the clay percentages in the respective depth-intervals, important changes occurred in the distribution of within-class variances over intervals. The correlations between the intervals can be used to reduce the number of intervals for classification. However, classes based only on clay percentages in the upper 1.2 m turned out to be already highly heterogeneous in the depth range 1.5-2.0 m.

The study showed that the suggested procedure may be a useful tool in constructing classifications to be applied in soil surveys. Further research is needed, however, with emphasis on the accuracy with which the classes are delineated in the field and on the choice of variables and weighting.

Samenvatting

Classificaties spelen een belangrijke rol in het denken en spreken. Dit geldt in het bijzonder waar het overdracht van wetenschappelijke informatie betreft. De bodemkunde vormt hierop geen uitzondering: bodemkundigen hebben steeds veel aandacht besteed aan het opstellen en verbeteren van classificaties. Bij de gebruikelijke, conventionele benadering van bodemclassificatie wordt, langs gedeeltelijk intuïtieve weg, getracht om zinvolle of bruikbare klassen te definiëren. Dit gebeurt aan de hand van kennis en theorie over variaties in bodemeigenschappen. Veelal volgt daarna een iteratief proces van corrigeren en bijstellen.

Numerieke classificatiemethoden worden sinds enkele tientallen jaren toegepast, vooral in biologie en menswetenschappen. Bij dergelijke methoden wordt gebruik gemaakt van een verzameling gegevens omtrent de in klassen te groeperen objecten, en wel volgens een vaste, exact omschreven procedure. Voordelen van numerieke methoden boven een conventionele benadering zijn dat de beschikbare informatie op meer intensieve en consistente wijze kan worden gebruikt, en dat het classificatieproces door automatisering kan worden versneld. Een indirect voordeel is bovendien dat de uitgangspunten bij het classificeren opnieuw nauwkeurig worden overwogen, omdat men gedwongen is de procedure expliciet en ondubbelzinnig te specificeren.

Aanleiding tot dit onderzoek waren enerzijds de concrete classificatieproblemen bij door de Stichting voor Bodemkartering uitgevoerde karteringen, anderzijds de lacunes in de literatuur over numerieke bodemclassificatie. De lacunes betreffen vooral de keuze uit het grote aantal mogelijke classificatiemethoden en hun toepassingsmogelijkheden in de bodemkartering. Met het oog hierop zijn eerst de belangrijkste algemene aspecten van bodemclassificatie in beschouwing genomen (Hoofdstuk 2): het doel van de classificatie, de keuze van de gegevens waarvan wordt uitgegaan, de eventuele voorbewerking van deze gegevens (b.v. schaalveranderingen), het gewenste type van classificatie en, wanneer eenmaal een classificatie gemaakt is, het evalueren daarvan en het toewijzen (identificeren) van bodemprofielen tot de klassen.

De bruikbaarheid van een classificatie bij de kartering hangt af van de vraag in hoeverre het mogelijk is de klassen tegen acceptabele kosten geografisch te omgrenzen, en in hoeverre er een leesbare kaart ontstaat, met eenheden die voldoende homogeen zijn wat betreft die bodemeigenschappen waarin de gebruikers van de kaart zijn geïnteresseerd. De bruikbaarheid hangt dus mede af van de te volgen karteringsmethode. Een veel toegepaste methode is de z.g. 'vrije kartering' (*sensu* Steur, 1961). De karteerder bouwt daarbij geleidelijk een beeld op van de geografische ligging van de klassen, en verricht de waarnemingen aan bodemprofielen dáár waar hij er de meeste informatie van verwacht, dus niet aselect of in een vast geografisch patroon. Deze methode geldt als relatief efficiënt, wanneer aan bepaalde voorwaarden is voldaan. Daarom werd de eis gesteld dat de classificatieprocedure in dit type van karteringen zou kunnen worden ingepast. De belangrijkste praktische consequenties hiervan zijn dat de classificatie in een vroeg stadium van het

veldwerk beschikbaar moet zijn, en dat zij identificatie van bodemprofielen in het veld mogelijk moet maken. Dit laatste houdt in dat de klassen worden gedefinieerd in termen van eigenschappen die in het veld zijn waar te nemen, en dat de definities van de klassen niet te ingewikkeld zijn.

Wanneer bodemprofielen worden voorgesteld als punten in een ruimte met de bodemvariabelen (b.v. het lutumgehalte op een bepaalde diepte) als onderling loodrechte assen, dan houdt het streven naar homogeniteit in, dat de klassegrenzen aan de verdeling van de punten in die ruimte worden aangepast. Deze aanpassing is er op gericht dat de klassen een zo klein mogelijke spreiding in eigenschappen krijgen, dat wil zeggen zo compact mogelijk worden. Het houdt dus niet in dat wanneer er duidelijke 'clusters' in de ruimte aanwezig zijn (gebieden met een relatief hoge puntendichtheid), deze niet doorsneden mogen worden door een klassegrens: dergelijke clusters kunnen langgerekte vormen hebben en te heterogeen zijn wat betreft één of meer variabelen.

In Hoofdstuk 3 is ingegaan op de keuze uit de vele mogelijke numerieke methoden. Er worden drie verschillende benaderingen van dit keuzeprobleem aan de orde gesteld, samen met een aantal van de methoden waartoe deze geleid hebben.

De eerste benadering is de heuristische. Men tracht daarbij intuitieve concepties omtrent het classificatieproces rechtstreeks te 'vertalen' in een numerieke procedure. Deze benadering overheerste in de eerste ontwikkelingsperiode van de numerieke classificatie. Terwijl wiskundigen zich veelal afzijdig hielden, werden vooral in de biologie en de menswetenschappen talloze methoden ontwikkeld, als antwoord op een grote verscheidenheid aan concrete classificatie problemen. Deze methoden leverden in een gegeven situatie een al dan niet bevredigend resultaat, maar in algemene zin bleven de merites grotendeels onbekend. Soms werden elementen van verschillende methoden samengevoegd tot een in logisch opzicht inconsistent geheel.

De tweede benadering van het keuzeprobleem is het vooraf definiëren van wiskundige eisen en vervolgens nagaan welke methoden aan die eisen voldoen. Hoewel deze benadering in principe aantrekkelijk lijkt, zal het in de praktijk vaak moeilijk zijn een logisch samenhangend stel eisen te formuleren, die een juiste afspiegeling vormen van de doelstelling van de classificatie.

Bij de derde benadering worden uit de doelstelling van de classificatie een doelfunctie en eventueel nevenvoorwaarden afgeleid. Vervolgens wordt een classificatie gezocht, waarbij de doelfunctie een zo hoog, dan wel zo laag mogelijke waarde bereikt, en waarbij bovendien aan de eventuele nevenvoorwaarden wordt voldaan. Deze benadering lijkt de voorkeur te verdienen boven de andere twee, o.a. omdat hierbij beter geanalyseerd kan worden waaraan eventuele tekortkomingen van een berekende classificatie zijn te wijten.

In 3.4 zijn doelfuncties, nevenvoorwaarden en optimalisatie-methoden besproken. Als doelfunctie is geselecteerd: het gemiddelde van de gekwadrateerde Euclidische afstand in de eigenschappen-ruimte van de profielen tot de (hypothetische) representant van de klasse waaraan elk profiel is toegewezen. Wanneer de centroide van een klasse, bestaande uit de gemiddelden voor de respectieve variabelen, wordt gebruikt als representant, dan is deze functie identiek aan de gepoolde binnen-klasse variantie, gesommeerd over de variabelen. Hoe kleiner de functiewaarde is, hoe homogener de klassen gemiddeld zullen zijn. Het gewenste aantal klassen kan als nevenvoorwaarde worden geïntroduceerd. Voor het minimaliseren van de doelfunctie werd een relocatiemethode gekozen. Dit is een iteratieve methode, waarbij een beginoplossing stapsgewijze wordt verbeterd door de profielen

steeds opnieuw toe te wijzen aan de klassen, totdat daarmee geen verdere verbetering meer mogelijk is.

In Hoofdstuk 4 is aangegeven hoe de relocatiemethode is ingebouwd in een classificatieprocedure die behalve op homogeniteit van de klassen, ook op hanteerbaarheid in het veld is gericht. De procedure werd toegepast op gegevens die waren verzameld bij een kartering van een gebied rond Stedum en Loppersum (Groningen). De gegevens bestonden uit in het veld geschatte waarden voor 20 opeenvolgende, 10 cm dikke lagen in 2212 profielen; ze betroffen de volgende eigenschappen: het percentage lutum, het percentage humus, de hoeveelheid $CaCo_3$, de graad van fysische rijping, de aanwezigheid van knip, en de hoeveelheid veen in de ondergrond.

In het eerste stadium van de procedure (schematisch weergegeven in fig. 6) werden aparte classificaties berekend voor elk van de zes eigenschappen. De variabelen bestonden uit de waarden voor de betreffende eigenschap in de respectieve lagen; de waargenomen verlopen van die eigenschap met de diepte vormden de te classificeren objecten. De representant van een klasse, die in dit geval kan worden beschouwd als een 'standaardverloop' (b.v. voor het percentage lutum), werd gevormd door de centroide van die klasse na enige afronding en vereenvoudiging om de identificatie van nieuwe verlopen in het veld te vergemakkelijken (fig. 13). Vervolgens vond een synthese plaats door alle zes classificaties met elkaar te combineren tot één classificatie. Elke mogelijke combinatie van standaardverlopen, met één verloop voor elke eigenschap, vormde daarbij de representant van een nieuwe klasse. Aangezien op deze wijze te veel klassen werden gevormd om in het veld te omgrenzen, zijn in een tweede stadium de klassen samengevoegd tot een beperkt aantal grotere eenheden. Net als in het eerste stadium werd hierbij de som van de gepoolde binnen-klasse varianties geminimaliseerd.

Vergeleken met het rechtstreeks berekenen van een classificatie op basis van alle variabelen tegelijkertijd, heeft de hier toegepaste procedure o.a. als voordeel dat het identificeren van bodemprofielen er sterk door wordt vergemakkelijkt. Bovendien zijn reeds verrichte identificaties eenvoudiger bij te werken, wanneer tijdens de kartering blijkt dat de classificatie beter moet worden aangepast aan de bodemkundige variaties in het gebied.

De geografische distributie van de berekende klassen werd in kaart gebracht, waarbij gebruik is gemaakt van een op computer en regeldrukker gebaseerde methode. Deze methode is naar aanleiding van het onderhavige onderzoek ontwikkeld, maar is algemeen toepasbaar wanneer men kwantitatieve gegevens van waarnemingspunten wil gebruiken voor het omgrenzen van klassen.

Bij het toepassen van de classificatieprocedure zoals beschreven in Hoofdstuk 4, dienen op een aantal punten keuzen te worden gedaan; elk van deze keuzen heeft in principe invloed op het eindresultaat. Door experimenten is informatie verkregen over deze effecten.

De classificatie moet worden gebaseerd op een steekproef van bodemprofielen uit het gebied waarvoor zij bestemd is. Om technische en economische redenen dient deze steekproef zo beperkt mogelijk gehouden te worden. Wanneer zij echter te klein is, wordt de classificatie teveel door toeval bepaald. Dan is de kans groot dat de klassegrenzen onvoldoende zijn aangepast aan de populatie als geheel. Uit de experimenten bleek dat 100 profielen te weinig was om er de classificaties voor de afzonderlijke eigenschappen op te

baseren. Een aantal van 600 profielen bleek voldoende, behalve misschien voor het classificeren van humus- en veen-verlopen.

Doordat het relocatieproces kan blijven steken in een lokaal minimum en zo niet de laagst mogelijke waarde van de doelfunctie bereikt, is het resultaat in principe óók afhankelijk van de keuze van de beginoplossing en de volgorde waarin de profielen worden toegewezen aan de klassen. Uit de experimenten bleek dat inderdaad enigszins verschillende classificaties resulteerden uit verschillende beginoplossingen en volgorden, maar dat de bereikte waarden van de doelfunctie slechts in onbelangrijke mate van elkaar afweken. Het bood dan ook nauwelijks voordelen om vooraf via een analyse van een kleinere steekproef een beginoplossing te construeren; uitgaande van een aselecte beginoplossing werden vrijwel even homogene klassen gevonden.

De keuzen waarmee men het eindresultaat doelgericht kan beïnvloeden, betreffen het aantal klassen en de gewichten van de variabelen. In het eerste stadium van de procedure kiest men, apart voor elk van de op de afzonderlijke eigenschappen te baseren classificaties, het aantal klassen en de gewichtsfactoren van de lagen waarin de profielen zijn verdeeld. In het tweede stadium kiest men het aantal klassen van de uiteindelijke classificatie en de gewichtsfactoren van de eigenschappen. Wanneer het aantal klassen wordt vergroot, zal in het algemeen de som van de gepoolde binnen-klasse varianties afnemen, terwijl de verdeling van deze varianties over de variabelen meer gelijkmatig zal worden. Naarmate voor een variabele een groter gewichtsfactor wordt gekozen, zullen bij gelijkblijvende gewichten voor de andere variabelen, de klassen wat betreft die variabele in het algemeen homogener worden, dat wil zeggen dat de gepoolde binnen-klasse variantie van die variabele zal afnemen.

Zowel het aantal klassen als de variabelen en de daarop eventueel toe te passen niet-lineaire transformaties en gewichtsfactoren dienen zo gekozen te worden dat voor die klassen voldoende nauwkeurige schattingen kunnen worden gemaakt met betrekking tot de bodemeigenschappen waarin de kaartgebruikers zijn geïnteresseerd. De effecten op het eindresultaat zijn echter moeilijk kwantitatief te voorspellen; zij hangen af van de frequentieverdeling binnen de steekproef waaruit de classificatie wordt berekend. Daarbij komt dat de homogeniteit van de uiteindelijke klassen wat betreft een bepaalde eigenschap niet alleen wordt beïnvloed door de gewichtsfactor, maar ook door het aantal klassen dat in het eerste stadium van de procedure voor die eigenschap is gedefinieerd. Om deze redenen is het aan te bevelen een gevoeligheidsanalyse uit te voeren, zodat men zich kan concentreren op de parameters met een belangrijke invloed en de diverse keuzen op elkaar kan afstemmen.

Alleen drastische schaalveranderingen van de lutumpercentages in de diverse lagen bleken belangrijke wijzigingen in de verdeling van de binnen-klasse varianties over de lagen tot gevolg te hebben. Van de correlaties tussen de lagen kan men gebruik maken door bepaalde lagen bij het classificeren buiten beschouwing te laten. Echter klassen die slechts gedefinieerd werden op basis van lutum percentages in de bovenste 1,2 m, bleken in het traject van 1,5 tot 2,0 m reeds zeer heterogeen te zijn.

Een algemene conclusie uit het onderzoek is dat de ontworpen procedure een nuttig hulpmiddel kan vormen bij het construeren van classificaties voor bodemkarteringen. Verder onderzoek is echter geboden, onder meer naar de nauwkeurigheid waarmee de klassen in het veld kunnen worden omgrensd en naar de keuze van variabelen en gewichten.

References

Anderson, A.J.B., 1971. Numeric examination of multivariate soil samples. J. Int. Ass. Math. Geol. 3: 1-14.

Arkley, R.J., 1971. Factor analysis and numerical taxonomy of soils. Proc. Soil Sci. Soc. Am. 35: 312-315.

Bakker, H. de, 1970. Purposes of soil classification. Geoderma 4: 195-208.

Bakker, H. de & J. Schelling, 1966. Systeem van bodemclassificatie voor Nederland. Pudoc, Wageningen.

Ball, G.H., 1966. A comparison of some cluster-seeking techniques. Rome Air Development Centre, Rome, N. Y. Technical Report RADC-TR-66-514.

Bodemkaart van Nederland, 1 : 50 000, 1964. Toelichting bij het kaartblad 43 West, Willemstad. Stichting voor Bodemkartering, Wageningen.

Bennema, J., 1974. Organic carbon profiles in Oxisols. Pedologie 24: 119-146.

Berry, B.J.L. & D.F. Marble, 1968. Spatial analysis; a reader in statistical geography. Prentice-Hall Inc., Englewood Cliffs, New Jersey.

Bezdek, J.C., 1974. Numerical taxonomy with fuzzy sets. J. Math. Biol. 1: 57-71.

Bidwell, O.W. & F.D. Hole, 1964a. An experiment in the numerical classification of some Kansas soils. Proc. Soil Sci. Soc. Am. 28: 263-269.

Bidwell, O.W. & F.D. Hole, 1964b. Numerical taxonomy and soil classification. Soil Sci. 97: 58-62.

Bidwell, O.W., L.F. Marcus & P.K. Sarkar, 1964. Numerical classification of soils by electronic computer. 8th Int. Congr. of Soil Sci., Bucharest, p. 933-941.

Bie, S.W., 1972. The relative efficacy of different procedures for soil survey in developing countries and elsewhere. Thesis, Univ. Oxford.

Bie, S.W. & P.H.T. Beckett, 1973. Comparisons of four soil surveys by air-photo interpretation of the Paphos area (Cyprus) Photogrammetria 29: 189-202.

Bock, H.H., 1974. Automatische Klassifikation. Vandenhoeck & Ruprecht, Göttingen.

Bolshev, L.N., 1969. Cluster analysis. Bull. Int. Stat. Inst. 43: 411-425.

Bonner, R.E., 1964. On some clustering techniques. IBM Journal 8: 22-32.

Boon van Ostade, A.H., 1969. Iteratieve cluster analyse. Thesis, Univ. Nijmegen.

Boulton, D.M. & C.S. Wallace, 1970. A program for numerical classification. Comp. J. 13: 63-69.

Bray, J.R. & J.T. Curtis, 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monographs 27: 325-349.

Bunge, W., 1966. Gerrymandering, geography and grouping. Geogr. Rev. 56: 256-263.

Burr, E.J., 1968. Cluster sorting with mixed character types. I. Standardization of character values. Austr. Comp. J. 1: 97-99.

Burr, E.J., 1970. Cluster sorting with mixed character types. II. Fusion strategies. Austr. Comp. J. 2: 98-103.

Burrough, P.A., P.H.T. Beckett & M.G. Jarvis, 1971. The relation between cost and utility in soil survey (I-III). J. Soil Sci. 22: 359-394.

Campbell, N.A., M.J. Mulcahy & W.M. McArthur, 1970. Numerical classification of soil profiles on the basis of field morphological properties. Austr. J. Soil Res. 8: 43-58.

Carmichael, J.W., J.A. George & R.S. Julius, 1968. Finding natural clusters. Syst. Zool. 17: 144-150.

Cattell, R.B. & M.A. Coulter, 1966. Principles of behavioural taxonomy and the mathematical basis of the taxonome computer program. Brit. J. Math. Statist. Psych. 19: 237-269.

Cipra, J.E., O.W. Bidwell & F.J. Rohlf, 1970. Numerical taxonomy of soils from nine orders by cluster and centroid-component analyses. Proc. Soil Sci. Soc. Am. 34: 281-287.

Cline, M.G., 1949. Basic principles of soil classification. Soil Sci. 67: 81-91.

Cole, A.J. & D. Wishart, 1970. An improved Jardine – Sibson algorithm for generating overlapping classes. Comp. J. 13: 156.
Cormack, R.M., 1971. A review of classification. J. Roy. Stat. Soc. A, 134: 321-367.
Cox, D.R., 1957. Note on grouping. J. Amer. Statist. Ass. 52: 543-547.
Crawford, R.M.M. & D. Wishart, 1968. A rapid classification and ordination method and its application to vegetation mapping. J. Ecol. 56: 385-304.
Cronbach, L.J. & G.C. Gleser, 1953: Assessing similarity between profiles. The Psychol. Bull. 50: 456-473.
Cuanalo, H.E.C. & R. Webster, 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.
Dagnelie, P., 1966. A propos des différentes méthodes de classification numérique. Rev. Stat. Appliquée 14: 55-75.
Dale, M.B., G.N. Lance & L. Albrecht, 1971. Extensions to information analysis. Austr. Comp. J. 3: 29-34.
Day, N.E., 1969. Estimating the components of a mixture of normal distributions. Biometrika 56: 463-475.
Demirmen, F., 1969. Multivariate procedures and Fortran IV program for elavuation and improvement of classifications. Computer Contribution 31, State Geol. Survey, Univ. Kansas, Lawrence.
Driessche, R. van den & R. Maignien, 1965. Application d'une méthode de la statistique approfonde à la pédologie. Pédologie 3: 79-88.
Eades, D.C., 1965. The inappropriateness of the correlation coefficient as a measure of taxonomic resemblence. Syst. Zool. 14: 98-100.
Edwards, A.W.F., 1971. Distances between populations on the basis of gene frequencies. Biometrics 27: 873-881.
Edwards, A.W.F. & L.L. Cavalli-Sforza, 1965. A method for cluster analysis. Biometrics 21: 362-375.
Emden, M.H. van, 1971. An analysis of complexity. Thesis, Amsterdam.
Emden, M.H. van, 1972. Interaction analysis, an application of information theory in phytosociology. In: Grundfragen und Methoden in der Pflanzensoziology (Ed. Tüxen), Junk, Den Haag.
Engleman, L. & J.A. Hartigan, 1969. Percentage points of a test for clusters. J. Am. Statist. Ass. 64: 1647-1648.
Farris, J.S., 1969. On the cophenetic correlation coefficient. Syst. Zool. 18: 279-285.
Feller, W., 1950. Introduction to probability theory and its applications. Vol. 1, Wiley, New York.
Firschein, O. & M. Fischler, 1963. Automatic subclass determination for pattern recognition applications. I. E. E. E. trans. on electr. comp., p. 137-141.
Fisher, L. & J.W. van Ness, 1971. Admissible clustering procedures. Biometrika 58: 91-104.
Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics VII: 179-188.
Fisher, W.D., 1958. On grouping for maximum homogeneity. J. Am. Statist. Ass. 53: 789-798.
Flake, R.H. & B.L. Turner, 1968. Numerical classification for taxonomic problems. J. Theor. Biol. 20: 260-270.
Forgy, E.W., 1963. Detecting 'natural' clusters of individuals. Western Psychol. Ass., Santa Monica, California.
Forgy, E.W., 1965. Cluster analysis of multivariate data: efficiency vs. interpretability of classifications. A. A. A. S. – Biometric Soc. meetings, Riverside, California.
Fortier, J.J. & H. Solomon, 1964. Clustering procedures. Dep. of Statistics. Stanford Univ., Techn. Report 7.
Friedman, H.P. & J. Rubin, 1967. On some invariant criteria for grouping data. J. Am. Statist. Ass. 62: 1159-1178.
Gabriel, K.R. & R.R. Sokal, 1969. A new statistical approach to geographic variation analysis. Syst. Zool. 18: 259-278.
Goodall, D.W., 1966. Numerical taxonomy of bacteria: some published data reexamined. J. Gen. Microbiol. 42: 25-37.
Gower, J.C., 1967. A comparison of some methods of cluster analysis. Biometrics 23: 623-637.
Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857-871.

Gower, J.C., 1972. Measures of taxonomic distance and their analysis. In: J.S. Weiner & J. Huizinga (Eds). The assessment of population affinities in man. Clarendon Press, Oxford.
Gower, J.C., 1974. Maximal predictive classification. Biometrics 30: 643-654.
Grigal, D.F. & H.F. Arneman, 1969. Numerical classification of some forested Minnesota soils. Proc. Soil Sci. Soc. Am. 33: 433-438.
Gruijter, J.J. de & S.W. Bie, 1975. A discrete approach to automated mapping of multivariate systems. Proc. Technical working session on Automated Cartography, Int. Cart. Ass. ITC, Enschede.
Hallsworth, E.G., 1965. The relationship between experimental pedology and soil classification. In: E.G. Hallsworth & D.V. Crawford (Eds). Experimental pedology. Butterworth, London, p. 354-374.
Haralick, R.M. & G.L. Kelly, 1969. Pattern recognition with measurement space and spatial clustering for multiple images. Prof. of the IEEE 57: 654-665.
Harding, E.F., 1967. The number of partitions of a set of N points in K dimensions induced by hyperplanes. Proc. Edinb. Math. Soc. 15 (series II): 285-289.
Harris, B., A. Farhi & J. Dufour, 1972. Aspects of a problem in clustering. Univ. Pennsylvania, Inst. for environmental studies. Report.
Hartigan, J.A., 1967. Representation of similarity matrices by trees. J. Am. Statist. Ass. 62: 1140-1158.
Hartigan, J.A., 1975. Clustering algorithms. Wiley, New York.
Hole, F.D. & M. Hironaka, 1960. An experiment in ordination of some soil profiles. Proc. Soil. Sci. Soc. Am. 24: 309-312.
Hotelling, H., 1931. The generalization of Student's ratio. Ann. Math. Statist. 2: 360-378.
Howard, R.N., 1966. Classifying a population into homogeneous groups. In: J.R. Lawrence (Ed.). Operational research and the social sciences. Tavistock Publications, London.
Hughes, R.E. & D.V. Lindley, 1955. Application of biometric methods to problems of classification in ecology. Nature 175: 806-807.
Huizinga, J., 1962. From DD to D^2 and back. The quantitative expression of resemblance. Proc. K. Ned. Akad. Wet. Series C 65: 380-391.
Jancey, R.C., 1966. Multidimensional group analysis. Aust. J. Bot. 14: 127-130.
Jardine, N. & R. Sibson, 1968. The construction of hierarchic and non-hierarchic classifications. Comp. J. 11: 177-184.
Jardine, N. & R. Sibson, 1971. Mathematical taxonomy. Wiley, New York.
Jensen, R.E., 1969. A dynamic programming algorithm for cluster analysis. J. Ops. Res. Soc. Am. 7: 1034-1057.
Johnson, L.A.S., 1970. Rainbow's end: the quest for an optimal taxonomy. Syst. Zool. 19: 203-239.
Johnston, R.J., 1970. Grouping and regionalizing: some methodological and technical observations. Economic Geography 46: 293-305.
Jones, K.S. & D. Jackson, 1967. Current approaches to classification and clump finding at the Cambridge Language Research Unit. Comp. J. 10: 29-37.
Kamping, G. & G. Rutten, 1969. De bodemgesteldheid van het ruilverkavelingsgebied Stedum-Loppersum. Rapport nr. 786, Stichting voor Bodemkartering, Wageningen.
Kendrick, W.B., 1965. Complexity and dependence in computer taxonomy. Taxon 14: 141-154.
Koontz, W.L.G. & K. Fukunaga, 1971. A nonparametric valley-seeking technique for cluster analysis. 2nd Int. Joint Conf. on Artificial Intelligence, Brit. Comp. Soc., London.
Kruskal, J.B., 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29: 1-27.
Kruskal, J.B., 1964b. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29: 115-129.
Krzanowski, W.J., 1971. A comparison of some distance measures applicable to multinomial data, using a rotational fit technique. Biometrics 27: 1062-1068.
Kuiper, F.K. & L. Fisher, 1975. A Monte Carlo comparison of six clustering procedures. Biometrics 31: 777-783.
Lamp, J., 1972. Untersuchungen zur numerischen Taxonomie von Böden – durchgeführt an einem Bodenareal der Hohen Geest Schleswig-Holsteins. Thesis, Kiel.

Lance, G.N., 1973. Hierarchical classificatory methods. In: Ralston, Wilf & Enslein (Eds). Mathematical methods for computer, Vol. III. Wiley, New York.
Lance, G.N. & W.T. Williams, 1965. Computer programs for monothetic classification (association analysis). Comp. J. 8: 246-249.
Lance, G.N. & W.T. Williams, 1966. A generalized sorting strategy for computer classifications. Nature 212: 218.
Lance, G.N. & W.T. Williams, 1967a. Mixed-data classificatory programs. I. Agglomerative systems. Austr. Comp. J. 1: 15-20.
Lance, G.N. & W.T. Williams, 1967b. A general theory of classificatory sorting strategies. I. Hierarchical systems. Comp. J. 9: 373-380.
Landwehr, J.M., 1972. Approximate conficence regions from cluster analysis. Annual Meeting Am. Stat. Ass., Montreal, Canada.
Leeper, G.W., 1963. Introduction to soil science. 4th ed. Melbourne Univ. Press.
Lerman, I.C., 1970. Les bases de la classification automatique. Gauthier-Villars, Paris.
Macnaughton-Smith, P. 1963. The classification of individuals by the possession of attributes associated with a criterion. Biometrics 19: 364-366.
Macnaughton-Smith, P., 1965. Some statistical and other numerical techniques for classifying individuals. H.M.S.O., London.
Macnaughton-Smith, P., W.T. Williams, M.B. Dale & L.G. Mockett, 1964. Dissimilarity analysis: a new technique of hierarchical subdivision. Nature 202: 1034-1035.
MacQueen, J., 1966. Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. on Math. Stat. and Prob.: 281-297.
Marriott, F.H.C., 1971. Practical problems in a method of cluster analysis. Biometrics 27: 501-514.
Michener, C.D. & R.R. Sokal, 1957. A quantitative approach to a problem in classification. Evolution 11: 130-162.
Monmonier, M.S., 1972. Contiguity-biased class-interval selection: a method for simplifying patterns on statistical maps. Geogr. Rev. 62: 203-228.
Moore, A.W. & J.S. Russell, 1967. Comparison of coefficients and grouping procedures in numerical analysis of soil trace element data. Geoderma 1: 139-156.
Moore, A.W., J.S. Russell & W.T. Ward, 1972. Numerical analysis of soils: a comparison of three soil profile models with field classification. J. Soil Sci. 23: 193-209.
Morrison, D.G., 1967. Measurement problems in cluster analysis. Management Sci. 13: B775-B780.
Muir. J.W., H.G.M. Hardie, R.H.E. Inkson & A.J.B. Anderson, 1970. The classification of soil profiles by traditional and numerical methods. Geoderma 4: 81-90.
Naarding, W.H., 1970. Samenstelling en gebruik van bodemkaarten ten behoeve van cultuurtechnische werken. Cultuurtechnisch tijdschrift 10: 54-92.
Needham, R.M., 1966. The termination of certain iterative processes. Rand Corporation, RM-5188-PR, Santa Monica, California.
Norris, J.M., 1971. The application of multivariate analysis to soil studies. I. Grouping of soils using different properties. J. Soil Sci. 22: 69-80.
Norris, J.M., 1972. The application of multivariate analysis to soil studies. III. Soil variation. J. Soil Sci. 23: 62-75.
Norris, J.M. & M.B. Dale, 1971. Transition matrix approach to numerical classification of soil profiles. Proc. Soil Sci. Soc. Am. 35: 487-491.
Norris, J.M. & J. Loveday, 1971. The application of multivariate analysis to soil studies. II. The allocation of soil profiles to established groups: a comparison of soil survey and computer method. J. Soil Sci. 22: 395-400.
Orloci, L., 1967a. An agglomerative method for classification of plant communities, J. Ecol. 55: 193-205.
Orloci, L., 1967b. Data centering: a review and evaluation with reference to component analysis. Syst. Zool. 16: 208-212.
Orloci, L., 1972. On information analysis in phytosociology. In: Grundfragen und Methoden in der Pflanzensoziology (Ed. Tüxen), Junk, Den Haag.
Pankhurst, R.J., 1975. Biological identification with computers. Academic Press, London.
Pearson, K., 1926. On the coefficient of racial likeness. Biometrika 18: 105-117.

Pielou, E.C., 1969. The classification of communities. In: An introduction to mathematical ecology. Wiley, New York, p. 236-249.
Prusinkiewicz, Z., 1969. Application of multivariate statistical analysis and computers in investigations of the genetic homogeneity of glacial deposits. Zesz. Nauk. UAM Geografia 8 – Nadbitka Poznań: 149-165.
Raj, D., 1968. Sampling theory. McGraw-Hill, New York.
Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Statist. Ass. 66: 846-850.
Rayner, J.H., 1966. Classification of soil by numerical methods. J. Soil Sci. 17: 79-92.
Rogers, D.J. & T.T. Tanimoto, 1960. A computer program for classifying plants Science 132: 1115-1118.
Rohlf, F.J., 1970. Adaptive hierarchical clustering schemes. Syst. Zool. 19: 58-82.
Rubin, J., 1967. Optimal classification into groups: an approach for solving the taxonomy problem. J. Theor. Biol. 15: 103-144.
Russell, J.S. & A.W. Moore, 1968. Comparison of different depth weightings in the numerical analysis of anisotropic soil profile data. 9th Int. Congr. of Soil Sci., Adelaide, Vol. 4, p. 205-213.
Sarkar, P.K., 1965. Numerical taxonomy of soils. Thesis. Kansas State Univ.
Sarkar, P.K., O.W. Bidwill & L.F. Marcus, 1966. Selection of characterististics for numerical classification of soils. Proc. Soil Sci. Soc. Am. 30: 269-272.
Schelling, J., 1970. Soil genesis, soil classification and soil survey. Geoderma 4: 165-193.
Schnell, P., 1964. Eine Methode zur Auffindung von Gruppen. Biom. Zeitschr. 6: 47-48.
Sclove, S.L., 1973. Population mixture models and clustering algorithms. Dep. of Statistics, Stanford Univ., Techn. Report 71.
Scott, A.J., 1969. On the optimal partitioning of spatially distributed point sets. In' A.J. Scott (Ed.). Studies in regional science. Pion, London.
Scott, A.J. & M.J. Symons, 1971. Clustering methods based on likelihood ratio criteria. Biometrics 27: 387-397.
Sebestyen, G.S., 1962. Decision-making processes in pattern recognition.
Shepard, D., 1969. A two-dimensional interpolation function for computer mapping of irregularly spaced data. Report No 15, Lab. for computer graphics and spatial analysis, Harvard Univ., Cambridge, Mass.
Siegel, S., 1956. Nonparametic statistics for the behavioral sciences. McGraw-Hill, New York.
Smirnov, E.S., 1960. Taxonomic analysis of a genus. Zharnal Obshenye Biologii 21: 89-103.
Sneath, P.H.A., 1957. The application of computers to taxonomy. J. Gen. Microbiol. 17: 201-226.
Sneath, P.H.A., 1962. The construction of taxonomic groups. In: Microbial classification. 12th Symp. of the Soc. for Gen. Microb., p. 289-332.
Sneath, P.H.A. & R.R. Sokal, 1973. Numerical taxonomy. Freeman, New York.
Soil Survey Staff, 1975. Soil taxonomy – a basic system of soil classification for making and interpreting soil surveys. Agr. Handb. No 436, U.S. Dep. of Agr., Washington, D.C.
Sokal, R.R., 1961. Distance as a measure of taxonomic similarity. Syst. Zool. 10: 70-79.
Sokal, R.R. & C.D. Michener, 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 38: 1409-1438.
Sokal, R.R. & F.J. Rohlf, 1962. The comparison of dendrograms by objective methods. Taxon 11: 33-40.
Sokal, R.R. & P.H.A. Sneath, 1963. Principles of numerical taxonomy. Freeman, New York.
Spence, N.A., 1968. A multifactor regionalization of British counties on the basis of employment data for 1961. Regional Studies 2: 87-104.
Spence, N.A. & P.J. Taylor, 1970. Quantitative methods in regional taxonomy. In: Progress in geography – international reviews of current research. Vol. 2. London.
Steur, G.G.L., 1961. Methods of soil surveying in use at the Netherlands Soil Survey Institute. Auger and Spade 11: 59-77.
Switzer, P., 1970. Numerical classification. In: Geostatistics (Ed. D.F. Merriam), Plenum Press, New York.
Taylor, P.J., 1969. The location variable in taxonomy. Geogr. Analysis 1: 181-195.

Ward, J.H., 1963. Hierarchical grouping to optimise an objective function. J. Am. Statis Ass. 58: 236-244.
Watanabe, S., 1965. Une explication mathématique du classement des objects. In: Dockx & Bernays (Eds). Information and prediction in science. Academic Press, London.
Watanabe, S., 1969a. Automatic feature extraction in pattern recognition. In: A. Grasselli (Ed.). Automatic interpretation and classification of images. Academic Press, London, p. 131-136.
Watanabe, S., 1969b. Knowing and guessing. Wiley, New York.
Watanabe, S., 1969c. Methodologies of pattern recognition. Academic Press, New York.
Webster, R., 1971. Wilks's criterion: a measure for comparing the value of general purpose soil classification. J. Soil. Sc. 22: 254-260.
Webster, R. & P.A. Burrough, 1972a. Computer-based soil mapping of small areas from sample data. I. Multivariate classification and ordination. J. Soil Sci. 23: 210-221.
Webster, R. & P.A. Burrough, 1972b. Computer-based soil mapping of small areas from sample data. II. Classification smoothing. J. Soil Sci. 23: 222-234.
Wilks, S.S., 1932. Certain generalizations in the analysis of variance. Biometrika 24: 471-494.
Williams, W.T., 1969. The problem of attribute – weighting in numerical classifications. Taxon 18: 369-374.
Williams, W.T., 1971. Principles of clustering. Ann. Rev. Ecol. Syst. 2: 303-326.
Williams, W.T. & H.T. Clifford, 1971. On the comparison of two classifications of the same set of elements. Taxon 20: 519-522.
Williams, W.T., H.T. Clifford & G.N. Lance, 1971. Group-size dependence: a rationale for choice between numerical classifications. Comp. J. 14: 157-162.
Williams, W.T. & M.B. Dale, 1965. Fundamental problems in numerical taxonomy. Adv. Bot. Res. 2: 35-68.
Williams, W.T. J.M. Lambert & G.N. Lance, 1966. Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. J. Ecol. 54: 427-445.
Williams, W.T., G.N. Lance, M.B. Dale & H.T. Clifford, 1971. Controversy concerning the criteria for taxonometric strategies. Comp. J. 14: 162-165.
Wishart, D., 1969a. An algorithm for hierarchical classifications. Biometrics 25: 165-170.
Wishart, D., 1969b. User manual for CLUSTAN IA. Computing laboratory. Univ. St. Andrews, Scotland.
Wishart, D., 1969c. Mode analysis: a generalization of nearest neighbour which reduces chaining effects. In: A.J. Cole (Ed.). Numerical taxonomy. Academic Press, London.
Wishart, D., 1970. Some problems in the theory and application of the methods of numerical taxonomy, Thesis, Univ. of St. Andrews, Scotland.
Yamane, T., 1967. Elementary sampling theory. Prentice-Hall Inc., Englewood Cliffs, New Jersey.
Zadeh, L.A., 1965. Fuzzy sets. Information and Control 8: 338-353.