J.J. de Gruijter

Netherlands Soil Survey Institute (Stiboka), Wageningen

Numerical classification of soils and its application in survey

pudoc

Centre for Agricultural Publishing and Documentation

Wageningen – 1977

ISBN 90 220 0608 5

The author graduated on 11 march 1977 as Doctor in de Landbouwwetenschappen at the Agricultural University, Wageningen, the Netherlands, on a thesis with the same title.

This publication will also be issued as Soil Survey Papers No 12 of the Soil Survey Institute, Wageningen.

Abstract

Gruijter, J.J. de (197[illegible]). 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.

Acknowledgments

I am greatly indebted to Dr Ir J. Schelling for his guidance and scientific support of the present study. The many valuable comments on earlier drafts from Prof. Dr Ir J. Bennema and Prof. Dr Ir L.C.A. Corsten are gratefully acknowledged. Many thanks are due to the Board of the Netherlands Soil Survey Institute, who made all necessary facilities available.

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*.

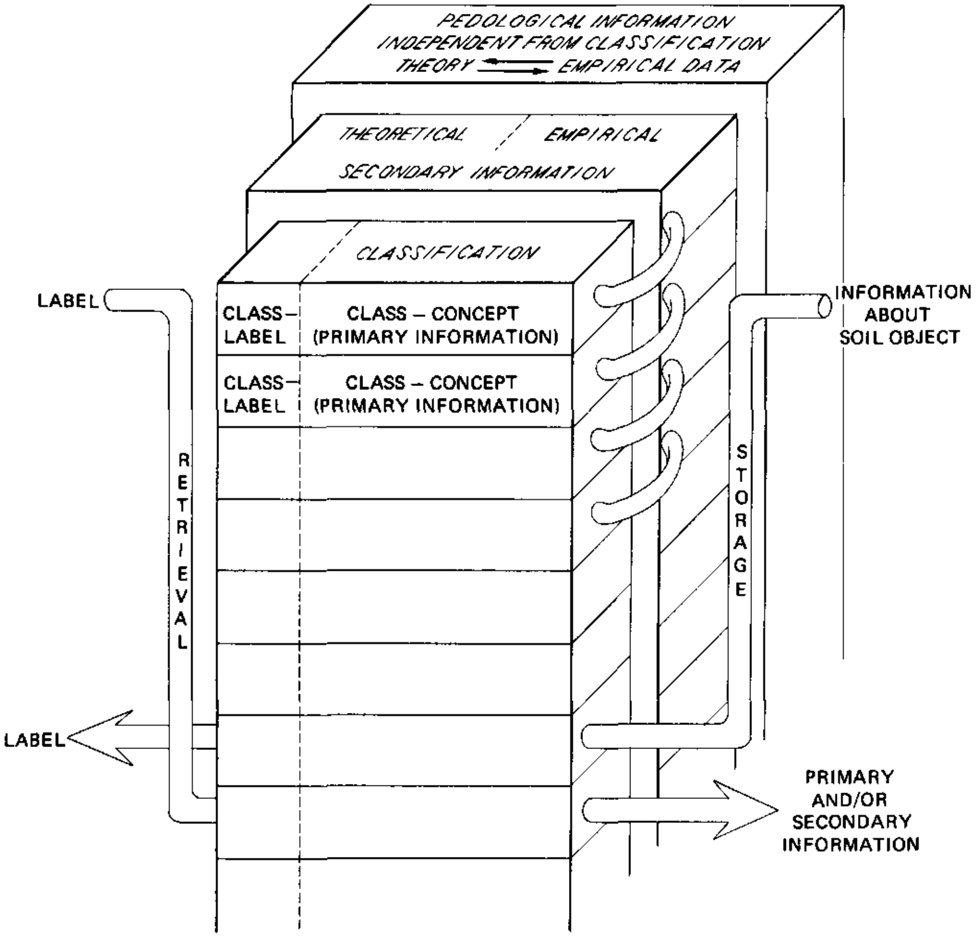

Fig. 1. Storage and retrieval by a classification.

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 heterogeneity 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 Orloci (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: $A \succ B$) if and only if each class of A is a subset of a class of B.

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

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

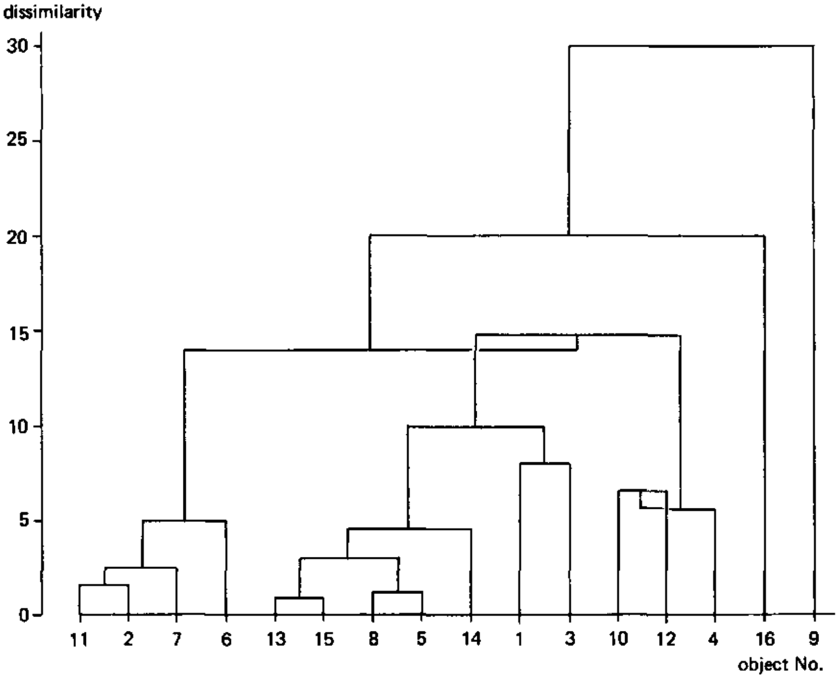

Fig. 2. Fictive dendogram for 16 objects.

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) $\text{tr}(W)$

This measure has a simple geometrical interpretation: $\text{tr}(W)/n$ is the mean squared Euclidean distance between each object and the centroid of the class to which it belongs. Of the three criteria $\text{tr}(W)$ 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) $\text{tr}(W^{-1}\ B)$

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